International Journal of Applied Engineering and Technology ISSN: 2277-212X (Online)
An Open Access, Online International Journal Available at http://www.cibtech.org/jet.htm
2016 Vol. 6 (4) October-December, pp.1-8/Dawande and Thakre
Review Article

MINING ONLINE SOCIAL MEDIA DATA: A SURVEY

***Dipali R. Dawande and K.S. Thakre**
Department of Information Technology, Sinhgad College of Engineering Pune
**Author for Correspondence*

ABSTRACT
Mining online social media is the process of representing, analyzing, and extracting actionable patterns and trends from raw social media data. Social media is a very popular way of expressing opinions and interacting with other people in the online world. People are becoming more interested in and relying on social network for information, news and opinion of other users on diverse subject matters. Mental illness has a deep impact on people, families, and by extension, society as a whole. Mining online social media data can allow analyzing person with mental disorders which communicate with others sufferers via online communities, providing a crucial resource for studies on textual signs of psychological health problems. This survey defines various techniques for mining online social data for detecting useful knowledge from massive datasets like trends, patterns and rules. Information retrieval, statistical modeling and machine learning are the areas where data mining techniques are used. The paper then reviews the related work and methods such as data mining, machine learning, social network analysis. The paper summarize the prominent issues with the current research and highlight future directions on learning constantly extending knowledge from social media data.

Keywords: *Mental Illness, Information Retrieval*

INTRODUCTION
Social media is designed as a group of internet based applications that build on the ideological and technological foundations of web 2.0 and that allow the creation and exchange of user generated content. Everyday vast amount of user generated content are created on social media sites i.e. facebook, Twitter, Google. Now, days the world is being a small place owing to the substantial influence of social media. The people are connected from various parts of the world, ages, and nationalities and it allows them to share their opinions, experiences, feelings, hobbies, pictures, and videos (Ferrara *et al.*, 2013). Social media has opened the door for private and public organizations from all domains to promote, benefit, analyze, learn, and improve their organizations based on the data provided in social media. Thus, social media is significant for academia and industry and quite perceptible in the amount of research done by these two sectors, seeking answers to pivotal questions. Most research on social network mining focuses on discovering the treasure of knowledge behind the data for improving people's life. In contrast, much less attention has been drawn to remedy the problems incurred from various social network applications. Social networks are important sources of online interactions and contents sharing. Social network platforms enable fast information exchange between users regardless of the location. The activities on social network are now followed by many organizations, individuals and even government of countries (Ferrara *et al.*, 2013).

The growing popularity of Online Social Networks (OSN) had generated a large amount of communication records that could be easily accessed and analyzed to study human social behavior. This represented a unique opportunity to understand properties of social networks that were impossible to assess in the past. It also allows the effective collection of large-scale data which gives rise to major computational challenges. However, for users to discover valuable, accurate and useful knowledge from social network data, the application of efficient data mining techniques has made it possible.

In the review of this survey, the research has been initiated with the security issues of the social networks in order to prevent threats & attacks to social networks (Shuai *et al.*, 2016). Then, it extends with the research that describes the implicit social graph which is formed by users' interactions with contacts and groups of contacts in which users explicitly add other individuals as their "friends" (Ferrara *et al.*, 2013). The next research on social networking and information seeking had explored whether social networks

moderate physiological indicators of emotion. As the use and advancement in the social network increased, it became essential to analyze and observe the user activity by various methods hence, on the concept of user level sentiment analysis come to light for research. Sentiment analysis has become one of the key emerging technologies in the effort to help people navigate the large amount of user generated content available online. Internet addiction could be conceptualized as a maladaptive pattern of Internet use behavior which is associated with several psychological and social problems, to solve such issues the research has been cause on psychometric properties of the internet addiction. The research has seen the personality and the individual differences i.e. self presentation and belonging on social network. The continuous use of social network and its belongings leads in order to detect mental health disorders for a particular individual, a variety of predictive models utilizing heterogeneous medical data have been emerged to study the concept, hence, this survey discuss various data mining techniques that can be used in various real time data mining applications.

