A Survey of Finding Trends in Data Mining Techniques for Social Media Analysis

Sri Lanka Journal of Social Sciences and Humanities
Volume 1 Issue 2, August 2021: 37-50
ISSN: 2773 692X (Online), 2773 6911 (Print)

Published by Faculty of Social Sciences and Languages, Sabaragamuwa University of Sri Lanka
Website: https://www.sab.ac.lk/sljssh
DOI: http://doi.org/10.4038/sljssh.v1i2.36
SLJSSH

Nanayakkara, A.C.[1,*], Kumara, B.T.G.S.[2] and Rathnayaka, R.M.K.T.[3]

[1] Center for Computer Studies, Sabaragamuwa University of Sri Lanka, Belihuloya, 70140, Sri Lanka.
[2] Department of Computing and Information Systems, Sabaragamuwa University of Sri Lanka, Belihuloya, 70140, Sri Lanka.
[3] Department of Physical Sciences and Technologies, Sabaragamuwa University of Sri Lanka, Belihuloya, 70140, Sri Lanka.

Received: 09 December, 2020, **Revised:** 02 March, 2021, **Accepted:** 03 April, 2021.
How to Cite this Article: Nanayakkara, A.C., Kumara, B.T.G.S. and Rathnayaka, R.M.K.T. (2021). A survey of finding trends in data mining techniques for social media analysis. *Sri Lanka Journal of Social Sciences and Humanities, 1*(2), 37-50.

Abstract

Social media have become very popular in the last few decades. Users rely on social network sites like Twitter, Facebook, YouTube, and LinkedIn for both information and entertainment needs. Social media analytics with data mining technology could be an analysis axis centered on extracting trends, patterns, and rules from the social media pool, to serve the people and organizations to have optimum choices concerning many disciplines. The traditional media analytical techniques appear obsolete and inadequate to gratify this immense array of unstructured social media knowledge characterized by three key problems namely; size, noise, and dynamism, predominantly shifting from the batch scale to the streaming one. The objective of this study is to investigate the data mining techniques that were used by social media networks during the years 2010 and 2020. The effort is a systematic review of content analysis in studies within the field of social media analytics that was published in principal databases. 125 articles were reviewed in this paper. Content analysis was implemented based on their approach, tools utilized, language, the dataset used, country, year, and nature of the experiment. The review discovered that 22 data mining techniques were employed with social media data while frequently used in Artificial Neural Network (ANN), Bayesian networks (BN) and Support Vector Machine (SVM), K-means Clustering, and Neuro-Fuzzy Logic Approach. The study has focused to assist the involved analyzers and educators to capture the research trends and problems associated with the Social media analytics process with future research initiatives.

Keywords: Data Mining, Data Mining Techniques, Social Media, Social Media Analysis

INTRODUCTION

Today, the use of social networks is growing ceaselessly and rapidly. These networks have become a substantial tool for unstructured data that belong to a host of domains, including education, business, science, and health, significantly (Smita & Sharma, 2014). The increasing reliance on social networks calls for a comprehensive tool has been the need for the hour in catering to such requirements in terms of analysis, understanding, and decision making on the generated data that is likely to facilitate reforming the unstructured data and place them within a systematic pattern. This situation has staged the presence of 'Data Mining' (Pal, 2011). Data mining is the process of digging meaningful insights from large data sets by using various statistical, machine learning, databases, and Artificial Intelligence methods and dispersing the knowledge into a form that is advantageous for various real-world applications (Tiwari & Kumar, 2020). It facilitates predicting future trends and behaviors, making proactive decisions, increasing profits, and cut costs, to answer business questions that consume too much time to answer (Fawzy et al., 2016).

Online Social networking (OSN) is a term used to describe web-based services that allow individuals to create a public/semi-public profile within a domain that lets its participants connect, build relationships, and collaborate on social issues i.e. Facebook, Twitter, LinkedIn, YouTube, etc. (Tiwari & Kumar, 2020). This has improved the concept and the technology of Web 2.0, by enabling the formation and exchange of User-Generated Content (Kaplan & Haenlein, 2010). There are various tools and techniques of data mining that have been utilized for OSN mining for decades. According to Nandi & Das, (2013), there is a strong motivation for efficiently propagating the right information to the right people via OSNs and which has become a research area of increasing importance. Different data mining techniques have been studied to process and analyze several types of data patterns, where the most popular data mining tasks were classification, summarization, association rules mining, and clustering in the selected period. (Tiwari & Kumar, 2020).

The problem statement which stimulates the study is "How were the data mining techniques being evolved concerning the social media research during the last decade?" To answer this broad question it is categorized as the core research objective and supplementary research questions. The main objective of this study is to summarize and explore the trends, applications, and hidden gaps in the usage of data mining techniques concerning social media analysis.

* **Corresponding author:** Tel.: +94 (71) 861 3988; Email: chethana@ccs.sab.ac.lk
https://orcid.org/0000-0002-1152-1552

Accompanied supplementary research questions (RQs) are given below.

RQ1: Were there significant differences among the data mining techniques used during the last decade in which are adequate to identify as separate cohorts?

The role of this question is to identify the historical development of data mining techniques since its childhood.

RQ2: Was there a substantial difference among the applications of data mining techniques in different domains of applications during the last decade?

This question aims to identify the available application domains which have been used data mining techniques for their day-to-day operations.

RQ3: Was there a substantial dissimilarity among the data mining techniques used and studied during the last decade?

The goal of this question is to find and categorize prevailing data mining techniques according to their main usage.

RQ4: Is there substantial segregation of methods in predictive data mining techniques used and studied during the last decade?

One Hundred and twenty-five (125) research articles were reviewed in compiling this paper. Content analysis has been implemented based on their approach, tools utilized, the dataset used, country, year, and the application. After a careful review of these articles, we found that 22 data mining techniques have been used with social media data However, still, it is in its infancy and it needs more effort by academia and industry to overcome all the issues and challenges with social media analysis.

This paper presents a comprehensive study for the data mining techniques, their applicability with social media Analysis and the challenges with them. The rest of the paper is organized as follows. Section two describes the applied methodology. Section three explains Literature Review while section four discusses the results and challenges. Finally, the study is concluded in Section Five by stating the Conclusions and future research initiatives.

LITERATURE REVIEW

Evolution of Data mining (RQ1): The labor-intensive illustration of patterns from data has happened for centuries. The origination of Data mining is rooted in four family lines: classical statistics, databases, artificial intelligence, and machine learning. Information Harvesting, Knowledge Mining, Knowledge Discovery in Databases (KDD), Data Dredging, Data Pattern Processing, Data Archaeology, Database Mining, and Knowledge Extraction are some of the synonyms for the term "data mining" (Smita & Sharma, 2014). In 1936, as the Computer age started where computer storage, processing power, and software have been increased, users have been focusing to store more transactional data and yield optimum use of their data to generate the best decisions (Pal, 2011).

Data has been reformed to information that is adequate to respond to many requirements and even to forecast the future of business opportunities and challenges. In the 1970s it made it possible to store and query terabytes in petabytes of data with sophisticated database management systems. The term KDD was found in 1989. The term “Data Mining” seemed around 1990 in the database community (Moussa et al., 2016).

As data sets have grown in magnitude and density, direct practical data analysis has gradually been improved with indirect, automated data processing, such as neural networks, cluster analysis, genetic algorithms, decision trees, and support vector machines. Even Though there are a large number of data mining algorithms under various techniques (Injadat et al., 2016), further studies are yet to see the world, gradually evolving with more innovative tools and techniques.

Figure 1: Technology evolution towards data mining

Source: Moussa et al., 2016.

Applications of Data mining (RQ2): With the quick developments in the field of data mining and its importance in various application sectors, many efforts have been taken to apply it to routine activities such as banking, insurance, sales and marketing, education, telecommunication, fraud detection, finance, medical, and so on (Adedoyin-Olowe et al., 2013; S.Neelamegam, 2015; Pushpam & Jayanthi, 2017; Sharma, 2014). Some of the vital application sectors are listed below.

Data Mining in the Education Sector: Educational Data Mining (EDM) is an emerging interdisciplinary research field that concerns developing methods that discover knowledge from data originating from educational Environments (Katare & Dubey, 2017). It can be used to predict students' future learning behavior and make accurate decisions on both students and the institution (Sasikala & Seenuvasan, 2007). With the outcomes, the institution can concentrate on what to teach and how to teach. Learning patterns of the students can be captured and used to develop techniques to teach those (Nandal et al., 2017). Among all data mining in higher education is a novel research field and this area of research is gaining popularity because of its potentials to enhancement of educational institutes and benefit all the stakeholders in the educational system. When considering the Literature there is a number of efforts in Educational data mining by using supervised classification algorithms like (Decision Tree, SVM, ANN, k-nearest Neighbors, and Bayesian networks), Regression Analysis, and K-Means Clustering. (Jacob et al., 2016).

