Neurocomputing 214 (2016) 654–670

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Data mining techniques in social media: A survey

MohammadNoor Injadat [a], Fadi Salo [a], Ali Bou Nassif [b,*]

[a] Department of Electrical and Computer Engineering, University of Western Ontario, 1151 Richmond St, London, Ontario, Canada N6A 3K7
[b] Department of Electrical and Computer Engineering, University of Sharjah, Sharjah, United Arab Emirates

ARTICLE INFO

Article history:
Received 25 September 2015
Received in revised form
3 June 2016
Accepted 14 June 2016
Communicated by Dr. Guan Ziyu
Available online 23 June 2016

Keywords:
Data mining
Social media
Social media networks analysis
Survey

ABSTRACT

Today, the use of social networks is growing ceaselessly and rapidly. More alarming is the fact that these networks have become a substantial pool for unstructured data that belong to a host of domains, including business, governments and health. The increasing reliance on social networks calls for data mining techniques that is likely to facilitate reforming the unstructured data and place them within a systematic pattern. The goal of the present survey is to analyze the data mining techniques that were utilized by social media networks between 2003 and 2015. Espousing criterion-based research strategies, 66 articles were identified to constitute the source of the present paper. After a careful review of these articles, we found that 19 data mining techniques have been used with social media data to address 9 different research objectives in 6 different industrial and services domains. However, the data mining applications in the social media are still raw and require more effort by academia and industry to adequately perform the job. We suggest that more research be conducted by both the academia and the industry since the studies done so far are not sufficiently exhaustive of data mining techniques.

1. Introduction

Undoubtedly, the world is shrinking into a small village owing to the tangible influence of social media. It connects people from different parts of the world, ages, and nationalities and allows them to share their opinions, experiences, feelings, hobbies, pictures, and videos. This has opened the door for public and private organizations from all domains to promote, benefit, analyze, learn, and improve their organizations based on the data provided in social media. Thus, the significance of social media for academia and industry is quite conspicuous in the amount of research done by these two sectors, seeking answers to pivotal questions.

The structure of the social media data is unorganized and is displayed in different forms such as: text, voice, images, and videos [1]. Moreover, the social media provides an enormous amount of continuous real time data that makes traditional statistical methods unsuitable to analyze this massive data [2]. Therefore, the data mining techniques can play an important role in overcoming this problem.

In spite of the large number of empirical research about data mining techniques and social media, a scant number of studies compare data mining techniques in terms of accuracy, performance, and suitability. For instance, it was observed that the accuracy of certain machine learning techniques is calculated in various methods which makes it difficult to find answers to the suitability of the data mining techniques.

Many researchers have selected their data mining techniques based solely on expert judgment (A31, A56). Few surveys have been conducted in this area without giving full justification for using data mining techniques in social media [3,4]. However, some studies discussed certain areas in the used data mining techniques in social media. In [5], Vuori, et al., discussed the information gathering and knowledge and information sharing through social media for companies. In [6], Rafeeque, et al., the work and challenges related to short text analysis have been reviewed. Akin to this study, [7], Tsytsarau, et al., reviewed the opinion mining and sentiment analysis development, providing a summary about the proposed methods of contradiction analysis. In [8], Gole, et al., discussed mining big data in social media and its challenges as a result of big data features such as: Volume, Velocity, Variety, Veracity and Value.

To the best of our knowledge, there is no previous study that systematically concentrates on the implemented data mining techniques in social media research, which has triggered the idea of the present survey. The review presented in this paper discusses the published research in the period from January 1, 2003 to January 7, 2015. The goal of this study is to probe the available articles with regards to: (I) the data mining techniques used to extract social media data, (II) the research area that requires mining data from social media, (III) a comparison between machine learning and non-machine learning data mining techniques,

[*] Corresponding author.
E-mail addresses: minjadat@uwo.ca (M. Injadat), fsalo@uwo.ca (F. Salo), anassif@sharjah.ac.ae (A.B. Nassif).

http://dx.doi.org/10.1016/j.neucom.2016.06.045
0925-2312/© 2016 Elsevier B.V. All rights reserved.

(IV) a comparison between different data mining techniques, and (V) the strength and weakness of the recommended data mining techniques in social media.

This manuscript is divided into five sections. Section 2 explains the implemented methodology. Section 3 describes our findings. Section 4 discusses the limitation of this review. Finally, Section 5 presents our findings, recommendations, and future work.

2. Methodology

In this review, we conducted a survey based on the Systematic Literature Review (SLR) proposed by Kitchenham and Charters [9] methodology which consists of: planning, conducting, and reporting phases where each phase consists of several stages. At the planning phase we created a review protocol which consists of six stages: specifying research questions, designing the search strategy, identifying the study selection procedures, specifying the quality assessment rules, detailing the data extraction strategy, and synthesizing the extracted data. Fig. 1 shows the review protocol stages.

The research questions have been specified based on the objectives of this review. At the next stage, we designed the search strategy referring to the first stage to retrieve the required and related articles. We also identified the search terms and article selection process, which is required for an accurate search. Stage three covered the selection criteria which specify the inclusion and exclusion rules; we also included more related articles from the references in the articles we used to enrich our literature resources related to the research questions. Stage four included the quality questions to filter the related articles. In stage five, we described the extraction strategy used to obtain the required data which could answer the research questions. Finally, in the last stage, we identified the methodologies used to synthesize the extracted data.

As indicated by Kitchenham and Charters [9], the review protocol is considered to be a critical element of any SLR. Therefore, to avoid researcher bias and to ensure the quality of the review protocol, regular meetings have continued between the authors.

The following Sections 2.1–2.6 will illustrate in detail the review protocol followed in this review.

2.1. Research questions

Summarizing and providing evidence of implementing the data mining techniques in social media is our main goal in this work. Thus, we identified the following five research questions (RQs):

1. RQ1: Which data mining techniques have been used in Social Media?
   The role of this question is to specify the data mining techniques that were implemented in mining social network data.
2. RQ2: In which research areas have data mining techniques been applied?
   The aim of this question is to identify the domains where the data mining techniques were applied and the research objectives among these domains. The most frequent domain will be identified as well as any new domains suggested.
3. RQ3: Do machine learning perform better than non- machine learning in data mining techniques?
   RQ3 compares machine learning and non-machine learning methods implemented in mining social media in term of accuracy. Few articles made a comparison between machine learning and non-machine learning methods. As mentioned in [10,11], only statistical techniques were considered as non-machine learning, whereas the other computational techniques are considered as machine learning methods.
4. RQ4: Is there any comparison that has been performed among different data mining techniques?
   The aim of RQ4 is to specify the data mining technique with high performance. The results produced by the answer of this question will be considered as evidence of the recommended techniques.
5. RQ5: What are the strengths and weaknesses of the implemented data mining techniques in social media?

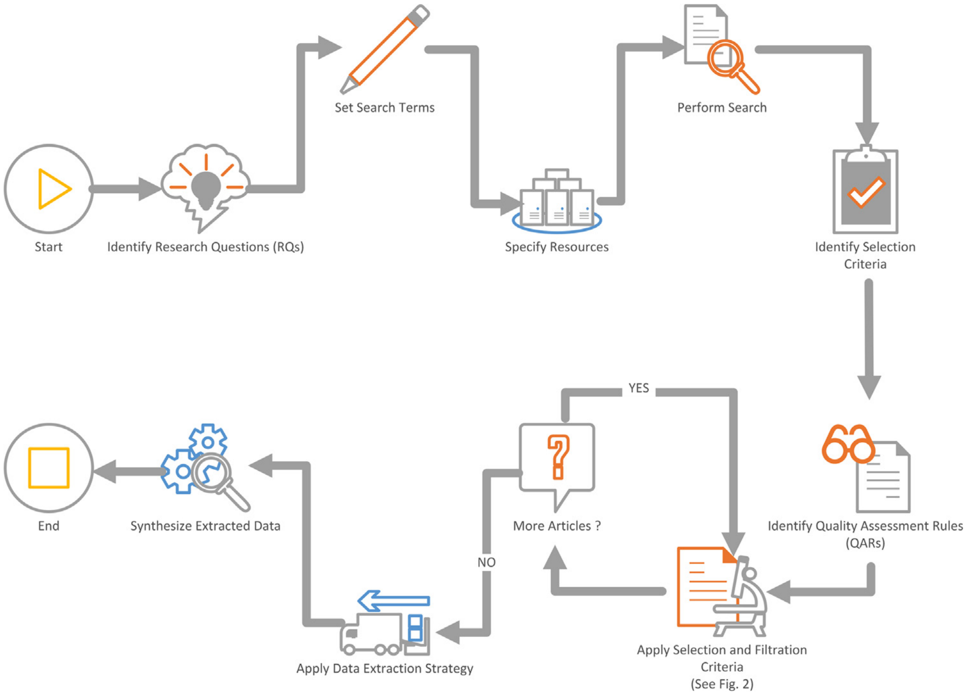
Start
Identify Research Questions (RQs)
Set Search Terms
Specify Resources
Perform Search
Identify Selection Criteria
Identify Quality Assessment Rules (QARs)
Apply Selection and Filtration Criteria (See Fig. 2)
More Articles ?
YES
NO
Apply Data Extraction Strategy
Synthesize Extracted Data
End

Fig. 1. Review protocol stages.

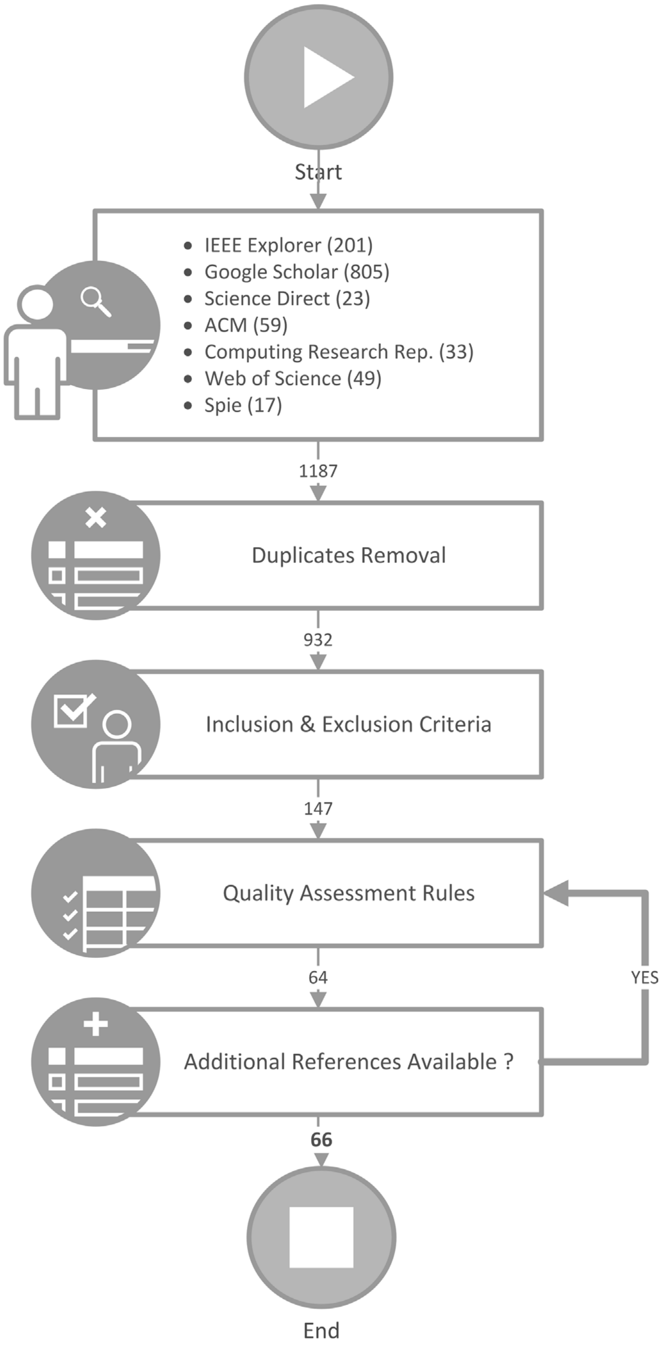

Fig. 2. Search and selection process.

