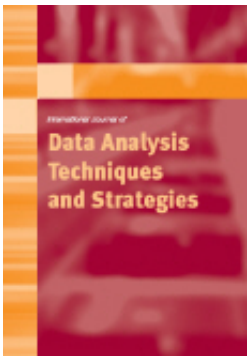


International Journal of Data Analysis Techniques and Strategies

ISSN online: 1755-8069 - ISSN print: 1755-8050
https://www.inderscience.com/ijdats

Text mining on social media data: a systematic literature review

Sarah Bukhari, Muhammad Ramzan

DOI: 10.1504/IJDATS.2024.10062936

Article History:

Received:	26 May 2022
Last revised:	22 November 2022
Accepted:	23 December 2022
Published online:	19 March 2024

Text mining on social media data: a systematic literature review

Sarah Bukhari*

Department of Information Technology,
Bahauddin Zakariya University,
Multan, Pakistan
Email: bukhari_sarah@yahoo.com
*Corresponding author

Muhammad Ramzan

Government Primary School Education Department,
105/WB, Vehari Road, Pakistan
Email: ramzan0713@gmail.com

Abstract: Text mining is the process of getting meaningful information from unstructured data. In this paper, a precise writing overview was directed to research text mining via online media information. Thus, a comprehensive deliberate writing audit (SLR) was completed to explore online media as a hotspot for the perception of text mining. For this reason, 40 articles were chosen from different notable sources after a concentrated SLR cycle of looking, sifting, and implementing the incorporation and avoidance models. As a result, the text mining strategies via web-based media information were featured regarding online media as a wellspring of data. A detail SLR which features the need of message mining methods on most recent online media information, cover more kinds of web-based media which were not shrouded in past work and furthermore present qualities and shortcomings of text mining strategies utilised in web-based media.

Keywords: social media; text mining; text mining techniques; role of social media; social media types.

Reference to this paper should be made as follows: Bukhari, S. and Ramzan, M. (2024) 'Text mining on social media data: a systematic literature review', *Int. J. Data Analysis Techniques and Strategies*, Vol. 16, No. 1, pp.82–104.

Biographical notes: Sarah Bukhari is working as an Assistant Professor in Information Technology Department, BZU, Multan, Pakistan. She received her PhD from University of Malaya, Malaysia. She supervised MS students and involved in various research, leading to publication of a number of academic papers in the areas of information systems specifically is social network analysis, data mining, bio-informatics, information seeking behaviour, networking and social media.

Muhammad Ramzan is working as a Primary School Teacher in Government School Education Department, 105/WB, Vehari, Pakistan. He did his MS from University, NFC_IET, Multan, Pakistan. He received his MCS from the Comsats University Vehari, Pakistan. He specialised in the area of text mining and social media.

1 Introduction

Text mining is a cycle to extract information in view of text. The wellspring of text mining should be coming from online media clients' exercises; e.g., posting, fans page, bunch, hashtag, tweet, and so on as crude information and data, the text mining sources can be gold data to be worked for scholarly dissecting purposes (Kaburuan et al., 2014). Text mining has been an extraordinary methodology in the writing for archive arrangement. It depends on recognisable proof and investigation of many texts for removing helpful data from unstructured information. Thusly, by applying text mining to a corpus of texts, we can separate valuable data and secret examples in text content as well as uncovering information. To this end, in this paper we concentrate on the use of text mining to a bunch of genuine occasions on Twitter to consequently recognise their validity (Hassan, 2018). With the quick ascent of text-based substance over the web, including articles and recorded archives, concentrating on present day procedures to acquire discernment into what is composed there by individuals is inescapable. Individuals regularly use language vernaculars over the web to post a status. Thusly, understanding the vernacular in which a text scrap is composed with is significant for knowledge what has been composed. With the significance of inferring a (genuine) positive judgement and to have a profound comprehension of the semantics of the normal language in internet-based informal organisations like Twitter, current innovations, like Lexica, corpora, and ontologies should be appropriately built and utilised (El-Jawad and Hodhod, 2017).

Online social media is an interconnected stage that assists with building social connection between normal interest individuals and furnishes the office to be associated with other interdisciplinary regions like business, governmental issues, sports, monetary organisations (Agrawal and Kaushal, 2016). Twitter, Facebook and so forth are one of the most popular persons to person communication sites. Here client speaks with one another by join various networks and conversation gatherings. Web-based media has the capacity to tackle dexterity issues among clients and furthermore it further develops the social mission's viability. It does this by spreading the necessary data. Unstructured and semi-organised language is utilised by clients to speak with one another. The spellings and definite linguistic development of a sentence is not given need by the clients. This might prompt various sorts of ambiguities, as lexical, syntactic, and semantic and so on the most requesting undertaking to perform is to separate legitimate examples with data exhaustively from unstructured structure. Text mining is an answer of previously mentioned issues (Dastanwala, 2016).

Web is a rich source of data and communication. That turns out to be more famous by its minimal expense, straightforward entry and accessibility. Subsequently, a rich number of uses are created by the utilisation of web among them the web-based media sites are additionally a critical commitment of web. In these sites, a colossal number of clients are conveying each other in text design. Hence, some of the time the maltreatment and other abuse is likewise conceivable by the different nature people groups. Furthermore, the manual examination of each imparted text is mind-boggling task. Subsequently, the text characterisation strategies are utilised for observing the abuse and maltreatment of the person-to-person communication sites (Dasondi, 2016).

Outcomes of the past work are deficient for our inspirations on account of the going with reasons: by far most of the work is based on Twitter, with little care in regards to picture sharing stages like Instagram and evidently, no previous assessment of complex, multi-name, different evened out extraction and portrayal in internet-based media has been made. In this paper, we revolve around the task of portraying the assessments about different online media applications like twitter, Facebook, YouTube, WhatsApp, etc. as opposed to just Twitter. The paper is organised as follows. First, we discuss the methods applied for the SLR. Then, the results of the review are discussed, and finally, we conclude with a summary of the paper, and highlight the limitations of the research and make recommendations for future works.

2 Research methodology

A SLR is pointed toward introducing and assessing the writing identified with the examination theme by using an exhaustive and auditable system (Hamid et al., 2016). This examination took on the SLR, an approach proposed (Hamid et al., 2016) to comprehend the job of text mining via social media information. This SLR includes a few discrete exercises, which are requested into three essential stages: arranging the audit, leading the survey and detailing the survey. The authenticity and immovable nature of the exact study procedure are confirmed by ensuing every movement as suggested in the three stages. In addition, Zhitomirsky-Geffet et al. (2009) proposed a catalogue BOW for online bibliography that is based on hierarchical concept index where matching topics get easily returned instead of long entries.