Data mining in Health Care: The essence of data mining is in the identification of patterns, relations, and models which gives support for predictions and the decision-making process for diagnoses and treatment planning in the field of health care (Moghaddasi, Hamid, Hoseini, Azamossadat, Asadi, Farkhondeh, Jahanbakhsh, 2012). In 2010 Sony and Gandhi highlighted the importance and usage of data mining in medicine and public health. Namely Prevention of data overload, Evidence-based medicine and prevention of hospital errors, Policy-making in public health, Early detection and management of diseases and pandemics, More value for money and cost savings, Non-invasive diagnosis, decision support, and early detection of Adverse drug events (ADEs) (Sony et al,2010). The automation of the medical field facilitates generating a plethora of electronic data about the health sector. With increased access to a large number of health records, healthcare providers have focused on optimizing the efficiency and quality of their organizations by using data mining techniques. In 2015 Lee and his group evaluated The interaction effect of Electronic health records (EHR) and hospitalist care on length of stay(LOS) by using generalized linear models with log-link normal distribution after controlling for patient and hospital characteristics (Lee et al., 2015). Data mining has already proven effective in areas such as predictive medicine, customer relationship management, detection of fraud and abuse, management of healthcare, malnutrition in children, and measuring the effectiveness of certain treatments (Sarker et al., 2016).

Data Mining in Agriculture: Substantial research interest has been witnessed in data mining in the agricultural background recently as lots of data mining techniques have been used in agriculture (Baskar et al., 2010). Applying data mining in agribusiness is a novel exploration field that can yield hidden patterns in huge data set to solve complex agriculture problems and predict the future trends of agricultural processes such as soil water boundaries for a specific soil type can be assessed by knowing the behavior of comparative soil types (Smita & Sharma, 2014). According to (Ramesh & Vardhan, 2013) Different Data Mining techniques have been used for Crop yield prediction, such as K-Means, K-Nearest Neighbor (KNN), Artificial Neural Networks(ANN), and Support Vector Machines(SVM) with respect to four parameters specifically year, rainfall, production and area of sowing. (Ami & Vinita, 2016) states that yield prediction is a very important agricultural problem that remains to be solved and it can be solved by employing data mining techniques like clustering and classification via selecting the most appropriate method for the task. (Medar et al., 2019) analyzed result of Multiple Linear Regression, Regression Tree, K-nearest Neighbor and Artificial Neural Network on Groundnut data of previous 8 years and they have done prediction based on soil, environmental and abiotic attributes. KNN algorithm had been given the best result compared to other algorithms for Groundnut crop yield prediction.

Data mining in Marketing & Finance: Markets are notable clients of data mining strategies. A retailer can use point-of-sale records of customer purchases to develop products and promotions which appeal to specific customer segments (Jeyapriya & Selvi, 2015). Numerous supermarkets offer free loyalty cards to clients that give them access to scaled-down prices not available to nonmembers. The cards make it simple for stores to follow who is purchasing what, when they are buying it, and at what cost. The stores would then be able to utilize this information, after analyzing it, for different purposes e.g. deciding when to put items on sale or when to sell them at the maximum. Moreover, they use data mining techniques to perform market basket analysis, sales forecasting, database marketing, and merchandise planning and distribution (Choudhery & Leung, 2017; Zhai et al., 2011). In Market basket analysis using data mining techniques, the ultimate goal is to find the products that customers frequently purchase together. The stores can use this information by putting these products close to each other and making them more visible and accessible for customers at the time of shopping (Maheshwari et al., 2016). In the banking field, data mining is used to predict credit card fraud, to estimate risk, to analyze the trend and profitability (Moin & Ahmed, 2012). In the financial markets, data mining techniques such as neural networks are used in stock forecasting, price prediction, and so on (S. P. Singh & Campus, 2013).

Data mining in Elections: It is a trending topic to predict election results through community ideas. Social media sites including Twitter, Facebook, and YouTube played a significant role in raising funds and getting candidates' messages to voters. (Anjaria & Guddeti, 2014) employed several supervised machine-learning techniques to classify the Twitter data for the case study of United States Presidential Elections 2012 and Karnataka State Assembly Elections (India) 2013. Conver describes several methods for predicting the political alignment of Twitter users based on the content and structure of their political communication in the run-up to the 2010 U.S. midterm elections (Conver et al, 2011). Anticipating the political behavior of people is very helpful for the election candidates to evaluate the possibility of their success and to be acknowledged about the public motivations to select them. In 2013 Amin and his group were able to design a participation anticipating system in the 11th presidential election of the Islamic Republic of Iran by using KNN, Classification Tree, and Naïve Bayes (Sangar et al., 2013). These demanding examples highlight the potential of data mining applied to social media data to predict outcomes at a national level.

Data mining in Telecommunication: The telecommunications field implement data mining techniques because of telecommunication industry generates massive, dynamic, competitive, and heterogeneous data from various operational systems which can be used for solving business problems that required urgent handling. These data include call detail data, customer data, and network data. Data Mining methods and business intelligence technology are widely used for handling business problems in this industry (Pal & Patel, 2014). To handle this large amount of data and to discover useful information from this data the automatic or semi-automatic method should be used as it simplifies the work and already there are some available Data Mining techniques in the domain. In 2013 Mohsin & Vidya pointed out the need for data mining under five criteria such as fraud detection, retain customer satisfaction, studying customer behavior, and detecting the highest profitable products among all (Moshi et al., 2013). Rahul and Usharani have invented highly sophisticated, customized, and advanced decision support systems by using data mining techniques that are well-predicting customers' churn behavior in advance (J & T., 2011).

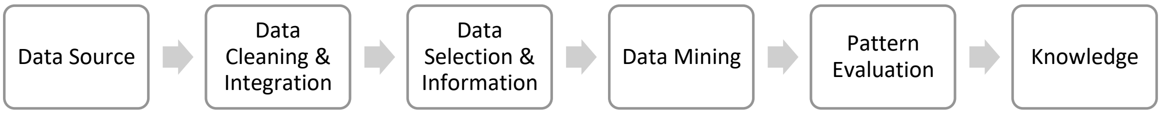
Data mining in Social Media: The fast evolution of social media platforms and their related applications have reformed the way Billions of people interact with each other on the Web. (Pushpam & Jayanthi, 2017). The mining of OSNs is an in-depth process of study due to its unpredictable nature of data (dynamic, wide-ranging, stormy, and scattered). This was made probable at the moment thanks to advances in data science and artificial intelligence in fields like pattern recognition, information fusion, data visualization, and knowledge discovery. This advancement of social media has bloomed on a plethora of end-user, or user-centered applications that required innovative and efficient techniques for data pre-processing, processing, analyzing, visualization, and predicting (Camacho et al., 2020). We can study and forecast several aspects like social media usage, online behaviors, sharing of content, connections between individuals, online buying behavior, etc., and these patterns can give substantial data to organizations, governments, and non - benefit associations, to plan their systems or present new projects, items or services (Balan & Rege, 2017). Even today, some industrial applications apply data mining techniques to OSNs mining to give better satisfaction to its users such as

Trending YouTube, Google Analytics, and YouTube Analytics. Social media analysis (SMA) is rapidly spanning the fields of machine learning, natural language processing, sentiment analysis, and evolutionary computation. This is not a novel research arena but the recent explosion of some OSNs and their advances in deep learning have reinvigorated the field and opened a new set of challenges and possibilities.

There exist a plethora of applications that attempt to extract valuable hidden insights from vast amounts of semi-structured and unstructured social media data to enable informed and insightful decision making (Pushpam et al., 2013, Zamani Alavijeh et al., 2015, Adedoyin-Olowe et al., 2013, Sebei et al., 2019). In my Results and Discussion section, I have included some of them. It was identified almost 25 most popular data mining techniques were among researchers during the last decade. It was pointed out that a large number of applications have been designed for analyzing opinions towards a product (user sentiment analysis) (Bahrami et al., 2011, Naqvi et al., 2018) predicting elections results (Sangar et al., 2013, Anjaria & Guddeti, 2014), detecting vaccination communities (Donzelli et al., 2018, Hourcade et al., 2018), studying how fake news spread through social networks (Hussain et al., 2018, Chakraborty et al., 2020), user behavior modeling (Blanco et al., 2020, Meldrum et al., 2017), decomposing and discovering social events (J. Singh et al., 2019, Li et al., 2013), topic modeling (Jelodar et al., 2020, Patil & Algur, 2019), etc. Conversely, the problem is still open, new hybrid techniques are required for information fusion, classification, pattern recognition, and disambiguation. This has made the areas that work in this field very diverse, spanning from computer science to network science, social sciences, mathematical sciences, medical and biological sciences, financial, management, and political sciences.

Data Mining Algorithms (RQ3): 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 examined over numerous iterations to locate the ideal parameters for making a model. These parameters are then applied across the whole data sets to extract noteworthy patterns. Selecting the optimum algorithm for a particular analytical task can be a challenge. Various algorithms can be utilized to perform the same task, each algorithm creates different outcomes, and few algorithms create more than one sort of result. More algorithms are available in the field to fulfill the needs of users (Pal, 2011).