This question will prove the suitable practice of the selected data mining techniques in social media such as text mining, media mining, content-based mining, context-aware mining, graph data mining, and multimedia mining.

2.2. Search strategy

The search strategy that we followed in this survey is explained in detail as follows:

2.2.1. Search terms

To construct the search terms we followed the following procedure [9]:

1. The main terms have been concluded from the research questions.
2. We defined new terms which replace the main terms: such as jargon, alternative spellings, and synonyms.
3. The top ten data mining algorithms were selected from published papers and books [12,13].
4. We used Boolean search operators (ANDs and ORs) to limit the search results in addition to “” for specific phrases.

We included in our search terms the top ten data mining techniques identified by [12,13]. Fig. 1 shows the stages of the review protocol.

The search terms used to retrieve the related publications are as follows. Note that different search terms have been used to get more related publications. The last search date was conducted on January 9, 2015.

- “data mining” AND “techniques” OR “technique” AND “social media”.
- “data mining” AND “machine learning” AND “social media”.
- “social media” AND “fuzzy” AND “data mining”.
- “social media” OR “social network” AND (“C4.5” OR “J48” OR “K-Means” OR “SVM” OR “support vector machines” OR “Apriori” OR “EM” OR “expectation maximization” OR “PageRank” OR “AdaBoost” OR “KNN” OR “k-NN” OR “k-nearest neighbors” OR “Naive Bayes” OR “CART”).

2.2.2. Survey resources

The following digital libraries were searched for the required articles:

- IEEE Explorer
- Google Scholar
- Science Direct
- ACM Digital Library
- Computing Research Repository
- Web of Science
- Spie

The first search process included journals, and Tier I social network related conferences, such as International Conference on Advances in Social Networks Analysis and Mining (ASONAM), ACM Conference on Online Social Networks (COSN), International World Wide Web Conference (WWW), and International Conference on Data Engineering (ICDE), from the above mentioned digital libraries. The search terms considered cover any part of the articles (metadata) and were restricted to articles published between January, 2003 and 2015, because the most popular social networks (Facebook, Twitter, LinkedIn, and MySpace) began after 2002 [14].

2.2.3. Search phases

We used the specified search terms to retrieve the primary related articles from these digital libraries. Moreover, a quick scan of the reference from the paper we selected helped to enrich the resources to answer the research questions. The inclusion criteria are explained in detail in Section 2.3.

The Google document platform was used to share and manage the search results and documents among authors. Based on the inclusion criteria, 147 relevant publications were chosen as candidate publications: 83 journal papers, 64 conference papers. Fig. 2 illustrates the breakdown of the identified articles at each search and selection phase.

2.3. Study selection

We obtained 1187 articles in the first search process. Because many articles did not provide sufficient information to answer the research questions, we performed another filtration step (see Fig. 2).

The filtration process was conducted individually by the authors and the results were discussed in scheduled meetings to ensure the accuracy and to resolve any differences. The selection and filtration steps are explained below:

1. Step 1: remove the duplicated articles obtained by authors and/or different libraries.
2. Step 2: apply inclusion and exclusion criteria to the candidate papers to avoid any irrelevant articles.
3. Step 3: apply the quality assessment rules to include the qualified articles that give the best answers to the research questions.
4. Step 4: search for additional related articles from the article references obtained from step 3 and repeat step 3 on the extra articles.

The inclusion and exclusion criteria applied in this survey are defined below:

Inclusion criteria:

- Use data mining techniques in social media.
- Use machine learning and non-machine learning data mining techniques in social media.
- Comparative studies that compare among data mining techniques.
- Comparative studies that compare between data mining and non-data mining techniques.
- Consider the latest edition of the article of the same research (if different versions are available).
- Consider only articles published between January 2003 and 2015.

Exclusion criteria:

- Exclude articles that include data mining that is not related to social media.
- Exclude articles that do not include data mining but are related to social media.
- Exclude non-journal and non-conferences articles.

Finally, after applying all filtration steps, 66 articles were considered as the resources for this review. The selected articles are listed in Appendix (A), Table A1.

2.4. Quality Assessment Rules (QARs)

The QARs were applied in the selected studies to evaluate article suitability in accordance with the research questions. Ten QARs were identified, and each one is worth 1 mark out of 10. Each QAR is scored as follows: “fully answered”=1, “above average”=0.75, “average”=0.5, “below average”=0.25, “not answered”=0. The overall score of the article will be the summation of the marks obtained for the 10 QARs. If the result was 5 or higher, the article was considered; otherwise it was excluded.

1. QAR1: Are the research objectives clearly defined?
2. QAR2: Is the data mining background clearly addressed?
3. QAR3: Are the data mining techniques used clearly defined?
4. QAR4: Is the design of the experiment suitable and acceptable?
5. QAR5: Is the study performed on sufficient social media data?

Table 1
Data extraction form.

Article ID
Data extractor
Data checker
Publication year
Authors
Article source
Article title
Article type
Domain
RQ1
RQ2
RQ3
RQ4
RQ5

Table 2
Selected articles' types distribution.

Article type	Freq.
Case study	4
Experiment	60
Survey	2
Grand total	66

Table 3
Candidate articles' quality distribution.

Calcification criteria	Freq.	%
Between 0 and 2.5	53	36
Between 2.75 and 4.75	28	19
Between 5 and 6.75	35	24
Between 7 and 8.5	22	15
Between 8.75 and 10	9	6
Grand total	147	100

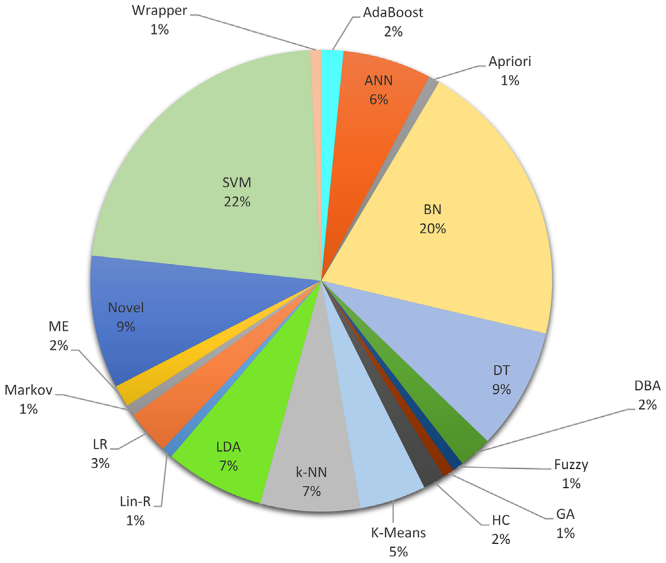

Fig. 3. Data mining techniques among selected papers.

6. QAR6: Is the data mining technique measured and reported?
7. QAR7: Is the proposed data mining technique compared with other techniques?
8. QAR8: Are the conclusions of the experiment clearly identified and reported?
9. QAR9: Are the methods used to analyze the results appropriate?
10. QAR10: Does the experiment enrich academia or industry?

Table 4
Data mining techniques frequencies among articles.

Technique	Frequencies	Technique	Frequencies
AdaBoost	2	k-NN	9
ANN	8	LDA	9
Apriori	1	Lin-R	1
BN	26	LR	4
DT	11	Markov	1
DBA	3	ME	2
Fuzzy	1	Novel	12
GA	1	SVM	29
HC	2	Wrapper	1
K-Means	6		

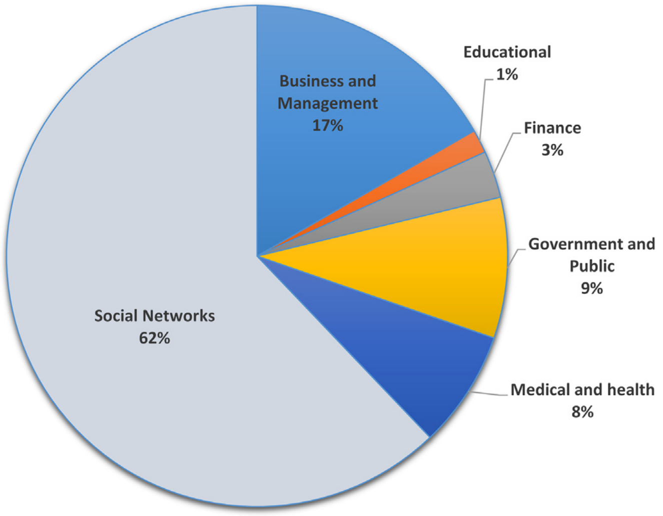

Fig. 4. Domains among articles.

The scores that resulted from applying the QARs on the selected articles are shown in Appendix (A), Table A2.

2.5. Data extraction strategy

In this stage, we explored the articles selected to extract the information required to answer the research questions. Therefore, we have designed an extraction form (see Table 1) to extract the needed data [9].

Based on the extraction form, two authors played the role of extraction and checking. In case of a disagreement between the extractor and checker, group meetings were conducted between all authors to resolve any issue.

Some difficulties occurred during the extraction process. For instance, different terminology was used for the same data mining technique such as C4.5 algorithm is the new name of the J48 technique [15]; however, the WEKA tool (which is commonly used by researchers) is still using the old name J48 (A26). Moreover, some articles used different abbreviations of the same technique such as: KNN, K-NN, Nearest Neighbor (A12, A34), Naïve Bayes, Naive Bayes, NB (A2, A37). Furthermore, many researchers were comparing between their techniques and other common techniques without mentioning technique names or, if mentioned, the reason behind picking certain technique (A31, A42, A53, A55).

Not all selected articles answered all the five RQs. Appendix (A), Table A3 illustrate the RQs that were answered by each selected study.

2.6. Synthesis of extracted data

To synthesize the data extracted from the selected articles, we used different procedures to aggregate evidence that will answer the RQs. The following explains the synthesis procedure we followed in detail:

For RQ1 and RQ2, we used the narrative synthesis method [9] were the extracted information was tabulated according to RQ1 and RQ2.

For the data extracted (quantitative) in RQ3 and RQ4, which came from different articles that have various accuracy calculation techniques, we used binary outcomes to measure the results, which are demonstrated in a comparable way [9].