Online Social Media Issues

- Social networking sites are virtual communities where user can create individual public profiles, interact with real life friends, and meet other people based on shared interests.
- The findings indicate that SNS are predominantly used for social purposes, mostly related to the maintenance of established offline networks.
- On the internet people engage in a variety of activities some of which may be potentially to be addictive.
- Rather than becoming addicted to the medium, some users may develop an addiction to specific activities they carry out online.
- There are five different types of internet addiction i.e. computer addiction, information overload, net compulsion, cyber-sexual addiction and cyber-relationship addiction that need to be control.

Types of Data Mining Techniques

The data mining techniques that had been applied by researchers in the area of social media are listed below.

- AdaBoost
- Artificial Neural Network (ANN)
- Apriori
- Bayesian Networks (BN)
- Decision Trees (DT)
- Density Based Algorithm (DBA)
- Fuzzy
- Hierarchical Clustering (HC)
- K-Means
- k-nearest Neighbors (k-NN)
- Markov
- Support Vector Machine (SVM)

Adaboost: AdaBoost, short for "Adaptive Boosting", is a machine learning meta-algorithm. It can be used in conjunction with many other types of learning algorithms to improve their performance. AdaBoost is adaptive in the sense that subsequent weak learners are tweaked in favor of those instances misclassified by previous classifiers. AdaBoost is sensitive to noisy data and outliers. In some problems it can be less susceptible to the over fitting problem than other learning algorithms. AdaBoost (with decision trees as the weak learners) is often referred to as the best out-of-the-box classifier (Shuai *et al.*, 2016; Ferrara *et al.*, 2013). When used with decision tree learning, information gathered at each stage of the AdaBoost algorithm about the relative 'hardness' of each training sample is fed into the tree growing algorithm such that later trees tend to focus on harder-to-classify examples.

Artificial Neural Network (ANN): Neural Networks are a computational approach which is based on a large collection of neural units loosely modeling the way the brain solves problems with large clusters of biological neurons connected by axons. Each neural unit is connected with many others, and links can be

enforcing or inhibitory in their effect on the activation state of connected neural units. Each individual neural unit may have a summation function which combines the values of all its inputs together. There may be a threshold function or limiting function on each connection and on the unit itself such that it must surpass it before it can propagate to other neurons. These systems are self-learning and trained rather than explicitly programmed and excel in areas where the solution or feature detection is difficult to express in a traditional computer program.

Apriori: Apriori is an algorithm for frequent item set mining and association rule learning over transactional databases. It proceeds by identifying the frequent individual items in the database and extending them to larger and larger item sets as long as those item sets appear sufficiently often in the database. The frequent item sets determined by Apriori can be used to determine association rules which highlight general trends in the database: this has applications in domains such as market basket analysis.

Bayesian Networks: A Bayesian network, Bayes network, belief network, Bayes(ian) model or probabilistic directed acyclic graphical model is a probabilistic graphical model (a type of statistical model) that represents a set of random variables and their conditional dependencies via a directed acyclic graph (DAG). For example, a Bayesian network could represent the probabilistic relationships between diseases and symptoms. Given symptoms, the network can be used to compute the probabilities of the presence of various diseases.

Decision Trees: Decision tree builds classification or regression models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. A decision node (e.g., Outlook) has two or more branches (e.g., Sunny, Overcast and Rainy). Leaf node (e.g., Play) represents a classification or decision. The topmost decision node in a tree which corresponds to the best predictor called root node. Decision trees can handle both categorical and numerical data.

Density Based Algorithm: Density-based spatial clustering of applications with noise (DBSCAN) is a data clustering algorithm. It is a density-based clustering algorithm: given a set of points in some space, it groups together points that are closely packed together (points with many nearby neighbors), marking as outliers points that lie alone in low-density regions (whose nearest neighbors are too far away). DBSCAN is one of the most common clustering algorithms and also most cited in scientific literature.

