Text Mining in Online Social Networks: A Systematic Review

Huda N Alhazmi
hnhazmi@uqu.edu.sa
Umm Al-Qura University, College of Computer and Information System, Makkah, Saudi Arabia

Summary
Online social networks contain a large amount of data that can be converted into valuable and insightful information. Text mining approaches allow exploring large-scale data efficiently. Therefore, this study reviews the recent literature on text mining in online social networks in a way that produces valid and valuable knowledge for further research. The review identifies text mining techniques used in social networking, the data used, tools, and the challenges. Research questions were formulated, then search strategy and selection criteria were defined, followed by the analysis of each paper to extract the data relevant to the research questions. The result shows that the most social media platforms used as a source of the data are Twitter and Facebook. The most common text mining technique were sentiment analysis and topic modeling. Classification and clustering were the most common approaches applied by the studies. The challenges include the need for processing with huge volumes of data, the noise, and the dynamic of the data. The study explores the recent development in text mining approaches in social networking by providing state and general view of work done in this research area.

Key words:
Online social networks; text mining; sentiment analysis; topic modeling; NLP; clustering; classification.

1. Introduction

Today, technology is deeply integrated with most activities in our modern life. With web technologies, digitalization, and communication, sharing and expressing ideas or opinions become much easier and faster [1, 2]. Social networking sites, including Twitter, Facebook, YouTube, Instagram, etc. create new ways for communication among individuals. These platforms can be a powerful source of valuable knowledge [1]. With the rapid increasing of the data sharing over such platforms, different mechanisms emerge to study the content on social media for public opinion mining [3]. Several social applications are introduced for businesses analytics, monitoring activities, and promoting products and services [4]. A large amount of data has opened opportunities to develop more efficient algorithms in machine learning and data mining to analyze this content [5] [6].

In most social networking platforms, the most common way people used to communicate is text. People write to share knowledge, information, and comments. Such unstructured data create analysis challenges; however, it could be a vital source for decision making in different domain [7]. Mining the text becomes a core field in data analysis. Hence, text mining is the process of exploring and extracting non-trivial patterns and interesting topics from natural textual data [1]. Text mining includes techniques of data mining, machine learning, and computational linguistics [8].

This study provides a systematic review of text mining applications in social networks analysis research, we review articles that have been published between 2013 and 2021. The study presents a detailed assessment to the recent tools and techniques used. The result might help researchers and developers to design their future research and to select the most appropriate approach.

The rest of the study is structured as follows. Section II presents the research background. Section III details methodology used for the review. Section IV discusses the results and finding. Section V concludes the review and section VI states the challenges and future directions.

2. Research Background

2.1 Online Social Networks (ONSs)

Social networking is a global phenomenon that has changed how people interact and communicate. It touches almost every aspect of our life such as social life, education, politics, healthcare, and communication. Online social similar interests and share comments, emails, images, videos, and blog posts. Moreover, it provides people with a chance to express their thought or opinion confidently.

The concept of social media founded in 1978 when the bulletin board system (BBS) was introduced. BBS have developed as a personal website in 1995 to share information. The concept of blogger appeared in 1999 where users can communicate using their blog [10]. LinkedIn and WordPress were launched in 2003, but the dramatic change was founded in 2004 when Facebook is launched as social networking for college students. Consecutively, in 2005, YouTube jumped into the seen then Twitter in 2006 that allows users to post a short message known as Tweet [10]. Online social networks allow users to generate profiles including attributes such as age, location, interests, etc. Popular platforms such as Facebook, Twitter, YouTube, and LinkedIn include tools that facilitate the communications and the interactions among their users [11].

Manuscript received March 5, 2022
Manuscript revised March 20, 2022
https://doi.org/10.22937/IJCSNS.2022.22.3.50

OSNs are groups of people who share information and common interest in an online setting [12].

2.2 Text Mining

Text mining has developed across different scientific disciplines such as statistics, computer science, linguistics, and library science [13]. Text mining focus on automated analysis of textual data as a form of natural language, its techniques deal with the unstructured text [14]. Despite the absent of a unifying definition of text mining, there is an agreement on the general process of analysis [15]. Furthermore, text mining is interconnected with Natural Language Processing (NLP) that is related to the analysis of natural languages [16]. Due to the need of using automatic tools for analyzing the textual data and extracting relevant information, software solutions are available for analyzing social media applications. Text mining tools are used to identify and analyze posts, likes, followers in online social networks to explore people's reactions and behavior. Moreover, it shows the variation in views and opinions regard different topics [17]. The fundamental process of text mining includes data collection, preprocessing, content analysis, finding and integration.

2.3 Preprocessing

Unstructured text may affect the analysis which leads to inaccurate output. Preprocessing is an essential phase that leads to efficient implementation [18]. The general idea of text mining techniques is transferring unstructured text data into structured data that can be used by analysis algorithms [13].