In RQ5, the strengths and weaknesses of the data mining techniques have the same meaning but are written in different ways. Therefore, to unify these points, we followed the reciprocal translation method [9] which is considered as one of the

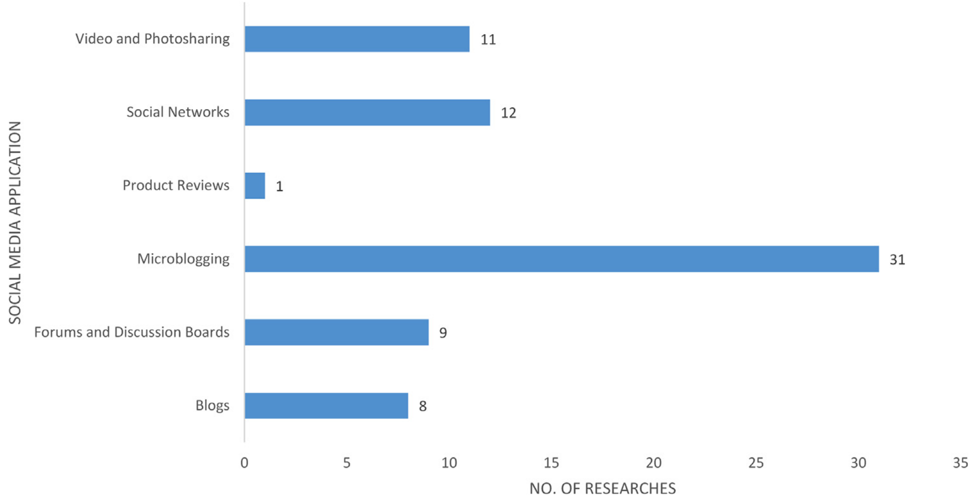

Fig. 5. Popularity of various social media application in researches.

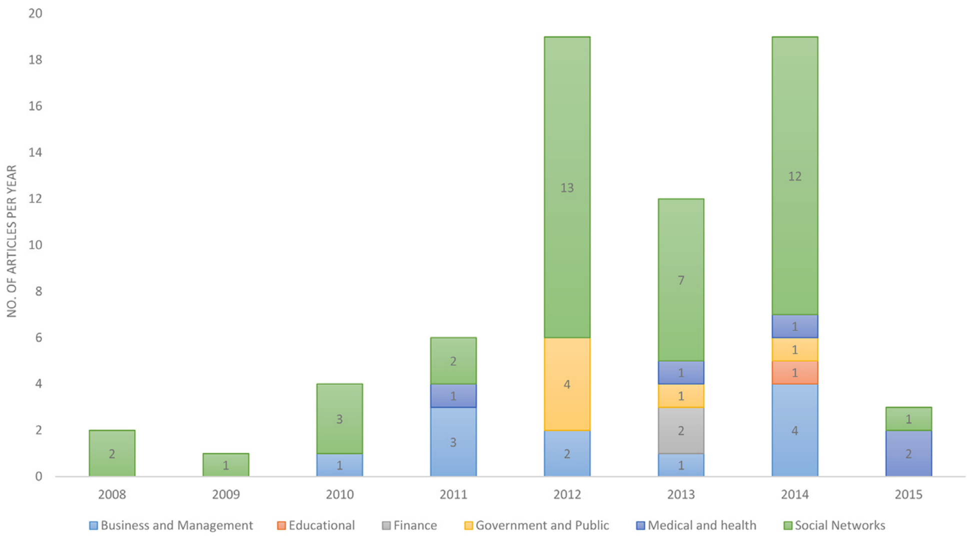

Fig. 6. Domains distribution per year.

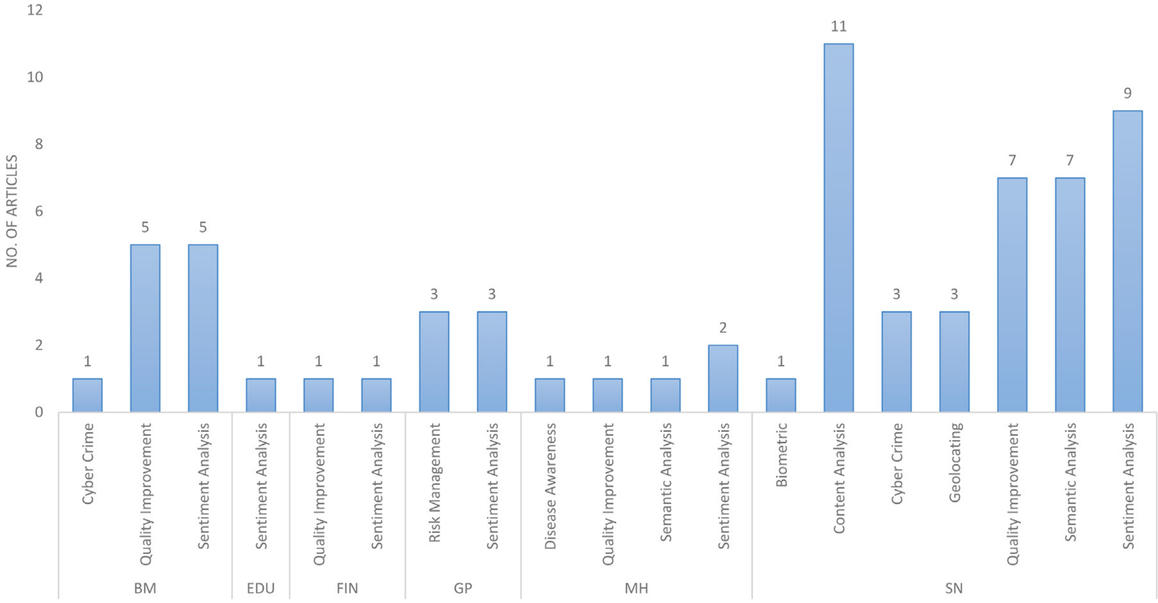

Fig. 7. Research objective among domains.

techniques that can be used for synthesizing the qualitative data.

3. Results and discussion

In this section, we will discuss the results obtained from this review. The first subsection gives an overview of the selected articles. The result of each RQ will be discussed in detail in the next five Sections 3.1–3.5.

The total number of the selected studies was 66 articles (see Appendix (A), Table A4) that implemented data mining techniques used in social media. The selected articles were retrieved only from journals published between January 2003 and 2015. Appendix (A), Table A4 shows the number of articles and the percentage grouped by publisher name. The types of articles considered in this survey are: experiment, case study, and survey. Table 2 shows the distribution of the selected articles among the three types.

With regards to the quality of the selected articles, we applied a quality assessment criterion to stream the articles based on the marks gained. The articles with grade five or greater (out of ten) were taken into consideration (see Table 3).

3.1. Types of data mining techniques (RQ1)

We identified 19 data mining techniques that had been applied by researchers in the area of social media. The list of these techniques is below.

- AdaBoost
- Artificial Neural Network (ANN)
- Apriori
- Bayesian Networks (BN)
- Decision Trees (DT)

Table 5
Strengths and weakness.

DM Tech.	Strength	Article ID	Weakness	Article ID
SVM	One of the best techniques for solving classification problems.	A31, A41, A48, A49, A53, A55, A56, A66	Suffer from problem with sparse context links.	A34
	Perform well with high dimensional feature space and small training set size.	A66		
	Suitable for offline clustering	A60		
ANN	Self-Organizing Map (SOM): High level capabilities that greatly facilitated the high-dimensional data analysis.	A4, A14, A44	Median SOM: Induce maps of lesser quality than maps obtained by the kernel version.	A14
	SOM: Has visual benefits.	A4		
DT	Random Forest (RF): Effective in giving estimates of what variables are important in the classification.	A1		
	RF: Robust technique and perform well with variety of learning tasks.	A33		
BN	Very effective for text clustering.	A3, A15		
	Simple classification algorithm.	A3, A41		
	Very efficient in terms of computation time.	A41		
k-NN	One of the simplest and most discriminative classifiers in pattern recognition.	A29	• Inferior performance on small datasets. • Performance will degrade for data with high dimensions. • Dependent on the chosen feature and distance measure.	A43
Fuzzy	Specialized in modeling with vague modes of social reasoning and takes into account the stochastic component of human reasoning.	A18	Requires expertise in semantic web and fuzzy systems to manually handle the semantic fuzzy rule through an offline process.	A18
K-Means	k-Medoids: Less sensitive to outliers.	A16	Requires the number of clusters as an input.	A21, A64
	• Uses as few clusters as possible and captures statistically and commercially important cluster characteristics. • Suitable for fix number of groups with unknown characteristics based on variables that one defines.	A20	When the number of clusters increases, the quality of discovered clusters quickly deteriorates.	A21
	Performs well at finding a very small number of clusters.	A21	Often converge to a local minima.	A32
	SK-Means: • Efficient in terms of speed. Works well with high-dimensional datasets. • Can be efficiently parallelized and converges to local maxima quickly. • Can be a model which allows it to be re-used in future classifications.	A35		
DBA	Density-Based Spatial Clustering of Application with Noise (DBSCAN): Does not require pre-specified number of clusters and noise filtering.	A10	DBSCAN: Includes all the density-reachable points to a cluster.	A10
	• Groups data based on their density connectivity. • Treats noises as outliers which would not be involved in any cluster. • Capable of detecting arbitrary-shaped clusters.	A42, A60	Unsuitable some real world applications, because there is no assumption about the number of clusters with fixed topics.	A42,A60
LDA	• Characterizing documents in addition to data clustering.	A40	Often converge to a local minima.	A32
	• Useful to develop multimedia applications.		Suffer from problem with sparse context links.	A34
	• Designed to exploit term-frequency.			
Wrapper			• Web Wrapper: Requires high level automation strategies. • Wrapper maintenance becomes unsuitable if the pool of Web pages largely increases.	A65
HC			Does not scale the growing of data size, because it relies on a fully specified similarity matrix.	A64

- Density Based Algorithm (DBA)
- Fuzzy
- Genetic Algorithm (GA)
- Hierarchical Clustering (HC)
- K-Means
- k-nearest Neighbors (k-NN)
- Linear Discriminant Analysis (LDA)
- Linear-Regression (Lin-R)
- Logistic Regression (LR)
- Markov
- Maximum Entropy (ME)
- Novel
- Support Vector Machine (SVM)
- Wrapper

Fig. 3 shows that SVM, BN, and DT are the most applied techniques in the area of social media with a percentage of 51% of the selected articles. Novel techniques with the percentage of 9% were not considered as the one of the highest; because each article has its dedicated novel technique. Table 4, includes detailed information about the frequencies of data mining techniques used by the selected articles in this review.

Appendix (A), Fig. A1 shows further demonstration about the findings, it illustrates the distribution of the data mining techniques per year during the considered period. Based on the figure, it can be clearly seen that the number of data mining techniques adopted by researchers in the social media area has increased dramatically in 2012 and 2014 with 39 and 35 techniques respectively. The number dropped slightly to 24 techniques in 2013. Moreover, it is worthwhile to mention that many novel techniques have arisen between 2012 to early 2015 with a total number of 12 new techniques.

