Pattern mining approaches used in social media data

Article

Accepted Version

Full Text
Accepted for the International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems

Pattern Mining Approaches used in Social Media Data

Jyotismita Chaki
School of Information Technology and Engineering, Vellore Institute of Technology, Vellore, India

Nilanjan Dey
Department of Information Technology, Techno India College of Technology, West Bengal, India

BK Panigrahi
Department of Electrical Engineering, IIT Delhi, India

Fuqian Shi
Cancer Institute of New Jersey, Rutgers University, NJ, USA

Simon James Fong
Department of Computer and Information Science, University of Macau, Macau, SAR, China

R. Simon Sherratt
Department of Biomedical Engineering, the University of Reading, RG6 6AY, UK

Abstract
Social media conveys a reachable platform for users to share information. The inescapable practice of social media has produced remarkable volumes of social data. Social media gathers the data in both structured-unstructured and formal-informal ways as users are not concerned with the exact grammatical structure and spelling when interacting with each other by means of various social networking websites (Twitter, Facebook, YouTube, LinkedIn, etc.). People are increasingly involved in and dependent on social media networks for data, news and opinions of other handlers on a variety of topics. The strong dependence on social media network sites contributes to enormous data generation characterized by three issues: scale, noise, and variety. Such problems also hinder social network data to be evaluated manually, resulting in the correct use of statistical analytical methods. Mining social media data can extract significant patterns that can be advantageous for consumers, users, and business. Pattern mining offers a wide variety of methods to detect valuable knowledge from huge datasets, such as patterns, trends, and rules. In this work, data was collected comprised of users' opinions and sentiments and then processed using a significant number of pattern mining methods. The results were then further analyzed to attain meaningful information. The aim of this paper is to deliver a summary and a set of strategies for utilizing the ubiquitous pattern mining approaches, and to recognize the challenges and future research guidelines of dealing out social media data.

Keywords:

Social media, pattern mining, classification, clustering, pre-processing, feature extraction, feature selection.

1. Introduction

Pattern mining research has effectively created abundant tools, techniques, and algorithms to tackle bulky amounts of data to resolve real-life issues by extracting significant patterns and acquiring discerning knowledge. Data produced from social media sites is distinct from traditional feature value data used in typical data mining.[1] Social media data is mainly user-produced content on social media sites. Social media data is noisy, vast, unstructured, distributed, and dynamic. An additional characteristic of social media data is its interpersonal nature that can complicate pattern analysis. These attributes challenge pattern mining tasks to discover new efficient methods and algorithms mainly in the aspect of text mining, opinion mining and sentiment analysis. Pattern mining of social media can aid researchers understand innovative phenomena because of the usage of social media and enhance business intelligence to deliver improved facilities and create innovative prospects. Some pattern mining methods have been premeditated particularly to detect rules and patterns

based on interpersonal attributes.[2] Pattern mining methods can assist to locate secret groups in a social networking site, find influential people in the enormous blogosphere, detect user sentiments for active planning, create approval systems for tasks ranging from creating new friends to purchasing certain products, guard user secrecy and safety, or develop and strengthen faith amongst users. Moreover, the open contact with data offers large volumes of information to enrich performance and enhance pattern mining methods. Improvements in the pattern mining relies on large datasets and social media are a source of data for corporate and academic pattern mining researchers to test and produce innovative pattern mining methods.

The whole pattern mining process can be considered as three main serial of data pre-processing, feature extraction and classification.[3] The data pre-processing aids decrease of noise, filtering, data standardization, etc. The feature extraction and selection stage characterize the data, resulting in a feature vector utilized to predict or detect data by clustering or classification. Classification is concerned with the matter of distinguishing unknown data as one from a group of previously learnt data. In clustering, the aim is to study the structural relations amongst data.

While some surveys have focused on pattern mining used in social media data, no specific survey has been reported to date on social media data derived from pattern mining techniques used on social media data while also incorporating spam detection, community detection, influence analysis, recommendation and link prediction techniques. We believe that this review can bring a beneficial summary of present work, and deliver probable future study guidelines in the area of social media data.

The rest of this paper is prepared in this fashion: In section 2, we briefly presented social media data and its structure, in sections 3-12 we review associated techniques. Lastly, in section 13 conclusions and future research guidelines of pattern mining in social media data are conferred.

2. Social Media Data Collection

Social media sites deliver a prevailing foundation for data collections since such sites store statistical information about individuals, their associations with other users in the same site and their reviews, ranking, or rating information reliant on the nature of the sites. There are nine important topics that need to be considered regarding the collection of the data (1) reliability, (2) correctness, (3) confidentiality, (4) structure, (5) language, (6) region, (7) type of data content, (8) venue (e.g. Facebook, Twitter), and (9) time-frame.[4]

2.1. *Source of Social Media Data*

There are various types of social media sites from which one can collect data from such as some online social networking (e.g., Facebook, LinkedIn, Myspace etc.), blogging (e.g., Business insider, Huffington post, Engadget etc.), microblogging (Twitter, Plurk, Tumblr etc.), wikis (e.g., Wikipedia, Wikihow, Wikitravel etc.), social news (e.g., Reddit, Digg, Slashdot etc.), social bookmarking (e.g., StumbleUpon, Delicious etc.), media sharing (e.g., YouTube, UstreamTV, Flicker etc.), opinion, reviews and rating (e.g., Yelp, Cnet, Epinions etc.), answers (e.g., WikiAnswers, Yahoo!Answers etc.), etc.[5] In general, data can be collected by various means including communicating with the site manager/administrator to obtain the dataset, or downloading the dataset that is formed for academic purposes, etc. It is best to interpret social media data as a foundation of raw data, which contains (and is not restricted to) the following metrics: likes, shares, comments, conversions, impressions, mentions, and most significantly, clicks – the click metadata is vital for understanding further about what people are concerned with. Table 1 provides a list of the famous social media networks.

Table 1. List of famous social media networks

Social network name	Date of Foundation	Country of Foundation	Monthly active user (reported on August 2020)	Purpose
Facebook	February, 2004	America	2.7 billion	Social Networking
WhatsApp	January, 2010	America	1.5 billion	Messaging, Audio and Video calling
Tumblr	February, 2007	America	371 million	Social Networking, Microblogging
Instagram	October, 2010	America	1 billion	Social Networking, Media Sharing
Twitter	March, 2006	America	330 million	Microblogging
Skype	August, 2003	America	100 million	Audio and Video call, Instant messaging
Viber	December, 2010	America	260 million	Messaging
Line	March, 2011	Japan	84 million	Messaging
Snapchat	September, 2011	America	301 million	Social Networking
Pinterest	March, 2010	America	291 million	Media Sharing
LinkedIn	December, 2002	America	303 million	Social Networking
VKontakte	October, 2006	Russia	400 million	Social Networking
Reddit	June, 2005	America	330 million	Social News
Google+	December, 2011	America	395 million	Social Networking
YouTube	February, 2005	America	1.9 billion	Media Sharing
Flicker	February, 2004	America	90 million	Media Sharing
Myspace	August, 2003	America	20 million	Social Networking
Vine	June, 2012	America	200 million	Media Sharing
Quora	June, 2009	America	100 million	Social Networking
Digg	November, 2004	America	8 million	Social News
StumbleUpon	November, 2001	America	35 million	Social bookmarking
Odnoklassniki	March, 2006	Russia	45 million	Social networking

2.2. *Acquiring Social Media Data*

The three main techniques proposed in the literature to collect data from social media are discussed here: Network traffic analysis, Ad-hoc applications and Crawling the user graph. Network traffic analysis arrests packet streams from a social media network link and then examines request-response pairs concerning user communications with social media.[6] From these pairs, it is probable to deduce information about the users who are browsing other users' pages. Additionally, information about the users can be gained by the analysis of the response consignment. Ad-hoc applications exploits a group of application programming interfaces (APIs), to deliver services and games to the social media users. In this application, a user does not interrelate directly with the application servers as the social media network architecture offers an interface layer between user and application. Generating ad-hoc applications to obtain data from social media permits to assemble information about the users in a dual way. The APIs generally let the application to access information on the users' profile who are registered to the application. Moreover, the study of the log on the application servers permits to extract information about the active user behavior. Crawling is the most widespread solution for data collection in social media and comprises on asking the social media for publicly offered information about users and can be demonstrated as a directed graph $G=(N, P)$, where N is the group of nodes (users) and P is the group of edges (social associations between users). Each node has incoming and outgoing links. The main objective is to collect information about the users by developing a technique to visit every user in the network. Crawling the social media network graph is a recursive procedure that initiates from a group of initial users and moves forward by determining new users at each step, thus enriching the collection of data procedure in terms of data representativeness and duration.

Table 2 displays the broadly utilized data collection methods in selected articles.

Table 2. The broadly utilized data collection methods in selected articles

Web crawler	User generated	API	References
✔	✘	✘	8, 11, 12, 18, 19, 23, 24, 37, 47, 48, 54, 59, 61, 66, 69, 73, 76, 77, 79, 81, 85, 86, 89, 90, 92, 96, 99, 100, 104, 105, 112, 115, 232, 233, 235, 238, 239
✘	✔	✘	14, 22, 23, 25, 26, 41, 46, 51, 53, 56, 58, 70, 71, 72, 74, 78, 88, 98, 101, 102, 110, 111, 231, 237
✘	✘	✔	15, 16, 17, 20, 45, 49, 50, 57, 65, 80, 81, 93, 95, 97, 107, 109

2.3. *Issues related to Social Media Data Collection*

Social media data mining is an emergent area where there are more complications than prepared solutions. Mining social media data is the job of mining user-produced content with social relations. Some of the data collection issues are:[7]

2.3.1. *Large Data Inconsistency*

Social media data is indubitably huge. Sometimes a system may become inconsistent while storing the multisource, multidimensional, and multisite social media data to aggregate information with sufficient statistics for effective mining.

2.3.2. *Gaining sufficient sample data*

Using a social media data APIs typically only offers a restricted volume of data can be attained each day. This may deprive knowledge of a population's distribution. Moreover, it is not guaranteed that much data can produce any accurate patterns. If the application does not attract numerous subscribers, the obtainable dataset is partial and unserviceable for analysis purposes.