The arranging action centres on fostering the audit convention. It clarifies the work process; in what way the audit is led through the analyst. The arranging stage includes the ID of the examination questions, the hunt methodology and the assessment of the assets, the consideration and rejection measures, the worth appraisal of the assets, and the technique for investigation. The subsequent stage implements the characterised convention in the arranging stage, while the clarification of the last description is expounded in the last stage. Notwithstanding, the philosophy by utilised in this audit was adjusted in view of crafted by Hamid et al. (2016). Hamid et al. (2016) likewise led a SLR named ‘Job of web-based media in data looking for conduct of global understudies. An orderly writing survey’ and worked on the system for simple reference. Figure 1 portrays the exercises of each stage Hamid et al. (2016), while every one of the SLR stages will be depicted in the accompanying subcategories.

2.1 Phase 1: planning

2.1.1 Identify the need

Recognising the requirement, the presentation segment referenced that there is a need to concentrate on message mining via online media information and the job of web-based media. Subsequently, three exploration questions were created to help the writing audit process.

2.1.2 Research questions

Research questions the examination questions, which are explicitly tended to through the review are by means of the following:

RQ1 Which text mining techniques have been used in social media?

RQ2 How to develop a systematic taxonomy that reflecting the text mining techniques used in social media?

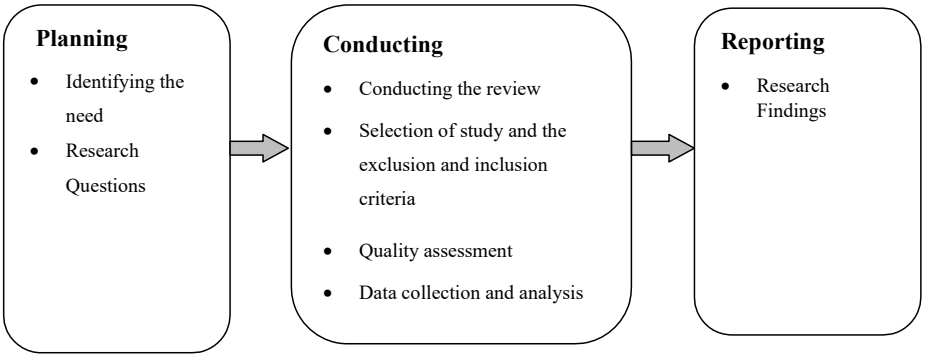
RQ3 What are the strengths and weaknesses of the implemented text mining techniques in social media?

2.2 Phase 2: conducting

2.2.1 Conduct the review

To get a feeling of the present status of the exploration on data looking for conduct, the job of web-based media and global understudies, both scholastic diaries and gathering procedures were analysed through logical information bases.

Figure 1 Activities in systematic literature review



In the first place, the survey started with a quest for tracking down significant information from various sources. In this review, two logical web-based information bases were utilised in view of the rules and the importance and accessibility of the inquiry terms. Table 1 displays the chose logical internet-based datasets and the purposes behind utilising them. Besides, extra examinations from non-ISI diaries or meetings were incorporated assuming still up in the air to be important and helpful for the review. The two internet-based information bases recorded in Table 1 were looked through utilising the recorded pursuit articulations displayed in Table 2. Microsoft word was utilised along with Mendeley programming to deal with the recovered articles. Microsoft word was utilised to aid the improvement of summery of articles, while Mendeley saved the articles as indicated by the year/writer/time they were distributed in and dealt with the references throughout the composition of the SLR.

Table 1 Selected scientific online databases

Online databases	*Reason to choose*
IEEE	Covers the professional association for the advancement of technology
Google Scholar	It includes most peer reviewed online academic journals and books, conference papers, thesis and dissertation
Science Direct	It covers the engineering, physical sciences, health sciences, social science, health science and the humanities.

Table 2 Search expressions used in the study

S. no.	*Search expression*
1	Text mining
2	Social media data
3	Text mining techniques
4	Facebook
5	Twitter
6	Instagram
7	YouTube

2.2.2 Study selection and the inclusion and exclusion criteria

The determination of studies was done through five stages, as portrayed in Table 3. The legitimisation for this confirmation methodology was to see and pick the papers that arranged with the complaints of the SLR. It was not unforeseen that the picked articulations (information bases) would return every one of the papers connected with the subject. Thusly, a few circuit and excusal standards were used to improve the outcomes. The Mendeley reference chief assisted with dealing with the copy references all the more productively and to create an incorporated document.

The incorporation and avoidance models were utilised to guarantee that main the significant articles were remembered for the SLR cycle. As online media began to turn into a peculiarity in 2000, the analysts chose to incorporate the time span from the year 2016 to 2021 as summed up in Table 4. In the meantime, data looking for conduct point is utilised throughout the previous forty years and to help the assertions we include the ancient references. The consideration and prohibition measures are introduced in Table 4.

2.2.3 Quality evaluation

In this step, every one of the involved papers was outlined. This development was done throughout the information mining improvement and was utilised to guarantee that the involved articles completed a basic commitment with systematic literature review association. The going with three main quality evaluation rules were applied to each of the included papers in basically the same manner as summed up in the overview under to come by the exact results (Dyba and Dingsoyr, 2008):

- Rigour: has an exhaustive and suitable methodology been claimed to enter research strategies in the review?
- Credibility: these are the discoveries first rate and significant?

- Relevance: it describes how valuable are the discoveries to the advanced education research local area?

Quality edge:

1 Is this research is founded on paper (or it simply an 'examples learned' story dependent upon well-qualified assessment)?

2 Are the examination points have a reasonable assertion?

3 Is there a sufficient depiction of the setting in which the exploration had done?

Table 3 Inclusion phases

Phase	*Phase description*
P1	Selection of studies-based on the conducted search
P2	Screening: inclusion-based on the inclusion criteria
P3	Screening: exclusion-based on the exclusion criteria
P4	Screening: exclusion-based quality assessment criteria
P5	Confirmation

Table 4 Inclusion/exclusion criteria

Inclusion criteria	Directly answer any at least one exploration question Focus on the information mining via web-based media information Published in years: 2016-2021
Exclusion criteria	Prohibit irrelevant books or upward introductions Exclude that is not connected with the exploration of the field Papers, when just dynamic. No whole text were accessible Papers that did not qualify the consideration rules

Rigour:

4 Were the examination configurations suitable to indicate the points of exploration?

5 Were the enrolment systems suitable with the points of exploration?

6 Were there benchmark collections with which we analyse medicines?

7 Was the information gathered as that tended to the examination problems?

8 Were the information examinations adequately thorough?

Credibility:

9 Has the connection among scientist and members been considered to a satisfactory degree?

10 Is there an unmistakable assertion of discoveries?

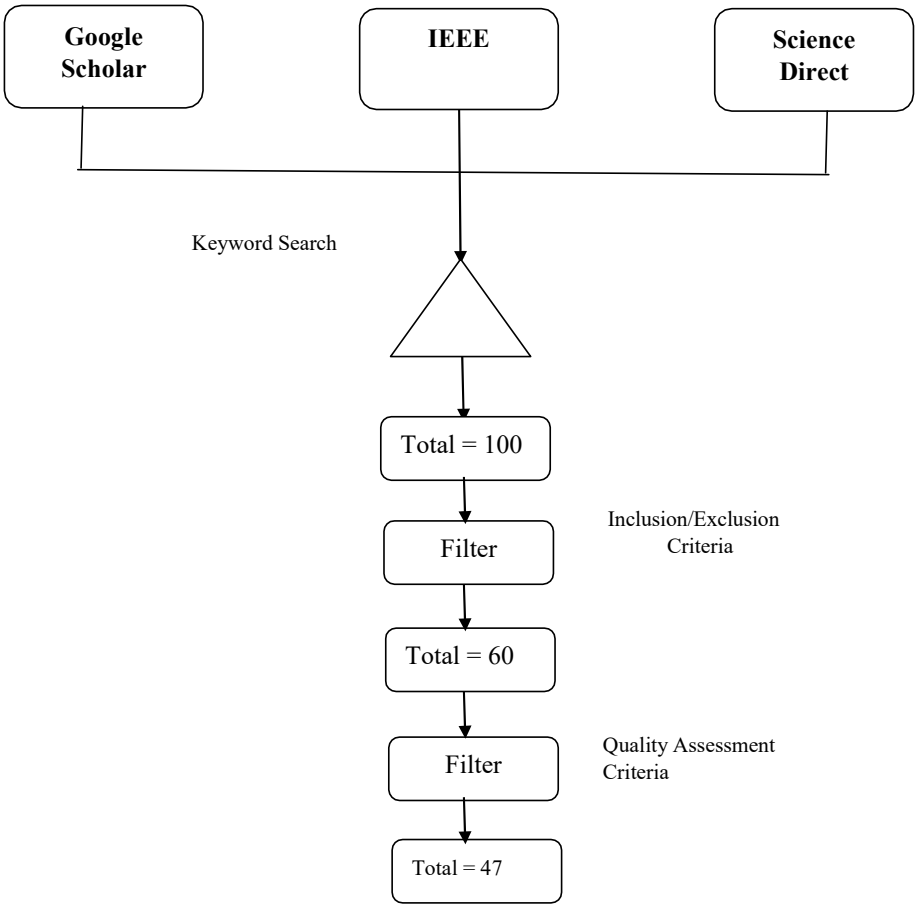
Relevance:

11 Is the investigation of significant worth for exploration or practice?

The hidden three estimates formed the base quality edge, to bar those research articles, which did not qualify the destinations of survey. The rules 4–8 present the problem of caution, which measured the examination method used, information game plan

contraptions and frameworks, and additionally the dependability of the revelations. Rules 9 and 10 were connected with the reliability of the research articles, which guaranteed the genuineness, and sincerity of the disclosures. The last rule, rule 11, indicates the importance of the work to the significant level preparation research area. The possibility of the picked articles was endeavoured by the usage of an appraisal tool dependent upon the depicted rules in the outline given above.

Figure 2 Publication collection method flow



2.2.4 Data collection and analysis

An information assortment structure was intended to gather the most pertinent data from the chose papers to work with the method involved with examining the assembled information. This structure, which from here on out will be alluded to as the information outline, is introduced in Table 5. The planning was done after a further review of the picked studies by the subject matter experts, prior to comprehension was touched on all of the problems in the last data, which has been accomplished.

2.3 Phase 3: reporting

2.3.1 Findings

These fragment reports the outcomes of the systematic literature review, which was driven. As included in Segment 2, the purpose of the decision association is to recognise anything that number pertinent papers as could be took into account the systematic literature review. The procedure for coordinating the chase and decision of the related articles by databases is present in Figure 1. The fundamental chase is coordinated by the picked openings through using the described hunt enunciations and the pursuit cycle that had been described ahead of time. The basic inquiry yield returned an outright 100 papers and 40 papers have been chosen, as summarised in Table 6.

Table 5 Data schema

Essential information	It includes the title and the author(s) of the paper
Publication	It refers to whether the distribution is a diary or meeting continuing
Year	It refers to the distribution years 2016 to 2021 of the articles
Objectives	It refers to the goals that the paper attempts to satisfy
Filed	It recognises the flood of the paper, regardless of whether it is identified with data looking for conduct or online media or worldwide understudies
Focus	It manages the focal point of the paper. It centres around the job of web-based media, giving data to defeat the issues of global understudies identified with concentrate abroad
Future work	It proposes the future work and the difficulties identified with the exploration questions

After the assurance of papers reliant upon the described guidelines (see Table 5), the quality assessment procedure is executed. The possibility of the picked articles was endeavoured by the usage of an evaluation tool subject to the portrayed guidelines in Table 5. Each included review was surveyed dependent upon its quality limit, thoroughness, believability and significance. Then, at that point, the solicitations on the quality measures depended upon a no or yes rule, 'yes' showing the meaning of the paper as indicated by excellence and 'no' displaying the pointlessness of the paper in friendly event the quality. Considering the 'yes' and 'no' scales, diverse excellence appraisal orders, as depicted in the given list, that were tended to for every included research article. As obviously basically every one of the research articles were inside the edge of huge worth to be melded, this presented that the importance and watchfulness quality assessment rules had extra thought. Unusually, authenticity had less consolidation. Something like 80% of all the quality appraisal standards were presented by the 'Yes' answers.

2.3.2 Result and discussion

Based on the review, there is a diversity of text mining techniques, which are applied on social media data, as Table 6 painted it. This table presents that most of the researchers debated the text mining techniques on social media data.

These text-mining techniques, for the purpose of discussion are categorised in different social media types as Twitter, Facebook, Instagram, YouTube and purely social media data.

RQ1 Which text mining techniques have been used in social media?

RQ2 How to develop a systematic taxonomy that reflecting the text mining techniques used in social media data?

RQ3 What are the strengths and weaknesses of the implemented text mining techniques in social media?

Table 6 Text mining techniques in social media

Social media type	*Text mining technique*	*Text mining algorithm/tools*	*References*
Twitter	Text classification and NLP	Decision tree, k-nearest neighbours, logistic regression, SGD classifier, random forest, SVM linear, naive Bayes	Elsayed et al. (2019), Ahmad et al. (2022), Choudhary (2018), Permana et al. (2022), Hidayat and Parwanto (2022) and Malhotra and Malhotra (2018)
Twitter	Text classification, Natural language processing	Different software's are used like API, REST API, and JSON. Python 3.x along with the Natural Language Toolkit (nltk) and tweepy libraries	Ardra et al. (2017) and Dessai and Usgaonkar (2022)
Twitter	Sentiment analysis 'syuzhet' package in R		Choudhary (2018)
Twitter	Text classification		Dasondi (2016)
Twitter	Text mining using fuzzy keyword match, support vector machine and Twitter latent Dirichlet allocation (LDA)	Fuzzy keyword match, support vector machine and Twitter latent Dirichlet allocation (LDA)	Dastanwala (2016)
Twitter	Text mining, text retrieval	A novel Twitter-text-mining model was effectively constructed.	Kaburuan et al. (2014)
Twitter	Incremental mining technique with set of frequent word item (SFWI)	CP-tree algorithm.	Sa et al. (2017)
Twitter	Clustering technique	Profiling social media users (PSMU) algorithm	Vasanthakumar et al. (2019)