3.2. Data mining techniques research areas (RQ2)

From the selected articles, we identified six general domains which applied various techniques in nine different research areas to mine the flow of big data gathered from social media. The list of these domains follows:

- Business and Management (BM)
- Education (EDU)
- Finance (FIN)
- Government and Public (GP)
- Medical and Health (MH)
- Social Networks (SN)

Fig. 4 shows that social networks and business and management were the most active domains used by data mining techniques, with a percentage of 79% among all domains. Government and public with a percentage of 9% represents the third active domain. Appendix (A), Table A5, includes detailed information about all domains.

For further analysis of Table 2, we investigated the experiments of the selected articles and plotted Fig. 5 which demonstrates the popularity of various types in social media application researches. Some experiments were conducted to mine and analyze one or more social media applications' data. Microblogging applications such as Twitter was the most popular application for researchers with 31 experiments followed by social networks such as (Facebook) with 12 experiments. Appendix (A), Table A6, includes detailed information about the frequencies of social media applications used by the selected articles in this review.

Fig. 6 demonstrates further information about the findings by illustrating the distribution of the domains applying data mining techniques per year. Based on the figure, it can be clearly seen that the number of publications has increased dramatically in 2012 and 2014 with 19 articles in 5 domains for both periods. In 2013, the number went down to 12 articles in 5 domains. The social network data analysis remains the most active domain among the considered period.

Among the selected articles, we identified 9 active research objectives adopted data mining techniques. The list of these research objectives follows:

- Biometric
- Content Analysis
- Cyber Crime
- Disease Awareness
- Geolocating
- Quality Improvement
- Risk Management
- Semantic Analysis
- Sentiment Analysis

Fig. 7 illustrates the distribution of these research areas. The sentiment analysis and quality improvement were the most active areas among articles with a frequencies of 21 and 14 respectively.

3.3. Machine learning versus non-machine learning methods in mining social media data (RQ3)

Data mining techniques are the process of extracting hidden knowledge from the data [16]. This can be done in many ways such as KNN, K-Means, and SVM as machine learning methods. Also the statistical methods in some cases are considered as non-machine learning methods which used to discover patterns. As Berson, et al. mentioned [11], “statistical techniques are driven by the data and are used to discover patterns and build predictive models”.

Out of the 66 papers identified, only three papers contain either experimental or theoretical knowledge about non-machine learning methods. Two of these papers (A11, A19) integrated non-machine learning methods with machine learning methods to improve the result of their proposed solution. The third paper (A53) mentioned that text mining techniques that depend on machine learning methods are different than non-machine learning methods because of: (i) in traditional quantitative analysis methods, conclusions are derived from the population sample, whereas machine learning methods allow the researcher to derive conclusions from the entire population, (ii) traditional quantitative methods require the researcher to analyze the data using a theoretical platform, while machine learning methods give the researcher the ability to extract the actual meaning of the mined data contained in natural language text. (iii) Machine learning methods investigate the textual data without human interaction, whereas traditional quantitative methods need the researcher to interpret the data before analyzing.

However, we disagree with the authors of paper A53 because the definition of data mining consists of three concepts [17]: Statistics, Data (Big or Small), and Machine Learning and Lifting. Thus, data mining includes all statistics (Descriptive and non-inferential parts of the classical statistics) and Exploratory Data Analysis (EDA) for the data using the power of computers for the purpose of lifting and learning the patterns of the data [17].

Consequently, machine learning data mining techniques and non-machine learning data mining techniques such as traditional quantitative methods in statistics are complementary to each other in data mining

3.4. Data mining techniques versus other data mining techniques (RQ4)

This RQ compares different data mining techniques that have been used in the selected articles. Since most of the articles based their findings on either weak statistical analysis or without using any statistics, we built our comparison based on their judgments, which relied on the experiment they made or by referring to their article references. For instance, papers (A31, A53) indicate that the SVM technique is one of the best categorization and feature selection techniques available relying on references published in 1998 and 2003; however, the paper was published in 2013. Further details are provided in Section 5.

After reviewing the papers selected, we found that many papers have common findings on the same data mining techniques. For instance, papers (A31, A45, A53, A59) found that SVM outperforms other techniques such as Naïve Bayes. In contrast, papers (A41, A51) claimed that Naïve Bayes and MLP are performed better than SVM. Some other papers (A3, A20, A35) claimed that K-Means performed better than other techniques such as C4.5. Finally, (A42, A60) found that the DBA technique outperforms other techniques in terms of working with noisy data.

3.5. Strengths and weaknesses of data mining techniques (RQ5)

This part of the review represents a good source of information where the best practices of the primary data mining techniques could be implemented. Table 5 summarizes the data mining techniques that could be implemented in the social media area. In addition to the traditional data mining techniques, Appendix (A), Table A7, summarizes the description and the main features of the novel techniques proposed by the researchers.

4. Limitations of this review

This study is restricted to journal and Tier 1 social network-related conferences papers in the field of data mining techniques and social media. By applying our search filtration strategy, we obtained a large number of articles, the majority of which were found to be irrelevant. The reason behind considering a small number of papers is to ensure that the papers selected fully match our research objectives. Nevertheless, including more related papers would have enriched our conclusions.

We considered only the data mining techniques that were recommended by more than one paper, as mentioned in Section 3.4. In addition, we applied rigorous quality assessment criteria to select the related articles that could provide synthesized results.

One more limitation is that having public social media datasets with clear description has a challenging task because the nature of social media data is unstructured with different data types such as text, images, and videos [18]; this makes social media datasets complex and in heterogeneous format [2].

5. Conclusions, recommendations, and future work

Our survey explored journal and Tier I conference papers that applied data mining techniques in social media between the period 2003 and 2015; 66 articles were selected to answer the five RQs of this review. Our conclusions are summarized as follows:

- RQ1: the most frequent data mining techniques used in social media articles are SVM, BN, and DT.
- RQ2: social network data analysis and business and management were the most active domains that requiring mining of social media data. In contrast, sentiment analysis and quality improvement were the most active research objectives in these domains.
- RQ3: machine learning data mining techniques and non-machine learning data mining techniques are both required for data mining purposes.
- RQ4: SVM and BN are the most recommended techniques to mine social media data used by most of the papers.
- RQ5: data mining techniques have various strengths and weaknesses which make the selection of certain techniques dependent on the type of the informative data required.

An immediate recommendation is that the area of social media still calls for more profound research that takes into account accurate implementation of data mining techniques in the academic and industrial sectors. A thorough investigation of the literature written in this area reveals that a significant number of the studies have not applied any statistical tests.

Quite understandably, research in the social media domain should house a twin-focus method which incorporates accurate result recording of experiments and appropriate statistical analysis.

The systematic literature review conducted in this study reveals that quite a few articles applied statistical tests, such as ANOVA, MANOVA, and *t*-test; these parametric statistical tests require normally distributed data [11]. Apparently, the majority of the studies reviewed failed to meet this condition and, therefore, the data provided can hardly be held reliable.

Our study also found that very few surveys and case studies have shed light on data mining techniques in social media from the software engineering perspective. By way of illustration, most of the published papers in the health domain were conducted by health researchers, who barely provide any information about the method utilized in their papers.

In addition to the method-related gap, another one still holds as far as other domains are concerned. The domains of Education, Customer Relationship Management (CRM), and Human Resource Management (HRM), among others, have not yet been explored by software engineers. This is a gap that we recommend future research could bridge by investigating CRM and HRM using data mining techniques. Such studies are anticipated to yield a more generic view and understanding of data mining techniques.

Acknowledgment

MohammadNoor Injadat and Fadi Salo would like to thank the University of Western Ontario for supporting this research.

Dr. Ali Bou Nassif would like to thank the University of Sharjah for supporting this research.

Appendix A.

See Appendix Tables A1–A7 and Fig. A1.

Table A1
Selected articles.

ID	Title	Year	Refs.
A1	#tag: Meme or Event?	2014	[19]
A2	@Phillies tweeting from philly? Predicting twitter user locations with spatial word usage	2012	[20]
A3	A framework for building web mining applications in the world of blogs: A case study in product sentiment analysis	2012	[21]
A4	A Novel Data-Mining Approach Leveraging Social Media to Monitor Consumer Opinion of Sitagliptin	2015	[22]
A5	A probabilistic generative model for mining cybercriminal networks from online social media	2014	[23]
A6	A semantic triplet based story classifier	2012	[24]
A7	An algorithm for local geoparsing of microtext	2013	[25]
A8	An interests discovery approach in social networks based on semantically enriched graphs	2012	[26]
A9	An Unsupervised Feature Selection Framework for Social Media Data	2014	[27]
A10	Analyzing and visualizing web opinion development and social interactions with density-based clustering	2011	[28]
A11	Analyzing the political landscape of 2012 Korean presidential election in twitter	2014	[29]
A12	Ano´nimos: An LP-Based Approach for Anonymizing Weighted Social Network Graphs	2012	[30]
A13	Ant colony based approach to predict stock market movement from mood collected on Twitter	2013	[31]
A14	Batch kernel SOM and related Laplacian methods for social network analysis	2008	[32]
A15	Bayesian filters for mobile recommender systems	2011	[33]
A16	Big Data for Big Business? A Taxonomy of Data-driven Business Models used by Start-up Firms	2014	[34]
A17	BTM: Topic Modeling over Short Texts	2014	[35]
A18	Building dynamic social network from sensory data feed	2010	[36]
A19	Business Intelligence from Social Media A Study from the VAST Box Office Challenge	2014	[37]
A20	Classifying ecommerce information sharing behavior by youths on social networking sites	2011	[38]
A21	Clustering memes in social media	2013	[39]
A22	Collaborative filtering based on collaborative tagging for enhancing the quality of recommendation	2010	[40]
A23	Collaborative visual modeling for automatic image annotation via sparse model coding	2012	[41]
A24	Confucius and its intelligent disciples: integrating social with search	2010	[42]
A25	Content Feature Enrichment for Analyzing Trust Relationships in Web Forums	2013	[43]
A26	Content Matters: A study of hate groups detection based on social networks analysis and web mining	2013	[44]
A27	Co-training over Domain-independent and Domain-dependent features for sentiment analysis of an online cancer support community	2013	[45]
A28	Data-Mining Twitter and the Autism Spectrum Disorder: A Pilot Study	2014	[46]
A29	Decision Fusion for Multimodal Biometrics Using Social Network Analysis	2014	[47]
A30	Detecting Deception in Online Social Networks	2014	[48]
A31	Enhancing financial performance with social media: An impression management perspective	2013	[49]
A32	Enriching short text representation in microblog for clustering	2012	[50]
A33	Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics	2011	[51]
A34	Exploring Context and Content Links in Social Media: A Latent Space Method	2012	[52]
A35	Gaining customer knowledge in low cost airlines through text mining	2014	[53]
A36	Intelligent Social Media Indexing and Sharing Using an Adaptive Indexing Search Engine	2012	[54]
A37	Latent Co-interests' Relationship Prediction	2013	[55]
A38	Learning by expansion: Exploiting social media for image classification with few training examples	2012	[56]
A39	Learning Stochastic Models of Information Flow	2012	[57]
A40	Mining Crowdsourced First Impressions in Online Social Video	2014	[58]
A41	Mining Social Media Data for Understanding Students' Learning Experiences	2014	[59]
A42	Mining spatio-temporal information on microblogging streams using a density-based online clustering method	2012	[60]
A43	Nearest-neighbor method using multiple neighborhood similarities for social media data mining	2012	[61]
A44	Network-Based Modeling and Intelligent Data Mining of Social Media for Improving Care	2015	[62]
A45	OMG U got flu? Analysis of shared health messages for bio-surveillance	2011	[63]
A46	Optimizing an organized modularity measure for topographic graph clustering: A deterministic annealing approach	2010	[64]
A47	Predicting Time-sensitive User Locations from Social Media	2013	[65]
A48	Resource discovery through social tagging: a classification and content analytic approach	2009	[66]
A49	Rumors Detection in Chinese via Crowd Responses	2014	[67]
A50	Search engine reinforced semi-supervised classification and graph-based summarization of microblogs	2015	[68]
A51	Sentimental causal rule discovery from Twitter	2014	[69]
A52	Social Network Analysis in Enterprise	2012	[70]
A53	Spreading Social Media Messages on Facebook: An Analysis of Restaurant Business-to-Consumer Communications	2013	[71]
A54	Studying user footprints in different online social networks	2012	[72]
A55	The Information Ecology of Social Media and Online Communities	2008	[73]
A56	The potential of social media in delivering transport policy goals	2014	[74]
A57	The social media genome: modeling individual topic-specific behavior in social media	2013	[75]
A58	Topic-sensitive influencer mining in interest-based social media networks via hypergraph learning	2014	[76]
A59	Twitter, MySpace, Digg: Unsupervised Sentiment Analysis in Social Media	2012	[77]
A60	Unsupervised and supervised learning to evaluate event relatedness based on content mining from social-media streams	2012	[78]
A61	Using explicit linguistic expressions of preference in social media to predict voting behavior	2013	[79]
A62	Using inter-comment similarity for comment spam detection in Chinese blogs	2011	[80]
A63	Using Sentiment to Detect Bots on Twitter: Are Humans more Opinionated than Bots?	2014	[81]
A64	Using social media to enhance emergency situation awareness	2012	[82]
A65	Web data extraction, applications and techniques: A survey	2014	[83]
A66	What's in twitter: I know what parties are popular and who you are supporting now!	2012	[84]

