

Big Data, Twitter Sentiment Analysis and Cloud Resource Scheduling: A Review

Nirmal Godara and Sanjeev Kumar

Department of Computer Science & Engineering

Guru Jambheshwar university of Science & Technology, Hisar, Haryana

Abstract - Big data is associated with large volumes of structured and unstructured information and creates challenges for conventional processing approaches. Social networking platforms generate large amounts of textual information, including tweets, reviews and blog posts, which can be analysed to understand opinions and sentiments. Cloud computing provides an environment for data storage and processing, while cloud resource scheduling is required to allocate tasks and virtual machines efficiently. This review examines research on Twitter sentiment analysis, text preprocessing, cloud security and cloud scheduling, with particular attention to studies published up to 2018. The literature includes machine-learning, lexicon-based and semi-supervised approaches for sentiment analysis; encryption and privacy-preserving techniques for cloud data; and heuristic, metaheuristic and hybrid approaches for task and workflow scheduling. The reviewed studies address different objectives including sentiment classification, preprocessing, confidentiality, access control, cost, deadlines, energy efficiency, resource provisioning and workflow execution. The comparison highlights the approaches and research objectives reported in the literature up to 2018.

Keywords: Big data; Twitter sentiment analysis; text preprocessing; cloud computing; cloud security; workflow scheduling; task scheduling; machine learning.

1. Introduction

Big data has become an important area of computing because data are generated from social networking, online transactions, mobile devices, sensors and other sources. Social networking produces large quantities of textual information in the form of tweets, reviews and blog posts. The analysis of such information can provide an understanding of users' opinions and feelings. Twitter sentiment analysis is challenging because tweets are short and contain textual variations, while machine-learning techniques can be used to analyse textual sentiment [1], [7], [14], [16]. Sentiment analysis generally involves collecting textual data, preprocessing the raw messages, extracting useful features and applying a classification method. Preprocessing is important because tweets may contain inconsistencies and redundancy. The reviewed literature includes removal of hashtags and URLs, stop words, non-English words, punctuation, symbols and numbers, as well as treatment of emoticons. Feature extraction can then transform preprocessed tweets into feature vectors, with bigram and n-gram models used in sentiment-analysis studies [1], [4], [19].

Machine-learning methods reported in the literature include Naive Bayes, Support Vector Machine, Maximum Entropy, Random Forest and Multilayer Perceptron. Studies also consider lexicon-based and semi-supervised approaches. Zhao and Xiaolin examined six preprocessing approaches, two feature models and four classifiers on five Twitter datasets and reported changes in classification accuracy and F1 measure for different preprocessing operations [19]. Song et al. proposed a Naive Bayes-based classification approach for Twitter sentiment analysis [14], while Ghiassi and Lee developed a domain-transferable lexicon set using a supervised machine-learning approach [7]. Cloud computing is another important environment for storing and processing large-scale data. Scheduling determines how tasks and virtual machines are mapped to available computing resources. Scheduling approaches may be centralized or decentralized, static or dynamic, preemptive or non-preemptive, immediate or batch, and heuristic or metaheuristic [5], [9], [21], [23]. The reviewed scheduling literature considers parameters such as time, cost, makespan, scalability, efficiency, resource consumption and performance [21].

Workflow scheduling is particularly important when tasks have dependencies. The literature includes approaches based on hyper-heuristics, gravitational search, ant colony optimization, genetic algorithms, energy-efficient scheduling and resource provisioning [10]–[13], [24]–[26]. Cloud security is also considered through encryption, access control and encrypted data processing. Searchable ciphertext-policy attribute-based encryption, homomorphic re-encryption and enhanced homomorphic encryption are among the techniques represented in the reference collection [15], [17], [22].

This review brings together these research directions to summarize the development of big-data analysis and cloud resource management up to 2018. The emphasis is on the methods, objectives and research areas reported in the selected literature, followed by a comparison of representative studies.