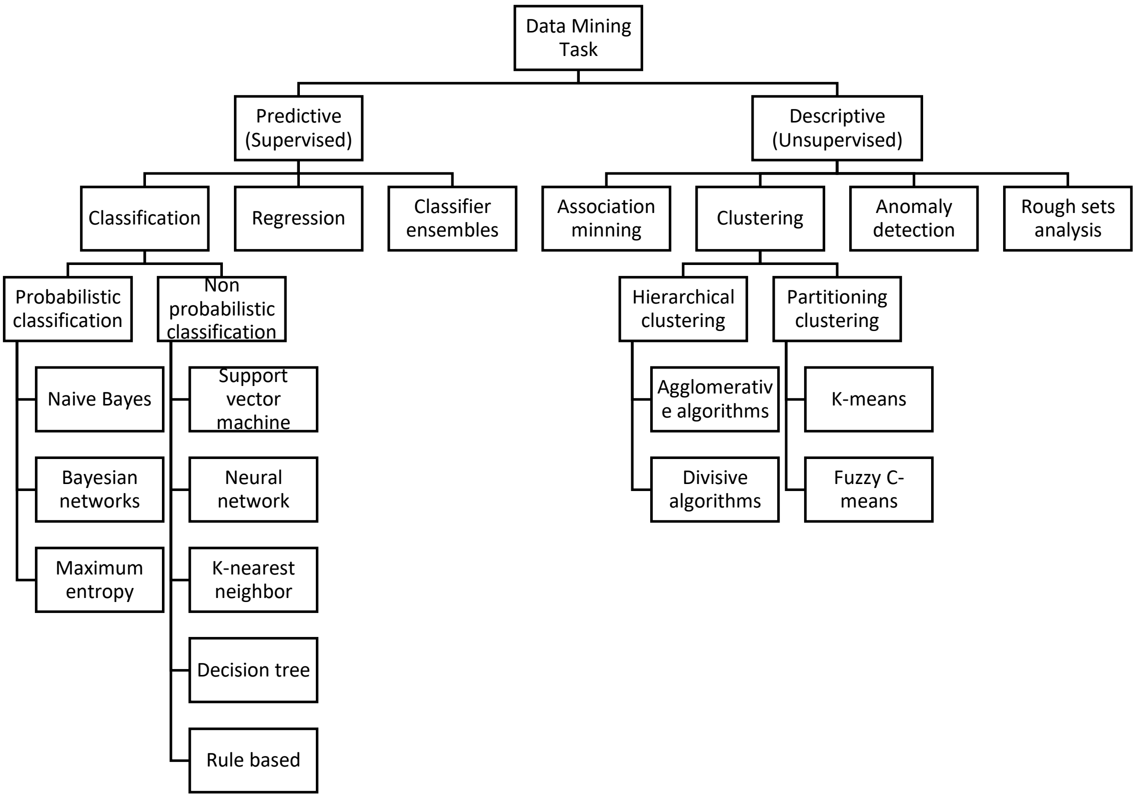
Figure 2: Steps in the data mining process



Source: Sebei et al., 2019

Several extensions have been engaged with data mining for mining massive, unpredictable, and fast-spreading behavior of data such as spatial data, graph data, web content data, and social media data, and new big data mining. According to Nandi & Das, (2013) data mining is collecting relevant information from unstructured data. Hence it helps to achieve specific objectives. The final output of a data mining effort is typically either to create a descriptive model or a predictive model. A descriptive model presents, in a concise form, the core characteristics of the data set. The purpose of a predictive model is to allow the data miner to predict an unknown (often future) value of a specific variable; the target variable (Pushpam & Jayanthi, 2017). The categorization of data mining techniques is shown in figure 3.2.

Figure 3: Classification of Data mining task



Source: Pushpam & Jayanthi, 2017.

As per fig: 3.2, there are two main types of data mining tasks i.e. Supervised machine learning and unsupervised machine learning. Supervised approaches (Predictive) rest on a priori knowledge of the data (e.g. class labels) while unsupervised algorithms (Descriptive) are used to characterize data without any prior knowledge as to what kinds of patterns will be revealed by the algorithm. The decision of selecting an approach whether supervised or unsupervised is depended on the data set and the problem domain.

Classification is a common supervised approach and is appropriate when the data set has labels or a small portion of the data has labels (Tiwari & Kumar, 2020). Classification algorithms begin with a set of training data which includes class labels for each data element. The algorithm learns from the training data and builds a model that will automatically categorize new data elements into one of the distinct classes provided with the training data. Classification rules and decision trees are examples of supervised Classification techniques.

Clustering is a habitual unsupervised data mining technique that is advantageous when handling data sets without labels (Zhai et al., 2011). Clustering algorithms do not depend on labeled training data to generate a model, unlike classification algorithms. As an alternative, clustering algorithms determine which elements in the data set are similar to each other based on the similarity of the data items (Tiwari & Kumar, 2020).

METHODOLOGY

This study examines related research articles relevant to social media analytics in principle databases during the years 2010 and 2020 as the most popular social networks (Facebook, Twitter, LinkedIn, and YouTube) began after 2002. The different publishing databases were chosen due to the availability of certain journals and accessibility of the abstract and full text for the selected articles. One single search engine was used for the data acquisition i.e. Google.

Search Query: The following keywords were used for the search queries to discover the relevant research studies.

1. “data mining” AND “techniques” OR “algorithms” AND “Online Social Network"
2. “data mining” AND “machine learning” AND “social media”
3. “social media” AND “data mining”
4. Online “social media” OR “social network” AND “data mining”

Survey Resources: The following digital libraries were searched for the required articles:

1. IEEE Xplore
2. Google Scholar
3. Science Direct
4. Research Gate
5. Elsevier
6. Publons
7. Academia

Finally, after applying all filtration steps, One Hundred and twenty-five (125) articles were considered as the resources for this review.

Data Extraction: Data extraction had been an arduous effort. Almost three hundred research papers were explored to select the most relevant research articles. Figure 2.0 indicates the format used in data extracting which includes the considered criteria: (author, approach, tools utilized, experiment, language, the dataset used, country, year, and title).

Figure 04: Data Extraction form

Year	Author	Approach	Tools	Experiment	Dataset	Country
2010	Fengkun Liu a, Hong Joo Lee	Recommender System	CF (Collaborative filtering)	Exploits association among users by way of item recommendation.	Cyworld	Korea
2010	Dwi AP Rahayu, Shonali Krishnaswamy, Oshadi Alahakoon	Product Ratings and Reviews	RnR Reviews and Ratings	It employs user-input-oriented-system to develop relative new reviews.	TripAdvisor.com	
2010	Hongning Wang, Yue Lu, Chengxiang Zhai	Aspect Rating Analysis	Latent Aspect Rating Analysis (LARA)	Used to determine every reviewer’s latent score on each aspects and the relevant influence on users when making final decision.	Hotel reviews from Trip Advisor	USA
2010	Swit Phuvipadawat, Tsuyoshi Murata	Topic Detection and Tracking (TDT)	Hot stream	Used for detecting and tracking breaking news in Twitter	Twitter	Japan
2010	Lujun Fang and Kristen LeFevre	Classification	Decision tree	Use dataset of Facebook and suggested Decision Tree outperforms Brute Force	Facebook	USA
2010	Tatsuya Fujiska, Ryong Lee, Kazutoshi Sumiya	Clustering	*K*-means	Micro-blog data gathers from Twitter	Twitter	Japan

Source: Compiled by author, 2020

RESULTS AND DISCUSSION

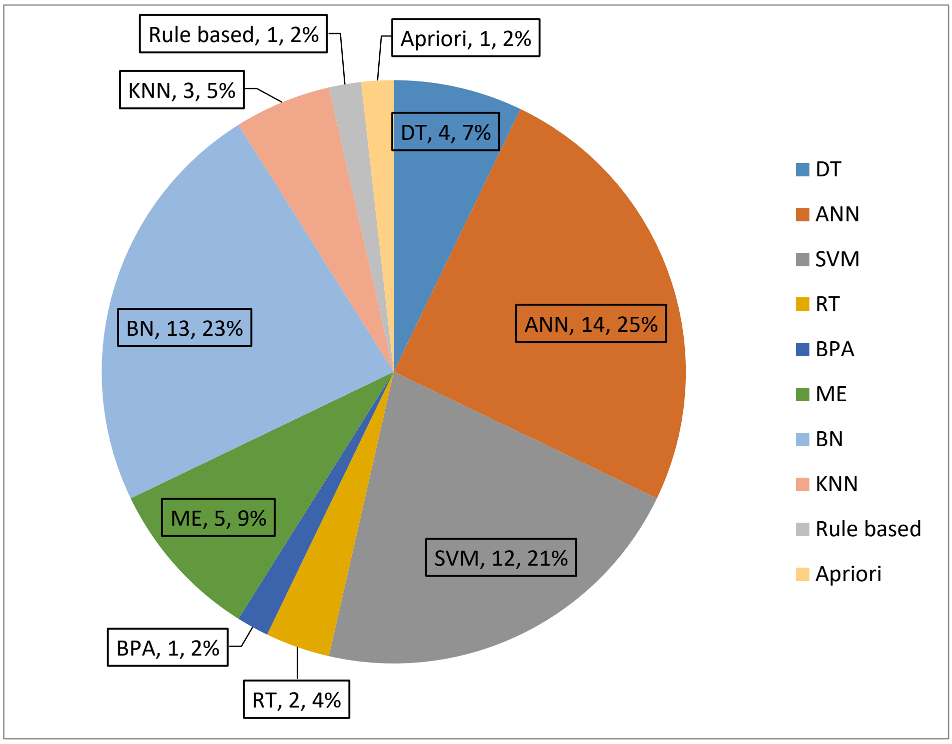
In this section, we will discuss the results obtained from this review after analyzing data in graphical mode. One hundred and twenty-five (125) research articles have been reviewed for this study in the application of data mining techniques in social media. The selected articles were retrieved only from journals published between January 2010 and December 2020. This survey identified that 22 data mining techniques have been applied by researchers in the area of social media analysis, within the sample. They are explained under the categories of predictive and descriptive data mining methods.