Table A2
Qars marks for the selected articles.

ID	QAR1	QAR2	QAR3	QAR4	QAR5	QAR6	QAR7	QAR8	QAR9	QAR10	Total
A1	0.75	0.25	0.25	0.75	0.75	1	1	0.75	0.75	0.75	7
A2	0.75	0	0.25	0.5	0.75	0.75	0	0.75	0.5	0.75	5
A3	1	0.75	0.75	0.75	0.75	0.5	0.25	0.75	0.25	0.5	6.25
A4	1	0.75	1	0.75	1	0.75	1	0.5	0.25	0.75	7.75
A5	1	1	0.5	1	1	1	1	1	0.75	0.75	9
A6	0.75	0.25	1	0.5	0.75	0.25	0	0.5	0.25	0.75	5
A7	1	0.5	0.75	0.75	0.5	0.75	0.5	0.75	0.25	0.5	6.25
A8	0.75	0	0.25	0.5	0.75	0.5	0.5	0.75	0.25	0.75	5
A9	1	0.75	1	0.75	1	1	1	1	0.5	0.75	8.75
A10	1	1	1	0.75	0.75	1	1	0.75	0.25	0.5	8
A11	1	1	1	0.75	0.75	0.75	0.75	0.75	0.25	0.5	7.5
A12	1	0.75	1	0.75	0.75	0.5	1	0.75	0.25	0.5	7.25
A13	0.75	0.5	0.5	0.5	0.5	0.25	0.25	0.75	0.25	0.75	5
A14	0.75	0.25	1	0.75	0.5	0.75	0.25	0.75	0.75	0.75	6.5
A15	0.75	0.5	0.75	0.5	0.75	0.25	0.25	0.5	0.25	0.75	5.25
A16	0.75	0.75	0.5	0.75	0.5	0.75	0	0.75	0.5	0.5	5.75
A17	1	0.75	0.75	0.75	1	1	0.5	0.5	0.25	0.75	7.25
A18	1	0.75	1	0.75	1	0.5	0.5	0.5	0.25	0.5	6.75
A19	1	0.75	0.75	1	0.75	1	1	1	0.75	0.75	8.75
A20	0.75	0.75	0.5	1	1	1	0	1	1	1	8
A21	0.5	0.5	0.5	0.75	0.75	1	1	0.75	0.5	0.75	7
A22	1	1	0.75	0.75	0.5	1	0	0.75	0.5	0.75	7
A23	0.75	0	0.25	0.5	0.75	0.75	0.75	0.75	0.5	0.75	5.75
A24	1	0.25	0.25	1	0.25	0.75	1	0.75	0.75	1	7
A25	0.75	0.25	0.25	1	0.25	0.75	1	0.75	1	0.75	6.75
A26	0.75	1	0.25	0.75	0.75	1	1	0.75	0.75	0.75	7.75
A27	0.75	0.25	0.25	0.5	0.75	0.75	0.5	0.5	0.5	0.75	5.5
A28	0.75	0	0.25	0.5	0.75	0.75	1	0.75	0.5	0.75	6
A29	1	0.75	0.75	1	1	0.75	0.75	0.75	0.25	0.75	7.75
A30	0.75	0.5	0.5	0.75	0.75	0	0	0.75	0.25	0.75	5
A31	1	1	1	0.75	1	1	1	1	1	1	9.75
A32	0.75	0.75	0.75	1	1	1	0.5	0.75	0.5	0.5	7.5
A33	1	0.5	0.5	1	0.75	0.75	0.75	1	0.5	1	7.75
A34	1	0.75	0.75	1	0.75	1	1	1	0.75	0.75	8.75
A35	1	1	1	1	0.75	1	0.75	0.75	0.25	1	8.5
A36	1	0.25	0.25	0.75	0.75	0.75	0	1	0.5	0.5	5.75
A37	1	1	0.75	0.75	0.75	0.75	1	0.75	0.5	0.5	7.75
A38	0.75	0.75	0.25	0.75	0.5	0.75	0.5	0.5	0.5	0.5	5.75
A39	0.75	0.75	0.75	0.75	0.5	0.25	0.25	0.75	0.25	0.75	5.75
A40	1	1	1	1	0.75	1	0	0.75	1	0.5	8
A41	1	1	1	1	0.75	1	0.75	1	0.5	0.75	8.75
A42	1	1	1	1	0.75	0.75	0	0.75	0.5	0.75	7.5
A43	0.75	0.75	0.75	0.5	0.75	0.75	0.5	0.5	0.5	0.5	6.25
A44	1	0.75	0.75	0.75	0.5	0.75	0.75	0.75	0.5	0.75	7.25
A45	1	1	1	1	0.75	1	1	1	0.75	0.75	9.25
A46	0.75	0	0.75	0.75	0.75	0.75	0.5	0.75	0.5	0.75	6.25
A47	0.75	0.75	0.5	0.5	0.75	1	0.5	0.75	0.5	0.75	6.75
A48	1	0.75	0.75	0.5	0.75	0.75	0	0.75	0.75	0.5	6.5
A49	0.75	0.5	0.75	0.75	0.5	0.5	0	0.5	0.25	0.75	5.25
A50	0.75	0	0.25	0.75	0.5	1	0.75	0.75	0.75	0.75	6.25
A51	1	0.5	0.5	0.5	0.75	0.5	0.75	0.5	0.25	0.5	5.75
A52	1	0.5	0.75	0.5	0.75	0.75	0	0.5	0.25	0.5	5.5
A53	1	1	0.75	0.75	1	1	1	1	1	0.75	9.25
A54	0.5	0.75	0.25	0.5	0.5	0.75	0	0.75	0.5	0.75	5.25
A55	1	1	0.75	1	0.75	0.75	0	0.75	0.5	0.75	7.25
A56	1	1	0.75	0.75	0.75	0.75	0	0.75	0.25	0.5	6.5
A57	0.5	0.5	0.25	0.75	0.75	0.5	0.25	0.5	0.5	0.5	5
A58	1	0.75	0.75	0.75	0.75	1	0.75	1	0.75	0.5	8
A59	1	1	1	1	1	1	1	0.75	0.75	0.5	9
A60	1	1	0.75	1	0.75	1	1	0.5	0.25	0.75	8
A61	0.75	0.5	0.25	0.5	0.75	0.75	0	0.75	0.75	0.5	5.5
A62	0.75	0.75	0.5	0.75	0.5	0.75	0.75	0.75	0.75	0.5	6.75
A63	0.75	0.5	0.25	0.75	0.75	0.25	0.5	0.25	0.25	0.75	5
A64	1	1	0.75	0.5	0.75	0.5	0.5	0.5	0.25	0.5	6.25
A65	1	1	1	0.75	1	0.75	0	0.5	0.25	0.5	6.75
A66	0.75	0.75	0.5	0.75	0.5	0.5	0.75	0.5	0.5	0.75	6.25

Table A3
RQS answered by articles.

ID	RQ1	RQ2	RQ3	RQ4	RQ5
A1	1	1	0	1	0
A2	1	1	0	0	0
A3	1	1	0	0	1
A4	1	1	0	0	1
A5	1	1	1	1	0
A6	1	1	0	0	0
A7	1	1	0	0	0
A8	1	1	0	0	0
A9	1	1	0	0	0
A10	1	1	0	1	1
A11	1	1	1	0	0
A12	1	1	0	1	1
A13	1	1	0	0	0
A14	1	1	0	0	1
A15	1	1	0	0	0
A16	1	1	0	1	1
A17	1	1	0	1	1
A18	1	1	0	0	1
A19	1	1	1	0	0
A20	1	1	0	0	1
A21	1	1	0	1	1
A22	1	1	0	0	0
A23	1	1	0	1	0
A24	1	1	0	1	0
A25	1	1	0	1	0
A26	1	1	0	1	0
A27	1	1	0	0	0
A28	1	1	0	1	0
A29	1	1	0	1	1
A30	1	1	0	0	0
A31	1	1	0	1	0
A32	1	1	0	1	1
A33	1	1	0	1	0
A34	1	1	0	1	1
A35	1	1	0	1	1
A36	1	1	0	0	1
A37	1	1	0	1	1
A38	1	1	0	0	0
A39	1	1	0	0	0
A40	1	1	0	0	1
A41	1	1	0	1	0
A42	1	1	0	1	1
A43	1	1	0	1	1
A44	1	1	0	1	1
A45	1	1	0	1	0
A46	1	1	0	1	0
A47	1	1	0	1	0
A48	1	1	0	0	1
A49	1	1	0	0	0
A50	1	1	0	1	0
A51	1	1	0	1	0
A52	1	1	0	0	1
A53	1	1	1	0	1
A54	1	1	0	1	0
A55	1	1	0	0	1
A56	1	1	0	0	1
A57	1	1	0	0	0
A58	1	1	0	1	0
A59	1	1	0	1	0
A60	1	1	0	0	1
A61	1	1	0	0	0
A62	1	1	0	1	0
A63	1	1	0	1	0
A64	1	1	0	1	1
A65	1	1	0	0	1
A66	1	1	0	1	1

Table A4
Articles percentage per journal.