2. Literature Review

2.1 Twitter Sentiment Analysis

Twitter sentiment analysis has been studied using machine-learning, lexicon-based and semi-supervised approaches. Kathuria and Sharma presented a comparative study of approaches for sentiment analysis of Twitter data and considered machine-learning methods such as SVM and MLP along with lexicon-based methods. The literature review reports attempts to improve machine-learning precision through bigram and trigram representations [1]. Kunal et al. studied textual dissection of live Twitter reviews using Naive Bayes. Their work is listed as a 2018 Procedia Computer Science publication [3]. Cirqueira et al. reviewed preprocessing approaches for sentiment analysis of Brazilian Portuguese social media and identified preprocessing as an important research area for improving analysis accuracy for particular platforms [4]. Ghiassi and Lee proposed a domain-transferable lexicon set for Twitter sentiment analysis using supervised machine learning. The work focuses on the domain dependence of features and the use of a Twitter-specific lexicon for vectorizing tweets [7]. Wu et al. reviewed Twitter spam detection and compared approaches for identifying spam in Twitter data [8]. These studies illustrate that Twitter analysis involves not only sentiment classification but also preprocessing, domain information and identification of unwanted content. Song et al. proposed a classification approach based on Naive Bayes for Twitter sentiment analysis. The reference describes the work as a 2017 publication and the literature review discusses Multinomial Naive Bayes together with weighting and feature-selection techniques [14]. Trupthi et al. investigated sentiment analysis on Twitter using the streaming API, showing the use of continuously obtained Twitter information for sentiment analysis [16]. Ghadeer et al. investigated preprocessing operators for Arabic sentiment analysis [18]. Jianqiang and Xiaolin compared text preprocessing methods for Twitter sentiment analysis and evaluated the effect of preprocessing on classification performance [19]. Patel et al. presented a literature survey on sentiment analysis of Twitter data using machine-learning approaches [20]. The reviewed works collectively cover preprocessing, feature extraction, supervised classification and domain/language-specific sentiment analysis.

2.2 Cloud Security and Encryption

Cloud computing introduces security requirements for stored and processed information. The reference collection includes studies addressing secure integration, encryption, encrypted data processing and searchable encrypted data. Stergiou et al. studied secure integration of IoT and cloud computing, reflecting the relationship between connected devices and cloud environments [6]. Li et al. proposed searchable ciphertext-policy attribute-based encryption with revocation in cloud storage [15]. The approach combines searchable encrypted storage with attribute-based access control and revocation. El Makkaoui et al. proposed Cloud-RSA, an enhanced homomorphic encryption scheme [17]. Ding et al. studied encrypted data processing with homomorphic re-encryption, addressing processing over encrypted data [22]. The literature therefore covers several security requirements rather than a single encryption problem. These include protection of cloud data, access control, search over encrypted data, re-encryption and processing while data remain encrypted. Such techniques are relevant to cloud environments in which data storage and computation have to be considered together with security.

2.3 Cloud Task and Workflow Scheduling

Cloud scheduling is concerned with assigning tasks to resources and virtual machines. The scheduling literature distinguishes independent and dependent tasks and identifies scheduling types such as centralized/decentralized, static/dynamic, preemptive/non-preemptive and immediate/batch. Heuristic and metaheuristic approaches are also discussed, with metaheuristics providing general search strategies for different optimization problems [5], [9], [21], [23]. Almezeini and Hafez presented a review on scheduling in cloud computing [5], while Bittencourt et al. discussed scheduling in distributed systems from a cloud-computing perspective [9]. Rodriguez and Buyya presented a taxonomy and survey of scheduling algorithms for scientific workflows in IaaS cloud computing environments [21]. Banga and Rana reviewed heuristic-based independent task scheduling techniques in cloud computing [23]. These studies provide classifications and surveys of the scheduling problem. Nayak and Tripathy used the Analytical Hierarchy Process for deadline-sensitive lease scheduling in cloud computing [2]. Alkhanak and Lee proposed a hyper-heuristic cost-optimization approach for scientific workflow scheduling [10]. Choudhary et al. proposed a GSA-based hybrid algorithm for bi-objective workflow scheduling [11], and Vinothina and Sridaran presented an ACO approach for workflow scheduling in cloud computing [12]. These studies represent different optimization approaches and objectives within workflow scheduling. Genetic and energy-related approaches are also represented. Garg and Er reported a genetic-based workflow scheduling algorithm in cloud computing [13]. Ghose et al. studied energy-efficient scheduling of scientific workflows in a cloud environment [24]. Kaur and Mehta investigated resource provisioning and workflow scheduling using an augmented Shuffled Frog Leaping Algorithm [25]. Liu et al. proposed a deadline-constrained coevolutionary genetic algorithm for scientific workflow scheduling [26]. The reviewed scheduling