Predictive Data Mining techniques applied to SMA (RQ4): Predictive data mining studies utilize past statistics and business intelligence to find patterns and predict trends. It takes specific variables or values in the data set to forecast unknown or future values of other variables of interest. Any effort to quantify the probable future based on past measures is involved by predictive analytics. Predictive data mining methods possess several algorithms that become very useful to mine OSNs data (Pushpam & Jayanthi, 2017).

The names and the percentage of applied predictive data mining techniques were available in my sample are given below in Figure 4.0.

Figure 05: Summary of predictive data mining applications

Predictive data mining technique
Decision Tree (DT)
Artificial Neural
Support Vector Machine (SVM)
Regression tree
Back Propagation algorithm
Maximum entropy (ME)
Bayesian networks (BN)
k-nearest Neighbors (k-NN)
Rule-based
Apriori Algorithm



Source: Compiled by author, 2020

Fig. 4.0 shows that Artificial Neural Network (ANN), Bayesian Networks (BN), and Support Vector Machine (SVM) are the most applied techniques in the area of social media during the year 2010 to 2020 with a percentage of 69 % from among the selected articles. Maximum Entropy (ME), Decision Tree (DT), and k-nearest Neighbors (k-NN) techniques have been used at the intermediate level while rule-based, apriori, RT and BPA were used marginally during the period concerned among the selected articles.

A chronological summary of some of the recent articles in classification based predictive data mining on OSNs is given in table 01 below.

Table 01: Predictive data mining applied to Social media data

Reference	Application	Techniques	Dataset	Country
(Fang & LeFevre, 2010)	Privacy wizards on social networks	DT	Facebook data	United State of America (USA)
(Bozkir et al., 2010)	Identification of user patterns	DT, ANN, SVM	Facebook	Turkey
(Surma & Furmanek, 2010)	identify the proper the target group in business	Classification and Regression tree	Biznes.net	Poland
(Priyadarshini, 2010)	Functional Analysis of ANN	ANN, Back Propagation algorithm	Medical data	India
(Rushdi Saleh et al., 2011)	Product and service	SVM	IMDB, Epinions, Amazon	Spain
(Castillo et al., 2011)	Information credibility on social media	DT	Twitter data	Spain
(Conovor, 2011)	Topic Detection and Tracking (TDT)	SVM	Twitter 2010 U.S. elections	USA
(Quercia et al., 2012)	Tracking "Gross Community Happiness"	ME	Twitter data	United Kingdom (UK)

(Liang & Dai, 2013)	Opinion Mining on Social media	BN	Twitter data	Taiwan
(Dalal & Zaveri, 2013)	Automatic Classification of Unstructured Blog Text	BN & ANN	Blog text	India
(Ren & Kang, 2013)	Predicting the delicate human emotions	BN	weblog	China
(Moraes et al., 2013)	The film, product (GPS, BOOK, Camera)	SVM, ANN	Amazon Benchmark dataset	Brazil
(Duwairi & Qarqaz, 2014)	Education, sports, political news	BN, SVM, KNN	Facebook data	Arabia
(Anjaria & Guddeti, 2014)	Election prediction	BN, ME, ANN	Twitter data	India
(Chen et al., 2014)	Understanding Students' Learning Experiences	BN	Twitter- Purdue dataset	USA
(Habernal et al., 2015)	Product, movie	ME, SVM	Social media data	Czech
(Shankar et al., 2016)	Text Filtering in OSN	SVM	Social media data	India
(Gauri et al., 2015)	Educational practices, Problems, and Issues mining	BN	students twitter data	India
(Yin et al., 2015)	Emergency Situation Awareness	BN and SVM	Twitter data	Australia
(Besiashvili et al., 2017)	Spam Filtration	ANN	Spam emails	Georgia
(Jeyapriya & Selvi, 2015)	Product review	BN	Amazon, Epinions, Cnet	India
(Remya et al., 2015)	educational data mining	BN	Student comments	India
(Gao et al., 2015)	Explore the emotion causes	Rule-based	Weibo	China
(Tang et al., 2015)	Review rating prediction (film and restaurant)	ANN	rotten tomatoes and yelp	China
(Le et al., 2016)	Tourism, Lodging Business in Philadelphia	ANN	Logistics data	Japan
(R. Kumar & Sharma, 2016)	Cloud Architecture	Hybrid Neuro-Fuzzy Approach in Cloud	Social media dataset	USA
(Alfaro et al., 2016)	Electoral campaigns	SVM, KNN	Personal weblog	Spain
(Dey et al., 2016)	Sentiment Analysis of Review Datasets	KNN and BN	IMDb, hotel reviews	India
(Tripathy et al., 2016)	Movie reviews	SVM, BN, ME,	IMDB	India
(Sarker et al., 2016)	Social Media Mining for Toxicovigilance	SVM, ME	Twitter data	USA
(Severyn et al., 2016)	Products (Apple iPad, Motorola Xoom, fiat500, etc)	SVM	YouTube data	Italy
(Kumar et al., 2017)	emoji detection, spelling correction, and emoticon detection	SVM classifier	Twitter data	Canada
(Balan & Rege, 2017)	Usage patterns of social media by small businesses	IBM Watson Analytics	Twitter data	Spain
(Choudhery & Leung, 2017)	Prediction of Box Office Revenue	polynomial regression model	Twitter data	Canada
(D. H. Pham & Le, 2018)	Hotel review	ANN	Trip Advisor	Vietnam
(Hourcade et al., 2018)	factors affecting consumers' perceptions	Convolutional neural networks (CNN)	Youtube data	Japan
(Kalra et al., 2019)	Classify n number of videos	Random Forest Classifier	YouTube data	India
(Jagiello et al., 2019)	Citizen science, behavior identification	Detrended correspondence analysis	YouTube data	Poland

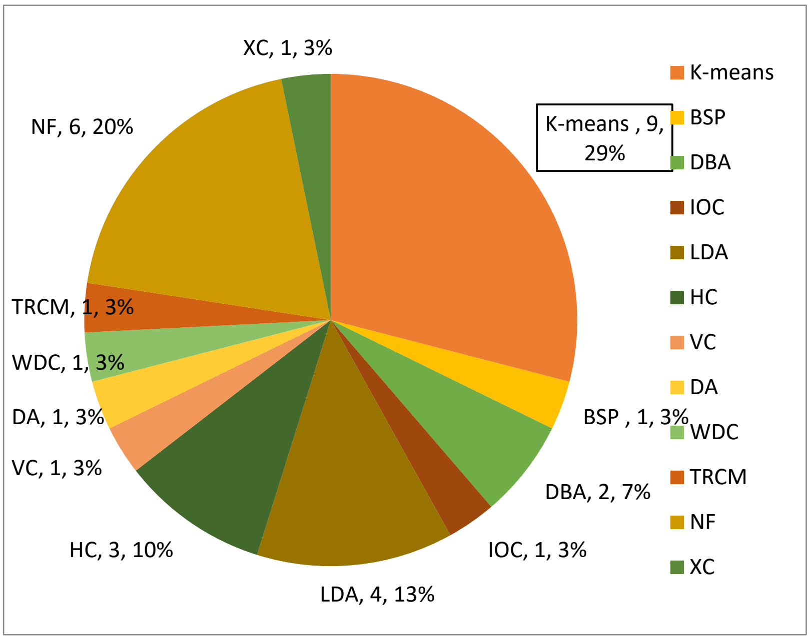
Source:

Descriptive Data Mining techniques applied to SMA (RQ5): Descriptive data mining methods search and summarize historical data to identify new insights for future events. This effort has focused on reports of what had happened. Table 4.1 explains some of the descriptive data mining studies performed by various researchers during the selected period.

The names and the percentage of applied descriptive data mining techniques were available in my sample are given below in Figure 4.1.

Figure 06: Summary of Descriptive data mining applications

Descriptive data mining
K-means Clustering
Business system planning (BSP)
Density Based Algorithm (DBA)
Incremental online Clustering(IOC)
Hierarchical Agglomerative Clustering (HAC)
Latent Dirichlet Allocation(LDA)
Vertex Clustering (VC)
Divisive Algorithm (DA)
Web Document Clustering (WDC)
Transaction-based Rule change Mining(TRCM)
Neuro Fuzzy Logic (NF)
Xmeans Clustering (XC)



Source: Compiled by author, 2020

Fig. 6 has shown that K-means Clustering, Neuro-Fuzzy Logic Approach (NF), Latent Dirichlet Allocation (LDA), and Hierarchical agglomerative clustering (HC) were the most applied descriptive techniques in the area of social media during the year 2010 to 2020 with a percentage of 72 % from among the selected articles. BSP (Business System Planning), VC (Vertex Clustering), DA (Divisive Algorithm), TRCM (Transaction-based rule Change Mining), and WDC (Web Document Clustering) techniques were the least applied methods among the studies concerned during the period considered.