Text preprocessing usually includes features extraction and selection. The extraction can be classified into morphological analysis and syntactical analysis. The morphological analysis focuses on individual words in documents and consists of tokenization, removing the stop-word, and stemming. Tokenization is the process of splitting the documents into a sequence of words by removing the punctuations [18]. Removing stop-words such as pronouns and function words like 'the', 'and', or 'it' improve the effectiveness and efficiency of text processing [19]. Stemming is the technique that reduces word variability by reducing the word to the root form [20]. The syntactical analysis consists of processes related to grammatical structure of the language such as part-of-speech (POS) tagging and parsing. The POS tagging is used to tag the words on their grammatical function such as noun, verb, or adjective to get the grammatical knowledge [21]. Parsing is used for inspecting the grammatical structure of a sentence.

Feature selection is the process of eliminating unrelated information from the text. A common approach is representing text documents as a matrix such that each text document is formatted as a vector space. Each row in the matrix denotes to the document and each column denotes to the term as a word or a phrase [22]. To calculate the feature vector, two common methods have been proposed: term frequency (TF) and inverse document frequency (IDF). TF measures the frequency of a term in a group of documents, the occurrences of a term related to the topic identify the information about that topic. IDF identifies the least frequent terms in the document. Whereas TF-IDF is used to identify the terms that discriminate documents from each other [23].

3. Methodology

The methodology has been developed based on the objective of the study, that is, to explore the recent text mining techniques applied in social networks analysis. The papers in this review were explored to address the following:

- The type and the of data that are utilized for extracting insights for decision-making.
- The text mining techniques that most used in social networking.
- The text mining algorithms that most applied in social networking.
- The tools used to perform the analysis.

The review was performed in steps, each of these steps is described in the following subsections.

3.1 Search and Selection Method

The aim of search strategy was to identify the most relevant works of text mining in social networks analysis field. Several platforms were searched such as Google Scholar, ResearchGate, Academia, and Direct science for articles at the topic level, then the studies were checked at the title-level and abstract-level, followed by full-text extraction. The primary research studies that explicitly use text mining in social networks were included, whereas surveys and reviews were excluded. Also, papers that provide a general overview of text mining applications on social media without description to the applied algorithms were excluded. The selection criteria resulted in 32 studies.

3.2 Data Extraction

Selected studies further analyzed to extract data related to the research focus, the area to which the study applied, size and type of data, social media platforms, and the applied algorithms. Then more investigation was applied to identify the text mining technique used and the text mining tools or software. Table I summarized the extracted data for each study. Therefore, the following processing steps are applied.

- The papers were categorized into five clusters based on the size of the textual data used for analysis which are: less than 10k, more than 10k and less than 100k, more than 100k and less than 1M, and 1M or more. They further grouped into clusters based on the type of the data such as tweets, posts, reviews, publications, and others.
- The publications were categorized in five clustered according to the text mining application employed in the publications included in this review.
- The collection of the publications also categorized based on the social media networks used.
- The investigation was performed on each publication to determine the employed algorithm. Most of the articles coded as clustering, classification, hyper, or dictionary. On the other hand, some articles coded as others.

4. Result and Discussion

4.1 Text Mining Techniques

Various text mining techniques have been applied in the social networks field by the selected studies. Table II summarized the papers aligned with each technique. Interestingly, we found that sentiment analysis and topic modeling are the most applications applied as shown in Fig. 1.

4.1.1 Sentiment Analysis (SA)

SA is important technique in text mining where the number of research in SA has increased exponentially. The emotions of people can be recognized and classified using sentiment analysis techniques. Additionally, exploring the underlying opinions referred to as opinion mining [24]. A large amount of online data has opened opportunities for growing in the field of SA and developing more efficient algorithms. The sentiments discovered in the reviews and feedback can be generally classified into emotion and polarity classification. Emotion classification is a process of identifying a set of labels, and polarity classification is a process of classifying positive and negative sentiments [25]. The sentiment analysis method consists of various phases: data collection, then preprocessing and cleaning, including removing stop words, punctuations, and duplicate data. Finally, the classification and analysis are carried out to determine the feeling expressed in the text. Techniques for features extraction and selection are used to reduce mistakes and achieve a greater level of accuracy in social media information. Sentiment analysis includes two approaches lexicon-based and machine learning. In the first approach, the frequency of the sentiment is calculated using dictionary. In the second approach, supervised and unsupervised techniques are used [26].

Twitter conversations related to smartphones were investigated to discover valuable information related to supply chain management [27]. The research found that the social media analysis associated with the conceptual model of smartphone supply chain management is more efficient than the traditional methods. Sentiment mining also explored in other research contexts, such as understanding the users' views and experiences of products [28, 29] or identifying emerging trends in the context of public health [30]. Opinion mining used to identify the customer satisfaction level, authors in [31] used the hierarchical clustering technique to group the customers' reviews on the intensity of the opinions expressed by the customers on various product features.