Table 6 Text mining techniques in social media (continued)

Social media type	*Text mining technique*	*Text mining algorithm/tools*	*References*
Twitter	Information extraction Text categorisation	Orange data mining software, meaning cloud and linguistic inquiry and word count	Roundtree (2018)
Twitter	Text mining using Weka tool	Weka tool is used	Hassan (2018)
Twitter	Sentiment analysis		Razavi and Rahbari (2020), El-Jawad and Hodhod (2017), Costales et al. (2022), Hidayat and Parwanto (2022) and Fakhrur et al. (2019)
Twitter	Text mining using TF-IDF algorithm	TF-IDF algorithm	Sutar (2017)
Facebook	Two text-mining techniques are string similarity indexes and corpus-based indexes		Agrawal and Kaushal (2016)
Facebook	Sentiment analysis for text preprocessing	Tools used in this research are apache spark and Apache Hadoop	Sriyanong et al. (2018)
Facebook	Sentiment analysis		Reguera et al. (2017)
Facebook	Clustering	Rapid Miner tool	Salloum et al. (2017)
Facebook	Logistics regression, naïve Bayes, support vector machine (SVM) and decision tree		Permana et al. (2022)
Instagram	Natural language processing		He et al. (2017)
Instagram	Clustering is applied	K-means algorithm	Fiallos et al. (2018)
Instagram	Classification	Naïve Bayes algorithm	Bayes (2019)
Instagram	Text mining using the facial recognition API		Zhou et al. (2016)
Instagram	Classification	Linear SVM classifiers based on text-based N-gram, logistic regression, decision	Hosseinmardi et al. (2015)
Web	Clustering	K-means	Cakir (2016)
Web	Preprocessing steps for sentiment analysis in Brazilian Portuguese social media		Cirqueira et al. (2018)
Web	Text mining	Used Kkokkoma Korean analyser	Kim and Ha (2016)

Table 6 Text mining techniques in social media (continued)

Social media type	*Text mining technique*	*Text mining algorithm/tools*	*References*
Web	Classification	CNN, LSTM, LSTM-CNN and SVM	Chen et al. (2018)
Web	Text mining using SVM	SVM model	Chen and Chuang (2019)
Web	Classification	Word emotion refinement algorithm based on word emotion association network (WEAN)	Jiang et al. (2017)
Web	Text mining	Naive Baves, random forest, support vector machine (SVM), K-nearest neighbour (KNN)	Jeelall and Cheerkoot-Jalim (2020)
YouTube	Classification	Naive Bayes, K-nearest neighbour, support vector machine (SVM) and one collective classifier namely, bagging	Zaman and Sharmin (2017)
YouTube	Machine learning and text mining techniques	Latent Dirichlet allocation probabilistic model	Vlachos and Tan (2018)
YouTube	Classification using gender identification model	The classification model that recognises the author's gender of a given data	Zahir et al. (2019)
Social media (mix)	Sentiment analysis		Nusantara (2019) and El Haddaoui et al. (2021)
Social media (mix)	Text mining		Kibtiah et al. (2020)
Social media (mix)	Text mining using SVM	SVM	Purnomo et al. (2016)
WhatsApp	Text mining using S2NOW algorithm	S2NOW algorithm is used.	Johari et al. (2021)
WeChat	Text mining and sentiment analysis techniques	Latent Dirichlet allocation (LDA)	Wu et al. (2018)
Telegram	Sentiment analysis techniques		Razavi and Rahbari (2020) and Reguera et al. (2017)

Figure 3 Overall representation of research (see online version for colours)