studies show that cloud scheduling is a multi-dimensional problem. Different studies address cost, deadlines, energy, resource provisioning, workflow dependencies and overall scheduling performance. The choice of scheduling method is related to the characteristics of the tasks and the objective being considered.

3. Comparison of Selected Studies

Reference	Year	Method / Approach	Research Area	Main Focus	Citation
[1] Kathuria & Sharma	2018	SVM, MLP, lexicon-based, n-gram	Twitter sentiment	Comparative study of sentiment-analysis approaches	[1]
[3] Kunal et al.	2018	Naive Bayes	Twitter sentiment	Textual dissection of live Twitter reviews	[3]
[4] Cirqueira et al.	2018	Preprocessing review	Sentiment analysis	Preprocessing for Brazilian Portuguese social media	[4]
[7] Ghiassi & Lee	2018	Twitter-specific lexicon + supervised ML	Twitter sentiment	Domain-transferable sentiment lexicon	[7]
[19] Jianqiang & Xiaolin	2017	Preprocessing + classifiers	Twitter sentiment	Comparison of preprocessing effects	[19]
[14] Song et al.	2017	Naive Bayes	Twitter sentiment	Classification with feature/weighting techniques	[14]
[5] Almezeini & Hafez	2018	Review	Cloud scheduling	Review of cloud scheduling	[5]
[9] Bittencourt et al.	2018	Survey / perspective	Distributed/cloud scheduling	Scheduling in distributed systems	[9]
[10] Alkhanak & Lee	2018	Hyper-heuristic	Workflow scheduling	Cost optimization	[10]
[11] Choudhary et al.	2018	GSA-based hybrid	Workflow scheduling	Bi-objective scheduling	[11]
[12] Vinothina & Sridaran	2018	ACO	Workflow scheduling	Workflow scheduling in cloud	[12]
[13] Garg & Er	2017	Genetic algorithm	Workflow scheduling	Genetic-based workflow scheduling	[13]
[24] Ghose et al.	2017	Energy-efficient scheduling	Scientific workflow	Energy efficiency	[24]
[25] Kaur & Mehta	2017	Augmented Shuffled Frog Leaping	Resource/workflow scheduling	Resource provisioning and workflow scheduling	[25]
[26] Liu et al.	2017	Coevolutionary genetic algorithm	Workflow scheduling	Deadline-constrained scheduling	[26]
[15] Li et al.	2017	Searchable CP-ABE + revocation	Cloud security	Searchable encrypted cloud storage	[15]
[17] El Makkaoui et al.	2017	Enhanced homomorphic encryption	Cloud security	Cloud-RSA encryption	[17]
[22] Ding et al.	2017	Homomorphic re-encryption	Cloud security	Encrypted data processing	[22]

The comparison indicates that the literature covers three closely related areas: analysis of social-media text, protection of cloud data, and allocation of cloud resources. Twitter studies mainly address preprocessing, lexical representation and classification. Security studies focus on encrypted storage, access control and encrypted computation. Scheduling studies use review, heuristic, metaheuristic and hybrid approaches for objectives including cost, deadlines, energy and resource provisioning.