4.1.2 Topic Modeling (TM)

TM is one of the growth techniques in text analytics. The most used approach in topic modeling is Latent Dirichlet Allocation (LDA). LDA is an unsupervised generative model that categorizes topics in documents [19]. LDA differs from supervised learning approaches that required be trained to classify the document by specific attributes. The generation process based is on the distribution of the underlying topics. Topics can be clustered based on common words where a probability weight is assigned to each word to indicate to its relative importance to that topic [32].

Study [33] applied the LDA algorithm to analyze the Facebook posts of breast cancer patients to detect the themes related to the quality of life of these patients. Then they compare the extracted topics with the topics of the self-administered questionnaires. They concluded that there is a good match between the topics extracted from the two sources, social media and self-administered questionnaires. Topic modeling also has been applied to identify a new emerging trend, in [34] authors used LDA to categorize articles in terms of their topics, and [35] used LDA to outline a research literature analysis in marketing. Unsupervised machine learning BTM cluster is utilized by [36] to discover hidden themes regarding risk behavior. They investigated context of prescription to discover the emerging trends regarding the patients' opioid analgesic abuse behavior.

4.1.3 Natural Language Processing (NLP)

NLP is concerned with the relations between natural languages and computers [37]. It focuses on processing and analyzing the unstructured textual information. NLP methods help to transform unclear text data into clear and

Table 1: Extracted Data

Study	*Research focus*	*Research Area*	*Text mining Techniques*	*Data Type*	*Data Size*	*Data Source*	*Tools Used*	*Mining Approach*
[26]	Detect spam in real-time.	NA	Text categorization	Tweets	400000	Twitter	NA	Classification Support Vector Machine, Random Forest, Neural Network
[27]	Identifying the influence of social media information of smartphone on supply chain management	Supply chain Management	Sentiment analysis	Tweets	NA	Tweets	Sentword	Pre-defined dictionaries to analyze a specific emotion towards smartphone brand
[28]	Compare feedback of students with respect to various teaching features	Education	Opinion Mining	Students' comments	12866	Module Evaluation Survey (MES)	Rapid miner	Classification: SVM, K Nearest Neighbor, Naïve Bayes, and Neural Network classifier.
[29]	Sentiment analysis of tweets and Facebook comments on commercial products	NA	Sentiment categories	Tweets and posts	4000	Twitter Facebook	Python package	classification models crowd lexicon with a Decision Tree classifier
[30]	Analyze the sentiment of vaccine-related Tweets for public health agencies	Health	Sentiment Analysis	Tweets	32597	Twitter	R package	Regression models to Analyze the sentiment of vaccine-related Tweets
[31]	Identify the customer satisfaction level	Business	Sentiment analysis	Reviews	7086	Amazon	Python package	Hierarchical Clustering
[33]	Detected new emerging topics to complete self administered questionnaires.	Health	Topic modeling	Posts	70092	Cancerdusein.org and Facebook	NA	LDA to detect the different topics
[34]	Identify trends of topics that discover the gap of knowledge structural engineering	Academic research	Topic Modeling	Articles	11027	Journal of Structural Engineering	NA	Clustering: LDA topic modeling technique
[35]	Identify major academic branches and research trends in design research	Academic research	Topic modeling	Publications	1560	NA	R packages	Clustering: K-means algorithm LDA algorithm
[36]	Detect emerging topics related to opioid analgesic abuse behavior in the context of prescription	Health	Topic modeling	Tweets	11 M	Twitter	NA	Unsupervised machine learning BTM Cluster
[39]	Detect the spread of fake news in the digital media	NA	NLP	News	20800	Kaggle	Python	Artificial Neural Network (ANN) classification model
[41]	Extracting fashion attributes from Instagram posts	Fashion domain	Text extracting	Posts	3000	Instagram	NLTK's	Classification: Weak supervised and generative modeling
[42]	Understand the important role that the social media play in decision making at the industry.	Businesses	Text extracting	Tweets and posts	874	Tweets Facebook	SPSS and Nvivo 9	Classification and clustering
[43]	Identify major academic research trends in design research	Academic Research	Topic modeling	Publications	20218	Web of Science (WOS)	NA	Clustering: K-means algorithm and LDA algorithm
[44]	Improve the quality of information services in teaching practice	Education	Topic-sentiment analysis	Students' reviews	171430	icourses web site	NA	Clustering: Latent Dirichlet Allocation (LDA)