Fuzzy: Fuzzy logic is an approach to computing based on "degrees of truth" rather than the usual "true or false" (1 or 0) Boolean logic on which the modern computer is based. Fuzzy logic seems closer to the way our brains work. We aggregate data and form a number of partial truths which we aggregate further into higher truths which in turn, when certain thresholds are exceeded, cause certain further results such as motor reaction. A similar kind of process is used in neural networks, expert systems and other artificial intelligence applications. Fuzzy logic is essential to the development of human-like capabilities for AI, sometimes referred to as artificial general intelligence: the representation of generalized human cognitive abilities in software so that, faced with an unfamiliar task, the AI system could find a solution.

K-Means: K-means is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed a priori. The main idea is to define k centroids, one for each cluster. These centroids should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest centroid. When no point is pending, the first step is completed and an early groupage is done. At this point we need to re-calculate k new centroids as barycenters of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new centroid. A loop has been generated. As a result of this loop we may notice that the k centroids change their location step by step until no more changes are done. In other words centroide do not move any more.

k-nearest Neighbors: Clustering and the Nearest Neighbor prediction technique are among the oldest techniques used in data mining. Most people have an intuition that they understand what clustering is -

namely that like records are grouped or clustered together. Nearest neighbor is a prediction technique that is quite similar to clustering - its essence is that in order to predict what a prediction value is in one record look for records with similar predictor values in the historical database and use the prediction value from the record that it "nearest" to the unclassified record. One of the improvements that is usually made to the basic nearest neighbor algorithm is to take a vote from the "K" nearest neighbors rather than just relying on the sole nearest neighbor to the unclassified record.

Markov: A Markov algorithm is a string rewriting system that uses grammar-like rules to operate on strings of symbols. Markov algorithms have been shown to be Turing-complete, which means that they are suitable as a general model of computation and can represent any mathematical expression from its simple notation. Markov algorithms are named after the Soviet mathematician Andrey Markov, Jr.

SVM: In machine learning, support vector machines (SVMs, also support vector networks (Shuai *et al.,* 2016)) are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis.

Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall.

Applications

There are six general domains which applied various techniques in different research areas to mine the flow of big data gathered from social media. The list of these domains follows:

- Business and Management
- Education
- Finance
- Government and Public
- Medical and Health
- Social Networks

There are some active research objectives adopted data mining techniques. The objectives in detail are as follows:

- *Biometric:* Biometrics is automated method of recognizing a person based on the physiological and behavioral characteristics. Data mining techniques are used in the biometric applications, mostly password based applications including in face recognition and fingerprint recognition. It is also for the national security (link analysis) and cyber security.
- *Content Analysis:* It is a research technique used to make replicable and valid inferences by interpreting and coding textual material. By systematically evaluating texts (e.g., documents, oral communication, and graphics), qualitative data can be converted in to qualitative data. The content analysis used in secure political and military intelligence, analysis traits of individuals, infer cultural aspects and change, providing legal and evaluating evidence etc.
- *Cyber Crime:* Cybercrime, or computer crime, is crime that involves a computer and a network. The computer may have been used in the commission of a crime, or it may be the target. Debarati Halder and K. Jaishankar define cybercrimes as: "Offences that are committed against individuals or groups of individuals with a criminal motive to intentionally harm the reputation of the victim or cause physical or mental harm, or loss, to the victim directly or indirectly, using modern telecommunication networks such as Internet (Chat rooms, emails, notice boards and groups) and mobile phones (SMS/MMS)". Such crimes may threaten a nation's security and financial health.
- *Disease Awareness:* As the use of social networking sites has been largely increased the kind of mental health diseases also been noted. The application can make an individual aware about various diseases among which the mental illness through the social media addiction can be easily analyzed by mining online social data (Shuai *et al.,* 2016).

 Geo Locating: The data mining techniques are used in geo locating for the tasks such as Linking Numeric and Geographic Distributions; investigates the link between numeric and geographic distributions of mapped data. Use Map Analysis to Characterize Data Groups; use of "data distance" to derive similarity among the data patterns in a set of map layers. Discover the "Miracle" in Mapping Data Clusters**;** use of "clustering" to identify inherent groupings of similar data patterns.