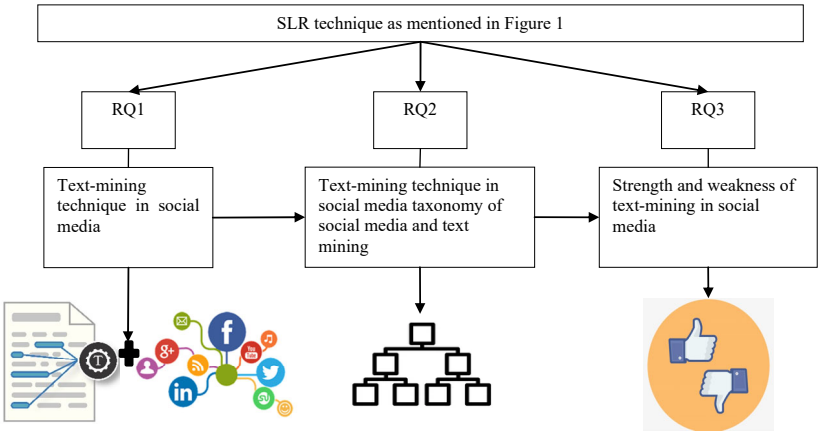


Figure 4 Taxonomy of text mining techniques used in social media

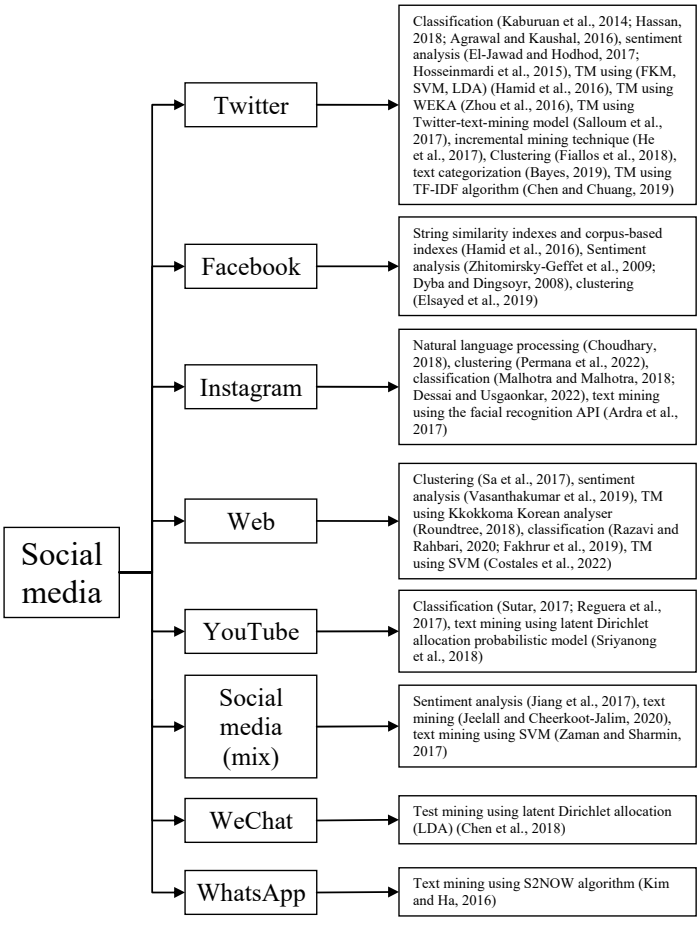


Table 7 Strength and weakness of text mining techniques in social media

Social media type	*Text mining technique*	*Strength*	*Weakness*	*References*
Twitter	NLP and ML algorithms	Good to extract knowledge from extremely unstructured data and classification of data.	Bad to review the computational shift accomplished in greater datasets by unstructured text examination techniques.	Elsayed et al. (2019)
Twitter	Sentiment analysis	Helpful for people to making decisions and improving performances	Not a 100% true survey	Ardra et al. (2017)
Twitter	Sentiment analysis 'syuzhet' package in R.	Helpful for people to making decision and improving selling.	This work cannot be used to predict and for actual rating of product	Choudhary (2018)
Twitter	Text classification data model with JAVA technology	With the increasing size of dataset, the performance of the classifier is improved.	The classifier is not appropriate for characterisation of the text in various classes. It is great just for two classes.	Dasondi (2016)
Twitter	Fuzzy keyword match, support vector machine and Twitter latent Dirichlet allocation (LDA)	The performance of Fuzzy keyword match is better than SVM and LDA for this scenario.	This work not used for online business, increasing customers, and presenting offers for related groups	Dastanwala (2016)
Twitter	An original Twitter-text-mining model was effectively built.	This model has multiple functions like Extractor, cleaner, analyser, calculator and predictor, especially for Indonesian community	This model is not capable for numerical words screening	Kaburuan et al. (2014)
Twitter	Incremental mining technique with set of frequent word item (SFWI) representing using CP-tree algorithm.	Gradual mining strategy with SFWI addressing utilising CP-tree calculation had been affirmed achieved to build the productivity of memory utilisation Around 1.84 occasions extra proficient than without steady procedure utilising FP-growth calculation. Furthermore, the time cycle could be speedier around 1.66 occasions with steady procedure.	It is not great when it is assumed the foundations of memory utilisation and effectiveness of time process for enormous text information.	Sa et al. (2017)

Table 7 Strength and weakness of text mining techniques in social media (continued)

Social media type	*Text mining technique*	*Strength*	*Weakness*	*References*
Twitter	Profiling social media users (PSMU) algorithm	The proposed PSMU calculation customs remarkable bunches and profiles the Twitter users definitively by accomplishing 97.53 % precision.	This algorithm does not provide profiling the social media users founded on location tagged posts.	Vasanthakumar et al. (2019)
Twitter	Data analysis to classify tweets and determine sentiment, I applied two software packages: linguistic inquiry and word count and meaning cloud	Meaning cloud researches unstructured criticism utilising text examination and semantic handling to perceive named elements and allocate text to bunch in a predefined scientific classification. LIWC applies text word references revalidated and intended to classes of sentiments, contemplations, inspirations and character.	The study scheme also inherits confines of text mining, which prevents the adjacent reading of all content.	Roundtree (2018)
Twitter	Software Weka is used. Different algorithms are applied. DT gave best result.	A decision tree not requires the normalisation of data.	A small variation in the data can cause a big change in the arrangement of the decision tree causing uncertainty.	Hassan (2018)
Twitter	Different deep learning and machine learning algorithms and a hybrid model have been used.	This hybrid model gives best result on English tweets than decision trees, naive Bayes, neural networks, random forrest and recurrent neural networks.	This hybrid model cannot examine the Arabic language tweets.	El-Jawad and Hodhod (2017)
Facebook	Different Indexes used in This article are Jaccard, dice, Ochiai, overlap, similarity matching, TF-IDF, latent semantic analysis	LSA is capable of assuring decent results, much better than Jaccard, dice, Ochiai, overlap, similarity matching, TF-IDF.	The suggested work displays the potential points but not whole showcase about irrelatively between comments and posts.	Agrawal and Kaushal (2016)
Facebook	Tools used in this research are apache spark and Apache Hadoop	This framework is capable of assuring goof results, this framework has less computation time than a single machine.	This framework is not considered for using the further big data technologies.	Sriyanong et al. (2018)

Table 7 Strength and weakness of text mining techniques in social media (continued)

Social media type	*Text mining technique*	*Strength*	*Weakness*	*References*
Facebook	Datasets are taken from FEDER and Facebook. R Language is used also. Text Blob Python library is used to perform sentiment analysis	We can realise that which words are recurrent in the decalogue by this tool and which words are less recurrent in Facebook.	The small number of groups analysed in the Facebook dataset.	Reguera et al. (2017)
Facebook	Face pager, Rapid Miner tool.	By this model, we can see what principle points are considered as the interesting issues that were talked about across all news channels given by Facebook posts.	By this model, we can see what essential subjects are considered as the intriguing issues that were inspected across all news channels given by Facebook posts.	Salloum et al. (2017)
Instagram	This review contributions a work process strategy for utilising normal language preprocessing, feeling investigation and text mining procedures to assess online literary substance.	The benefit of our work process strategy from different strategies is that our technique will lead opinion investigation and text digging for chosen classifications as opposed to removing the total informational collection. By applying this, organisations can accentuation on the intrigued classes for profound examination to get point-by-point discernments.	The shortcoming of this work process technique is that it needs a few human exertions for information investigation and clarification to perceive significant gatherings or subjects from the information test that is selected randomly from the dataset.	He et al. (2017)
Instagram	K-means clustering algorithm and the TF-IDF (term frequency-inverse document frequency) model is applied.	The procedure of this work can be suitable in areas like, market research, digital marketing, social media, social studies, opinion polls and other fields.	This approach could include test with other hashtags correlated to fields other than tourism.	Fiallos et al. (2018)
Instagram	Naïve Bayes algorithm	Naive Bayes algorithm can be applied to categorise comments on Instagram and a good option for this task – once again, it showed its competencies.	This model shows weak performance to remove the stop words and classification of comments in each post.	Bayes (2019)

Table 7 Strength and weakness of text mining techniques in social media (continued)