Study	*Research focus*	*Research Area*	*Text mining Techniques*	*Data Type*	*Data Size*	*Data Source*	*Tools Used*	*Mining Approach*
[45]	Classifies aspects and opinion words related to social domain.	Entertainment	Opinion mining	Reviews	2000	IMPD	NA	Classification: Naïve Bayes classifier
[46]	Investigate the level of employees' engagement prediction	Business	Information extraction	Reviews	130000	IBM internal social media	NA	Classification: Naïve Bayes Multinomial method
[47]	Prediction of students' academic performance	Education	Prediction	Tweets	1064371	Twitter	NA	Classification: unsupervised learning of word embeddings
[48]	Detect the views of stakeholders to helping in policy-making decisions	Business	Sentiment analysis.	Movie reviews	7086	Kaggle.	Python packages	Classification:Linear SVC, DT, and Naïve Bayes
[49]	Analyze users' views about drug and cosmetic products to understand their experiences with these products	Health	Sentiment Analysis	Tweets and posts	6216	Twitter Facebook	NA	Classification: Naïve Bayes Lexicon-based
[50]	Understand how hotels are perceived by consumers.	Business	Sentiment Analysis	Reviews	11043	Tripadvisor	R packages	Natural processing languages (NPL)
[51]	Detect tourist behaviors and sentiments	Tourism	Sentiment analysis	Tweets	11532	Twitter	RStudio and Knime	Classification
[52]	Classification of customer reviews	NA	Sentiment analysis	Review.	1940	NA	NA	Machine learning (SVM) combined with domain specific lexicons.
[53]	Builds a dynamic representation of words that captures their contextual semantics	NA	Sentiment analysis	Tweets	4469	Twitter	NA	Lexicon based approach to detect the sentiment at both entity-level and tweet-level.
[54]	Solve high dimensional data mining problem	Business	Sentiment analysis	Reviews	NA	Amazon	NA	Deep Neural Network
[55]	Analyze shared ideas on the teaching profession	Education	Opinion mining	Tweets	35718	Twitter	KNIME and NodeXL	Association rule mining: Apriori algorithm was used
[56]	Understand student issues and problems in their educational experiences	Education	Topic modeling and sentiment analysis	Tweets	25284	Twitter	NA	Classification: Naєıve Bayes Multi-Label Classifier
[57]	Investigate the performance of SVM for polarity detection	NA	Sentiment Analysis	Tweets and IMDP reviews	5110	Twitter and IMPD	Weka	Classification: SVM classifier
[58]	Address the need for better coherence and understanding of actions in online discourse.	NA	NLP	Discussion threads	25000	Different social networking sites	NA	Coherence analysis and speech act classification
[59]	Sentiment classification to determine the public reaction towards the news and its effects.	NA	Sentiment Analysis	Tweets	NA	Twitter	NA	A hybrid method that includes Hybrid K, clustering and boosting.
[60]	Study issues of text mining in social network for medical purposes.	Health	Prediction	Tweets	NA	Twitter	NA	Classification: Bayes' Classifier, SVM, and Neural Network:

Study	*Research focus*	*Research Area*	*Text mining Techniques*	*Data Type*	*Data Size*	*Data Source*	*Tools Used*	*Mining Approach*
[61]	Developed a Warning system for tsunami to the public.	Crisis management	Real-time tweets analysis	Tweets	NA	Twitter	NA	Integrate the classification and geo-parsing (GEO)

precise data. These methos applied on natural language data to extract meaningful information [37]. NLP is concerned with Natural Language Generation (NLG) and Natural Language Understanding (NLU). NLG applications include a syntactic realizer to ensure that grammatical rules are correct in the generated text. NLU consists of some components such as lexical analyzer, syntax analyzer, and semantic analyzer [38].

In NLP, different techniques are used to deal with the textual data such as feature extractors and word embeddings. Authors in [39] examine the performance of three feature extractors techniques, naming, BERT embeddings, Glove embeddings, and TD-IDF vectorizer using Artificial Neural Network (ANN) on two fake news datasets. Their result revealed that BERT embeddings register the best performance4.1.2 Information Extraction

4.1.4 Information Extraction (IE)

IE is used to extract meaningful information from a text. IE is the process of identifying the object by extracting relevant attributes then establish the relationship between them. Extracted entities and keywords are stored in a database for further processing. IE helps to mine some informed pattern to take a decision. IE includes various tasks to remove the noise from the extracted patterns [40]. Several techniques such as document ranking, matching, and clustering are used in text mining for information extraction.

Authors in [41] proposed a supervised and generative model to extract fashion attributes from Instagram posts. Their result showed that word embeddings beat a baseline method. They concluded that combining weak supervision signals using generative models is more practical with unlabeled data.

Reviews and customers' opinions are valuable information for the organizations to understand their customers. In [42] researchers perform competitive analysis in three pizza chains, they extracted their data from Twitter and Facebook.

Their findings suggest that extracting and analyzing information from social media can help organizations in maintaining their relations with the customers.

Twitter is the most social media source that used to extract the data in the reviewed studies in our dataset, followed by Facebook, however, other social network sources show a low percentage. Data extracted from social media sources have high commercial value. Due to that, extracting the data from most of social networks such as Google, Facebook, and Thomson Reuters is very difficult. In contrast, Twitter provides opportunities to researchers to extract the data to get valuable insights [37].