Publication venue	Type	Freq.	%	Publication venue	Type	Freq.	%
ACM TRANSACTIONS ON INTELLIGENT SYSTEMS AND TECHNOLOGY	Jour.	2	3	IEEE TRANSACTIONS ON LEARNING TECHNOLOGIES	Jour.	1	2
AI MAGAZINE	Jour.	1	2	IEEE TRANSACTIONS ON MULTIMEDIA	Jour.	2	3
CAMBRIDGE SERVICE ALLIANCE BLOG	Jour.	1	2	IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE	Jour.	1	2
CORNELL HOSPITALITY QUARTERLY	Jour.	1	2	IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS	Jour.	2	3
DECISION SUPPORT SYSTEMS	Jour.	1	2	IEEE/ACM INTERNATIONAL CONFERENCE ON ADVANCES IN SOCIAL NETWORKS ANALYSIS AND MINING	Conf.	20	27
ELECTRONIC COMMERCE RESEARCH AND APPLICATIONS	Jour.	1	2	INDUSTRIAL MANAGEMENT & DATA SYSTEMS	Jour.	1	2
EXPERT SYSTEMS WITH APPLICATIONS	Jour.	4	6				
FRONTIERS OF COMPUTER SCIENCE IN CHINA	Jour.	1	2	JOURNAL OF BIOMEDICAL SEMANTICS	Jour.	1	2
GEOINFORMATICA	Jour.	1	2	JOURNAL OF INFORMATION SCIENCE	Jour.	1	2
IEEE COMPUTATIONAL INTELLIGENCE MAGAZINE	Jour.	1	2	KNOWLEDGE-BASED SYSTEMS	Jour.	1	2
IEEE COMPUTER GRAPHICS AND APPLICATIONS	Jour.	1	2	NEUROCOMPUTING	Jour.	6	9
IEEE INTELLIGENT SYSTEMS	Jour.	2	3	ONLINE INFORMATION REVIEW	Jour.	1	2
IEEE INTERNATIONAL CONFERENCE ON DATA ENGINEERING (ICDE)	Conf.	1	2	PROCEEDINGS OF THE IEEE	Jour.	1	2
IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS	Jour.	2	3	PROCEEDINGS OF THE VLDB ENDOWMENT	Jour.	1	2
IEEE TRANSACTIONS ON INSTRUMENTATION AND MEASUREMENT	Jour.	1	2	TRANSPORT POLICY	Jour.	1	2
IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING	Jour.	4	6	TSINGHUA SCIENCE AND TECHNOLOGY	Jour.	1	2
				Grand total		66	100

Table A5
Domains frequencies among articles.

Domain	Frequency	%
Business and management	11	17
Education	1	2
Finance	2	3
Government and public	6	9
Medical and health	5	8
Social networks	41	62
Grand total	66	100

Table A6
Popularity of various social media application in researches.

Social media application	Frequency	Ref
Blogs	8	A6, A12, A18, A22, A31, A37, A48, A55
Forums and Discussion boards	9	A4, A5, A6, A24, A25, A27, A31, A44, A52
Microblogging	31	A1, A2, A5, A7, A9, A11, A13, A15, A17, A21, A28, A30, A32, A35, A39, A41, A42, A45, A47, A49, A50, A51, A54, A57, A59, A60, A61, A62, A63, A64, A66
Product reviews	1	A33
Social networks	12	A8, A10, A12, A14, A15, A18, A26, A32, A46, A53, A54, A59
Video and photo sharing	11	A9, A12, A15, A23, A29, A34, A36, A38, A40, A43, A58

Table A7
Novel techniques features.

ID	Novel Tech.	Features	Compared with
A12	Ano´nimos	● Applied to preserve linear properties by generation of inequalities corresponding to decisions made by the algorithm during its execution. ● Preserve multiple linear properties in a single anonymized graph	
A17	Biterm Topic Model (BTM)	● Capture the topics within short texts by explicitly modeling word co-occurrence patterns in the whole corpus. ● Discover more prominent and coherent topics than the state-of-the art competitors.	Outperforms the online LDA in terms of effectiveness
A37	Interest-based Factor Graph Model (I-FGM)	● Proposed to take both network topology and node features into consideration. ● Makes the most of the strong inference abilities of the probability model and the graph model.	
A58	Topic-Sensitive Influencer Mining (TSIM)	● Aims to find the influential nodes in the networks. ● Improves the performance significantly in the applications of friends' suggestion and photo recommendation.	Outperforms LDA in terms of friends' suggestion and photo recommendation.
A34	Latent Space Method	● Discovers the latent semantic space from both context and content links in multimedia information networks. ● Solve the problem with sparse context. ● The learned latent semantic space can be applied for many applications, such as multimedia annotation and retrieval.	Extends the traditional LSI algorithm by low-rank approximation.
A32	Novel	● The proposed framework performs language knowledge integration and feature reduction simultaneously. ● Improves the short texts clustering performance. ● Scales linearly with the number of short texts and the number of integrated languages.	
A64	Online Incremental Clustering Algorithm	● Provide useful situation awareness information through a set of tightly integrated components. ● Enhance timely situation awareness across a range of crisis types.	Resolves the weaknesses in K-means and EM.
A10	Scalable Distance-Based Clustering (SDC)	● Proposed SDC technique for Web opinion clustering. ● Ensures that a required density must be reached in the initial clusters and uses scalable distances to expand the initial clusters. ● Does not require a predefined number of clusters. ● Able to filter noise.	
A29	Decision Fusion for Multimodal Biometrics	● Reduces the false acceptance rate for both single biometric traits and multimodal biometrics when the social network analysis is employed. ● Independently classify an actor from the relationship among actors.	
A9	Unsupervised Feature Selection Framework (LUFS)	● Exploit link information effectively in comparison with the state-of-the-art unsupervised feature selection methods.	
A43	Neighborhood Similarity Measure	● Encodes both the local density information and semantic information. ● Enhances the scalability to conduct approximated nearest neighbor search. ● Enhance the robustness on diversified genres of images.	Outperforms the k-NN methods using the labeled data only.
A8	Semantic Social Graph (SSG)	● Discovers the implicit semantic relations between entities in text messages. ● Enriches graph representation of entities contained in text messages generated by a user.	Significantly outperforms Naive Bayes classifier in accuracy and reliability

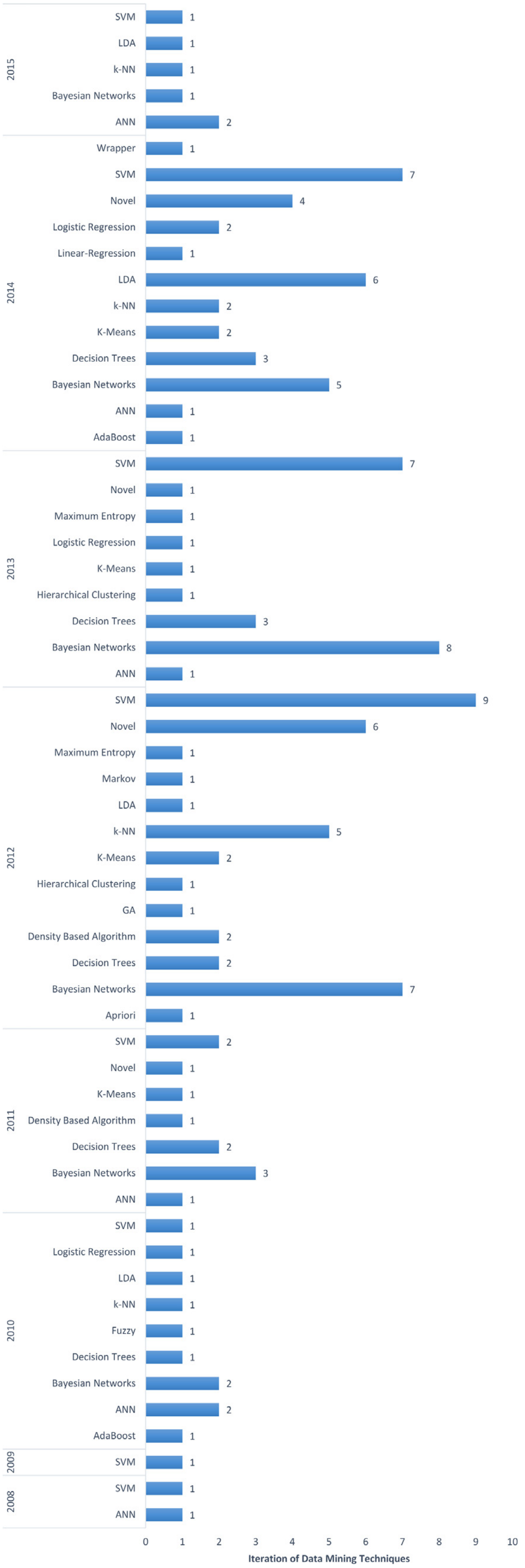

Fig. A1. Data mining techniques iteration per year.

References

[1] A.L. Kavanaugh, E.a. Fox, S.D. Sheetz, S. Yang, L.T. Li, D.J. Shoemaker, et al., Social media use by government: from the routine to the critical, Gov. Inf. Q 29 (2012) 480–491, http://dx.doi.org/10.1016/j.giq.2012.06.002.

[2] H. Chen, R.H.L. Chiang, V.C. Storey, Business intelligence and analytics: from big data to big impact, Mis Q 36 (2012) 1165–1188.

[3] M. Zuber, A survey of data mining techniques for social network analysis, Int. J. Res. Comput. Eng. Electron. 3 (2014) 1–8.

[4] S. Yu, S. Kak, A survey of prediction using social media, arXiv Prepr. arXiv1203.1647, 2012, pp. 1–20. 〈http://arxiv.org/abs/1203.1647〉.

[5] V. Vuori, J. Väisänen, The use of social media in gathering and sharing competitive intelligence, in: ICEB 2009 Proceedings, 2009, pp. 1–8.

[6] P.C. Rafeeque, S. Sendhilkumar, A survey on short text analysis in web, in: Proceedings of the 2011 Third International Conference Advances Computing, 2011, pp. 365–371. doi: 〈http://dx.doi.org/10.1109/ICoAC.2011.6165203〉.

[7] M. Tsytsarau, T. Palpanas, Survey on mining subjective data on the web, Data Min. Knowl. Discov. 24 (2012) 478–514, http://dx.doi.org/10.1007/s10618-011-0238-6.

[8] S. Gole, B. Tidke, A survey of big data in social media using data mining techniques, in: 2015 Int. Conf. Adv. Comput. Commun. Syst. (ICACCS-2015), 2015, pp. 1–5. doi: 〈http://dx.doi.org/10.1109/ICACCS.2015.7324059〉.