A chronological summary of some of the recent studies in clustering-based descriptive data mining on OSNs is given in table 02 below.

Table 02: Descriptive data mining techniques applied to Social media data

Reference	Application	Techniques	Dataset	Country
(Fujisaka & Lee, 2010)	User Behavior Patterns Identification	K-means	Twitter data	Japan
(B, 2011)	Social Network Analysis	BSP	Social media data	India
(Tsagkalidou et al., 2011)	People emotion-laden reactions and attitude	K-means	Twitter data	Greece
(Yang & Ng, 2011)	Web Opinion Development and Social Interactions with DBC	DBA	MySpace.com	China
(Becker et al., 2011)	Topic Detection and Tracking (TDT) Real-World Event Identification	IOC, NB	Twitter data	USA
(Zhai et al., 2011)	Opinion Mining	EM-based & constrained-LDA	Insurance dataset	USA
(Weng et al., 2011)	Topic Detection and Tracking (TDT)	Event Detection with Clustering	Twitter data	Singapore
(M. C. Pham et al., 2011)	Recommender System	HAC	DBLP digital library, Epinion	German
(Papadopoulos et al., 2012)	Community detection (hierarchical clustering)	VC	Flickr data	Greece
(Li & Liu, 2012)	Sentiment Analysis	K-means	IMDB film reviews	Australia
(Long et al., 2012)	Churn Analysis of OSN users	K-Means, DT	Pengyou data	China
(Tsagkalidou et al., 2011)	Effect-aware community detection (cultural, social, economic, political events)	DA	Twitter data	Greece

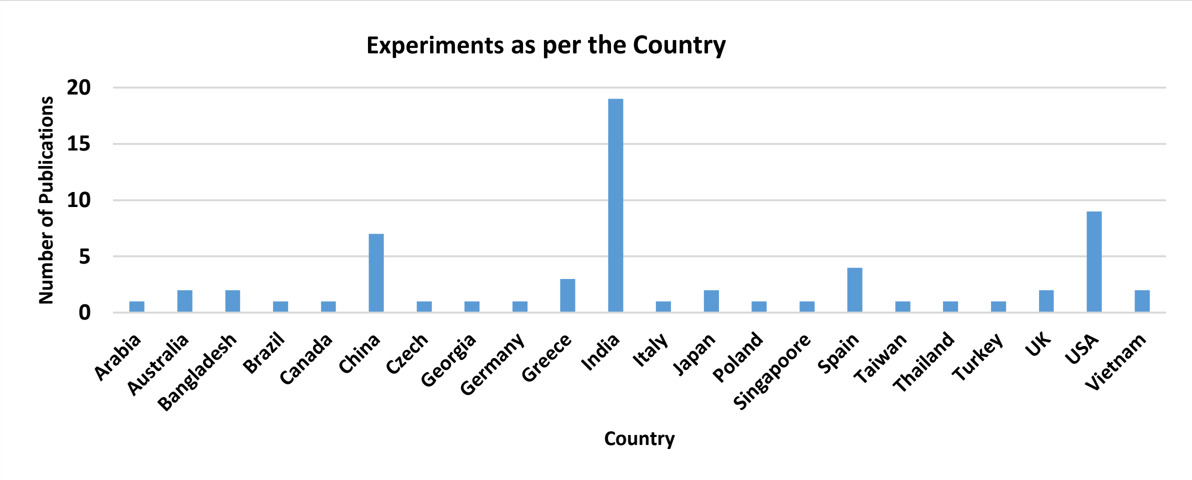
(Mehta et al., 2013)	Web mining	WDC	Social media data	India
(Aiello et al., 2013)	Topic Detection and Tracking (TDT) detect real-world events	LDA	Twitter data	USA
(Adedoyin-Olowe et al., 2013)	Temporal Analysis of Evolving Concepts, Association rule mining	TRCM	Twitter data	UK
(Roohi, 2013)	A FRAMEWORK FOR ANALYSIS	FL	Social media data	India
(De & Kopparapu, 2013)	Harness Ideas from an Ideas Portal	HAC	Company ideas	India
(Archambault et al., 2013)	Exploring Topic and Sentiment in Microblogging Data	HAC	Twitter (us cities and election 2012 dataset)	USA
(Memon et al., 2015)	Travel Recommendation	DBA	Flicker data	China
(Haque & Rahman, 2014)	Sentiment Analysis	FL, ANN	Twitter data	Bangladesh
(Kalyani et al., 2015)	Dynamic Sentiment Detection, Users mood swing analyzer	K-means	Facebook data	India
(Sharma, 2014)	Clustering in Data Mining: A Brief Review	ANN, FL	Social media data	India
(Science & Engineering, 2016)	Web Usage Mining for Web Personalization	FL and K-Mean	Server, Proxy, Client Log files	India
(Singh, 2017)	Educational data mining	K-means Clustering	University student's data	India
(Soni et al., 2016)	Association Rule Mining	K-means Clustering	University student's data	India
(Phu et al., 2017)	English sentiment classification in a distributed environment	FL in HADOOP Map Reduce in Cloudera	Facebook data	Vietnam
(Huang et al., 2017)	Topic Sentiment Analysis in Microblogging	LDA	Weibo data	China
(García-Pablos et al., 2018)	Aspect Based Sentiment Analysis	LDA	SemEval-2016	Spain
(Chandra Pandey et al., 2017)	Twitter sentiment analysis	K-means and cuckoo search method (CSK)	Twitter data	India
(Boonjing & Pimchangthong, 2018)	Positive Customer Reaction to Advertising in Social Media	K-means	Social media data	Thailand
(H. Kumar & Kaur, 2018)	Clustering and Ranking of Social Media Users	XMeans, Expectation Maximization and HAC	Facebook data	India
(Asimuzzaman et al., 2018)	Sentiment Analysis of Bangla Microblogs	ANN, FL	Bangla tweets	Bangladesh
(Boonjing et al., 2018)	semantic-based clustering algorithm	K-means clustering	Twitter data	Korea
(Hoffman et al., 2018)	community detection in networks	Cohen's κ, a similarity measure	Social media data	USA
(J. Singh et al., 2019)	Topic modeling algorithms for Event Detection	LDA, Hierarchical clustering, K-Means	Twitter data	India
(Patil & Algur, 2019)	Trend Detection in Clustering	K-Means Clustering	Twitter data	India

(Jelodar et al., 2020)	latent-topic detection and sentiment analysis	fuzzy lattice reasoning and classification	Youtube comment data	China

Source:

Regional domination of data mining in social media research studies during last decade (RQ6): When considering applied literature (selected 125 research papers), it was identified that India had been the leading country (19 research works) that had carried out research studies on data mining techniques in social media analysis, during the last decade. United States of America (USA) had become the second-highest contributor in the domain and then China, Spain, and Greece indicated a considerable role.

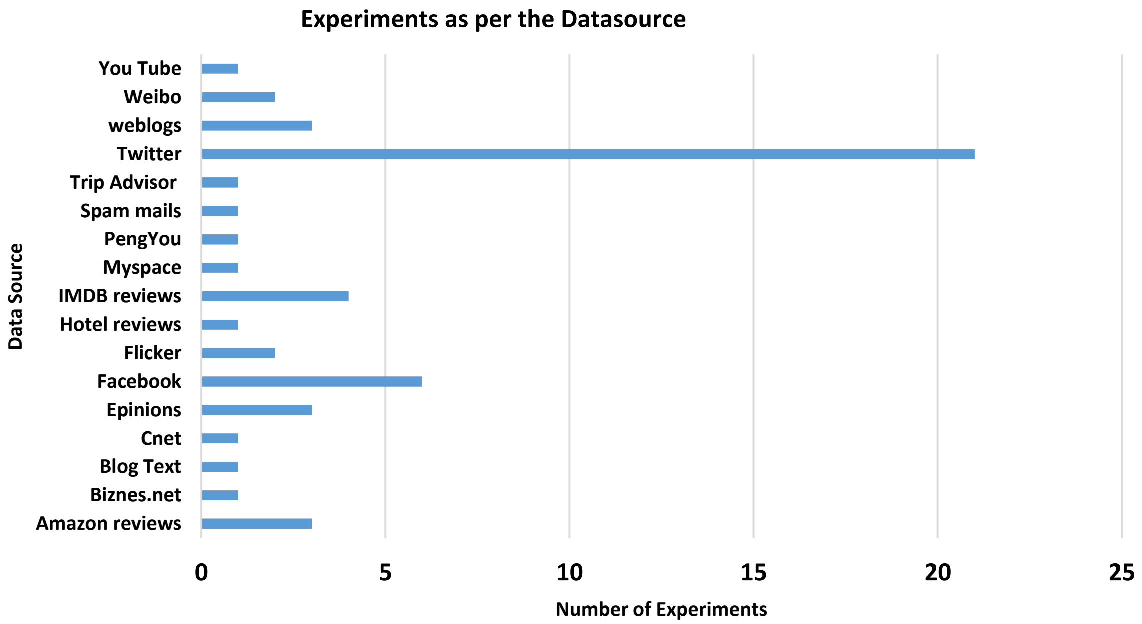
Figure 07: Research contribution according to the country



Source:

Highly discussed social media platform among research community (RQ7): According to the underneath bar chart, it reveals that most of the researchers have been focused on Twitter data set analysis surpassing all other social media data. The reason behind that may be the unlimited facilities given by Twitter API and the invented accessible methods of data extraction, preprocessing, analyzing, and visualization. Facebook data analytics has become the second focused area and then Internet Movie Database (IMDB), Amazon reviews, weblogs, Epinions, and Weibo get ranked. Rendering the above chart highlights that YouTube data analytics is still in its infancy, and it can be further developed with additional research works in enhanced technologies.