4. Research Trends and Observations

The literature up to 2018 shows increasing use of machine-learning and optimization approaches for large-scale data and cloud environments. In Twitter sentiment analysis, preprocessing and feature representation remain important because the quality and form of textual input affect classification. The studies include supervised, semi-supervised, lexicon-based and Naive Bayes approaches [1], [3], [4], [7], [14], [19]. In cloud security, the literature moves beyond basic encryption toward searchable encryption, attribute-based access control, revocation and homomorphic processing [15], [17], [22]. In cloud scheduling, the literature includes conventional reviews as well as optimization-based approaches, with different studies focusing on cost, deadline constraints, energy efficiency, workflow scheduling and resource provisioning [5], [10]–[13], [21], [24]–[26]. The reviewed literature consequently reflects the need to consider data analysis, security and resource management as important components of cloud-based data processing. The approaches differ in their specific objectives and methods, and the scheduling problem in particular involves multiple parameters including time, cost, makespan, efficiency, scalability and resource consumption [21].

5. Conclusion

This short review examined research published up to 2018 on Twitter sentiment analysis, cloud security and cloud task/workflow scheduling. The literature reports machine-learning, lexicon-based and semi-supervised approaches for sentiment analysis; encryption, searchable encryption and homomorphic techniques for cloud security; and heuristic, metaheuristic and hybrid approaches for cloud scheduling. The comparison shows that the research community addressed different requirements of large-scale data processing through complementary methods involving classification, security, resource allocation, cost optimization, deadline management, energy efficiency and workflow scheduling. These studies provide a broad foundation for understanding big-data processing and cloud resource management up to 2018.

References

- [1] A. Kathuria and A. Sharma, “A Comparative Study of Various Approaches for Sentiment Analysis of Twitter Data,” 2018.
- [2] S. C. Nayak and C. Tripathy, “Deadline sensitive lease scheduling in cloud computing environment using AHP,” *Journal of King Saud University–Computer and Information Sciences*, vol. 30, no. 2, pp. 152–163, Apr. 2018.
- [3] S. Kunal, A. Saha, A. Varma, and V. Tiwari, “Textual dissection of live Twitter reviews using naive Bayes,” *Procedia Computer Science*, vol. 132, pp. 307–313, 2018.
- [4] D. Cirqueira, M. F. Pinheiro, A. Jacob, F. Lobato, and Á. Santana, “A literature review in preprocessing for sentiment analysis for Brazilian Portuguese social media,” in *2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, pp. 746–749, 2018.
- [5] N. Almezeini and A. Hafez, “Review on scheduling in cloud computing,” *International Journal of Computer Science and Network Security*, vol. 18, no. 2, pp. 108–111, Jan. 2018.
- [6] C. Stergiou, K. E. Psannis, B. G. Kim, and B. Gupta, “Secure integration of IoT and cloud computing,” *Future Generation Computer Systems*, vol. 78, pp. 964–975, 2018.
- [7] M. Ghiassi and S. Lee, “A domain transferable lexicon set for Twitter sentiment analysis using a supervised machine learning approach,” *Expert Systems with Applications*, vol. 106, pp. 197–216, 2018.
- [8] T. Wu, S. Wen, Y. Xiang, and W. Zhou, “Twitter spam detection: Survey of new approaches and comparative study,” *Computers & Security*, vol. 76, pp. 265–284, 2018.
- [9] L. F. Bittencourt, A. Goldman, E. R. Madeira, N. L. da Fonseca, and R. Sakellariou, “Scheduling in distributed systems: a cloud computing perspective,” *Computer Science Review*, vol. 30, pp. 31–54, 2018.
- [10] E. N. Alkhanak and S. P. Lee, “A hyper-heuristic cost optimization approach for Scientific Workflow Scheduling in cloud computing,” *Future Generation Computer Systems*, vol. 86, pp. 480–506, 2018.
- [11] A. Choudhary, I. Gupta, V. Singh, and P. K. Jana, “A GSA based hybrid algorithm for bi-objective workflow scheduling in cloud computing,” *Future Generation Computer Systems*, vol. 83, pp. 14–26, 2018.
- [12] V. Vinothina and R. Sridaran, “An Approach for Workflow Scheduling in Cloud Using ACO,” in *Big Data Analytics*, Springer, Singapore, pp. 525–531, 2018.
- [13] E. J. Garg and G. B. Er, “Research Paper on Genetic Based Workflow Scheduling Algorithm in Cloud Computing,” *International Journal of Advanced Research in Computer Science*, vol. 8, no. 5, May 2017.
- [14] J. Song, K. T. Kim, B. J. Lee, S. Y. Kim, and H. Y. Youn, “A novel classification approach based on Naïve Bayes for Twitter sentiment analysis,” *TIIS*, vol. 11, no. 6, pp. 2996–3011, 2017.