[9] B. Kitchenham, S. Charters, Guidelines for performing Systematic Literature Reviews in Software Engineering, Tech. Rep., EBSE-2007-01, Keele Univ. Univ. Durham, 2007. doi: 〈http://dx.doi.org/10.1145/1134285.1134500〉.

[10] D. Hand, Statistics and data mining: intersecting disciplines, ACM SIGKDD Explor. Newsl., vol. 1, 1999, pp. 16–19. doi: 〈http://dx.doi.org/10.1145/846170.846171〉.

[11] A. Berson, S.J. Smith, Building Data Mining Applications for CRM, McGraw-Hill, Inc, New York, NY, USA, 2002.

[12] X. Wu, V. Kumar, The Top Ten Algorithms in Data Mining, CRC Press, 2009, ISBN: 9781420089646.

[13] X. Wu, V. Kumar, J. Ross Quinlan, J. Ghosh, Q. Yang, H. Motoda, et al., Top 10 algorithms in data mining, Knowl. Inf. Syst. 14 (2008) 1–37, http://dx.doi.org/10.1007/s10115-007-0114-2.

[14] D.M. Boyd, N.B. Ellison, Social network sites: definition, history, and scholarship, J. Comput. Commun. 13 (2007) 210–230, http://dx.doi.org/10.1111/j.1083-6101.2007.00393.x.

[15] M.G. Smith, L. Bull, Feature construction and selection using genetic programming and a genetic algorithm, in: Genetic Program, Springer, 2003, pp. 229–237.

[16] U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, others, Knowledge discovery and data mining: towards a unifying framework., in: KDD, 1996, pp. 82–88.

[17] B. Ratner, Statistical and Machine-learning Data Mining: Techniques for Better Predictive Modeling and Analysis of Big Data, CRC Press, 2011, ISBN: 9781439860915.

[18] D. Pohl, A. Bouchachia, H. Hellwagner, Social media for crisis management: clustering approaches for sub-event detection, Multimed. Tools Appl. (2013) 1–32, http://dx.doi.org/10.1007/s11042-013-1804-2.

[19] D. Kotsakos, P. Sakkos, I. Katakis, D. Gunopulos, #tag: Meme or Event?, in: 2014 IEEE/ACM Int. Conf. Adv. Soc. Networks Anal. Min., 2014, pp. 391–394. doi: 〈http://dx.doi.org/10.1109/ASONAM.2014.6921615〉.

[20] H.W. Chang, D. Lee, M. Eltaher, J. Lee, Phillies tweeting from philly? Predicting twitter user locations with spatial word usage, in: Proc. 2012 IEEE/ACM Int. Conf. Adv. Soc. Networks Anal. Mining, ASONAM 2012, 2012, pp. 111–118. doi: 〈http://dx.doi.org/10.1109/ASONAM.2012.29〉.

[21] E. Costa, R. Ferreira, P. Brito, I.I. Bittencourt, O. Holanda, A. MacHado, et al., A framework for building web mining applications in the world of blogs: a case study in product sentiment analysis, Expert Syst. Appl. 39 (2012) 4813–4834, http://dx.doi.org/10.1016/j.eswa.2011.09.135.

[22] A. Akay, A. Dragomir, A. Novel Data-Mining, Approach leveraging social media to monitor consumer opinion of sitagliptin, IEEE J. Biomed. Heal. Inform. 19 (2015) 389–396, http://dx.doi.org/10.1109/JBHI.2013.2295834.

[23] R.Y.K. Lau, Y. Xia, Y. Ye, A probabilistic generative model for mining cyber-criminal networks from online social media, IEEE Comput. Intell. Mag. 9 (2014) 31–43, http://dx.doi.org/10.1109/MCI.2013.2291689.

[24] B. Ceran, R. Karad, A. Mandvekar, S.R. Corman, H. Davulcu, A semantic triplet based story classifier, in: Proc. 2012 IEEE/ACM Int. Conf. Adv. Soc. Networks Anal. Mining, ASONAM 2012, 2012, pp. 573–580. doi: 〈http://dx.doi.org/10.1109/ASONAM.2012.97〉.

[25] J. Gelernter, S. Balaji, An algorithm for local geoparsing of microtext, Geoinformatica 17 (2013) 635–667, http://dx.doi.org/10.1007/s10707-012-0173-8.

[26] A. Al-Kouz, S. Albayrak, An interests discovery approach in social networks based on semantically enriched graphs, in: Proc. 2012 IEEE/ACM Int. Conf. Adv. Soc. Networks Anal. Mining, ASONAM 2012, 2012, pp. 1272–1277. doi: 〈http://dx.doi.org/10.1109/ASONAM.2012.219〉.

[27] J. Tang, H. Liu, An unsupervised feature selection framework for social media data, IEEE Trans. Knowl. Data Eng. 4347 (2014) 2914–2927, http://dx.doi.org/10.1109/TKDE.2014.2320728.

[28] C.C. Yang, T.D. Ng, Analyzing and visualizing web opinion development and social interactions with density-based clustering, IEEE Trans. Syst. Man Cybern. Part A Syst. Hum. 41 (2011) 1144–1155, http://dx.doi.org/10.1109/TSMCA.2011.2113334.

[29] M. Song, M.C. Kim, Y.K. Jeong, Analyzing the political landscape of 2012 korean presidential election in twitter, IEEE Intell. Syst. 29 (2014) 18–26, http://dx.doi.