2.3.3. *Noise removal errors*

Social media data can contain a large portion of noisy data. While storing this data for analysis purposes, two important situations can be created: (1) instinctively eliminating noise can deteriorate the research problem because the elimination can also remove valuable information, and (2) the definition of noise may be relative and complicated since it relies on the task at hand.

3. Preprocessing Techniques for Social Media Data

Many techniques of pattern analysis are dependent on the quality and excellence of the data and not on the quantity of the data.[8] Social media data collected for pattern mining is not directly prepared. The nature of the collected data will also affect the hardware and database used.

Prior to the discussion of social media data preprocessing techniques, it is important to recognize probable data complications:[9] (1) Missing data: Once a part of information occurred but was not involved in the raw data collected for whatsoever cause. Issues occur when: a) numeric blank data or a missing value is mistakenly replaced by 'zero', and b) textual missing word may alter the whole meaning of a sentence. (2) Incorrect data: Once a part of the information is wrongly specified or is incorrectly construed (e.g. a system presenting a currency value is in $ when in fact it is in £ or pretending text is in UK English rather than US English). (3) Unreliable data: Parts of the information may be incompatibly specified (e.g., different formats for dates: 2018/7/10, 10/7/2018 or 7/10/2018, a mixture of text cases, mixing English and Bengali in a text message, or placing Bengali quotes in an English text). Data preprocessing prepares the raw data suitable for pattern mining.

The procedure of data preprocessing may include eliminating typographical errors or authenticating and correcting data against a recognized list of entities. Generally, unstructured noisy raw data may comprise of quotes, misspelled words, extra spaces, program codes, special characters, additional line breaks, foreign words, etc.[10] Therefore, to accomplish high-quality pattern mining, it is required to conduct data preprocessing at the first step to: eliminating duplicates, spell checking and correction, changing the text case, finding and replacing text, eliminating non-print characters and spaces from text, correcting times and dates, correcting numbers, outliers and number signs, changing and reorganizing rows, columns and table data, negation handling, stemming, expanding abbreviation, tokenization, etc.

3.1. *Cleaning Raw Data*

A typical method of cleaning social media data is to put the data into a spreadsheet, worksheet or spreadsheet-like table, and then reformat the content.[11] For example, Google Refine3 is an independent desktop application for social media data cleaning and transcoding to several formats.

3.2. *Tagging and Parsing Unstructured Data*

Since social media data is created by humans, and hence is unstructured, a method is needed to convert it into structured data to gain meaningful information.[12] Thus, unstructured social media data need to be preprocessed, tagged and then parsed so as to analyze/quantify the social media data. A solo data set can deliver some motivating understandings. However, merging several data sets and processing the unstructured data may be more difficult to understand, it can permit researchers to reply to questions that were not possible before.

3.3. *Storing data*

To extract eloquent information from the collected social media data, it is very important to store data in a database. Every social platform has specific rules about how their own data can be kept and used. Databases related to social media data can comprise of (1) Flat file: a 2-dimensional database comprising records that do not have any structured interrelationship and can be found sequentially. (2) Relational database: prepared as a group of formally defined tables to identify relations between deposited data, permitting more complex relationships between the data items (e.g., SQL database). (3) noSQL databases: a type of database management system (DBMS) defined by its non-observance to the broadly used relational database management system (RDBMS) model.[13]

3.4. *Sampling data*

Often, processing the complete dataset is expensive. With the enormous progress of social media, processing bulky amounts of data is very difficult.[14] Therefore, it is best to encourage sampling. A small random subgroup of data can be carefully chosen and processed as a substitute of the complete data. The selection procedure should assure that the sample is demonstrative of the statistical distribution that rules the data, thus confirming that outcomes gained on the subset sample are more or less the same as gained on the complete dataset. The following are three major sampling methods are generally used:[15] (1) Random sampling: samples are carefully chosen consistently from the dataset by using some probability distributions. (2) Sampling with or without substitution: In sampling with substitution, a sample can be chosen several times in the data. In sampling without substitution, samples are eliminated from the selection pool once chosen. (3) Stratified sampling: the dataset is first divided into several bins; then a fixed number of samples are chosen from each bin utilizing random sampling.

Table 3 shows the broadly utilized pre-processing methods in selected articles.

Table 3. The broadly utilized pre-processing methods in selected articles

P1	P2	P3	P4	P5	P6	References
✓	✗	✗	✗	✗	✗	8, 58, 99, 107
✗	✓	✗	✗	✗	✗	11, 20, 24, 49, 52, 93, 94, 95, 96, 97, 99, 105, 109
✗	✗	✗	✓	✗	✗	92
✗	✗	✗	✗	✓	✗	94
✗	✗	✗	✗	✗	✓	96

*P1: Lemmatization, P2: Filter, P3: Data Cleaning, P4: Graph construction, P5: Remove duplicate, P6: Conversion of unstructured data into structured data

4. Dimensionality Reduction

The feature vector that results from the preprocessing stage is usually not suitable to be processed by a subsequent stage due to its high-dimensionality. This creates difficulties for traditional pattern mining tasks because of the curse of dimensionality. Additionally, with a huge number of features, learning architectures tend to over fit creating performance debasement on unobserved data. High dimensional data can increase the computational costs and memory storage requirements for pattern analysis.

Dimensionality reduction is one of the influential tools to address the earlier discussed issues.[16] It can be primarily classified into two main segments: feature extraction and feature selection. Feature extraction substitutes the high-dimensional features into a new low dimensional feature space by merging the original features in a linear or nonlinear fashion and is favored when the original input data does not comprise any comprehensible features for a given learning algorithm. Feature selection specifically chooses a subgroup of appropriate features for model creation and preserves physical senses of the original features. Thus, it provides good interpretability and readability models by preserving some of the original features.

4.1. *Feature Extraction*

Various kinds of features used for social media data pattern mining are classified as text-based and visual-based methods.[17] Text-based features are further divided into Morphological features, Frequent features, and Implicit features. There are three sorts of morphological features, i.e. syntactic, semantic and lexicon structural. Syntactic feature use part of speech (POS) tagging, dependency depth (DD) feature, chunk labels, N-gram word, etc. Semantic features highlight semantic orientation (SO) and contextual information (CI). The SO method makes utilization of point wise mutual information (PMI) and latent semantic analysis (LSA) that gives a polarity score to every phrase or word. Conversely, the CI technique is used to include text at the sentence level. Lexicon structural feature comprises of special character regularities, word circulations and word level lexical features. Frequent features, additionally known as hot features, use the apriori association rule (AAR) which is widely used in text mining. Implicit features are not evident in the review which comprises of adjectives and adverbs. For example, the adjective "light" expresses a weight feature, but it requires some level of domain information for analyses.

There are abundant visual features proposed for various social media visual recognition techniques that can be divided into three groups: low-level, middle-level and appealing features. Low-level features signify the visual information straight from the pixel values of the useful regions, common patterns or spreading of colors include gist descriptor, color histogram (CH), bag-of-visual-words (BoVW), local binary pattern (LBP), etc.[18] The image is preserved as a 2-dimensional signal. One main issue of the low-level feature is the deficiency of coincidence amongst the information that can be mined from the visual data and the proper explanation of the data. The middle-level feature implements a group of perception made on the low-level features; therefore, the image can be denoted at the semantical level include senti adjective noun pairs (sentiANP), classemes, attribute etc.[19] The third type is appealing features that are devotedly premeditated for mining the intellectual human acuity on the visual appearance. The appealing features can be grouped into two categories: aesthetic (AF) and principles-of-art (PoA). AF features support understanding the visual illustration at an abstract level, such as "lovely", and includes luminosity, dark channel, symmetry, sharpness, colorfulness, white balance, color harmony, eye sensitivity, etc. PoA features have been shown to be supportive in image emotion recognition including balance, harmony, emphasis, gradiation, variety, etc.

4.2. *Feature Selection*

Feature selection methods in social media data have an important role for detecting pertinent features and growing recognition accuracy which are congregated into four main groups: Statistical, Natural Language Processing (NLP) or heuristic-based, Clustering based, and Hybrid. Statistical methods are further classified into three sub categories, univariate, multivariate and hybrid.[20] Univariate approaches, known as feature filtering approaches, receipts feature distinctly, include chi-square, information gain (IG), log likelyhood, occurrence frequency, minimum frequency thresholds, etc. Univariate methods have computational proficiency, but they overlook feature relations. Multivariate methods contemplate assembly of features and utilize a wrapper method for feature selection, include recursive feature elimination, decision tree models, genetic algorithms, etc. Hybrid methods combine univariate multivariate and additional approaches for attaining accuracy and efficacy. NLP based methods, primarily used in emotion mining, work on three elementary principles: (1) noun phrases, noun, adverbs, adjectives generally used as features, (2) subjective words can act as features, (3) A is feature of product B in phrases corresponding 'A of B' or 'B has A'. This method has attained high accuracy, but has low recall with reliance on accurateness of POS tagging. Clustering methods require limited parameters to tune. The main limitation of this method is that only the most important features can be taken out and it is tough to extract insignificant features. Hybrid methods such as combination of syntactic and lexical features with a determined entropy, POS tagging with WordNet dictionary, combination of point-wise mutual information and association rules for recognizing features with a benefit of using HowNet dictionary, bootstrapping recursive learning approach with added linguistic rules for selecting low occurring features, combination of LBP and CH for visual media

recognition, combined (sentiANP and classemes) feature for detecting semantic concepts of visual emotions, etc., used in the literature of social media data feature selection.[21]

Some advantages and limitations of different feature extraction and selection methods are as follows: (1) POS tagging can help in identifying differences and similarities between words. It can also help in defining authorship in social media text as the use of words varies from person to person. But it is challenging to establish the PoS tags of a word in a given sense, mainly because most premise of this article don't use diacritics. And the same word can be written in various ways. (2) The main advantages of n-gram method are its scalability and simplicity. With larger value of none can store more social media context that enables small experiments to increase proficiently. The main drawback of N-gram model is that it ignores any clear representation of long-range dependence. For this cause, n-gram models did not have a major effect on social media linguistic theory, where part of the explicit objective is to model these dependencies. (3) The main advantage of PMI method is that by using this method one can estimate whether the two items in the social media is having a genuine association or not. (4) LSA is easy to understand, implement and use. LSA is capable of producing decent results on social media dataset with diverse topics. But the model is not humanly readable. Evaluation or debug is possible through searching similar words for each word in the latent space though. But otherwise not easy to interpret. (5) AAR is easy to understand and simple. It is unsupervised, so don't need any labeled social media data. But if the dataset is small, the algorithm can find many false associations among the social media data that happened simply by chance. (6) CH is easy to understand and implement. But two different images sharing the same color information will have the same histogram. (6) BoVW is easy to realize and implement. But it leads to high dimensional feature vector due to large size of vocabulary. It also doesn't influence co-occurrence statistics between words. (7) The main advantage of LBP is it is simple and low in complexity. But it is noise sensitive.