Figure 08: Social media highly discussed among researchers



Source:

Challenges with SMA (RQ8): SMA had to have a plethora of complications related to the nature of social media data, their assortment, and lack of advanced analysis and mining strategies. Some of the key challenges (Katare & Dubey, 2017; A. Kumar et al., 2014; Paidi, 2012) are categorized as follows.

Data Issues

- Social media data are created in enormous amounts and are highly dynamic and complex in their nature. Thus, they cannot be prepared effectively using conventional data processing applications or database management tools as well as desktop statistics and visualization packages.
- Data quality is influenced by noisy data, messy data, missing values, inaccurate values, and inadequate data size.
- To handle vast datasets, it needs to have distributed or parallel approaches.
- Efficient data cleaning and data analysis methods are needed to handle noise.
- Absence of data or inaccessible to data.

Technical Issues

- Due to Social media data filling in at a remarkable rate, it needs to have persistent refreshing of models made to deal with information volume and speed.
- It is very expensive to purchase and keep up refined programming software, servers, and storage devices that handle enormous measures of data.
- It needs to have a wide scope of data analysis tools to mine various types of information in data sets
- It is unreasonable to anticipate that one framework should mine a wide range of information, given the scope of data types and different objectives of data mining. Thus, explicit data mining frameworks should be made for mining explicit sorts of information.

Security Issues

- The proliferation of security and privacy concerns by people, associations, and governments.
- Safety of data security, integrity, and privacy.

Results Presentation and Visualization issues

- The demonstration of revealed information is a significant errand in the data mining process. It should be communicated in visual portrayals, or other expressive structures like trees, tables, rules, graphs, charts, crosstabs, matrices, or curves. Henceforth humans can simply recognize and put on that knowledge.
- Data mining systems can reveal many patterns. A large number of the patterns discovered may be unexciting to the given user, either because they speak to basic information or need curiosity.
- Many patterns in DM might be the aftereffect of irregular fluctuations; so numerous such patterns might be pointless.

CONCLUSION

In this information era, Data Mining is the most wanted one to extract valuable information, pattern, and correlation, the trend from the huge volume of data that is unstructured and in different formats. Managing and analyzing this kind of data is a great challenge for researchers. Subsequently, a great number of efforts have been invested in the area of Data Mining. Several extensions have been engaged with data mining for mining massive, unpredictable, and fast-spreading behavior of data. The extracted information is applied in many fields like education, medicine, agriculture, banking, sales, marketing, and social media to predict the future event or find the value of the target variable. Every single field needs different kinds of knowledge and utilizes a distinctive data repository.

This paper provides ideas of Data Mining including the evolution of data mining, its applications, types of data mining, algorithms and techniques, and challenges with SMA. The selected papers cover a wide number of fundamental research domains and areas of available data mining techniques. The survey discusses the most frequently used social media mining algorithms such as Classification and Clustering. It has been pointed out the popularity of clustering, Network graph mining, and hybrid approaches among the researchers have taken a growing trend. although the techniques and algorithms used to analyze the user-generated data (comments, blogs) are rapidly progressing, many of the issues in this field of study still need further work and remain unsolved. Because of the uniqueness of social media data – speed, size, dynamism, unstructured, messy, heterogeneous, and so on, researchers are welcome to accomplish more exploration of existing and forthcoming technologies.

REFERENCES

Adedoyin-Olowe, M., Gaber, M. M., & Stahl, F. (2013). *A Survey of Data Mining Techniques for Social Media Analysis*. http://arxiv.org/abs/1312.4617

Adedoyin-Olowe, M., Gaber, M. M., & Stahl, F. (2013). TRCM: A methodology for temporal analysis of evolving concepts in Twitter. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *7895 LNAI* (PART 2), 135–145. https://doi.org/10.1007/978-3-642-38610-7_13

Aiello, L. M., Petkos, G., Martin, C., Corney, D., Papadopoulos, S., Skraba, R., Goker, A., Kompatsiaris, I., & Jaimes, A. (2013). Sensing trending topics on Twitter. *IEEE Transactions on Multimedia*, *15*(6), 1268–1282. https://doi.org/10.1109/TMM.2013.2265080

Alfaro, C., Cano-Montero, J., Gómez, J., Moguerza, J. M., & Ortega, F. (2016). A multi-stage method for content classification and opinion mining on weblog comments. *Annals of Operations Research*, *236*(1), 197–213. https://doi.org/10.1007/s10479-013-1449-6

Anjaria, M., & Guddeti, R. M. R. (2014). A novel sentiment analysis of social networks using supervised learning. *Social Network Analysis and Mining*, *4*(1), 1–15. https://doi.org/10.1007/s13278-014-0181-9

Archambault, D., Greene, D., & Cunningham, P. (2013). TwitterCrowds: Techniques for Exploring Topic and Sentiment in Microblogging Data. 1–19. http://arxiv.org/abs/1306.3839

Asimuzzaman, M., Nath, P. D., Hossain, F., Hossain, A., & Rahman, R. M. (2018). Sentiment analysis of Bangla microblogs using the adaptive neuro-fuzzy system. ICNC-FSKD 2017 - 13th International Conference on Natural Computation, Fuzzy Systems, and Knowledge Discovery, 1631–1638. https://doi.org/10.1109/FSKD.2017.8393010

B, S. P. (2011). Improved BSP Clustering Algorithm for Social Network Analysis. Bonfring International Journal of Software Engineering and Soft Computing, 1(1), 15–20. https://doi.org/10.9756/bijsesc.1003

Balan, S., & Rege, J. (2017). Mining for social media: Usage patterns of small businesses. Business Systems Research, 8(1), 43–50. https://doi.org/10.1515/bsrj-2017-0004

Baskar, S. S., Arockiam, L., Arul Kumar, V., & Jeyasimman, L. (2010). Brief Survey of Application of Data Mining Techniques to Agriculture. Agricultural Journal, 5(2), 116–118. https://doi.org/10.3923/aj.2010.116.118

Becker, H., Naaman, M., & Gravano, L. (2011). Beyond trending topics: Real-world event identification on Twitter. *Icwsm*, 1–17. http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/viewPDFInterstitial/2745/320

Besiashvili, G., Bliadze, T., & Kochladze, Z. (2018). *Application of Adaptive Neural Networks for the Filtration of Spam*. *36*, 42–32. https://doi.org/10.29007/t3rl

Boonjing, V., & Pimchangthong, D. (2018). Data mining for positive customer reaction to advertising in social media. *Lecture Notes in Business Information Processing*, *311*, 83–95. https://doi.org/10.1007/978-3-319-77721-4_5

Bozkir, A. S., Güzin Mazman, S., & Akçapinar Sezer, E. (2010). Identification of user patterns in social networks by data mining techniques: Facebook case. *Communications in Computer and Information Science*, *96 CCIS* (3), 145–153.

Castillo, C., Mendoza, M., & Poblete, B. (2011). Information credibility on Twitter. *Proceedings of the 20th International Conference Companion on World Wide Web, WWW 2011*, *May 2014*, 675–684. https://doi.org/10.1145/1963405.1963500

Chandra Pandey, A., Singh Rajpoot, D., & Saraswat, M. (2017). Twitter sentiment analysis using the hybrid cuckoo search method. *Information Processing and Management*, *53*(4), 764–779. https://doi.org/10.1016/j.ipm.2017.02.004

Chen, X., Vorvoreanu, M., & Madhavan, K. P. C. (2014). Mining social media data for understanding students' learning experiences. *IEEE Transactions on Learning Technologies*, *7*(3), 246–259. https://doi.org/10.1109/TLT.2013.2296520

Choudhery, D., & Leung, C. K. (2017). Social media mining: Prediction of box office revenue. *ACM International Conference Proceeding Series*, *Part F129476*, 20–29. https://doi.org/10.1145/3105831.3105854

Conovor. (2011). *Conover_2011_Predicting_Political_Alignment_of_Twitter_Users.pdf*.