org/10.1109/MIS.2014.20.
[30] S. Das, Ö. Eğecioğlu, A. El Abbadi, Anónimos: an LP-based approach for anonymizing weighted social network graphs, IEEE Trans. Knowl. Data Eng. 24 (2012) 590–604, http://dx.doi.org/10.1109/TKDE.2010.267.
[31] S. Bouktif, M.A. Awad, Ant colony based approach to predict stock market movement from mood collected on Twitter, in: 2013 IEEE/ACM Int. Conf. Adv. Soc. Networks Anal. Min. Ant, 2013, pp. 837–845. doi: ⟨http://dx.doi.org/10.1145/2492517.2500282⟩.
[32] R. Boulet, B. Jouve, F. Rossi, N. Villa, Batch kernel SOM and related Laplacian methods for social network analysis, Neurocomputing 71 (2008) 1257–1273, http://dx.doi.org/10.1016/j.neucom.2007.12.026.
[33] M. Saravanan, S. Buveneswari, S. Divya, V. Ramya, Bayesian filters for mobile recommender systems, in: Proc. – 2011 Int. Conf. Adv. Soc. Networks Anal. Mining, ASONAM 2011, 2011, pp. 715–721. doi: ⟨http://dx.doi.org/10.1109/ASONAM.2011.51⟩.
[34] P.M. Hartmann, M. Zaki, N. Feldmann, A. Neely, Big data for big business? A taxonomy of data-driven business models used by start-up firms, Cambridge Serv. Alliance Blog (2014) 1–29 ⟨http://cambridgeservicealliance.blogspot.co.uk/2014/04/big-data-for-big-business_3.html⟩.
[35] X. Cheng, X. Yan, Y. Lan, J. Guo, B.T.M. Topic, Modeling over short texts, IEEE Trans. Knowl. Data Eng. 26 (2014) 2928–2941, http://dx.doi.org/10.1109/TKDE.2014.2313872.
[36] M.A. Rahman, A. El Saddik, W. Gueaieb, Building dynamic social network from sensory data feed, IEEE Trans. Instrum. Meas. 59 (2010) 1327–1341, http://dx.doi.org/10.1109/TIM.2009.2038307.
[37] B.I. Analytics, Business intelligence from social media a study from the VAST box office challenge, IEEE Comput. Graph. Appl. 34 (2014) 58–69, http://dx.doi.org/10.1109/MCG.2014.61.
[38] B.J. Jansen, K. Sobel, G. Cook, Classifying ecommerce information sharing behaviour by youths on social networking sites, J. Inf. Sci. 37 (2011) 120–136, http://dx.doi.org/10.1177/0165551510396975.
[39] E. Ferrara, M. JafariAsbagh, O. Varol, V. Qazvinian, F. Menczer, A. Flammini, Clustering memes in social media, in: Proc. 2013 IEEE/ACM Int. Conf. Adv. Soc. Networks Anal. Min. -ASONAM'13, 2013, pp. 548–555. doi: ⟨http://dx.doi.org/10.1145/2492517.2492530⟩.
[40] H.-N. Kim, A.-T. Ji, I. Ha, G.-S. Jo, Collaborative filtering based on collaborative tagging for enhancing the quality of recommendation, Electron. Commer. Res. Appl. 9 (2010) 73–83, http://dx.doi.org/10.1016/j.elerap.2009.08.004.
[41] M. Wang, F. Li, M. Wang, Collaborative visual modeling for automatic image annotation via sparse model coding, Neurocomputing 95 (2012) 22–28, http://dx.doi.org/10.1016/j.neucom.2011.04.049.
[42] X. Si, E.Y. Chang, Z. Gyöngyi, M. Sun, Confucius and its intelligent disciples: integrating social with search, in: Proc. VLDB Endow., vol. 3, 2010, pp. 1505–1516. doi: ⟨http://dx.doi.org/10.1145/1645953.1645955⟩.
[43] J. Piorkowski, L. Zhou, Content feature enrichment for analyzing trust relationships in web forums, in: 2013 IEEE/ACM Int. Conf. Adv. Soc. Networks Anal. Min. Content, 2013, pp. 1486–1487.
[44] I. Ting, S. Wang, Content Matters: A study of hate groups detection based on social networks analysis and web mining, in: 2013 IEEE/ACM Int. Conf. Adv. Soc. Networks Anal. Min., 2013, pp. 1196–1201. doi: ⟨http://dx.doi.org/10.1145/2492517.2500254⟩.
[45] P. Biyani, C. Caragea, P. Mitra, C. Zhou, J. Yen, G.E. Greer, et al., Co-training over domain-independent and domain-dependent features for sentiment analysis of an online cancer support community, in: 2013 IEEE/ACM Int. Conf. Adv. Soc. Networks Anal. Mining, ASONAM 2013, August 25– 28, 2013, 2013, pp. 413–417. doi: ⟨http://dx.doi.org/10.1145/2492517.2492606⟩.
[46] A. Beykikhoshk, T. Caelli, Data-mining twitter and the autism spectrum disorder: a pilot study, in: 2014 IEEE/ACM Int. Conf. Adv. Soc. Networks Anal. Min., 2014, pp. 349–356.
[47] P.P. Paul, M.L. Gavrilova, R. Alhajj, Decision fusion for multimodal biometrics using social network analysis, IEEE Trans. Syst. Man Cybern. Syst. 44 (2014) 1522–1533.
[48] J.S. Alowibdi, U.A. Buy, P.S. Yu, L. Stenneth, Detecting deception in online social networks, in: 2014 IEEE/ACM Int. Conf. Adv. Soc. Networks Anal. Min., 2014, pp. 383–390.
[49] D. Schniederjans, E.S. Cao, M. Schniederjans, Enhancing financial performance with social media: an impression management perspective, Decis. Support. Syst. 55 (2013) 911–918, http://dx.doi.org/10.1016/j.dss.2012.12.027.
[50] J. Tang, X. Wang, H. Gao, X. Hu, H. Liu, Enriching short text representation in microblog for clustering, Front. Comput. Sci. China 6 (2012) 88–101, http://dx.doi.org/10.1007/s11704-011-1167-7.
[51] A. Ghose, P.G. Ipeirotis, Estimating the helpfulness and economic impact of product reviews: mining text and reviewer characteristics, IEEE Trans. Knowl. Data Eng. 23 (2011) 1498–1512, http://dx.doi.org/10.1109/TKDE.2010.188.
[52] G. Qi, C. Aggarwal, Q. Tian, S. Member, Exploring context and content links in social media: a latent space method, IEEE Trans. Pattern Anal. Mach. Intell. 34 (2012) 850–862.
[53] B. Yee Liau, P. Pei Tan, Gaining customer knowledge in low cost airlines through text mining, Ind. Manag. Data Syst. 114 (2014) 1344–1359, http://dx.doi.org/10.1108/IMDS-07-2014-0225.
[54] C.H.C. Leung, A.W.S. Chan, A. Milani, J. Liu, Y. Li, Intelligent social media indexing and sharing using an adaptive indexing search engine, ACM Trans. Intell. Syst. Technol. 3 (2012) 1–27, http://dx.doi.org/10.1145/2168752.2168761.
[55] F. Tan, L. Li, Z. Zhang, Y. Guo, Latent co-interests' relationship prediction, Tsinghua Sci. Technol. 18 (2013) 379–386.
[56] S.Y. Wang, W.S. Liao, L.C. Hsieh, Y.Y. Chen, W.H. Hsu, Learning by expansion: exploiting social media for image classification with few training examples, Neurocomputing 95 (2012) 117–125, http://dx.doi.org/10.1016/j.neucom.2011.05.043.
[57] L. Dickens, I. Molloy, J. Lobo, Learning stochastic models of information flow, in: 2012 IEEE 28th Int. Conf. Data Eng., 2012, pp. 570–581.
[58] J. Biel, D. Gatica-perez, Mining crowdsourced first impressions in online social video, IEEE Trans. Multimed. 16 (2014) 2062–2074.
[59] X. Chen, M. Vorvoreanu, K. Madhavan, Mining social media data for understanding students' learning experiences, IEEE Trans. Learn. Technol. 7 (2014) 246–259 http://web.ics.purdue.edu/∼chen654/pub/XinChen_etal_IEEETrans_tlt-cs_Mining_Twitter.pdf/nhttp://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6697807.
[60] C.H. Lee, Mining spatio-temporal information on microblogging streams using a density-based online clustering method, Expert. Syst. Appl. 39 (2012) 9623–9641, http://dx.doi.org/10.1016/j.eswa.2012.02.136.
[61] S. Wang, Q. Huang, S. Jiang, Q. Tian, L. Qin, Nearest-neighbor method using multiple neighborhood similarities for social media data mining, Neurocomputing 95 (2012) 105–116, http://dx.doi.org/10.1016/j.neucom.2011.06.039.
[62] A. Akay, A. Dragomir, B.-E. Erlandsson, Network-based modeling and intelligent data mining of social media for improving care, IEEE J. Biomed. Heal. Inform. 19 (2015) 210–218.
[63] N. Collier, N.T. Son, N.M. Nguyen, OMG U got flu? Analysis of shared health messages for bio-surveillance, J. Biomed. Semant. 2 (2011) 1–10, http://dx.doi.org/10.1186/2041-1480-2-S5-S9.
[64] F. Rossi, N. Villa-Vialaneix, Optimizing an organized modularity measure for topographic graph clustering: a deterministic annealing approach, Neurocomputing 73 (2010) 1142–1163, http://dx.doi.org/10.1016/j.neucom.2009.11.023.
[65] A. Jaiswal, W. Peng, T. Sun, Predicting time-sensitive user locations from social media, in: 2013 IEEE/ ACM Int. Conf. Adv. Soc. Networks Anal. Min., 2013, pp. 870–877. doi: ⟨http://dx.doi.org/10.1145/2492517.2500229⟩.
[66] D.H.-L. Goh, A. Chua, C.S. Lee, K. Razikin, Resource discovery through social tagging: a classification and content analytic approach, Online Inf. Rev. 33 (2009) 568–583, http://dx.doi.org/10.1108/14684520910969961.
[67] G. Cai, H. Wu, R. Lv, Rumors detection in chinese via crowd responses, in: 2014 IEEE/ACM Int. Conf. Adv. Soc. Networks Anal. Min., 2014, pp. 912–917.
[68] Y. Chen, X. Zhang, Z. Li, J.-P. Ng, Search engine reinforced semi-supervised classification and graph-based summarization of microblogs, Neurocomputing 152 (2015) 274–286, http://dx.doi.org/10.1016/j.neucom.2014.10.068.
[69] R. Dehkharghani, H. Mercan, A. Javeed, Y. Saygin, Sentimental causal rule discovery from Twitter, Expert Syst. Appl. 41 (2014) 4950–5958, http://dx.doi.org/10.1016/j.eswa.2014.02.024.
[70] C.-Y. Lin, L. Wu, Z. Wen, H. Tong, V. Griffiths-Fisher, L. Shi, et al., Social network analysis in enterprise, Proc. IEEE 100 (2012) 2759–2776, http://dx.doi.org/10.1109/JPROC.2012.2203090.
[71] L. Kwok, B. Yu, Spreading social media messages on facebook: an analysis of restaurant business-to-consumer communications, Cornell Hosp. Q. 54 (2013) 84–94, http://dx.doi.org/10.1177/1938965512458360.
[72] A. Malhotra, L. Totti, W. Meira, P. Kumaraguru, V. Almeida, Studying user footprints in different online social networks, in: Proc. 2012 IEEE/ACM Int. Conf. Adv. Soc. Networks Anal. Mining, ASONAM 2012, 2012, pp. 1065–1070. doi: ⟨http://dx.doi.org/10.1109/ASONAM.2012.184⟩.
[73] T. Finin, A. Joshi, P. Kolari, A. Java, A. Kale, A. Karandikar, The information ecology of social media and online communities, AI Mag. 29 (2008) 77–92, http://dx.doi.org/10.1609/aimag.v29i3.2158.
[74] A. Gal-Tzur, S.M. Grant-Muller, T. Kuflik, E. Minkov, S. Nocera, I. Shoor, The potential of social media in delivering transport policy goals, Transp. Policy 32 (2014) 115–123, http://dx.doi.org/10.1016/j.tranpol.2014.01.007.
[75] P. Bogdanov, M. Busch, J. Moehlis, A.K. Singh, B.K. Szymanski, The social media genome: modeling individual topic-specific behavior in social media, in: Proc. 2013 IEEE/ACM Int. Conf. Adv. Soc. Networks Anal. Min., 2013, pp. 236–242. doi: ⟨http://dx.doi.org/10.1145/2492517.2492621⟩.
[76] Q. Fang, J. Sang, C. Xu, Y. Rui, Topic-sensitive influencer mining in interest-based social media networks via hypergraph learning, IEEE Trans. Multimed. 16 (2014) 796–812, http://dx.doi.org/10.1109/TMM.2014.2298216.
[77] G. Paltoglou, M. Thelwall, Twitter, myspace, digg: unsupervised sentiment analysis in social media, ACM Trans. Intell. Syst. Technol. 3 (2012) 1–19, http://dx.doi.org/10.1145/2337542.2337551.
[78] C.H. Lee, Unsupervised and supervised learning to evaluate event relatedness based on content mining from social-media streams, Expert Syst. Appl. 39 (2012) 13338–13356, http://dx.doi.org/10.1016/j.eswa.2012.05.068.
[79] S. O'Banion, L. Birnbaum, Using explicit linguistic expressions of preference in social media to predict voting behavior, in: 2013 IEEE/ACM Int. Conf. Adv. Soc. Networks Anal. Min., 2013, pp. 207–214. doi: ⟨http://dx.doi.org/10.1145/2492517.2492538⟩.
[80] J.H. Wang, M.S. Lin, Using inter-comment similarity for comment spam detection in Chinese blogs, in: Proc. – 2011 Int. Conf. Adv. Soc. Networks Anal. Mining, ASONAM 2011, 2011, pp. 189–194. doi: ⟨http://dx.doi.org/10.1109/ASONAM.2011.49⟩.
[81] J. Dickerson, V. Kagan, V. Subrahmanian, Using sentiment to detect bots on Twitter: are humans more opinionated than bots?, in: 2014 IEEE/ACM Int. Conf. Adv. Soc. Networks Anal. Min., 2014, pp. 620–627. ⟨http://jpdickerson.com/pubs/dickerson14using.pdf⟩.
[82] J. Yin, A. Lampert, M. Cameron, B. Robinson, R. Power, Using social media to

enhance emergency situation awareness, IEEE Intell. Syst. 27 (2012) 52–59, http://dx.doi.org/10.1109/MIS.2012.6.

[83] E. Ferrara, P. De Meo, G. Fiumara, R. Baumgartner, Web data extraction, applications and techniques: a survey, Knowl. Based Syst. 70 (2014) 301–323, http://dx.doi.org/10.1016/j.knosys.2014.07.007.

[84] A. Boutet, H. Kim, E. Yoneki, What's in twitter: I know what parties are popular and who you are supporting now!, in: Proc. 2012 IEEE/ACM Int. Conf. Adv. Soc. Networks Anal. Mining, ASONAM 2012, 2012, pp. 132–139. doi: ⟨http://dx.doi.org/10.1109/ASONAM.2012.32⟩.

MohammadNoor Injadat received the B.Sc. and M.Sc. degrees in computer science from Al al-Bayt University and University Putra Malaysia in Jordan and Malaysia in 2000 and 2002, respectively. He obtained a Master of Engineering in Electrical and Computer Engineering from University of Western Ontario in 2015. He is currently working toward his Ph.D. degree in Software Engineering at the Department of Electrical and Computer Engineering, University of Western Ontario in Canada. His research interests include data mining, machine learning, social network analysis, data analytics, and cloud computing. MohhammadNoor is a member of IEEE Computer Society.

Fadi Salo received the B.Sc. and M.Sc. degrees in computer science from Al-Ahliyya Amman University and University Putra Malaysia in Jordan and Malaysia in 1999 and 2005, respectively. He obtained a Master of Engineering in Electrical and Computer Engineering from University of Western Ontario in 2015. He is currently working toward his Ph.D. degree in Software Engineering at the Department of Electrical and Computer Engineering, University of Western Ontario in Canada. His research interests include data mining, text mining, machine learning, social network analysis, data analytics, and intrusion detection systems. Fadi is a member of IEEE Computer Society.

Ali Bou Nassif is currently an Assistant Professor at University of Sharjah, UAE. He obtained a Master's degree in Computer Science and a Ph.D. degree in Electrical and Computer Engineering from Western University in 2009 and 2012, respectively. Ali's research interests include the applications of statistical and artificial intelligence models in different areas such as software engineering, electrical engineering, e-learning and social media, as well as cloud computing and mobile computing. Ali is a registered professional engineer in Ontario, as well as a member of IEEE Computer Society and ACM Association for Computing Machinery.