Table 4 presents a selection of further works on feature extraction and selection techniques from the pattern mining of social media data literature.

Table 4. Review of Feature Extraction and Selection Techniques in the Social Media Data Literature

Technique	Field of usage
POS tagging	Analysis of social media text.[22]
DD feature	Social media friendship prediction.[23]
Chunk labels	Social media text as well as social media message analysis.[24]
N-gram word	Detect tension in tweet message.[25]
PMI feature	Tourism opinion in social text.[26]
LSA	Social media message analysis.[27]
AAR	Identification of user patterns, adverse drug reactions in social media.[28]
Implicit features	Creating friend suggestion in social media.[29]
Gist descriptor	Sentiment analysis from microblogging.[30]
CH	Social media logo information retrieval.[31]
BoVW	Modeling personality based on selfies.[32]
LBP	Predicting image and video interestingness from social media content.[33]
SentiANP	Contextual enrichment of remote-sensed events with social media.[34]
Classemes	Social media sentiment analysis.[35]
Aesthetic	Social media image popularity prediction.[36]
Principles-of-art	Social media image emotion perception.[37]
Chi–square	Tourism experiences analysis from social media.[38]
Log likely-hood	Deriving the maximum web-browsing likelihood estimate of a user's social profile.[39]

5. Classification

Classification is the organization or division of various social network features into groups through the discovery of a relation between features.[40] There are two steps to the classification method: (1) Classification procedures try to obtain a model, that is based on a assembly of training data, for the attribute class as the social media datasets function, and (2) it applies the previously obtained model to the new unknown social media datasets to determine the corresponding each data class.

5.1. *K-nearest Neighbor Classifier*

The k nearest neighbor (KNN) classifier suggests that comparable social media data occurs in close vicinity. In other terms, equivalent social media details are neighboring to each other.[41] First, the classifier installs social network training

and messaging details. The K value is then chosen by some experiment conducted. For each social media testing data, the distance to every training data point is calculated. Calculated distances are kept in a list and sorted. Thereafter, the first K points will be chosen. The test data is then allocated to the class dependent on the number of classes existing in the chosen points. For K=1, the query sample shall be allocated to the training class of its closest neighbor.

5.2. *Naïve Bayes*

It is a methodology of classification based on Bayes' theorem, with an independence assumption between predictors.[42] A Naive Bayes classifier believes, in simple terms, that the inclusion of a particular feature in a class is irrelevant to any other feature being present. Predicting class of test data set by using this classifier is simple and fast. For multi-class prediction, it also performs well. This classifier works better compared to other models like logistic regression when assuming independence holds, so less training data is needed.

5.3. *Support Vector Machine (SVM)*

SVM can handle high-dimensional social media data and produces a relatively detailed (geometrical) model.[43] An SVM predictor is focused on a kernel function K which defines a specific form of similarity measure between media data. Examples of kernel functions are linear, polynomial, RBF (radial basis function), or sigmoid kernel.

5.4. *Decision Tree (DT)*

By creating a decision tree, the decision tree classifier generates the classification model.[44] Every node in the tree defines a check on a social media data attribute, every branch going down from that node corresponds to one of the probable attribute values. Each leaf represents associated social media data class labels. Media data instances in the training set are categorized by exploring them from the tree's root down to a leaf, based on the result of the tests along the way. Beginning from the tree root node, each node divides the space of the data instance into two or more sub-spaces as per a testing condition of the attributes. Then moving down, the tree branch which corresponds to the media data attribute value creates a new node. The method is then replicated for the subtree rooted in the new node till all records have been identified in the training collection. Typically, the decision tree design process operates top-down by selecting an attribute test condition at each stage that best separates the media data.

5.5. *Deep Learning (DL)*

DL approaches derive data features and acquire incidental features while training that can distinguish between vast quantities of social media data.[45] Addition to this, if there are satisfactory number of samples then features of several variables which can affect classification, deep learning approaches learns to expose such factors even though learning descriptions of attributes are present. This can assist to handle vast intraclass variations and noisy data present on social media. The ability of DL to learn features from unrecognized social media information has the skill to enhance social media training feature. Addition to this, deep learning has the capability to make simpler features learned outside of the comparable information to training information, which is beneficial in handling social media data.

Neural networks (NN) are a non-linear approach and is noticed as a largely employed deep learning model for data pattern mining in social media.[46] NN contains a huge number of hugely inter-connected processing foundations called neurons, employed to solve any obvious problem related to social media data. As NN have the capability to derive considerable statistics from a large assembly of data, neurons are built for accurate application related to social media. Some deep learning model that can be used in social media are Deep Belief Networks (DBN) [49], Convolutional Neural Networks (CNN),[47] Recurrent Neural Networks (RNN),[48] Deep Boltzmann Machines or Restricted Boltzmann Machines (RBM),[49] autoencoder,[50] etc.

Finally, Table 5 presents some additional demonstrative works on pattern classification approaches in the field of social media.

Table 5. Review of Classification Techniques in the Social Media Data (SMD) Literature

Technique	Field of usage
KNN	Social media face identification.[51]
NB	Forecasting stock prices using social media analysis.[52]
SVM	Opinion mining from social media.[53]
DT	Construction of a profile of subjective well-being based on social media language in Facebook status updates.[54]
NN	Feedback generation on social media.[55]
DBN	Relationship extraction from interaction context in social media data.[56]
CNN	Aspect extraction for opinion mining.[57]
RNN	Mining e-cigarette adverse events in social media.[58]
RBM	Joint sentiment/topic modeling on social media text data.[59]

6. Clustering

Clustering is important for social media data as it instructs the important grouping of unidentified media data available by having references from training media datasets.[60] Cluster rules are based on the customer that can be used to meet their need. For example, the user can search for features of similar clusters, natural clusters, valuable media information clusters or exceptional information clusters. Clustering methods need to build standards that generate similarities between data points, and respectively each standard produces clusters that are distinct. The clustering method can be used for the determination of natural groupings in sentiment analysis and emotion mining, and thus portray a class summary in a collection of documents. Clustering doesn't really necessitate the pre-knowledge of a document class and thus does not involve a training procedure. So, human involvement is unobstructed and therefore saves a lot of time.

6.1. *Density-based Approaches*

Density-based approaches find clusters to be a clustered based on the area of some similarity and distinct from a less concentrated region in media data space.[61] Such methods provide reasonable accuracy and ability to merge two clusters, namely Density-Based Spatial Cluster of Noise Applications (DBSCAN), Order Points for Identifying Cluster Structure (OPTICS), etc.

6.2. *Hierarchical Approaches*

Clusters produced in this procedure generate a tree-type framework of hierarchical media data.[62] New clusters are created utilizing a previously developed one, including Balanced Iterative Reducing Clustering and Using Hierarchies (BIRCH), Clustering Using Representatives (CURE), etc. It is classified into two kinds: Divisive (top-down process) and Agglomerative (bottom-up process).

6.3. *Partitioning Approaches*

These approaches divide media data into a number of clusters and every division forms a single cluster.[63] This approach strengthens the cluster theory likeness feature by using distance as the key feature, like Clustering Large Applications based on Randomized Search (CLARANS), and K-means, etc.

6.4. *Grid-based Methods*

In this technique, the media data space is framed into a limited number of cells that generate a grid-like structure.[64] The entire clustering process completed on these grids are independent and fast of the number of media data include CLustering In Quest (CLIQUE), Statistical Information Grid (STING), wave cluster etc. Table 6 offers some additional representative works on clustering methods in the area of social media.

Table 6. Review of Clustering Techniques in the Social Media Data (SMD) Literature

Technique	Field of usage
OPTICS	Visualizing the gay community in Beijing with location-based social media.[65]
DBSCAN	Density-based place clustering in geo-social networks.[66]
CURE	Leveraging social media networks for classification.[67]
BIRCH	Sign prediction in social networks with positive and negative links.[68]
K-means	Discover social circles in ego networks.[69]
CLIQUE	Facilitating information seeking for hyperlocal communities using social media.[70]
CLARANS	Entrepreneurial team formation and search using social media data.[71]

7. Challenges related to pattern mining for Social Media

The type of data in social media creates a number of challenges to their collection, analysis and mining. First, social media data is collected in bulky amounts and is extremely dynamic and composite or complex in their nature. Hence, social media data cannot be processed effortlessly utilizing traditional pattern processing techniques or database management applications along with desktop statistics and imagining packages. Moreover, social media data holds both structured and unstructured features. While structured data includes user profile attributes, temporal, spatial, thematic data along with observational related data (e.g. number of comments, likes, mentions, retweets, etc.), unstructured data contains user-produced textual data. This excess information denotes a substantial challenge that demands huge computing volumes and erudite sampling, mining, and analysis approaches and also undependable due to their probable variation and incompleteness.

Privacy problems constantly exist when media data is collected.[72] Researchers and other users involved in collecting media data may face questions such as whether it is right to gather, process, utilize and report on social media data though these are truly "free or public" in principle. While sampling media data, it is hard to choose appropriate samples which are representative of the actual data. Other challenges consist of searching groups in social media, mining patterns in social media data and evaluating overlapping groups.

8. Social Media Spam Detection

Social spam is low-quality social media networks data which is comparable to e-mail spam, since unwanted bulk messages are not request for, or subscribe to by users.[73] Such spam is an annoyance to users and hampers them from consuming or searching for data that is relevant to them. Individual social networks are skilled to filter a substantial quantity of the spam they obtain, though they generally need huge quantity of resources (e.g. Personnel) and experience a delay before perceiving new kinds of spam.