 Quality Improvement: Quality Improvement is a formal approach to the analysis of performance and systematic efforts to improve it. Many quality improvement (QI) programs including six sigma, design for six sigma, and *kaizen* require collection and analysis of data to solve quality problems. Due to advances in data collection systems and analysis tools, data mining (DM) has widely been applied for QI in manufacturing.

 Risk Management: Risk management is the identification, assessment, and prioritization of risks followed by coordinated and economical application of resources to minimize, monitor, and control the probability and/or impact of unfortunate events. The data mining techniques used in the application of risk management like fraud detection, shopping bag analysis, insurance and text categorization.

Related Work

Social networks, in many forms, have existed since people first began to interact. Indeed, put two or more people together making the foundation of a social network. Therefore, in today's internet-everywhere world, online social networks have become entirely ubiquitous. Social media are computer-mediated technologies that allow individuals, companies, governments, and other organizations to create, share, and view information, ideas, career interests, and other forms of expression via virtual communities and networks.

Luo *et al.,* (2009) has studied threats to social networks and analyze the attackers targets and the methods how attackers perform the attacks. The authors separate social networks into two parts: user and social networking site. Then, they discussed in details the counter measures against the threats to social networks and a security framework of social networks. Here the author has elaborated the security measures and issues using the security framework of social network via classification amongst user, social networking site, and technology underpinning. However, the paper does not ensure the proper techniques to the social network security.

Wise *et al.,* (2010) based on earlier research on social networking and information seeking, the author proposed that Facebook.com use could be conceptualized as serving two primary goals: passive social browsing (i.e., newsfeeds) and extractive social searching (i.e., friends' profiles). This paper explored whether these categories adequately reflect facebook use and whether they moderate physiological indicators of emotion. This paper attempted to explicate the use of online social networks like facebook through the different lenses of uses and gratifications, social information-seeking strategies, and self-report accounts of online social-networking use however, this concept is limited up to the facebook here not for all social networking sites.

Tan *et al.,* (2011) proposed User-Level Sentiment Analysis Incorporating Social Networks. This paper shows that information about social relationships can be used to improve user-level sentiment analysis. According to this approach the connected users may hold similar opinion; therefore, relationship information can complement what author can extract about a user's viewpoints from their utterances. Employing Twitter as a source for experimental data, and working within a semi-supervised framework, The author propose models that are induced either from the Twitter follower/followee network or from the network in Twitter formed by users referring to each other using "@" mentions. The general idea in this paper, to explore social network structures to help sentiment analysis, represents an interesting research direction in social network mining.

Arnaboldi *et al.,* (2013) has proposed Dynamics of Personal Social Relationships in Online Social Networks: a Study on Twitter. In this paper author analyze a data set of Twitter communication records, studying the dynamic processes that govern the maintenance of online social relationships. The methods to perform analysis, the author studied the time series of the direct tweets (replies and mentions) and of the non-direct tweets sent by each ego. For some performance indices (i.e., new users contacted per day

and total number of new users contacted) they counted the number of new alters contacted by ego each day until the network is active. The results indicate that human behavior in Twitter significantly differs from other social networks studied in literature in different research fields.

Shang & Jiao (2012) proposed Semi-Supervised Learning with Mixed Knowledge Information. This paper propose a novel semi-supervised learning (SSL) approach i.e. semi-supervised learning with Mixed Knowledge Information (SSL-MKI) that can simultaneously handle both sparse labeled data and additional pair wise constraints together with unlabeled data. It first construct a unified SSL framework to combine the manifold assumption and the pair wise constraints assumption for classification tasks. Then, it presents a Modified Fixed Point Continuation (MFPC) algorithm with an eigen value thresholding (EVT) operator to learn the enhanced kernel matrix. The author developed a two-stage optimization strategy and provides an efficient SSL approach.