Social media type	*Text mining technique*	*Strength*	*Weakness*	*References*
Instagram	The facial recognition API	This tool is helpful to catch potential network between drug related pages and drug users on Instagram. A significant invention of this study is in face image analysis and in multimedia data analysis.	This tool is not able to categorise drug users, dealers, sellers and glassmakers/legitimate and not able to categorise drug-users and non-drug-users.	Zhou et al. (2016)
Instagram	Linear SVM classifiers established on decision, logistic regression, text-based N-gram.	By this classifier, substantial number of media meetings containing cyber aggression and profanity that were not marked as cyberbullying, proposing that classifiers for cyberbullying, mere profanity detection and more sophisticated.	The weakness of this classifier is that it is considered only for media meetings that have at least one profanity word. A more valuable classifier is needed that can apply to all media meetings.	Hosseinmardi et al. (2015)
Web	Corpus, TF-IDF, Word2Vec, LDA and K-means clustering.	This design, on text mining upholds language autonomous investigates. This technique assists with further developing text mining competently in Turkish language that is a word request free language.	This classification algorithm cannot be able to categorise new data sources efficiently.	Cakir (2016)
Web	Preprocessing steps for sentiment analysis in Brazilian Portuguese social media.	This paper actually considers that there is unquestionably not a thorough and uniform framework for preprocessing of Social Media data, which rely upon SA in the Brazilian Portuguese language	The results of this literature review are restricted to only Brazilian Portuguese.	Cirqueira et al. (2018)
Web	LSTM, CNN, LSTM-CNN is used to categorise short texts of SinaWeibo.	This LSTM-CNN is capable of assuring goof results, this framework gives better result than LSTM and CNN.	This LSTM-CNN is not able to categorise traffic related microblogs into complete categories like traffic status, traffic accidents.	Chen et al. (2018)

Table 7 Strength and weakness of text mining techniques in social media (continued)

Social media type	*Text mining technique*	*Strength*	*Weakness*	*References*
Web	SVM model	SVM works relatively well when there is a small dataset; in this case, there is small dataset about food.	SVM does not work well for large dataset.	Chen and Chuang (2019)
Web	word emotion refinement algorithm, word emotion association network (WEAN)	The outcome indicates that, this structure can figure news event sentiment precisely.	This structure is not able to reflect word emotion pattern and emotion distance into text sentiment calculation.	Jiang et al. (2017)
YouTube	Support vector machine (SVM), naive Bayes, K-nearest neighbour and one collective classifier known as Bagging.	This bagging tool is capable of assuring goof results than naive Bayes, support vector machine (SVM), K-nearest neighbour and it proved in this study.	The bagging algorithm is restricted only for YouTube dataset instead of the additional social media dataset to discover account holder having spam.	Zaman and Sharmin (2017)
YouTube	machine learning and text mining techniques	This review offers the establishment for a strong sign on what the public wishes, fears, and thinks with respect to android robots, and their get-together in the public arena.	We cannot take into figured, the variety of perspectives over the long haul which could offer significant data on how open perception of androids is differing.	Vlachos and Tan (2018)
YouTube	A classification model that specifies the gender of the writer of a given text.	This gender identification classified model attains an ideal performance. The accuracy of the model is 92% while the average precision is about 98%, which is very high.	This arrangement model cannot apply to texts taken from other informal organisations, to gauge its ability to bargain gender identification from general online media like Arabic printed content.	Zahir et al. (2019)

Table 7 Strength and weakness of text mining techniques in social media (continued)

Social media type	*Text mining technique*	*Strength*	*Weakness*	*References*
WhatsApp	S2 NOW algorithm is used.	The algorithm helps the usage of the private and public main pair and would absolutely offer attractive Herington persuade in the upcoming time.	This plan cannot join the possibility of neighbourhood, recognising evidence and creation in the stream research work of combination and grouping of the WhatsApp visits of various individuals.	Johari et al. (2021)
WeChat	Sentiment analysis and text mining techniques	Helpful to find hot technology topics and keywords of the WeChat social media.	Opinion examination is not enough for the flawlessness of the feeling word reference and the foundation in innovation subjects.	Wu et al. (2018)

4 Conclusions

Prior research was utilised to represent the significance of leading this review for not many social media types and on the bygone eras. Then, at that point, a complete SLR was performed on some sort of social media and the job of web-based media. The main examination question found which information mining procedure is applied on which online media. These requirements found from the writing during looking through the catchphrases as issues, hardships and difficulties of online media and observe which text mining procedure is applied in which kind of web-based media. These data needs were ordered into various social media types like Facebook, YouTube, Twitter, Instagram and on broad information. There is a need to find which information mining procedure is applied in which online media in most recent social media types from 2016 to 2021.

Notwithstanding, further investigations should be directed. As examined in the presentation area, there is an absence of examination on most recent social media types in a most recent time from 2016 to 2021. Hence, there is a hole in the writing regarding the finding of data about most recent setting of web-based media. Therefore, our third examination question is about the shortcoming and strength of social media procedures on various sorts of online media. In the first place, there are studies via online media yet absence of studies on the continuous use of these social media types. Subsequently, it is important to lead such examinations to get the data of most recent social media types on which distinctive text mining strategies are applied.

Hence, more examination should be centred, for example, an investigation of most recent social media types from interpersonal organisations, in order to see how individuals, look for data from online organisations. Second, to find data via social media

an individual necessity to do specific things, where in the past times, an individual needed to play out specific exercises or some extra/various exercises to acquire data from libraries. Moreover, more investigations are expected to recognise the various practices of individuals, as referenced in the presentation, for example, as far as dynamic and uninvolved data looking for conduct via web-based media.

AI calculation and natural language handling procedures are utilised to get helpful data from unstructured social media (Elsayed et al., 2019). Various strategies of text mining and diverse soft products are utilised like API, REST API, JSON. Python 3.x alongside the natural language toolkit (nltk) and tweepy libraries to distinguish the conduct of youth. This procedure proposed may assist individuals with settling on choice and execution improvement (Ardra et al., 2017; Sriyanong et al., 2018). A programmed text grouping procedure is proposed. The proposed procedure is applied on the twitter dataset for plan and displaying of the suggested classifier. The suggested strategy assesses each word in both the accessible categories that are positive or negative. The suggested strategy is likewise encouraging for using the procedure for the large information climate for streamed information grouping (Dasondi, 2016).

Various boundaries are been utilised like fuzzy keyword match, support vector machine and Twitter latent Dirichlet allocation (LDA) machine learning ways to deal with recognise the interest group from a rundown of devotees. Tweets of the record proprietor to portion devotee will assist the record proprietor to payout assets successfully by sending proposals to the fitting client. This article set new bearings for online business, expanding clients, and introducing offers for related gatherings or related local area (Dastanwala, 2016). To investigate spam recognition by utilising various strategies on Facebook pages and posts, diverse Indexes utilised in this article are Jaccard, dice, Ochiai, overlap, similarity coordinating, TF-IDF, latent semantic analysis (Agrawal and Kaushal, 2016). The review presents a work process approach of utilising normal language preprocessing, message mining and feeling investigation procedures to examine online text-based substance. Specifically, the work process approach proposes to use the latent Dirichlet allocation (LDA) model to overhaul the number of gatherings. An obstacle with this work cycle approach is that it requires some human effort for data assessment and interpretation to perceive obvious classes or subjects from the model data self-assertively looked over the educational assortment (He et al., 2017).