The first problem to address social spam finding is the object classes should be realized as probable candidates as a spam category.[74] Spam can be inserted to social networks at various stages. The old understanding is to categorize pages or site as spam created on their content, specifically, resources that system manipulators distinguish as non-relevant. Secondly, one can place emphasis on spam posts. Lastly, someone can look at user profiles that have been created in order to inject international material into the framework. This type of accounts may or may not blend with genuine spam content, with the purpose of mask spamming activity. Flagging users as spammers is the method considered by some social networks about spam detection, for instance BibSonomy. This method is spontaneous and easy from an administrator's standpoint.

The social spam recognition procedures can be divided into three key parts:[75] (1) Mapping and Assembly: Mapping methods are utilized to adapt an explicit social media network entity into a structure demarcated standard model for the entity such as message model, profile model, or webpage model. Assembly is the procedure of penetrating every model entity for related matters and then choosing that model entity. For instance, if someone is working with a text entity and the message includes URLs, these URL-related web pages are retrieved and web page objects are created, which are then gathered along with the message object. For spam identification, this added detail is also important as it can deliver a rich foundation of information for the later phases. (2) Pre-filtering: With the purpose of decreasing classification cost, fast-

path methods of pre-filtering can be adopted to rapidly filter out earlier recognized or alike spam in incoming social media network matters. Some of these methods contain blacklists, hashing, similarity matching, etc. Entries are appended to these lists because of bad behavior or past spamming, and therefore it is probable that items that comprise such matters should be forbidden. Shingling and hashing can be utilized to rapidly compute likeness in contrast to earlier spammy items. The number of prior spammy entries of an article is verified against can be imperfect to evade high lookup charges. These methods may have limitations because of their delay time in sensing new spam, but they are meaningfully improving the time occupied to categorize an entity as non-spam or spam. 3) Classification: Supervised machine learning techniques like Bayes, Support Vector Machine, LogitBoost, etc., are utilized to categorize the incoming entity and related entity.

Table 7 presents additional works on spam detection methods in the area of social media.

Table 7. Review of Spam Detection Techniques in the Social Media Data (SMD) Literature

Technique	Field of usage
Mapping and assembly	Analysis of social spam data from homepage, blog and Twitter page.[76]
Pre-filtering	Spam detection from verbal social media text.[77]
Classification	Twitter and myspace social data spam detection using Naive Bayes and Support vector machine classifier.[78]

9. Community Detection

Specified a social media network S = (Ver, Edg), a social media community can be demarcated as a network subgraph containing a Social Media entities group SC ⊆ Ver which are related with a mutual component of attention.[79] This component can change as an actual person, a topic, an event, a place, a cause or an action. The variability of approaches that have seemed in literature for recognizing communities is bigger, as for every community description there are many approaches demanding to identify the corresponding communities. Here, the maximum significant five classes of such approaches are discussed.

9.1. *Cohesive Subgraph Discovery*

This approach is a description of the structural attributes that a network subgraph should please so as to be measured as a community. Zhao and Tung used a k-mutual friend subgraph to find a community.[80] A k-mutual-friend was a connected subgraph such that every edge was maintained by at least k pairs of edges establishing a triangle with that edge within the graph. The k-mutual-friend number of this subgraph is equal to k. When this type of subgraph structure is stated, approaches include the list of such assemblies in the network is under study. The k-cores, n-cliques, cliques, lambda sets and LS sets, are instances of these cohesive structures and consequently algorithmic organizations for counting such structures, for instance the Bron–Kerbosch algorithm and the effective k-core decomposition algorithm, fit in to this community identification approaches. Addition to this, approaches like the Clique Percolation Method (CPM) and the SCAN algorithm that led to the detection of subgraph assemblies with definite attributes, fall under this type of approaches.

9.2. *Vertex Clustering*

These methods initiate from the old data clustering study. Papadopoulos *et al.*, developed a representative way of forming a vertex clustering that can be explained by traditional data clustering approaches (for example hierarchical agglomerative clustering and k-means) is by implanting graph vertices in a vector space, where pairwise distances amongst vertices can be computed.[81] One more common technique is to utilize the graph spectrum for plotting graph vertices to points in a low-dimensional space, where the group assembly is weightier. Other vertex likeness procedures like the neighborhood overlap and structural equivalence have been utilized to calculate likenesses amongst graph vertices. Lastly, a noteworthy method, called Walktrap, utilized the random-walk based resemblance amongst vertices and amongst communities and utilized modularity in a hierarchical agglomerative clustering structure to make an ideal vertex clustering construction.

9.3. *Optimization of Community Quality*

There is a huge number of approaches that are created based on enhancing some graph-based measurement of the quality of the community. Cut-based measures and subgraph density, like conductance and normalized cut, are among the initial approaches that to be utilized for measuring some network division quality into clusters. Shi and Malik used normalized cut for segmenting a graph where the normalized cut criterion was measured for both the total dissimilarity between the different groups as well as the total similarity within the groups.[82] A research was enthused by the modularity measure. Approximate modularity maximization systems flourish in the state-of-art. Besides the important greedy optimization method, and speeded up forms of it, like max-heap based agglomeration and iterative heuristic schemes, more classy optimization approaches have been invented, for example, extremal optimization, simulated annealing and spectral optimization. Approaches directing at the optimization of local procedures of community quality, like subgraph and local modularity, also be a member of this class. Lastly, this class comprises approaches that deals with the hills and valleys in the spreading of network-based edge or node functions.

9.4. *Divisive*

These approaches depend on the detection of basic network elements (vertices and edges) which are located amongst communities. E.g., Girvan and Newman used a seminal algorithm which gradually eliminated the network edges built on some edge distance till communities occurred as detached graph components.[83] Numerous measures of edge distance have been developed, for example, random-walk, edge, and current-flow distance, in addition to information centrality and the edge clustering coefficient. A comparable principle is implemented by vertex removal approaches; such approaches eliminate vertices to disclose community structure. Lastly, min-cut/max-flow approaches accept a divisive perception: they attempt to recognize graph cuts that have the smallest size.

9.5. *Model-based*

This is a wide-ranging class of approaches that either ponder a procedure arranged on the social media network, that discloses its groups, or they study a fundamental structure of statistical nature that can produce the social media network partition into communities. Instances of dynamic procedures is label propagation where researchers initialized a node with a unique label and at every iteration of the algorithm, every node adopts a label that a maximum number of its neighbors have, with ties broken uniformly randomly. At the end of the algorithm, nodes having the same labels are grouped together as communities.[84] Other examples of dynamic processes are synchronization of Kuramoto oscillators, diffusion flow, better recognized as Markov Cluster Algorithm, and the popular spin model by. Furthermore, community identification can be detected as a statistical inference issue, presuming some fundamental probabilistic structure, like the planted partition structure, which produces the community structure and calculates the model constraints. Another model-based method based on the principle that a decent clustering is determined by a low encoding cost, therefore they achieve community identification by searching the cluster structure which outcomes in the lowest probable cluster encoding cost.

Table 8 presents additional works on community identification approaches in the area of social media.

Table 8. Review of Community Detection Techniques in the Social Media Data (SMD) Literature

Technique	Field of usage
Cohesive subgraph discovery	Community detection from social data.[85]
Vertex clustering	Community identification with edge content in social media, community detection in Twitter, Facebook LinkedIn and Google+ data.[86]
Community quality optimization	Social media community detection using common interests, friendship, religion, social media community detection using graph representation, social media community detection using exhaustive search, maximum matching, and greedy heuristics.[87]
Divisive	Analysis of heterogeneous relationships of various social data, social community detection using memes and behaviors.[88]
Model based	Community detection using mobile social network, social media community detection using family, friends, and colleagues.[89]

10. Influence Analysis

Social Influence can be demarcated as follows: Specified two persons P, Q in a social network, P applies the influence on Q, specifically, P has the effect of alteration the estimation of Q in an indirect or direct way.[90] The present assessment measurement for social influence comprise centrality measures, entropy measure, link topological ranking measures, etc.

10.1. *Centrality Measures*

Centrality is a significant perception in reviewing social media networks and calculates how central a person is located in a social media network. The normally utilized tools are network analysis and graph theory. Diverse centrality measurement have been suggested in the state-of-art to broadly be utilized in the analysis of social influence, as well as degree centrality, betweenness centrality, closeness centrality, Katz centrality, and eigenvector centrality.[91]

Newman and Wellman used degree centrality demarcated by calculating the link numbers occurred upon a node, specifically, the edge numbers that a node held.[92, 93] For a directed network, metrics are used for degree centrality, that is, a count of the edge numbers focused to the node specified as in-degree, and the number of edges that the node directs to others called out-degree. Opsahl *et al.* and Borgatti *et al.* used closeness centrality demarcated as the mean distance between two vertices in a network.[94, 95] In influence analysis, it can be treated as a measurement of efficacy of every node in terms of distributing data in the social media network. The bigger the node's closeness centrality, the well placed the node is in the social media network. Frantz *et al.* used closeness centrality which counted the number of times a node occurs as a link along the minimum distance path of two other nodes.[96] Under the supposition that item transmission tracks the minimum distance paths, a large closeness centrality node owns an improved influence throughout the social media network. The idea of closeness centrality has a widespread real time application, for example, transport, biology, and social networks. Okamoto *et al*. used eigenvector centrality as a metric of influence of a specified node in a social media network.[97] Relative scores are allocated to the entire network nodes built on a supposition that connects to high-scoring nodes donate more to the node score than connections to low-scoring nodes. Kiss and Bichler used Katz centrality as a simplification of degree centrality.[98] While Katz centrality counts the number of the entire nodes that can be linked over a path, degree centrality measures the number of direct neighbors.

10.2. *Topological Link Rank based Measures*

In social networks, nodes play a vital role, a connection to a highly significant (maximum neighbor) node is more appreciated compared to a less significant node (minimum neighbor). Search engines influence topological link rank thru PageRank (PR) and Hyperlink-Induced Topic Search (HITS) procedures.

PR is a measurement, developed by Google, that examines the quantity and quality of links to a website page to decide a relative score of the significance and importance of that page on a scale of 0 to 10.[99] Google characterizes a relation from page X to page Y as voting, from page X to page Y as voting. But Google focuses at much more than the massive number of votes, or links a page gets; for instance, it even analyzes the page casting the vote. Votes cast by pages which are itself "significant" weigh more heavily and contribute to "significant" other pages.