Lai *et al.,* (2013) has studied Psychometric Properties of the Internet Addiction Test in Chinese Adolescents. This study examined the psychometric properties of the Young's Internet Addiction Test (IAT) in 844 Hong Kong Chinese adolescents (37.7% boys) with mean age of 15.9 (standard deviation¼3.5) years. Methods Demographic items, Internet use habits, IAT, and the Revised Chen Internet Addiction Scale (CIAS-R) were administered. 3 percent of the participants were classified as addicted and 31.6% as occasional problematic Internet users. Confirmatory factor analysis results indicated that the 18-item second-order three-factor model has the best fit with authors data (Satorra–Bentler scaled w2¼160.56, df¼132, p<.05, normed fit index¼0.95, non-normed fit index¼0.99, comparative fit index¼0.99, root mean square error of approximation¼0.02).

Pourkazemi and Keyvanpour (2013) did a survey on community detection methods based on the nature of social networks. This survey shows several methods that have been proposed to detect communities, which represent the high importance of discovering communities for understanding social networks and detecting the useful and hidden patterns in the aforementioned network. Community detecting method for signed social network, Community detection methods in positive social network, The methods are Community detection methods for heterogeneous social network, Community detection methods in static social network, Community detection methods for dynamic social network, Community detection methods for directed social network.

Wen *et al.,* (2014) has proposed Exploring Social Influence on Location-Based Social Networks. In this paper, author studied the impact of social relations hidden in LBSNs, i.e., the social influence of friends. They propose a new social influence-based user recommender framework (SIR) to discover the potential value from reliable users (i.e., close friends and travel experts). They modeled the propagation of influence using diffusion-based mechanism. The performance of *SIR* framework is better than the state-of the- art user recommendation mechanisms in terms of accuracy and reliability.

Zhang *et al.,* (2015) *M-SEQ*: Early Detection of Anxiety and Depression via Temporal Orders of Diagnoses in Electronic Health Data. This paper proposes *M-SEQ*, an early detection framework for anxiety/depression using electronic health data from primary care visit sequences M-SEQ. First discovers a set of diagnosis codes that are discriminative of anxiety /depression, and then extracts each diagnosis pair from each patient's health record to represent the temporal orders of diagnoses. Further, it incorporates the extracted temporal order information with the existing representation to predict whether a patient is at risk of anxiety/depression.

Dinakar *et al.,* (2015) Sentiment Analysis of Social Network Content in this paper, the author performed sentiment analysis of online social activities of an individual which is then author analyzed to cluster behavioral and psychological tendencies of the individual.

Barnaghi *et al.,* (2016) proposed Opinion Mining and Sentiment Polarity on Twitter and Correlation between Events and Sentiment. This paper provides a positive or negative sentiment on Twitter posts using a well-known machine learning method for text categorization. The trained model is based on the Bayesian Logistic Regression (BLR) classification method. They used external Lexicons to detect subjective or objective tweets, added Unigram and Bigram features and used TF-IDF (Term Frequency-Inverse Document Frequency) to filter out the features.

Strength and Weakness

DM Technique	Strength	Weakness
ANN	One of the best techniques for solving classification Problems. Perform well with high dimensional feature space and small training set size. Suitable for offline clustering	Suffer from problem with sparse context links
BN	Very effective for text clustering. Simple classification algorithm. Very efficient in terms of computation time.	BN cannot be used to model the correlation relationships between random variables. Doing full Bayesian learning is extremely computationally expensive.
DT	Random Forest (RF): Effective in giving estimates of what variables are important in the classification. RF: Robust technique and perform well with variety of learning tasks.	The more decisions there are in a tree, the less accurate any expected outcomes are likely to be.
k-NN	One of the simplest and most discriminative classifiers in pattern recognition.	Inferior performance on small datasets. Performance will degrade for data with high diamensions. Dependent on the chosen feature and distance measure.
Fuzzy	Specialized in modeling with vague modes of social reasoning and takes into account the stochastic component of human reasoning.	Requires expertise in semantic web and fuzzy systems to manually handle the semantic fuzzy rule through an offline process.
K-Means	Uses as few clusters as possible and captures statistically and commercially important cluster characteristics. Suitable for fix number of groups with unknown characteristics based on variables that one defines. Performs well at finding a very small number of clusters.	When the number of clusters increases, the quality of discovered clusters quickly deteriorates. Often converge to a local minima
DBA	Density-Based Spatial Clustering of Application with Noise (DBSCAN): Does not require pre-specified number of clusters and noise filtering.	DBSCAN: Includes all the density-reachable points to a cluster.
HC	No apriori information about the number of clusters required. Easy to implement and gives best result in some cases.	Does not scale the growing of data size, because it relies on a fully specified similarity matrix