A sharp Twitter-text-mining model was effectively developed. It was endeavoured to scratch tweet-texts for expecting business exchanges and commitment pay in Indonesia. The model finally can tentatively gauge a total exchange trade and besides charge pay of business. Where the calculation really subject to brief assumption and basic assessment (Kaburuan et al., 2014). Message mining and sentiment analysis methodology are used to analyse the information in the WeChat official accounts through the technique for message mining, close by the assistance of concerning the message pop-ups in the WeChat official accounts as the enlightening assortment (Wu et al., 2018). To inspect a corpora of Instagram posts from the style locale, present a construction for taking out design credits from Instagram, and train a critical dressing classifier with slight organisation to pack Instagram posts dependent upon the associated message (Hammar, 2018). Instruments used in this assessment are Apache Glimmer and Apache Hadoop to plan and encourage a capable text preprocessing structure on a significant data establishment, which is proposed to help the text preprocessing task to diminish the computation time (Sriyanong et al., 2018).

Tools used in this research are apache spark and Apache Hadoop. This framework is not considered for using the further big data technologies (Sriyanong et al., 2018). Naïve Bayes algorithm is used. This model shows weak performance to remove the stop words and classification of comments in each post (Bayes, 2019). The facial recognition API tool is not able to categorise drug users, dealers, sellers and glass makers/legitimate and not able to categorise drug-users and non-drug-users (Zhou et al., 2016). The literature review consists of preprocessing steps for sentiment analysis in Brazilian Portuguese social media is restricted to only Brazilian Portuguese (Cirqueira et al., 2018).

This LSTM-CNN is not able to categorise traffic related microblogs into complete categories like traffic status, traffic accidents (Chen et al., 2018). SVM does not work well for large dataset (Chen and Chuang, 2019). The Bagging algorithm is restricted only for YouTube dataset instead of the additional social media dataset to discover account holder having spam (Zaman and Sharmin, 2017). A classification model that specifies the gender of the writer of a given text is applied. This arrangement model cannot apply to texts taken from other informal organisations, to gauge its ability to bargain gender identification from general online media like Arabic printed content (Zahir et al., 2019).

References

Agrawal, H. and Kaushal, R. (2016) 'Analysis of text mining techniques over public pages of Facebook', *IEEE 6th Int. Conf. Adv. Comput.*, pp.4–9, DOI: 10.1109/IACC.2016.12.

Ahmad, E., Khan, K.U. and Ahmad, E. (2022) 'SocialPulse: a tool for extracting interesting insights from social media', *EasyChair*, No. 9276.

Ardra, B.M., Varughese, M.S., Joseph, P.E. and Thomas, S.K.K. (2017) 'Analyzing the behavior of youth to sociality using social media mining', *Int. Conf. Intell. Comput. Control Syst.*, IEEE, pp.1231–1235.

Bayes, A.N. (2019) 'Implementation of naive Bayes algorithm for spam comments classification on Instagram', *2019 Int. Conf. Inf. Commun. Technol.*, pp.508–513.

Cakir, M.U. (2016) 'Text mining analysis in Turkish language using big data tools', *IEEE 40th Annu. Comput. Softw. Appl. Conf.*, DOI: 10.1109/COMPSAC.2016.203.

Chen, L.S. and Chuang, Y.J. (2019) 'A study of social media reviews effects on the success of crowdfunding projects', *2019 IEEE 10th Int. Conf. Aware. Sci. Technol. iCAST 2019 – Proc.*, pp.1–5, DOI: 10.1109/ICAwST.2019.8923411.

Chen, Y. et al. (2018) 'Texts with deep learning approaches', *IEEE Trans. Intell. Transp. Syst.*, Vol. PP, No. 8, pp.1–10.

Choudhary, M.M. (2018) 'Sentiment analysis of text reviewing algorithm using data mining', *2018 Int. Conf. Smart Syst. Inven. Technol.*, pp.532–538, DOI: 10.1109/ICSSIT.2018.8748599.

Cirqueira, D., Antonio, J., Lobato, F. and Santana, A. (2018) 'A literature review in preprocessing for sentiment analysis for Brazilian Portuguese social media', *IEEE/WIC/ACM Int. Conf. Web Intell.*, DOI: 10.1109/WI.2018.00008.

Costales, J.A., De Los Santos, C.M., Catulay, J.J.J.E. and Albino, M.G. (2022) 'Sentiment analysis for Twitter tweets: a framework to detect sentiment using naïve Bayes algorithm', *4th Int. Conf. Comput. Commun. Internet (ICCCI)*, pp.39–44, DOI: 10.1109/ICCCI55554.2022.9850257.

Dasondi, V. (2016) 'An implementation of graph based text classification technique for social media', *Symp. Colossal Data Anal. Networking*, IEEE.

Dastanwala, P.B. (2016) 'A review on social audience identification on Twitter using text mining methods', *IEEE*, pp.1917–1920.

Dessai, S. and Usgaonkar, S.S. (2022) 'Depression detection on social media using text mining', *3rd Int. Conf. Emerg. Technol. INCET 2022*, pp.3–6, 2022, DOI: 10.1109/INCET54531.2022.9824931.

Dyba, T. and Dingsoyr, T. (2008) 'Empirical studies of agile software development: a systematic review', *Inf. Softw. Technol.*, Vol. 50, pp.833–859, DOI: 10.1016/j.infsof.2008.01.006.

El Haddaoui, B., Chiheb, R., Faizi, R. and El Afia, A. (20121) 'A sentiment analysis: a review and framework foundations', *Int. J. Data Anal. Tech. Strateg.*, Vol. 13, No. 4, pp.336–355.

El-Jawad, M.H.A. and Hodhod, R. (2017) 'Sentiment analysis of social media networks using machine learning', *2018 14th Int. Comput. Eng. Conf.*, pp.174–176.

Elsayed, M., Abdelwahab, A. and Ahdelkader, H. (2019) 'A proposed framework for improving analysis of big unstructured data in social media', *Proc. – ICCES 2019 14th Int. Conf. Comput. Eng. Syst.*, pp.61–65, DOI: 10.1109/ICCES48960.2019.9068154.

Fakhrur, M., Abu, R., Idris, N. and Shuib, L. (2019) 'An enhancement of Malay social media text normalization for lexicon-based sentiment analysis', *2019 Int. Conf. Asian Lang. Process. (IALP), IEEE*, pp.211–215.

Fiallos, A., Jimenes, K. and Fiallos, C. (2018) 'Detecting topics and locations on Instagram photos', *Int. Conf. eDemocracy eGovernment*, pp.246–250.

Hamid, S., Bukhari, S., Ravana, S.D., Norman, A.A. and Ijab, M.T. (2016) 'Role of social media in information-seeking behaviour of international students: a systematic literature review', *Aslib J. Inf. Manag.*, Vol. 68, No. 5, pp.643–666, DOI: 10.1108/AJIM-03-2016-0031.

Hammar, K. (2018) 'Deep text mining of Instagram data without strong supervision', *2018 IEEE/WIC/ACM Int. Conf. Web Intell.*, pp.158–165, DOI: 10.1109/WI.2018.00-94.

Hassan, D. (2018) 'A text mining approach for evaluating event credibility on twitter', *Proc. – 2018 IEEE 27th Int. Conf. Enabling Technol. Infrastruct. Collab. Enterp. WETICE 2018*, pp.175–178, DOI: 10.1109/WETICE.2018.00039.