Liu *et al.* used the HITS algorithm to rate web pages which is based on links.[100] HITS Algorithm, invented by Jon Kleinberg, is a link analysis algorithm that scores web pages. This algorithm is utilized to discover and rate web pages related to a specific search by the web link structures. HITS use authorities and hubs to describe the recursive web-page relationships. The collection of highly applicable web pages is called Roots, provided a query to a Search Engine. They are Authorities with potential. Pages that are not really important but that lead to the Root pages are called Hubs. Thus, an Authority is a page linked to by many nodes, while a Node is a page linked to many authorities.

10.3. *Entropy Measures*

Entropy is an efficient tool to define the complication and ambiguity of social impact, thus it has been broadly utilized in social networks.[101] There are two concepts, interaction frequency entropy and friend entropy, to calculate social impact in mobile social networks. Graph entropy is also used to compute social impact in a social network. Transfer entropy is used to calculate the directed causality-based impact. With the purpose of recognizing peer influence, transfer entropy in online social networks is used.

Table 9 presents additional works on influence analysis methods in the area of social media.

Table 9. Review of Influence Analysis Techniques in the Social Media Data (SMD) Literature

Technique	Field of usage
Degree centrality	Wikipedia data analysis, analysis of text blog content.[102]
Closeness centrality	Identification of stakeholders from social media.[103]
Betweenness centrality	Detect the influential spreaders of information in Twitter or Facebook social media.[104]
Eigenvector centrality	Analyzing Twitter data.[105]
Katz centrality	Coverage of IPCC Working Group.[106]
PageRank algorithm	Analysis of social media question / answering domain.[107]
HITS algorithms	Analyzing domestic extremist groups on the Web.[108]
Frequency entropy	Geolocation prediction in social media.[109]
Transfer entropy	Examining the dynamics of individual and group behavior, characterizing patterns of information diffusion in social media, analysis of microblogging time series.[110]
Graph entropy	Sentiment analysis of social media data.[111]

11. Recommendations in Social Networks

To minimize information overloading, many social media websites use recommender systems to propose item recommendations articles that help users interact.[112, 113] Recommender systems typically employ four common recommendation methods, (1) collaborative filtering, (2) content-based filtering, (3) hybrid filtering, and (4) knowledge-based filtering. Collaborative filtering techniques gather and examine a wide range of information about the activities, preferences, or behaviors of users and anticipate what users want based on their resemblance with others.[114] Collaborative filtering does not rely on machine-queryable content and can therefore recommend complicated items, such as films, accurately without needing knowledge of the item itself. Collaborative filtering can be memory-based or model-based. Memory based collaborative filtering is divided into user-based and item-based. Memory based filtering finds the resemblances between user *A* and other users through rating information, then, the first *B* users are chosen as $\underline{A}$'s neighbors. In item-based collaborative filtering, a user's rating matrix is calculated from the resemblances between item *I* and other items, and then the items are sorted with the former M items being chosen as the neighbors of *I*. The Cosine similarity metric can aid creating a model to estimate the rating for the user. This model can be developed utilizing clustering techniques, Bayesian networks, latent semantic or association rule mining. Content based filtering techniques are based on an item description and a user preference profile.[115] Keywords are used to describe items and a user profile is developed to indicate the type of item that the user prefers. These algorithms recommend items similar to those currently selected or previously liked by the user. Particularly, different candidate items are equated to items previously rated by the user and the best matching items are recommend. Latest research has indicated that a hybrid approach that combines collaborative with content-based filtering can be very effective. Several studies have compared the performance of the hybrid method quantitatively with pure collaborative and content-based strategies. They illustrate that the hybrid methods can offer improved recommendations over pure methods. These techniques can also be utilized to resolve cold starting sparsity issues. Knowledge based filtering is a recommendation system that offers explicit information on items, user preferences and criteria for recommendations. It offers the same items that users have previously liked. Recommendation for items are related to the preferences of users.

There can be implicit and explicit feedback techniques in the recommendation system in social networks. Implicit feedback includes different forms of user experiences that are not inherently expected to deliver a deliberate device assessment but can nevertheless be used to infer the positive or negative opinion of the user.[116] Explicit feedback, on the

other hand, can be seen as an intentional and unambiguous quality evaluation by a user.[117] The assessment 's definition often depends on the environment and the context it is given in. For recommender systems, clear feedback is often a rank (numeric) provided by a user for one particular object. There are some challenges of using the implicit recommendation system: (1) In many cases, several types of user interactions must be viewed in parallel and the question arises of how to combine them, (2) Often there are feedback signals both explicit and implicit, but with different degrees of representation of the item space. Then it needs appropriate ways to integrate them, (3) When explicit feedback is present, it can be easier for the consumer to understand the logic of the suggestions given because they can, for example, be utilized more effectively in system-generated explications. Recommendations arising from implicit feedback signals cannot be that logical or apparent to the consumer, (4) Implicit feedback signals are widespread in most domains for the few very common products while feedback for rare pieces can be very limited.

12. Link Prediction in Social Networks

Link prediction is a vital task in the analysis of social networks that also has wide applications including bioinformatics and e-commerce.[118, 119, 120] Link prediction methods use node topology and social theory information to compute node pair resemblances. Node pair similarity predicts links created on a simple concept: the more the pair is similar, the more likely it is to be a link between them - users tend to establish associations with people in religion, education, interests and location. A node has characteristics in a social network (mail address, social network profile, publication record, etc.) which can be directly utilized to calculate the similarity between nodes. Since node attributes are typically textual then string-based and text-based metrics of similarity are generally utilized. Node-based metrics are beneficial for predicting links if characteristics and actions of users can be acquired. Link prediction metrics may be also based on topological information forming topology-based matrices. These system of measurement can be categorized into (1) neighbor-based metrics, (2) path-based metrics, and (3) random-walk-based metrics. It is known that in social networks that users tend to generate new relations with users that are friendlier to them. Many neighbor-based metrics have been designed to predict links such as Common neighbor, Jaccard coefficient, Sørensen Index, Salton Cosine Similarity, Hub promoted, Hub depressed, Leicht-Holme-Nerman, Adamic-Adar coefficient, or Preferential attachment. Path-based metrics calculate node pair similarities using methods such as Local path, Katz metric, Relationship strength similarity, FriendLink or, Vertex collocation profile. Random walk utilizes the probability of transition from a node to its neighbors to specify a random walker's destination from the recent node, such as (1) Hitting time, Commute time, Cosine similarity time, (4) SimRank, Rooted PageRank, or PropFlow.

Recently research has used classic social theories including triadic closure, community, homophily, weak and strong links, and structural equilibrium to address mining and analysis problems. Unlike previous metrics that only use topology and node, social theory-based link prediction metrics can enhance system performance by catching valuable added social communication information, particularly for large scale social media networks. Community information considers user behavior and interest, then predicts likely future links. Homophily has been exploited to obtain not only links amongst a user and their concerned facilities, but also links amongst users with common interests. The triadic closure process finds links by studying choices about friendship, i.e. whether people could choose new friends who are friends of friends and find that friends of friends tend to become friends.

13. Conclusions and Future Works

Pattern mining social media is an emerging research area. It provides challenging tasks and proposes many prospects for future examination to improve a rising necessity for secure and smarter applications. This article has reviewed the works in pattern mining approaches for social media data analysis, together with media data collection, pre-processing of media data, dimensionality reduction, classification, clustering, spam detection, community detection and influence analysis. The works discussed will aid in creating innovative pattern mining solutions to address social media data issues.

Pattern mining is the most desired in this digital world to extract useful pattern / information from enormous volume of unstructured data in various formats such as image, text, audio, video, graphics etc. Researchers face a huge challenge in handling and assessing these kinds of data. As a consequence, several researchers have concentrated on this field of pattern mining. Extracted knowledge is used in many areas including social media to forecast future events or to identify target variable values. Every field requires different kinds of information, and utilize various repositories of data. The data is unstructured, heterogeneous, complex and noisy in large databases. Data velocity and volume are inconsistent. It is therefore impractical to have one system to mine all these types of data. Specific pattern mining systems for the processing of different types of data should be built. Thus, there should be more numbers of pattern mining algorithms to process various data types. Each fraction of a second, enormous amount of data enters to digital world. Social media is

one of the main sources that provide this immense amount of data. Such data are strongly influenced by the characteristics of 5 V's, so it is a very big challenge to manage such results.

Spam recognition and the reliability of social media is an area where improved filtering techniques will lead to more reliable analytical outcomes. Much of the traffic on social media sites are initiated from cell phone devices which frequently give geolocation. Together with the geospatial information, analysis of social media data could produce beneficial new perceptions, enhance the predictive abilities and create a motivating area of research. Lastly, relationship does not mean causation and discovering causative mechanisms would be an additional exciting area of future research. While researchers can deal with data mining and machine learning methods, underlying social assemblies many not be apparent. Design with data scientists and social scientists is likely to be truly mutually beneficial.