Conclusion

The social networking sites are playing a vital role in everyone's life. In this paper, we summarized various studies that had taken place for understanding human behavior by collecting data from online social networking sites along with the different techniques that are used for mining online social data with its applications. The social media features are useful to accurately track relationships between individuals and their social behavior within a network. This paper has discussed the experimental aspect of reality mining dataset using network measures of social network analysis. Different data mining techniques have been used in social network analysis as covered in this survey.

REFERENCES

Arnaboldi V, Conti M and Passarella A (2013). Dynamics of Personal Social Relationships in Online Social Networks: a Study on Twitter. *COSN '13 Proceedings of the First ACM Conference on Online Social Networks*, Boston, Massachusetts, USA.

Barnaghi P *et al.*, (2016). Proposed Opinion Mining and Sentiment Polarity on Twitter and Correlation between Events and Sentiment, 2016. *IEEE Second International Conference on Big Data Computing Service and Applications* 978-1-5090-2251-9/16.

Dinakar S *et al.*, (2015). Sentiment Analysis of Social Network Content, 2015. *IEEE 16th International Conference on Information Reuse and Integration,* 978-1-4673-6656-4/15.

Ferrara E, Jafari Asbagh M, Varol O, Qazvinian V, Menczer F and Flammini A (2013). Clustering memes in social media. In: *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining - ASONAM '13,* 548–555. doi:10.1145/2492517.2492530

Jansen BJ, Sobel K and Cook G (2011). Classifying ecommerce information sharing behaviour by youths on social networking sites. *Journal of Information Science* **37** 120–136. oi:10.1177/0165551510396975.

Kim H-N, Ji A-T, Ha I and Jo G-S (2010). Collaborative filtering based on collaborative tagging for enhancing the quality of recommendation. *Electronic Commerce Research and Applications* **9** 73–83. doi:10.1016/j.elerap.2009.08.004.

Lai CM *et al.*, (2013). Psychometric Properties of the Internet Addiction Test in Chinese Adolescents. *Journal of Pediatric Psychology* **38**(7).

Luo W, Liu J, Liu J & Fan C (2009). An Analysis of Security in Social Networks 2009. In *Eighth IEEE International Conference on Dependable, Autonomic and Secure Computing*. 978-0-7695-3929-4/09.

Pourkazemi M and Keyvanpour MR (2013). A survey on community detection methods based on the nature of social networks. *3rd International Conference on Computer and Knowledge Engineering (ICCKE 2013).*

Shang F & Jiao LC (2012). Semi-Supervised Learning with Mixed Knowledge information. *KDD'12*, Beijing, China Copyright 2012 ACM 978-1-4503-1462-6/12/08.

Shuai H-H, Shen C-Y *et al.*, (2016). Mining Online Social Data for Detecting Social Network Mental Disorders. In *International World Wide Web Conference Committee (IW3C2),* Montréal, Québec, Canada. ACM 978-1-4503-4143-1/16/04.

Tan C, Lee L and Tang J (2011). User-Level Sentiment Analysis Incorporating Social Networks. *KDD'11,* San Diego, California, USA. Copyright 2011 ACM 978-1-4503-0813-7/11/08.

Wen YT *et al.*, (2014). Exploring Social Influence on Location-Based Social Networks, 2014. *IEEE International Conference on Data Mining*.

Wise K, Alhabash S and Park H (2010). Emotional Responses During Social Information Seeking on Facebook. *Cyberpsychology, Behavior and Social Networking* **13**(5).

Zhang J *et al.*, (2015). *M-SEQ*: Early Detection of Anxiety and Depression via Temporal Orders of Diagnoses in Electronic Health Data, 2015. *IEEE International Conference on Big Data (Big Data)*.