4.2 Type and Source of Data

Social media data is rich with opportunities to understand human behavior as individuals and society. Regarding the type of textual data, the review revealed that 45.45% of the reviewed research explore tweets dataset and 18.18% Facebook posts. Less frequently are the reviews around 15.14% and then publications and other text like blogs, websites with 21.23%. Remarkably, the reviewed research showed wide variability of the size of datasets that used as shown in Table III. The highest percentage of our corpus represents the research that used between 10k and 100k textual data, where the lowest percentage denotes to the research investigated 1M and more. On other hand, 9.09% of the articles do not indicate the size of data that used in their study. From Table III, it can be found that the data between 10k and 100k are the mostly used because the high volumes of data are very challenging.

4.3 Text Mining Algorithms

The most algorithms used in analyzing the text in social networking are classification and clustering. Classification is a supervised learning that learn from training process a set of rules. The classification method comprises quantitative approaches to automate NLP to classify each text to a certain category. The most common algorithms are K-Nearest Neighbour (KNN), Decision Trees (DT), Support Vector Machine (SVN), and Artificial neural networks (ANN) [37]. Clustering is an unsupervised algorithm that grouped the text in clusters. Different clustering techniques include different strategies that can be categorized in three types, naming, partitional, hierarchical, and semantic-based clustering [14].

The studies in this review cover a variety of text analysis algorithms. Most of the articles focus on classification or clustering. We noted that the number of articles in the dataset that employed clustering and classification algorithms increased recently. Clustering and classification are the most data mining techniques that extensively studied in the context of text [63].

Table 2: Text Mining Techniques

Text Mining Technique	*Study*	*Social Network*	*Mining Algorithms*
Sentiment analysis	[27], [28], [29], [30], [31], [45], [48], [49], [50], [51], [52], [53], [54], [55], [57], [63]	Tweets IMDP Tripadvisor Kaggle Amazon	Lexicon based approach Classification: SVM, Naïve Bayes, SVC Clustering: Decision Tree Association rule mining: Apriori algorithm Hyper approach Other approach
Topic Modeling	[33], [34], [35], [36], [43], [44], [56]	Twitter Facebook Web of Science Science Direct	Unsupervised machine learning: BTM, LDA Clustering: K-means algorithm
Information Extraction	[42], [46]	Twitter Facebook IBM internal social media	Classification: Naïve Bayes Other approach
NLP	[39], [58], [59]	Kaggle Twitter Different social networking sites	Classification: Artificial Neural Network (ANN) LAP-based text analytics Hybrid clustering approach
Other application	[41], [47], [60], [61], [62]	Twitter	Classification Hypered approach

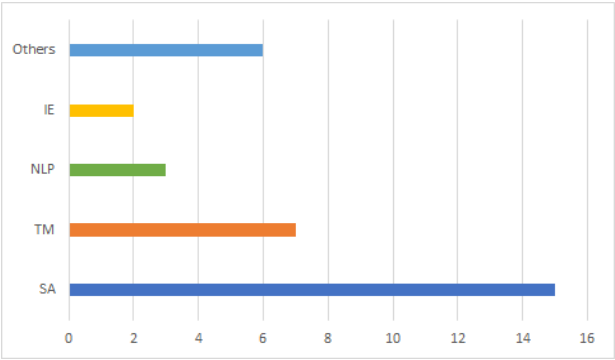

Fig. 1. Most applied text mining techniques in online networking

4.4 Clarity and Quality

To ensure the quality of studies, the text mining application has been documented in detail with transparency. The quality assessment was performed on the studies based on three components: 1) document the preprocessing in detail, 2) describe the text mining techniques in detail, and 3) the tool and the software used. Data cleaning and preprocessing is an essential step in the analysis. This step might affect the quality of the result. We found that 33.3 % of the studies describe the preprocessing step in detail, while 36.67 % reported the preprocessing without detail. On the other hand, 20 % do not indicate or mention this step. We also found that most of the articles mentioned using free packages such as R and python for implementation. Contrarily, some articles do not fully clear about their method, the software, or the packages they used.

Table 3: Data Size

Data size	*No. of paper*	*Percentage*
< 10K	11	30.3%
10K < and < 100K	12	33.33%
100K < and < 1M	4	12.12%
1M or more	2	6.06%
Not indicated	4	9.09%

5. Conclusion

Text mining applications have significantly affected the research in social networks analysis. Through exploring the research in online social networks, 32 research studies were analyzed to provide meaningful insights on the approaches applied to improve decision-making in different areas using social media platforms. The review reveals finding that answer the research questions, the most common text mining techniques were sentiment analysis and topic modeling. Clustering and classification are the most data algorithms that extensively studied. As data preprocessing is essential step in text mining that might affect the accuracy of the analysis, the researchers are recommended to describe these steps in detail. The diverse nature of text data in social media poses many challenges, including the massive volume of the data, noise and linguistics issues. Future works need to consider these challenging to reach more.

6. Challenges and Future Directions

The nature of social media text poses great challenges in developing text mining applications in social networking.

- A major challenge is the high volume of unstructured data which has a great potential in uncovering hidden patterns and valuable information.
- A common challenge of data extracted from social media is dealing with noisy and dynamic data and linguistic variations. Informal communication arises the level of noise and may contain misspelling, grammatical error, and varying writing styles. researchers need to apply preprocessing techniques to filter out irrelevant information. Many studies in this review have performed effective preprocessing steps, however, some studies unraveled this important stage.