References

1. S. Stieglitz, M. Mirbabaie, B. Ross and C. Neuberger, Social media analytics–Challenges in topic discovery, data collection, and data preparation, *International journal of information management.* ***39*** (2018) 156-168. https://doi.org/10.1016/j.ijinfomgt.2017.12.002
2. A. A. Alalwan, N. P. Rana, Y. K. Dwivedi and R. Algharabat, Social media in marketing: A review and analysis of the existing literature, *Telematics and Informatics.* **34** (2017) 1177-1190. https://doi.org/10.1016/j.tele.2017.05.008. [2]
3. S. Pajo, D. Vandevenne and J. R. Duflou, Automated feature extraction from social media for systematic lead user identification, *Technology Analysis & Strategic Management.* **29(6)** (2017) 642-654. https://doi.org/10.1080/09537325.2016.1220517. [5]
4. Z. Xiang, Q. Du, Y. Ma and W. Fan, Assessing reliability of social media data: lessons from mining TripAdvisor hotel reviews, *Information Technology & Tourism.* **18** (2018) 43-59. https://doi.org/10.1007/s40558-017-0098-z. [8]
5. J. Kim and M. Hastak, Social network analysis: Characteristics of online social networks after a disaster, *International Journal of Information Management.* **38** (2018) 86-96. https://doi.org/10.1016/j.ijinfomgt.2017.08.003. [11]
6. A. Rathore, P. Ilavarasan and Y. Dwivedi, Social media content and product co-creation: an emerging paradigm, *Journal of Enterprise Information Management.* **29** (2016) 7-8. https://doi.org/10.1108/JEIM-06-2015-0047. [14]
7. S. Lomborg and A. Bechmann, Using APIs for data collection on social media, *The Information Society.* **30** (2014) 256-265. https://doi.org/10.1080/01972243.2014.915276. [17]
8. W. Fan and M. D. Gordon, The power of social media analytics, *Commun. Acm.* **57** (2014) 74-81. https://doi.org/10.1145/2602574. [20]
9. M. Naaman, Social multimedia: highlighting opportunities for search and mining of multimedia data in social media applications, *Multimedia Tools and Applications.* **56** (2012) 9-34. https://doi.org/10.1007/s11042-010-0538-7. [21]
10. X. Chen, M. Vorvoreanu and K. Madhavan, Mining social media data for understanding students' learning experiences, *IEEE Transactions on Learning Technologies.* **7** (2014) 246-259. https://doi.org/10.1109/TLT.2013.2296520. [24]
11. W. He, S. Zha and L. Li, Social media competitive analysis and text mining: A case study in the pizza industry, *International Journal of Information Management.* **33** (2013) 464-472. https://doi.org/10.1016/j.ijinfomgt.2013.01.001. [27]
12. M. Khan, M. Dickinson and S. Kübler, Does size matter? text and grammar revision for parsing social media data, In *Proceedings of the Workshop on Language Analysis in Social Media.* (2013), pp. 1-10. [30]
13. M. A. Shareef, B. Mukerji, Y. K. Dwivedi, N. P. Rana and R. Islam, Social media marketing: Comparative effect of advertisement sources, *Journal of Retailing and Consumer Services.* **46** (2019) 58-69. https://doi.org/10.1016/j.jretconser.2017.11.001. [33]
14. D. Ruths and J. Pfeffer, Social media for large studies of behavior, *Science.* **346** (2014) 1063-1064. https://doi.org/10.1126/science.346.6213.1063. [36]
15. L. A. Palinkas, S. M. Horwitz, C. A. Green, J. P. Wisdom, N. Duan and K. Hoagwood, Purposeful sampling for qualitative data collection and analysis in mixed method implementation research, *Administration and policy in mental health and mental health services research.* **42** (2015) 533-544. https://doi.org/10.1007/s10488-013-0528-y. [37]
16. F. S. Tsai, Dimensionality reduction techniques for blog visualization, *Expert Systems with Applications.* **38** (2011) 2766-2773. https://doi.org/10.1016/j.eswa.2010.08.067. [40]
17. D. Barbieri, D. Braga, S. Ceri, E. Della Valle, Y. Huang, V. Tresp, A. Rettinger and H. Wermser, Deductive and inductive stream reasoning for semantic social media analytics, *IEEE Intelligent Systems.* **25** (2010) 32-41. https://doi.org/10.1109/MIS.2010.142. [43]
18. N. Cummins, S. Amiriparian, S. Ottl, M. Gerczuk, M. Schmitt and B. Schuller, Multimodal Bag-of-Words for cross domains sentiment analysis, In *2018 IEEE International Conference on Acoustics, Speech and Signal*

Processing (ICASSP). (Calgary, AB, Canada; Sept. 2018), pp. 4954-4958. https://doi.org/10.1109/ICASSP.2018.8462660. [46]

19. T. Niu, S. Zhu, L. Pang and A. El Saddik, Sentiment analysis on multi-view social data, In *International Conference on Multimedia Modeling.* Springer, Cham, eds. Q. Tian, N. Sebe, G. J. Qi, B. Huet, R. Hong, X. Liu (2016), pp. 15-27. https://doi.org/10.1007/978-3-319-27674-8_2. [49]
20. J. K. Lee and E. Kim, Incidental exposure to news: Predictors in the social media setting and effects on information gain online, *Computers in Human Behavior.* **75** (2017) 1008-1015. https://doi.org/10.1016/j.chb.2017.02.018. [52]
21. M. L. Tseng, Using social media and qualitative and quantitative information scales to benchmark corporate sustainability, *Journal of cleaner production.* **142** (2017) 727-738. https://doi.org/10.1016/j.jclepro.2016.05.062. [55]
22. M. R. Balusu, T. Merghani and J. Eisenstein, Stylistic variation in social media Part-of-Speech Tagging, In *Proceedings of the Second Workshop on Stylistic Variation* (New Orleans; Jun. 2018), pp. 11-19. https://doi.org/10.18653/v1/W18-1602. [58]
23. G. C. Kane, M. Alavi, G. J. Labianca and S. Borgatti, What's different about social media networks? A framework and research agenda, *MIS Quarterly, forthcoming.* **38** (2013) 274-304. https://doi.org/10.25300/MISQ/2014/38.1.13. [60]
24. N Padilla-Zea, S Aceto and D Burgos, Social Seducement: Empowering Social Economy Entrepreneurship. The Training Approach, *International Journal of Interactive Multimedia and Artificial Intelligence*. **5(7)** (2019) 135-150. [62]
25. P. Burnap, O. F. Rana, N. Avis, M. Williams, W. Housley, A. Edwards, J. Morgan and L. Sloan, Detecting tension in online communities with computational Twitter analysis, *Technological Forecasting and Social Change.* **95** (2015) 96-108. https://doi.org/10.1016/j.techfore.2013.04.013. [64]
26. F. W. Chen, A. Guevara Plaza and P. Alarcón Urbistondo, Automatically extracting tourism-related opinion from Chinese social media, *Current Issues in Tourism.* **20** (2017) 1070-1087. https://doi.org/10.1080/13683500.2015.1132196. [66]
27. L. Song, R. Y. Lau, R. C. Kwok, K. Mirkovski and W. Dou, Who are the spoilers in social media marketing? Incremental learning of latent semantics for social spam detection, *Electronic commerce research.* **17** (2017) 51-81. https://doi.org/10.1007/s10660-016-9244-5. [68]
28. C. C. Yang, H. Yang, L. Jiang and M. Zhang, Social media mining for drug safety signal detection, In *Proceedings of the 2012 ACM international workshop on Smart health and wellbeing* (2012), pp. 33-40. https://doi.org/10.1145/2389707.2389714. [70]
29. M. Roth, A. Ben-David, D. Deutscher, G. Flysher, I Horn, A. Leichtberg, N. Leiser, Y. Matias and R. Merom, Suggesting friends using the implicit social graph, In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining* (Jul. 2010), pp. 233-242. https://doi.org/10.1145/1835804.1835836. [72]
30. D. Cao, R. Ji, D. Lin and S. Li, A cross-media public sentiment analysis system for microblog, *Multimedia Systems.* **22** (2016) 479-486. https://doi.org/10.1007/s00530-014-0407-8. [74]
31. F. Wang, S. Qi, G. Gao, S. Zhao and X. Wang, Logo information recognition in large-scale social media data, *Multimedia Systems.* **22** (2016) 63-73. https://doi.org/10.1007/s00530-014-0393-x. [76]
32. S. C. Guntuku, L. Qiu, S. Roy, W. Lin and V. Jakhetiya, Do others perceive you as you want them to?: Modeling personality based on selfies, In *Proceedings of the 1st* ACM *international workshop on affect & sentiment in multimedia* (2015), pp. 21-26. https://doi.org/10.1145/2813524.2813528. [78]
33. C. H. Demarty, M. Sjöberg, B. Ionescu, T. T. Do, M. Gygli and N. Duong, Mediaeval 2017 predicting media interestingness task, In *Proc of MediaEval 2017* (2017). [79]
34. B. Bischke, D. Borth, C. Schulze and A. Dengel, Contextual enrichment of remote-sensed events with social media streams, In *Proceedings of the 24th ACM international conference on Multimedia* (Oct. 2016), pp. 1077-1081. https://doi.org/10.1145/2964284.2984063. [80]
35. A. M. El-Gazzar, T. M. Mohamed and R. A. Sadek, A hybrid SVD-HSV visual sentiment analysis system, In IEEE *2017 Eighth International Conference on Intelligent Computing and Information Systems (ICICIS)* (Cairo, Egypt; Jan. 2017), pp. 360-365. https://doi.org/10.1109/INTELCIS.2017.8260063. [82]
36. F. Gelli, T. Uricchio, M. Bertini, A. Del Bimbo and S. F. Chang, Image popularity prediction in social media using sentiment and context features, In *Proceedings of the 23rd ACM international conference on Multimedia* (Oct. 2015), pp. 907-910. https://doi.org/10.1145/2733373.2806361. [84]
37. S. Zhao, H. Yao, Y. Gao, G. Ding and T. S. Chua, Predicting personalized image emotion perceptions in social networks, *IEEE transactions on affective computing.* **9** (2016) 526-540. https://doi.org/10.1109/TAFFC.2016.2628787. [86]
38. A. M. Munar and J. K. Jacobsen, Motivations for sharing tourism experiences through social media, *Tourism management.* **43** (2014) 46-54. https://doi.org/10.1016/j.tourman.2014.01.012. [88]
39. F. Morone, B. Min, L. Bo, R. Mari and H. A. Makse, Collective influence algorithm to find influencers via optimal percolation in massively large social media, *Scientific reports.* **6** (2016) 30062. https://doi.org/10.1038/srep30062. [90]