Dalal, M. K., & Zaveri, M. A. (2013). Automatic Classification of Unstructured Blog Text. *Journal of Intelligent Learning Systems and Applications*, *05*(02), 108–114. https://doi.org/10.4236/jilsa.2013.52012

De, A., & Kopparapu, S. K. (2013). Unsupervised clustering technique to harness ideas from an Ideas Portal. *Proceedings of the 2013 International Conference on Advances in Computing, Communications and Informatics, ICACCI 2013*, 1563–1568. https://doi.org/10.1109/ICACCI.2013.6637413

Dey, L., Chakraborty, S., Biswas, A., Bose, B., & Tiwari, S. (2016). Sentiment Analysis of Review Datasets Using Naïve Bayes 'and K-NN Classifier. *International Journal of Information Engineering and Electronic Business*, *8*(4), 54–62. https://doi.org/10.5815/ijieeb.2016.04.07

Duwairi, R. M., & Qarqaz, I. (2014). Arabic sentiment analysis using supervised classification. *Proceedings - 2014 International Conference on Future Internet of Things and Cloud, FiCloud 2014*, 579–583. https://doi.org/10.1109/FiCloud.2014.100

Fang, L., & LeFevre, K. (2010). Privacy wizards for social networking sites. *Proceedings of the 19th International Conference on World Wide Web, WWW '10*, 351–360. https://doi.org/10.1145/1772690.1772727

Fujisaka, T., & Lee, R. (2010). Discovery of User Behavior Patterns from. *ICUIMC '10 - Proceedings of the 4th International Conference on Uniquitous Information Management and Communication*. http://dl.acm.org/citation.cfm?id=2108660

Gao, K., Xu, H., & Wang, J. (2015). A rule-based approach to emotion cause detection for Chinese micro-blogs. *Expert Systems with Applications*, *42*(9), 4517–4528. https://doi.org/10.1016/j.eswa.2015.01.064

García-Pablos, A., Cuadros, M., & Rigau, G. (2018). W2VLDA: Almost unsupervised system for Aspect Based Sentiment Analysis. *Expert Systems with Applications*, *91*, 127–137. https://doi.org/10.1016/j.eswa.2017.08.049

Haque, A., & Rahman, T. (2014). Sentiment Analysis by Using Fuzzy Logic. *International Journal of Computer Science, Engineering and Information Technology*, *4*(1), 33–48. https://doi.org/10.5121/ijcseit.2014.4104

Hoffman, M., Steinley, D., Gates, K. M., Prinstein, M. J., & Brusco, M. J. (2018). Detecting Clusters/Communities in Social Networks. *Multivariate Behavioral Research*, *53*(1), 57–73. https://doi.org/10.1080/00273171.2017.1391682

Hourcade, J. P., Mascher, S. L., Wu, D., Pantoja, L., Rosenthal, S., Jagiello, Z. A., Dyderski, M. K., Dylewski, Ł., Kalra, G. S., Kathuria, R. S., Kumar, A., Park, D., An, J., Jansen, B. J., Paek, H. J., Kim, K., Hove, T., Rothe, S., Höllerer, T., … Telg, R. W. (2018). Large-scale video classification with convolutional neural networks. Journal of Applied Communications, 25(2), 411–416. https://doi.org/10.14400/JDC.2020.18.5.411

Huang, F., Zhang, S., Zhang, J., & Yu, G. (2017). Multimodal learning for topic sentiment analysis in microblogging. Neurocomputing, 253, 144–153. https://doi.org/10.1016/j.neucom.2016.10.086

Injadat, M. N., Salo, F., & Nassif, A. B. (2016). Data mining techniques in social media: A survey. Neurocomputing, 214, 654–670. https://doi.org/10.1016/j.neucom.2016.06.045

Jacob, J., Jha, K., Kotak, P., & Puthran, S. (2016). Educational Data Mining techniques and their applications. *Proceedings of the 2015 International Conference on Green Computing and Internet of Things, ICGCIoT 2015*, 1344–1348. https://doi.org/10.1109/ICGCIoT.2015.7380675

Jagiello, Z. A., Dyderski, M. K., & Dylewski, Ł. (2019). What can we learn about the behavior of red and grey squirrels from YouTube? *Ecological Informatics*, *51*(February), 52–60. https://doi.org/10.1016/j.ecoinf.2019.02.006

Jelodar, H., Wang, Y., Rabbani, M., Ahmadi, S. B. B., Boukela, L., Zhao, R., & Larik, R. S. A. (2021). An NLP framework based on meaningful latent-topic detection and sentiment analysis via fuzzy lattice reasoning on youtube comments. *Multimedia Tools and Applications*, *80*(3), 4155–4181.

Jeyapriya, A., & Selvi, C. S. K. (2015). Extracting aspects and mining opinions in product reviews using a supervised learning algorithm. *2nd International Conference on Electronics and Communication Systems, ICECS 2015*, 548–552.

Kalra, G. S., Kathuria, R. S., & Kumar, A. (2019). YouTube Video Classification based on Title and Description Text. *Proceedings - 2019 International Conference on Computing, Communication, and Intelligent Systems, ICCCIS 2019*, *2019-Janua*, 74–79. https://doi.org/10.1109/ICCCIS48478.2019.8974514

Kalyani, Gupta, E., Rathee, G., Kumar, P., & Chauhan, D. S. (2015). Mood Swing Analyzer: A Dynamic Sentiment Detection Approach. *Proceedings of the National Academy of Sciences India Section A - Physical Sciences*, *85*(1), 149–157. https://doi.org/10.1007/s40010-014-0169-x

Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of Social Media. *Business Horizons*, *53*(1), 59–68. https://doi.org/10.1016/j.bushor.2009.09.003

Katare, A., & Dubey, S. (2017). A Comparative Study of Classification Algorithms in EDM using 2 Level Classification for Predicting Student's Performance. *International Journal of Computer Applications*, *165*(9), 35–40. https://doi.org/10.5120/ijca2017914023

Kumar, A., Tyagi, A. K., & Tyagi, S. K. (2014). Data Mining: Various Issues and Challenges for Future A Short discussion on Data Mining issues for future work. *International Conference on Advanced Developments in Engineering and Technology*, *4*(1), 1–8.

Kumar, H., & Kaur, H. (2018). Clustering and ranking social media users based on temporal analysis. *2017 International Conference on Infocom Technologies and Unmanned Systems: Trends and Future Directions, ICTUS 2017*, *2018-January*, 271–275. https://doi.org/10.1109/ICTUS.2017.8286018

Kumar, R., & Sharma, M. K. (2016). Advanced Neuro-Fuzzy Approach for Social Media Mining Methods using Cloud. *International Journal of Computer Applications*, *137*(10), 56–58. https://doi.org/10.5120/ijca2016908927

Le, T., Pardo, P., & Claster, W. (2016). Application of artificial neural network in social media data analysis: A case of lodging business in Philadelphia. *Studies in Computational Intelligence*, *628*(February), 369–376. https://doi.org/10.1007/978-3-319-28495-8_16

Lee, J., Kuo, Y. F., Lin, Y. L., & Goodwin, J. S. (2015). The combined effect of the electronic health record and hospitalist care on length of stay. *The American Journal of Managed Care*, *21*(3), e215–e221.

Li, G., & Liu, F. (2012). Application of a clustering method on sentiment analysis. *Journal of Information Science*, *38*(2), 127–139. https://doi.org/10.1177/0165551511432670

Liang, P. W., & Dai, B. R. (2013). Opinion mining on social media data. Proceedings - IEEE International Conference on Mobile Data Management, 2, 91–96. https://doi.org/10.1109/MDM.2013.73

Maheshwari, N., Pandey, N. K., & Agarwal, P. (2016). Market Basket Analysis using Association Rule Learning. International Journal of Computer Applications, 975–8887. https://pdfs.semanticscholar.org/320f/b44e15c5cf1dbb03f085f121d453e8fc3932.pdf

Medar, R., Rajpurohit, V. S., & Shweta, S. (2019). Crop Yield Prediction using Machine Learning Techniques. 2019 IEEE 5th International Conference for Convergence in Technology, I2CT 2019, 3(5), 1093–1097. https://doi.org/10.1109/I2CT45611.2019.9033611

Mehta, N., Kathuria, M., Singh, M., & Cse, A. P. (2013). Comparison of Conventional & Fuzzy Clustering Techniques: A Survey. International Journal of Advanced Research in Computer and Communication Engineering, 2(1). www.ijarcce.com

Memon, I., Chen, L., Majid, A., Lv, M., Hussain, I., & Chen, G. (2015). Travel Recommendation Using Geo-Tagged Photos in Social Media for Tourists. Wireless Personal Communications, 80(4), 1347–1362. https://doi.org/10.1007/s11277-014-2082-7

Moin, K. I., & Ahmed, Q. B. (2012). Use of Data Mining in Banking. International Journal of Engineering Research and Applications (IJERA), 2(2), 738–742. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.416.7821&rep=rep1&type=pdf

Moghaddasi, Hamid, Hoseini, Azamossadat, Asadi, Farkhondeh, Jahanbakhsh, M. (2012). Application of Data Mining in Health. Health Information Management, 9(2), 297-304.

Moraes, R., Valiati, J. F., & Gavião Neto, W. P. (2013). Document-level sentiment classification: An empirical comparison between SVM and ANN. Expert Systems with Applications, 40(2), 621–633. https://doi.org/10.1016/j.eswa.2012.07.059

Moussa, S., Fawzy, D., & Badr, N. (2016). The Evolution of Data Mining Techniques to Big Data Analytics : An Extensive The Evolution of Data Mining Techniques to Big Data Analytics : An Extensive Study with Application to Renewable Energy Data Analytics. August.