- The efficient of clustering and classification depend on the problems and the type of data. Most efficient model can be produced by combining text mining techniques with social networking data analytics approaches.

The results of this review provide implications for future research in text mining in social networking. Studies in this review have stated the efficiency of the performance of their system is affected because of the huge text have to analyze. Therefore, analysis models need to manage the size of data and maintain the efficiency of the performance. Most of the study used only one source of data for analysis. Integrate data from different sources can improve the efficiency of the models. Moreover, using variety of the sources for the data can improve the accuracy of the model.

References

[1] L. Sorensen, "User managed trust in social networking - Comparing Facebook, MySpace and Linkedin," *2009 1st International Conference on Wireless Communication, Vehicular Technology, Information Theory and Aerospace & Electronic Systems Technology*, 2009, pp. 427-431.

[2] M. Naaman, "Social Multimedia: Highlighting Opportunities for Search and Mining of Multimedia Data in Social Media Applications," *Multimedia Tools Appl*, vol. 56, no. 1, pp. 9-34, 2012.

[3] S. Rani and P.Kumar, "A sentiment analysis system for social media using machine learning techniques: Social enablement," *Digital Scholarship in the Humanities*, vol. 34, no. 4, 2018.

[4] D. Sudarsa, S. K. Pathuri, and L. J. Rao, "Sentiment Analysis for Social Networks Using Machine Learning Techniques," *International Journal of Engineering and Technology(UAE)*, vol. 7, no. 2, pp. 473-476, 2018.

[5] K. Jani, M. Chaudhuri, H. Patel, and M. Shah, "Machine learning in films: an approach towards automation in film censoring," *Journal of Data, Information and Management*, vol. 2, pp. 55-64, 2019.

[6] N. Naw, "Twitter sentiment analysis using support vector machine and K-NN classifiers," *Int. J. Sci. Res. Publ*, vol. 8, no. 10, pp. 407–411, 2018.

[7] M. Grčar, D. Cherepnalkoski, I. Mozetič, and N. K. Petra, "Stance and influence of Twitter users regarding the Brexit referendum," *Computational Social Networks*, vol.4, no. 1, 2017.

[8] G. Piatetsky-Shapiro, "Data mining and knowledge discovery 1996 to 2005: overcoming the hype and moving from "university" to "business" and "analytics"." *Data Min Knowl Disc*, vol. 15, no. 1, pp. 99–105, 2007..

[9] J Cao, K. Basoglu, H. Sheng, and P. Lowry, "A Systematic Review of Social Networks Research in Information Systems: Building a Foundation for Exciting Future Research," *Communications of the Association for Information Systems*, vol 36, pp. 227-758, 2015.

[10] H. B. Haq, H. Kayani, S. K. Toor, A. Mansoor, and A. Raheem, "The Impact Of Social Media: A Survey," *International Journal of Scientific & Technology Research*, vol. 9, pp. 341-348, 2021.

[11] D. M. Boyd and N. B. Ellison, "Social network sites: Definition, history, and scholarship," *Journal of Computer-Mediated Communication*, vol. 13, no.1, pp. 210-230, 2007.

[12] R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins, "Trawling the Web for emerging cybercommunities," *Computer Networks: The International Journal of Computer and Telecommunications Networking*, vol. 31, no. 11-16, pp. 1481-1493, 1999.

[13] G. Miner, J. Elder, and R. Nisbet, *Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications*. Waltham, MA: Academic Press, 2012.

[14] F. Liu and L. Xiong, "Survey on text clustering algorithm - Research present situation of text clustering algorithm," *IEEE 2nd International Conference on Software Engineering and Service Science,* pp. 196-199.

[15] U Fayyad, G. Piatetsky-Shapiro, and S. Smyth, "The KDD process for extracting useful knowledge from volumes of data," *Communications of the ACM*, vol. **39**, no. 11, pp. 27–34, 1996.

[16] M. A. Emran and K. Shaalan, "A Survey of Intelligent Language Tutoring Systems," 2014 International Conference on Advances in Computing, Communications and Informatics (ICACCI), 2014, pp. 393-399.

[17] Y. Zhao, "Analysing twitter data with text mining and social network analysis," *in Proceedings of the 11th Australasian Data Mining and Analytics Conference (AusDM 2013*), 2013, p. 23.

[18] G. Forman and E. Kirshenbaum, "Extremely fast text feature extraction for classification and indexing," *In Proceedings of 17th ACM Conference on Information and Knowledge Management*, Napa Valley California, USA 2008, pp. 26–30.

[19] D. Blei, A. Ng, and M.I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. **3**, no. 1, pp. 993–1022, 2003.

[20] M. Rüdiger, D. Antons, and O. Salge, "From text to data: on the role and effect of text pre-processing in text mining research," *Academy of Management Proceedings*, vol. 2017, no.1, 2017.