40. A. Zubiaga, E. Kochkina, M. Liakata, R. Procter, M. Lukasik, K. Bontcheva, T. Cohn and I Augenstein, Discourse-aware rumour stance classification in social media using sequential classifiers, *Information Processing & Management.* **54** (2018) 273-290. https://doi.org/10.1016/j.ipm.2017.11.009. [92]
41. S. Mei, H. Li, J. Fan, X. Zhu and C. R. Dyer, Inferring air pollution by sniffing social media, In *Proceedings of the 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (Beijing, China; Oct. 2014), pp. 534-539. https://doi.org/10.1109/ASONAM.2014.6921638. [93]
42. R. Sudrajat, M. Si, R. Rosadi, S. Si, M. Kom and H. Muhammad, Implementation of data mining in analyzing social media users personality with Naive Bayes classifier: A case study of instagram social media, *International Journal of Computer Science Issues (IJCSI).* **13(4)** (2016) 76. https://doi.org/10.20943/01201604.7682. [94]
43. P. Nakov, S. Rosenthal, S. Kiritchenko, S. M. Mohammad, Z. Kozareva, A. Ritter, V. Stoyanov and X. Zhu, Developing a successful SemEval task in sentiment analysis of Twitter and other social media texts, *Language Resources and Evaluation.* **50** (2016) 35-65. https://doi.org/10.1007/s10579-015-9328-1. [95]
44. S. Yordanova and D. Kabakchieva, Sentiment Classification of Hotel Reviews in Social Media with Decision Tree Learning, *International Journal of Computer Applications.* **158** (2017). https://doi.org/10.5120/ijca2017912806. [96]
45. H. Lin, J. Jia, Q. Guo, Y. Xue, Q. Li, J. Huang, L. Cai and L. Feng, User-level psychological stress detection from social media using deep neural network, In *Proceedings of the 22nd ACM international conference on Multimedia* (Nov. 2014), pp. 507-516. https://doi.org/10.1145/2647868.2654945. [97]
46. L. E. Sherman, A. A. Payton, LM. Hernandez, P. M. Greenfield and M. Dapretto, The power of the like in adolescence: effects of peer influence on neural and behavioral responses to social media, *Psychological science.* **27** (2016) 1027-1035. https://doi.org/10.1177/0956797616645673. [98]
47. S. Poria, E. Cambria, D. Hazarika and P. Vij, A deeper look into sarcastic tweets using deep convolutional neural networks, *arXiv preprint arXiv:1610.08815* (2016). [99]
48. J. Ma, W. Gao, P. Mitra, S. Kwon, B. J. Jansen, K. F. Wong and M. Cha, Detecting rumors from microblogs with recurrent neural networks, In *Ijcai* (2016) 3818-3824. [100]
49. N. Phan, D. Dou, B. Piniewski and D. Kil, Social restricted boltzmann machine: Human behavior prediction in health social networks, In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining.* **2015** (2015) 424-431. https://doi.org/10.1145/2808797.2809307. [101]
50. X. Ma, H. Wang, H. Li, J. Liu and H. Jiang, Exploring sharing patterns for video recommendation on YouTube-like social media, *Multimedia Systems.* **20** (2014) 675-691. https://doi.org/10.1007/s00530-013-0309-1. [102]
51. C. Otto, D. Wang and A. K. Jain, Clustering millions of faces by identity, *IEEE transactions on pattern analysis and machine intelligence.* **40** (2017) 289-303. https://doi.org/10.1109/TPAMI.2017.2679100. [103]
52. S. Coyne, P. Madiraju and J. Coelho, Forecasting Stock Prices Using Social Media Analysis, In *2017 IEEE 15th Intl Conf on Dependable, Autonomic and Secure Computing, 15th Intl Conf on Pervasive Intelligence and Computing, 3rd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/DataCom/CyberSciTech)* (Orlando, FL, USA; Apr. 2017), pp. 1031-1038. https://doi.org/10.1109/DASC-PICom-DataCom-CyberSciTec.2017.169. [105]
53. G. Wang, D. Zheng, S. Yang and J. Ma, FCE-SVM: a new cluster based ensemble method for opinion mining from social media, *Information Systems and e-Business Management.* **16(4)** (2018) 721-742. https://doi.org/10.1007/s10257-017-0352-0. [107]
54. L. Chen, T. Gong, M. Kosinski, D. Stillwell and R. L. Davidson, Building a profile of subjective well-being for social media users, *PloS one.* **12** (2017) e0187278. https://doi.org/10.1371/journal.pone.0187278. [109]
55. L. E. Sherman, L. M. Hernandez, P. M. Greenfield and M. Dapretto, What the brain 'Likes': neural correlates of providing feedback on social media, *Social cognitive and affective neuroscience.* **13(7)** (2018) 699-707. https://doi.org/10.1093/scan/nsy051. [111]
56. J. Leng and P. Jiang, A deep learning approach for relationship extraction from interaction context in social manufacturing paradigm, *Knowledge-Based Systems.* **100** (2016) 188-199. https://doi.org/10.1016/j.knosys.2016.03.008. [113]
57. S. Poria, E. Cambria and A. Gelbukh, Aspect extraction for opinion mining with a deep convolutional neural network, *Knowledge-Based Systems.* **108** (2016) 42-49. https://doi.org/10.1016/j.knosys.2016.06.009. [115]
58. J. Xie, X. Liu and D. Dajun Zeng, Mining e-cigarette adverse events in social media using Bi-LSTM recurrent neural network with word embedding representation, *Journal of the American Medical Informatics Association.* **25** (2017) 72-80. https://doi.org/10.1093/jamia/ocx045. [117]
59. M. Fatemi and M. Safayani, Joint sentiment/topic modeling on text data using a boosted restricted Boltzmann Machine, *Multimedia Tools and Applications.* **78** (2019) 20637-20653. arxiv.org/abs/1711.03736. [119]
60. M. Del Vicario, A. Bessi, F. Zollo, F. Petroni, A. Scala, G. Caldarelli, H. E. Stanley and W. Quattrociocchi, The spreading of misinformation online, *Proceedings of the National Academy of Sciences.* **113** (2016) 554-559. https://doi.org/10.1073/pnas.1517441113. [121]
61. C. C. Yang and T. D. Ng, 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** (2011) 1144-1155. https://doi.org/10.1109/TSMCA.2011.2113334. [124]
62. A. Shepitsen, J. Gemmell, B. Mobasher and R. Burke, Personalized recommendation in social tagging systems using hierarchical clustering, In *Proceedings of the 2008 ACM conference on Recommender systems* (Oct. 2008), pp. 259-266. https://doi.org/10.1145/1454008.1454048. [125]

63. H. Becker, M. Naaman and L. Gravano, Learning similarity metrics for event identification in social media, In *Proceedings of the third ACM international conference on Web search and data mining* (Feb. 2010), pp. 291-300. https://doi.org/10.1145/1718487.1718524. [126]
64. M. A. Smith, B. Shneiderman, N. Milic-Frayling, E. Mendes Rodrigues, V. Barash, C. Dunne, T. Capone, A. Perer and E. Gleave, Analyzing (social media) networks with NodeXL, In ACM *Proceedings of the fourth international conference on Communities and technologies* (Jun. 2009), pp. 255-264. https://doi.org/10.1145/1556460.1556497. [127]
65. B. Zhao, D. Z. Sui and Z. Li, Visualizing the gay community in Beijing with location-based social media, *Environment and planning A.* **49** (2017) 977-979. https://doi.org/10.1177/0308518X16685885. [128]
66. D. Wu, J. Shi and N. Mamoulis, Density-Based Place Clustering Using Geo-Social Network Data, *IEEE Transactions on Knowledge and Data Engineering.* **30** (2017) 838-851. https://doi.org/10.1109/TKDE.2017.2782256. [130]
67. C. R. Sunstein, *# Republic: Divided democracy in the age of social media*, Princeton University Press (2018). [132]
68. A. Javari and M. Jalili, Cluster-based collaborative filtering for sign prediction in social networks with positive and negative links, *ACM Transactions on Intelligent Systems and Technology (TIST).* **5** (2014) 24. https://doi.org/10.1145/2501977. [134]
69. J. Leskovec and J. J. Mcauley, Learning to discover social circles in ego networks, In *Advances in neural information processing systems* (2012) 539-547. [136]
70. Y. Hu, SD. Farnham and A Monroy-Hernández, Whoo. ly: Facilitating information seeking for hyperlocal communities using social media, In *Proceedings of the* ACM *SIGCHI Conference on Human Factors in Computing Systems* (Apr. 2013), pp. 3481-3490. https://doi.org/10.1145/2470654.2466478. [138]
71. H. Aldrich and P. Kim, Small worlds, infinite possibilities? How social networks affect entrepreneurial team formation and search, *Strategic Entrepreneurship Journal.* **1** (2007) 147-165. https://doi.org/10.1002/sej.8. [140]
72. M. Madden, A. Lenhart, S. Cortesi, U. Gasser, M. Duggan, A. Smith and M. Beaton, Teens, social media, and privacy, *Pew Research Center.* **21** (2013) 2-86. [142]
73. H. Gao, J. Hu, C. Wilson, Z. Li, Y. Chen and B. Y. Zhao, Detecting and characterizing social spam campaigns, In *Proceedings of the 10th ACM SIGCOMM conference on Internet measurement* (Nov. 2010), pp. 35-47. https://doi.org/10.1145/1879141.1879147. [144]
74. B. Wang, A. Zubiaga, M. Liakata and R. Procter, Making the most of tweet-inherent features for social spam detection on twitter, *In proc of WWW Workshop on Making Sense of Microposts* (2015). [145]
75. Y. Zhang and M. Pennacchiotti, Recommending branded products from social media, In *Proceedings of the 7th ACM conference on Recommender systems* (Oct. 2013), pp. 77-84. https://doi.org/10.1145/2507157.2507170. [147]
76. S. Sedhai and A. Sun, Semi-supervised spam detection in Twitter stream, *IEEE Transactions on Computational Social Systems.* **5** (2017) 169-175. https://doi.org/10.1109/TCSS.2017.2773581. [150]
77. H. Xie, X. Li, T. Wang, L. Chen, K. Li, FL. Wang, Y. Cai, Q. Li and H. Min, Personalized search for social media via dominating verbal context, *Neurocomputing.* **172** (2016) 27-37. https://doi.org/10.1016/j.neucom.2014.12.109. [153]
78. S. Sharmin and Z. Zaman, Spam detection in social media employing machine learning tool for text mining, In *2017 13th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)* (Jaipur; Apr. 2017), pp. 137-142. https://doi.org/10.1145/10.1109/SITIS.2017.32. [156]
79. C. Zhang, C. Fan, W. Yao, X. Hu and A. Mostafavi, Social media for intelligent public information and warning in disasters: An interdisciplinary review, *International Journal of Information Management.* **49** (2019) 190-207. https://doi.org/10.1016/j.ijinfomgt.2019.04.004. [159]
80. F. Zhao and A. K. Tung, Large scale cohesive subgraphs discovery for social network visual analysis, *Proceedings of the VLDB Endowment.* **6** (2012) 85-96. https://doi.org/10.14778/2535568.2448942. [161]
81. S. Papadopoulos, Y. Kompatsiaris, A. Vakali and P. Spyridonos, Community detection in social media, *Data Mining and Knowledge Discovery.* **24** (2012) 515-554. https://doi.org/10.1007/s10618-011-0224-z. [164]
82. J. Shi and J. Malik, Normalized cuts and image segmentation, *IEEE Transactions on Pattern Analysis and Machine Intelligence.* **22** (2000) 107. https://doi.org/10.1109/34.868688. [167]
83. S. Fortunato, V. Latora and M. Marchiori, Method to find community structures based on information centrality, *Physical review E.* **70** (2004) 056104. https://doi.org/10.1103/PhysRevE.70.056104. [170]
84. I. X. Leung, P. Hui, P. Lio and J. Crowcroft, Towards real-time community detection in large networks, *Physical Review E.* **79** (2009) 066107. https://doi.org/10.1103/PhysRevE.79.066107. [173]
85. M. Plantié and M. Crampes, Survey on social community detection, In *Social media retrieval.* Springer, London, eds. N. Ramzan, R. van Zwol, J. S. Lee, K. Clüver, X. S. Hua. (2013), pp. 65-85. https://doi.org/10.1007/978-1-4471-4555-4_4. [175]
86. F. Stahl, M. M. Gaber and M. Adedoyin-Olowe, A survey of data mining techniques for social media analysis, *Journal of Data Mining & Digital Humanities 2014* (2014). [178]
87. M. G. Gong, L. J. Zhang, J. J. Ma and L. C. Jiao, Community detection in dynamic social networks based on multiobjective immune algorithm, *Journal of Computer Science and Technology.* **27(3)** (2012) 455-467. https://doi.org/10.1007/s11390-012-1235-y. [181]