Nandal, R., Dhamija, P., & Sehrawat, H. (2017). a Review Paper on Prediction Analysis: Predicting Student Result based on Past Result. International Journal of Engineering and Technology, 9(2), 1204–1208. https://doi.org/10.21817/ijet/2017/v9i2/170902226

Nandi, G., & Das, A. (2013). A survey on using data mining techniques for online social network analysis. Int. J. Comput. Sci. Issues (IJCSI), 10(6), 162–167.

Paidi, A. N. (2012). Data Mining: Future Trends and Applications. International Journal of Modern Engineering Research, 2(6), 4657–4663.

Pal, J. K. (2011). Usefulness and applications of data mining in extracting information from different perspectives. Annals of Library and Information Studies, 58(1), 7–16.

Papadopoulos, S., Kompatsiaris, Y., Vakali, A., & Spyridonos, P. (2012). Community detection in social media performance and application considerations. Data Mining and Knowledge Discovery, 24(3), 515–554. https://doi.org/10.1007/s10618-011-0224-z

Patil, R. H., & Algur, S. P. (2019). Classification connection of Twitter data using k-means clustering. International Journal of Innovative Technology and Exploring Engineering, 8(6 Special Issue 4), 14–22. https://doi.org/10.35940/ijitee.F1004.0486S419

Pham, D. H., & Le, A. C. (2018). Learning multiple layers of knowledge representation for aspect-based sentiment analysis. Data and Knowledge Engineering, 114, 26–39. https://doi.org/10.1016/j.datak.2017.06.001

Pham, M. C., Cao, Y., Klamma, R., & Jarke, M. (2011). A clustering approach for collaborative filtering recommendation using social network analysis. Journal of Universal Computer Science, 17(4), 583–604. https://doi.org/10.3217/jucs-017-04-0583

Phu, V. N., Dat, N. D., Ngoc Tran, V. T., Ngoc Chau, V. T., & Nguyen, T. A. (2017). Fuzzy C-means for English sentiment classification in a distributed system. Applied Intelligence, 46(3), 717–738. https://doi.org/10.1007/s10489-016-0858-z

Priyadarshini, R. (2010). Functional Analysis of Artificial Neural Network for Dataset Classification. Special Issue of IJCCT, 1(2), 3–5.

Pushpam, C. A., & Jayanthi, J. G. (2017). Overview on Data Mining in Social Media. International Journal of Computer Sciences and Engineering, 5(11), 147–157. https://doi.org/10.26438/ijcse/v5i11.147157

Quercia, D., Ellis, J., Capra, L., & Crowcroft, J. (2012). Tracking "gross community happiness" from tweets. Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW, 965–968. https://doi.org/10.1145/2145204.2145347

Ramesh, D., & Vardhan, B. V. (2013). Data mining techniques and applications to agricultural yield data. International journal of advanced research in computer and communication engineering, 2(9), 3477-3480.

Ren, F., & Kang, X. (2013). Employing hierarchical Bayesian networks in simple and complex emotion topic analysis. *Computer Speech and Language*, *27*(4), 943–968. https://doi.org/10.1016/j.csl.2012.07.012

Roohi, F. (2013). Neuro-fuzzy approach to data clustering: a framework for analysis. *Neuro-Fuzzy Approach to Data Clustering: A Framework for Analysis*, *9*(9), 183–193.

Rushdi Saleh, M., Martín-Valdivia, M. T., Montejo-Ráez, A., & Ureña-López, L. A. (2011). Experiments with SVM to classify opinions in different domains. *Expert Systems with Applications*, *38*(12), 14799–14804. https://doi.org/10.1016/j.eswa.2011.05.070

S.Neelamegam, E. R. (2015). 59 An Overview of Classification Algorithm in Data mining. *International Journal of Advanced Research in Computer and Communication Engineering*, *4*(12), 255–257.

Sangar, A. B., Khaze, S. R., & Ebrahimi, L. (2013). Participation Anticipating in Elections Using Data Mining Methods. *International Journal on Cybernetics & Informatics*, *2*(2), 47–60. https://doi.org/10.5121/ijci.2013.2205

Sarker, A., O'Connor, K., Ginn, R., Scotch, M., Smith, K., Malone, D., & Gonzalez, G. (2016). Social media mining for toxicovigilance: Automatic monitoring of prescription medication abuse from Twitter. *Drug Safety*, *39*(3), 231–240. https://doi.org/10.1007/s40264-015-0379-

Science, C., & Engineering, S. (2016). *A Review on Web Usage Mining for Web Personalization Using Clustering Techniques*. *6*(5), 455–459.

Severyn, A., Moschitti, A., Uryupina, O., Plank, B., & Filippova, K. (2016). Multi-lingual opinion mining on YouTube. *Information Processing and Management*, *52*(1), 46–60. https://doi.org/10.1016/j.ipm.2015.03.002

Shankar, R., Shilpa, K. M., Patil, S., & Swamy, S. (2016). *A Survey on Sentimental Analysis in Different Indian Dialects*. *5*(4). https://doi.org/10.17148/IJARCCE.2016.54262

Sharma, M. (2014). Clustering in Data Mining A Brief Review. *International Journal of Core Engineering and Management*, *1*(5), 117–126.

Sharma, P. (2014). *Use of Data Mining in Various Field : A Survey Paper*. *16*(3), 18–21.

Singh, A. (2017). Mining of Social Media data of University students. Education and Information Technologies, 22(4), 1515–1526. https://doi.org/10.1007/s10639-016-9501-1

Singh, J., Kaur, I., & Singh, A. K. (2019). Event detection from Twitter data. 2019 4th International Conference on Information Systems and Computer Networks, ISCON 2019, 793–798. https://doi.org/10.1109/ISCON47742.2019.9036286

Singh, S. P., & Campus, N. (2013). It is challenging to make an investment decision for the stock available in the market. There are various models, machine learning and data mining techniques proposed, but the capability and intelligence of financial investors are what is required [11]. 3(2), 177–182.

Smita, S., & Sharma, P. (2014). Use of Data Mining in Various Field: A Survey Paper. IOSR Journal of Computer Engineering, 16(3), 18–21. https://doi.org/10.9790/0661-16351821

Soni, H. K., Sharma, S., & Mishra, P. K. (2016). Association Rule Mining: a Data Profiling and Prospective Approach. International Journal of Current Engineering and Scientific Research (Ijcesr), 3(1), 57–60. https://doi.org/DOI:10.21276/ijcesr

Soni, N., & Gandhi, C. (2010). Application of data mining to health care. *International Journal of Computer Science and its Applications*.

Surma, J., & Furmanek, A. (2010). Improving marketing response by data mining in the social network. *Proceedings - 2010 International Conference on Advances in Social Network Analysis and Mining, ASONAM 2010*, 446–451. https://doi.org/10.1109/ASONAM.2010.21

Tang, D., Qin, B., Liu, T., & Yang, Y. (2015). User modeling with neural network for review rating prediction. *IJCAI International Joint Conference on Artificial Intelligence*, *2015-January* (Ijcai), 1340–1346.

Tiwari, D., & Kumar, M. (2020). Social Media Data Mining Techniques: A Survey. In *Advances in Intelligent Systems and Computing* (Vol. 933). Springer Singapore. https://doi.org/10.1007/978-981-13-7166-0_18

Tripathy, A., Agrawal, A., & Rath, S. K. (2016). Classification of sentiment reviews using n-gram machine learning approach. *Expert Systems with Applications*, *57*, 117–126.

Tsagkalidou, K., Koutsonikola, V., Vakali, A., & Kafetsios, K. (2011). Emotional aware clustering on micro-blogging sources. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *6974 LNCS* (PART 1), 387–396. https://doi.org/10.1007/978-3-642-24600-5_42

Weng, J., Yao, Y., Leonardi, E., & Lee, B. S. (2011). Event detection in Twitter. *HP Laboratories Technical Report*, *98*, 1–21.

Yang, C. C., & Ng, T. D. (2011). Analyzing and visualizing web opinion development and social interactions with density-based clustering. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, *41*(6), 1144–1155. https://doi.org/10.1109/TSMCA.2011.2113334

Yin, J., Karimi, S., Lampert, A., Cameron, M., Robinson, B., & Power, R. (2015). Using social media to enhance emergency awareness: *IJCAI International Joint Conference on Artificial Intelligence*, *2015-January* (July 2015), 4234–4239.

Zamani Alavijeh, Z., Rohilla, P., Sharma, O., Radhakrishnan, S., Joshi, G., Sonawane, S., Khanday, H. A., Hashmy, R., April, M., Kateb, F., Kalita, J., Besiashvili, G., Bliadze, T., Kochladze, Z., Mewari, R., Singh, A., &

Srivastava, A. (2015). A Survey on Text Filtering in Online Social Networks. *International Journal of Computer Applications*, *4*(6), 39–44. https://doi.org/10.5120/20753-3149

Zhai, Z., Liu, B., Xu, H., & Jia, P. (2011). Clustering product features for opinion mining. Proceedings of the 4th ACM International Conference on Web Search and Data Mining, WSDM 2011, 347–354. https://doi.org/10.1145/1935826.1935884