[21] K. Yoshida, Y Tsuruoka, Y. Miyao, and J. Tsujii, "Ambiguous part-of-speech tagging for improving accuracy and domain portability of syntactic parsers*," In Proceedings of 20th International Conference on Artificial Intelligence*, Hyderabad, India, 2007, pp. 1783–1788.

[22] J. Hua, W. D. Tembe and E. R. Dougherty, "Performance of feature-selection methods in theclassification of high-dimension data," *Pattern Recognition,* vol. 42, no.3 , pp. 409–424, 2009.

[23] C. B. H. Shekar and G. Shoba, "Classification of documents using Kohonens self organizing map. International," *Journal of Computer Theory and Engineering (IACSIT)*, vol. 1, no. 5, pp. 610–613, 2009.

[24] J. Akaichi, Z. Dhouioui and M. J. López-Huertas Pérez, "Text mining facebook status updates for sentiment classification," *2013 17th International Conference on System Theory, Control and Computing (ICSTCC)*, 2013, pp. 640-645.

[25] E. Cambria, "Affective Computing and Sentiment Analysis," in *IEEE Intelligent Systems*, vol. 31, no. 2, pp. 102-107, 2016.

[26] H. Gupta, M. S. Jamal, S. Madisetty and M. S. Desarkar, "A framework for real-time spam detection in Twitter," *2018 10th International Conference on Communication Systems & Networks (COMSNETS)*, 2018, pp. 380-383.

[27] A. Akundi, B. Tseng, J. Wu, E. Smith, S. Mandapaka, and F.Aguirre, "Text Mining to Understand the Influence of Social Media Applications on Smartphone Supply Chain," *Procedia Computer Science*, 2018, PP. 87-94.

[28] V. Dhanalakshmi, D. Bino and A. M. Saravanan, "Opinion mining from student feedback data using supervised learning algorithms," *2016 3rd MEC International Conference on Big Data and Smart City (ICBDSC)*, 2016, pp. 1-5.

[29] T. Nicolas and D. Constantinos, "Opinion Mining From Social Media Short Texts: Does Collective Intelligence Beat

Deep Learning?," *Frontiers in Robotics and AI*, vol. 5, p. 138 , 2019.

[30] F. Gesualdo et al., "How do Twitter users react to TV broadcasts dedicated to vaccines in Italy?," *Eur J Public Health*, vol. 30, no. 3, pp. 510-515, 2020.

[31] G. Nath, R. Ghosh, and R. Nath, "Cluster Analysis of Customer Reviews: Summarizing Customer Reviews to Help Manufacturers Identify Customer Satisfaction Level," *Procedding of 7th international conference of business analytics and intellegenc*, India, 2019.

[32] C. Lewis and S. Young, "Fad or future? Automated analysis of financial text and its implications for corporate reporting," *Accounting and Business Research*, vol. 49, no. 5, pp. 587-615, 2009.

[33] M. D. Tapi Nzali, S. Bringay, C. Lavergne, C. Mollevi, and T. Opitz, "What Patients Can Tell Us: Topic Analysis for Social Media on Breast Cancer," *JMIR Med Inform*, vol. 5, no. 3, 2017.

[34] M. Ezzeldin and W. El-Dakhakhni, "Metaresearching Structural Engineering Using Text Mining: Trend Identifications and Knowledge Gap Discoveries," *Journal of Structural Engineering-asce*, vol. 146, pp. 04020061, 2020.

[35] A. Amado, P. Cortez, P. Rita, and S. Moro, "Research Trends On Big Data In Marketing: A Text Mining And Topic Modeling Based Literature Analysis," *European Research on Management and Business Economics (ERMBE)*, vol. 24, no. 1, pp. 1-7, 2017.

[36] J. Kalyanam, T. Katsuki, G. Lanckriet, and T. M.Mackey, "Exploring trends of nonmedical use of prescription drugs and polydrug abuse in the Twittersphere using unsupervised machine learning," *Addict Behav*, vol. 65, pp. 289-295, 2017.

[37] B. Batrinca, "Social media analytics: a survey of techniques, tools and platforms," *AI & SOCIETY,* vol. 30, no. 1, pp. 89–116, 2015.

[38] K. Sumathy and M. Chidambaram, "Text Mining: Concepts, Applications, Tools and Issues An Overview," *International Journal of Computer Applications*, vol. 80, pp. 29-32, 2013.

[39] B. Ahmed, G. Ali, A. Hussain, A. Baseer, and J. Ahmed, "Analysis of Text Feature Extractors using Deep Learning on Fake News", *Eng. Technol. Appl. Sci. Res.*, vol. 11, no. 2, pp. 7001–7005, Apr. 2021.

[40] J. Piskorski, R. Yangarber, "Information Extraction: Past, Present and Future," *in Multi-source, Multilingual Information Extraction and Summarization*, Berlin, Heidelberg, Springer, 2013, pp. 23-49.