88. L. Weng, F. Menczer and Y. Y. Ahn, Virality prediction and community structure in social networks, *Scientific reports.* **3** (2013) 2522. https://doi.org/10.1038/srep02522. [184]
89. J. Xie, S. Kelley and B. K. Szymanski, Overlapping community detection in networks: The state-of-the-art and comparative study, *Acm computing surveys (csur).* **45** (2013) 43. https://doi.org/10.1145/2501654.2501657. [187]
90. D. C. Cercel and S. Trausan-Matu, Opinion propagation in online social networks: A survey, In *Proceedings of the 4th* ACM *International Conference on Web Intelligence, Mining and Semantics (WIMS14)* (Jun. 2014), pp. 1-10. https://doi.org/10.1145/2611040.2611088. [190]
91. S. P. Borgatti, Centrality and network flow, *Social networks.* **27** (2005) 55-71. https://doi.org/10.1016/j.socnet.2004.11.008. [191]
92. M. E. Newman, Analysis of weighted networks, *Physical review E.* **70** (2004) 056131. https://doi.org/10.1103/PhysRevE.70.056131. [193]
93. B. Wellman, The development of social network analysis: A study in the sociology of science, *Contemporary Sociology.* **37** (2008) 221. https://doi.org/10.1177/009430610803700308. [194]
94. S. P. Borgatti, K. M. Carley and D. Krackhardt, On the robustness of centrality measures under conditions of imperfect data, *Social networks.* **28** (2006) 124-136. https://doi.org/10.1016/j.socnet.2005.05.001. [195]
95. T. Opsahl, F. Agneessens and J. Skvoretz, Node centrality in weighted networks: Generalizing degree and shortest paths, *Social networks.* **32** (2010) 245-251. https://doi.org/10.1016/j.socnet.2010.03.006. [196]
96. T. L. Frantz, M. Cataldo and K. M. Carley, Robustness of centrality measures under uncertainty: Examining the role of network topology, *Computational and Mathematical Organization Theory.* **15(4)** (2009) 303. https://doi.org/10.1007/s10588-009-9063-5. [197]
97. K. Okamoto, W. Chen and X. Y. Li, Ranking of closeness centrality for large-scale social networks, In *International Workshop on Frontiers in Algorithmics.* Springer, Berlin, Heidelberg, eds. F. P. Preparata, X. Wu, J. Yin (2008), pp. 186-195. https://doi.org/10.1007/978-3-540-69311-6_21. [198]
98. C. Kiss and M. Bichler, Identification of influencers—measuring influence in customer networks, *Decision Support Systems.* **46** (2008) 233-253. https://doi.org/10.1016/j.dss.2008.06.007. [199]
99. J. M. Kleinberg, Authoritative sources in a hyperlinked environment, *Journal of the ACM (JACM).* **46(5)** (1999) 604-632. https://doi.org/10.1145/324133.324140. [200]
100. Q. Liu, B. Xiang, N. J. Yuan, E. Chen, H. Xiong, Y. Zheng and Y. Yang, An influence propagation view of pagerank, *ACM Transactions on Knowledge Discovery from Data (TKDD).* **11(3)** (2017) 30. https://doi.org/10.1145/3046941. [202]
101. S. Peng, J. Li and A. Yang, Entropy-based social influence evaluation in mobile social networks, In *International Conference on Algorithms and Architectures for Parallel Processing.* Springer, Cham, eds. G. Wang, A. Zomaya, G. Martinez, K. Li (Dec. 2015), pp. 637-647. https://doi.org/10.1007/978-3-319-27119-4_44. [203]
102. P. Melville, V. Sindhwani and R. Lawrence, Social media analytics: Channeling the power of the blogosphere for marketing insight, *Proc. of the WIN.* **1** (2009) 1-5. https://doi.org/10.1.1.157.3485. [205]
103. K. Sedereviciute and C. Valentini, Towards a more holistic stakeholder analysis approach. Mapping known and undiscovered stakeholders from social media, *International Journal of Strategic Communication.* **5** (2011) 221-239. https://doi.org/10.1080/1553118X.2011.592170. [207]
104. S. Pei, L. Muchnik, J. S. Andrade Jr, Z. Zheng and H. A. Makse, Searching for superspreaders of information in real-world social media, *Scientific reports.* **4** (2014) 5547. https://doi.org/10.1038/srep05547. [209]
105. O. Oh, K. H. Kwon and H. R. Rao, An Exploration of Social Media in Extreme Events: Rumor Theory and Twitter during the Haiti Earthquake 2010, In *Icis.* **231** 7332-7336. [211]
106. S. O'Neill, H. T. Williams, T. Kurz, B. Wiersma and M. Boykoff, Dominant frames in legacy and social media coverage of the IPCC Fifth Assessment Report, *Nature Climate Change.* **5(4)** (2015) 380. https://doi.org/10.1038/nclimate2535. [213]
107. H. Kwak, C. Lee, H. Park and S. Moon, What is Twitter, a social network or a news media?, In ACM *Proceedings of the 19th international conference on World wide web* (Apr. 2010), pp. 591-600. https://doi.org/10.1145/1772690.1772751. [215]
108. Y. Zhou, E. Reid, J. Qin, H. Chen and G. Lai, US domestic extremist groups on the Web: link and content analysis, *IEEE intelligent systems.* **20** (2005) 44-51. https://doi.org/10.1109/MIS.2005.96 [217]
109. B. Han, P. Cook and T. Baldwin, Geolocation prediction in social media data by finding location indicative words, In *Proceedings of COLING 2012* (Apr. 2012), pp. 1045-1062. [219]
110. J. Borge-Holthoefer, N. Perra, B. Gonçalves, S. González-Bailón, A. Arenas, Y. Moreno and A. Vespignani, The dynamics of information-driven coordination phenomena: A transfer entropy analysis, *Science advances.* **2** (2016) e1501158. https://doi.org/10.1145/2072609.2072614. [221]
111. L. Zhu, A. Galstyan, J. Cheng and K. Lerman, Tripartite graph clustering for dynamic sentiment analysis on social media, In *Proceedings of the 2014 ACM SIGMOD international conference on Management of data* (Jun. 2014), pp. 1531-1542. https://doi.org/10.1145/2588555.2593682. [223]
112. S. Khater, D. Gračanin and H. G. Elmongui, Personalized recommendation for online social networks information: Personal preferences and location-based community trends, *IEEE Transactions on Computational Social Systems.* **4** (2017) 104-120. https://doi.org/10.1109/TCSS.2017.2720632. [225]
113. B. Yang, Y. Lei, J. Liu and W. Li, Social collaborative filtering by trust, *IEEE transactions on pattern analysis and machine intelligence.* **39** (2016) 1633-1647. https://doi.org/10.1109/TPAMI.2016.2605085. [226]

114. T. Ma, J. Zhou, M. Tang, Y. Tian, A. Al-Dhelaan, M. Al-Rodhaan and S. Lee, Social network and tag sources based augmenting collaborative recommender system, *IEICE transactions on Information and Systems*. **98** (2015) 902-910. https://doi.org/10.1587/transinf.2014EDP7283. [227]
115. S. Wei, X. Zheng, D. Chen and C. Chen, A hybrid approach for movie recommendation via tags and ratings, *Electronic Commerce Research and Applications*. **18** (2016) 83-94. https://doi.org/10.1016/j.elerap.2016.01.003. [229]
116. S. Mandal, and A. Maiti, Explicit feedback meet with implicit feedback in GPMF: a generalized probabilistic matrix factorization model for recommendation, *Applied Intelligence*. (2020) 1-24. https://doi.org/10.1007/s10489-020-01643-1. [230]
117. F. Villarroel Ordenes, S. Ludwig, K. De Ruyter, D. Grewal, and M. Wetzels, Unveiling what is written in the stars: Analyzing explicit, implicit, and discourse patterns of sentiment in social media, *Journal of Consumer Research*. **43** (2017), 875-894. https://doi.org/10.1093/jcr/ucw070. [234]
118. Y. Zhao, Y. J. Wu, E. Levina and J. Zhu, Link prediction for partially observed networks, *Journal of Computational and Graphical Statistics*. **26** (2017) 725-733. https://doi.org/10.1080/10618600.2017.1286243. [237]
119. K. Chakraborty, S. Bhattacharyya, and R. Bag, A Survey of Sentiment Analysis from Social Media Data, *IEEE Transactions on Computational Social Systems*. **7** (2020) 450-464. https://doi.org/10.1109/TCSS.2019.2956957
120. N. N. Daud, S. H. Ab Hamid, M. Saadoon, F. Sahran, and N. B. Anuar, Applications of link prediction in social networks: A review, *Journal of Network and Computer Applications*. **166** (2020) 102716. https://doi.org/10.1016/j.jnca.2020.102716