- [15] J. Li, Y. Shi, and Y. Zhang, "Searchable ciphertext-policy attribute-based encryption with revocation in cloud storage," *International Journal of Communication Systems*, vol. 30, no. 1, p. e2942, Jan. 2017.
- [16] M. Trupthi, S. Pabboju, and G. Narasimha, "Sentiment analysis on twitter using streaming API," in *IEEE 7th International Advance Computing Conference (IACC)*, pp. 915–919, 2017.
- [17] K. El Makkaoui, A. Ezzati, and A. Beni-Hssane, "Cloud-RSA: an enhanced homomorphic encryption scheme," in *Europe and MENA Cooperation Advances in Information and Communication Technologies*, Springer, Cham, pp. 471–480, 2017.
- [18] A. S. Ghadeer, I. Aljarah, and H. Alsawalqah, "Enhancing the Arabic sentiment analysis using different preprocessing operators," *Proceedings of the New Trends in Information Technology (NTIT'17)*, vol. 113, 2017.
- [19] Z. Jianqiang and G. Xiaolin, "Comparison research on text pre-processing methods on twitter sentiment analysis," *IEEE Access*, vol. 5, pp. 2870–2879, 2017.
- [20] A. P. Patel, A. V. Patel, S. G. Butani, and P. B. Sawant, "Literature Survey on Sentiment Analysis of Twitter Data using Machine Learning Approaches," *International Journal for Innovative Research in Science & Technology*, vol. 3, no. 10, 2017.
- [21] M. A. Rodriguez and R. Buyya, "A taxonomy and survey on scheduling algorithms for scientific workflows in IaaS cloud computing environments," *Concurrency and Computation: Practice and Experience*, vol. 29, no. 8, p. e4041, 2017.
- [22] W. Ding, Z. Yan, and R. H. Deng, "Encrypted data processing with homomorphic re-encryption," *Information Sciences*, vol. 409, pp. 35–55, 2017.
- [23] P. Banga and S. P. Rana, "Heuristic based independent task scheduling techniques in cloud computing: a review," *International Journal of Computer Applications*, vol. 166, no. 1, 2017.
- [24] M. Ghose, P. Verma, S. Karmakar, and A. Sahu, "Energy efficient scheduling of scientific workflows in cloud environment," in *IEEE 19th International Conference on High Performance Computing and Communications; IEEE 15th International Conference on Smart City; IEEE 3rd International Conference on Data Science and Systems (HPCC/SmartCity/DSS)*, pp. 170–177, 2017.
- [25] P. Kaur and S. Mehta, "Resource provisioning and work flow scheduling in clouds using augmented Shuffled Frog Leaping Algorithm," *Journal of Parallel and Distributed Computing*, vol. 101, pp. 41–50, Mar. 2017.
- [26] L. Liu, M. Zhang, R. Buyya, and Q. Fan, "Deadline-constrained coevolutionary genetic algorithm for scientific workflow scheduling in cloud computing," *Concurrency and Computation: Practice and Experience*, vol. 29, no. 5, p. e3942, 2017.