[41] K. Hammar, S. Jaradat, N. Dokoohaki, and M. Matskin, "Deep Text Mining of Instagram Data without Strong Supervision," *IEEE/WIC/ACM International Conference on Web Intelligence (WI)*,2018, pp. 158-165.

[42] W. He, S. Zha, and L. Li, "Social media competitive analysis and text mining: A case study in the pizza industry," *International Journal of Information Management*, vol. 33, no. 3, pp. 464-472, 2013.

[43] B. Nie and S. Shouqian, "Using Text Mining Techniques to Identify Research Trends: A Case Study of Design Research," *Applied Sciences,* vol. 7, pp. 401, 2017.

[44] K. Wang and Y. Zhang, "Topic Sentiment Analysis in Online Learning Community from College Students" *Journal of Data and Information Science*, vol.5, no.2, pp.33-61, 2020.

[45] J. Mir, A. Mahmood, and S. Khatoon, "Aspect Based Classification Model for Social Reviews", *Eng. Technol. Appl. Sci. Res.*, vol. 7, no. 6, pp. 2296–2302.

[46] A. Golestani, M. Masli, N. S. Shami, J. Jones, A. Menon and J. Mondal, "Real-Time Prediction of Employee Engagement Using Social Media and Text Mining," *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 2018, pp. 1383-1387.

[47] I. Smirnov, "Estimating educational outcomes from students' short texts on social media," *EPJ Data Sci.* vol. 9, pp. 27, 2020.

[48] A. Bhardwaj, "Sentiment Analysis and Text Classification for Social Media Contents Using Machine Learning Techniques," *Proceedings of the 2nd International Conference on IoT, Social, Mobile, Analytics & Cloud in Computational Vision & Bio-Engineering (ISMAC-CVB 2020),* November, 2020.

[49] H. Isah, P. Trundle and D. Neagu, "Social media analysis for product safety using text mining and sentiment analysis," *2014 14th UK Workshop on Computational Intelligence (UKCI)*, 2014, pp. 1-7.

[50] W. He, X. Tian, R. Tao, W. Zhang, W. Yan, and V. Akula, "Application of social media analytics: A case of analyzing online hotel reviews", *Online Information Review*, 2017.

[51] D. Flores-Ruiz, A. Elizondo-Salto, and M. Barroso-González, "Using Social Media in Tourist Sentiment Analysis: A Case Study of Andalusia during the Covid-19 Pandemic", *Sustainability*, vol. 13, no. 7, pp. 3836, 2021.

[52] C. Bhadane, H. Dalal, and H. Doshi, "Sentiment Analysis: Measuring Opinions," *Procedia Computer Science*, vol. 45, pp. 808-814, 2015.

[53] H. Saif, Y. He, M. Fernandez, and H. Alani, "Contextual semantics for sentiment analysis of Twitter," *Information Processing & Management*, vol. 52, no.1, pp. 5-19, 2016.

[54] Z. Hu, J. Hu, W. Ding and X. Zheng, "Review Sentiment Analysis Based on Deep Learning," *2015 IEEE 12th International Conference on e-Business Engineering*, 2015.

[55] S. Gündüzalp and G. Şener, "The Analysis of Opinions About Teaching Profession on Twitter Through Text Mining," *Research on Education and Media*, vol.12, no.1, pp.3-12, 2020.

[56] X. Chen, M. Vorvoreanu, and K. Madhavan, "Mining Social Media Data for Understanding Students' Learning Experiences," in *IEEE Transactions on Learning Technologies*, vol. 7, no. 3, pp. 246-259, July-Sept. 2014.

[57] M. Ahmad, and S. Aftab, "Analyzing the Performance of SVM for Polarity Detection with Different Datasets," *International Journal of Modern Education and Computer Science*, vol. 9, pp. 29-36, 2017.

[58] A. Abbasi, Y.Zhou, S. Deng, and P. Zhang, "Text analytics to support sense-making in social media: a language-action perspective," MIS Q, vol. 42, no. 2, pp. 427–464, 2018.

[59] M. Madhukar and S. Verma, "Hybrid Semantic Analysis of Tweets: A Case Study of Tweets on Girl-Child in India", *Eng. Technol. Appl. Sci. Res.*, vol. 7, no. 5, pp. 2014–2016, Oct. 2017.

[60] K. Wegrzyn-Wolska, L. Bougueroua and G. Dziczkowski, "Social media analysis for e-health and medical purposes," *2011 International Conference on Computational Aspects of Social Networks (CASoN)*, 2011, pp. 278-283.

[61] A. Zielinski, S. Middleton, L. Tokarchuk, and X. Wang, "Social Media Text Mining and Network Analysis for Decision Support in Natural Crisis Management," *International Conference on Information Systems for Crisis Response and Management ISCRAM*, Baden-Baden, Germany, 2013.

[62] R. Bekkerman, R. El-Yaniv, N. Tishby, and Y. Winter, "Distributional Word Clusters vs. Words for Text Categorization," *Journal of Machine Learning Research*, vol. 3, pp. 1183-1208, 2003.