He, W., Yan, G., Shen, J. and Tian, X. (2017) 'Developing a workflow approach for mining online social media data', *2017 IEEE SmartWorld, Ubiquitous Intell. Comput. Adv. Trust. Comput. Scalable Comput. Commun. Cloud Big Data Comput. Internet People Smart City Innov.*

Hidayat, I.H. and Parwanto, R.E. (2022) 'Sentiment analysis on the perception and mindset of the people of indonesia on the use of vaccines to deal with the COVID-19 pandemic using the text mining method', *International Conference on Information Management and Technology (ICIMTech)*, August, pp.57–61.

Hosseinmardi, H., Mattson, S.A., Rafiq, R.I., Han, R., Lv, Q. and Mishra, S. (2015) *Prediction of Cyberbullying Incidents on the Instagram Social Network*, arXiv Prepr. arXiv1508.06257, DOI: 10.1007/978-3-319-27433-1_4.

Jeelall, S. and Cheerkoot-Jalim, S. (2020) 'HealthMine: a tool for social media text mining in health', *3rd Int. Conf. Emerg. Trends Electr. Electron. Commun. Eng. ELECOM – Proc.*, pp.53–57, DOI: 10.1109/ELECOM49001.2020.9297002.

Jiang, D., Luo, X., Xuan, J. and Xu, Z. (2017) 'Sentiment computing for the news event based on the social media big data', *IEEE Access*, Vol. 5, No. c, pp.2373–2382, DOI: 10.1109/ACCESS.2016.2607218.

Johari, R., Kalra, S., Dahiya, S. and Gupta, K. (2021) 'S2NOW: secure social network ontology using WhatsApp', *Secur. Commun. Networks*, Vol. 2021, DOI: 10.1155/2021/7940103.

Kaburuan, E.R., Lindawati, A.S.L., Surjandy, S., Putra, M.R. and Utama, D.N. (2014) 'A model configuration of social media text mining for projecting the online-commerce transaction (Case: twitter tweets scraping)', *Int. Conf. Cyber IT Serv. Manag.*, pp.5–8.

Kibtiah, T.M., Miranda, E., Fernando, Y. and Aryuni, M. (2020) 'Terrorism, social media and text mining technique: review of six years past studies', *Int. Conf. Inf. Manag. Technol. (ICIMTech)*, August, IEEE, pp.571–576.

Kim, S-m. and Ha, Y-g. (2016) 'Automated discovery of small business domain knowledge using web crawling and data mining', in *2016 International Conference on Big Data and Smart Computing (BigComp)*, IEEE, pp.481–484.

Malhotra, R. and Malhotra, K. (2018) 'An analysis of the 2016 United States Presidential Election using Chanakya – a knowledge discovery platform for text mining', *Int. J. Knowl. Eng. Data Min.*, Vol. 5, Nos. 1–2, p.1, DOI: 10.1504/ijkedm.2018.10013307.

Nusantara, U.M. (2019) 'Sentiment analysis on official news accounts of Twitter media in predicting Facebook stock', *5th Int. Conf. New Media Stud.*, pp.74–79.

Permana, M.A., Thohir, M.I., Mantoro, T. and Ayu, M.A. (2022) 'Crime Rate Detection Based on Text Mining on Social Media Using Logistic Regression Algorithm', *IEEE 7th Int. Conf. Comput. Eng. Des.*

Purnomo, F., Ricky, M.Y. and City, A.S. (2016) 'Smart city's context awareness using social media', *Int. Conf. ICT Smart Soc.*, IEEE, pp.20–21.

Razavi, S.Z. and Rahbari, M. (2020) 'Understanding reactions to natural disasters: a text mining approach to analyze social media content', *2020 7th Int. Conf. Soc. Netw. Anal. Manag. Secur. SNAMS 2020*, DOI: 10.1109/SNAMS52053.2020.9336570.

Reguera, N., Subirats, L. and Armayones, M. (2017) 'Mining Facebook data of people with rare diseases', *Proceedings – IEEE Symposium on Computer-Based Medical Systems*, November, Vol. 2017, pp.588–593, DOI: 10.1109/CBMS.2017.124.

Roundtree, A.K. (2018) 'From engineers' tweets: text mining social media for perspectives on engineering communication', *IEEE Int. Prof. Commun. Conf.*, July, pp.6–15, DOI: 10.1109/ProComm.2018.00009.

Sa, D., Ramdhani, M.A., Rahman, A. and Darmalaksana, W. (2017) 'Incremental technique with set of frequent word item sets for mining large Indonesian text data', *Increm. Tech. with set Freq. word item sets Min. large Indones. text data. 2017 5th Int. Conf. Cyber IT Serv. Manag. (CITSM)*, IEEE, pp.1–6.

Salloum, S.A., Al-Emran, M. and Shaalan, K. (2017) 'Mining text in news channels: a case study from Facebook', *Int. J. Inf. Technol. Lang. Stud.*, Vol. 1, No. 1, pp.1–9.

Sriyanong, W., Moungmingsuk, N. and Khamphakdee, N. (2018) 'A text preprocessing framework for text mining on big data infrastructure', *2nd Int. Conf. Imaging, Signal Process. Commun.*, pp.169–173.

Sutar, S.G. (2017) 'Intelligent data mining technique of social media for improving health care', in *2017 International Conference on Intelligent Computing and Control Systems (ICICCS) 2017*, IEEE, 15 June, pp.1356–1360.

Vasanthakumar, G.U., Shashikumar, D.R. and Suresh, L. (2019) *Profiling Social Media Users, a Content-Based Data Mining Technique for Twitter Users*, DOI: 10.1109/ICAIT47043.2019.8987304.

Vlachos, E. and Tan, Z.H. (2018) 'Public perception of android robots: Indications from an analysis of YouTube comments', *IEEE Int. Conf. Intell. Robot. Syst.*, pp.1255–1260, DOI: 10.1109/IROS.2018.8594058.

Wu, F., Tong, Y., Huang, L., Miao, H. and Li, X. (2018) 'The application prospect analysis of technology based on WeChat official accounts', *PICMET 2018 – Portl. Int. Conf. Manag. Eng. Technol. Manag. Technol. Entrep. Engine Econ. Growth, Proc.*, pp.1–9, DOI: 10.23919/PICMET.2018.8481831.

Zahir, J., Oukaja, Y.M. and Mousannif, H. (2019) 'Author gender identification from Arabic YouTube comments', *Proceedings – 15th International Conference on Signal Image Technology and Internet Based Systems, SISITS 2019*, November, pp.672–676, DOI: 10.1109/SITIS.2019.00109.

Zaman, Z. and Sharmin, S. (2017) 'Spam detection in social media employing machine learning tool for text mining', *13th Int. Conf. Signal-Image Technol. Internet-Based Syst.*, DOI: 10.1109/SITIS.2017.32.

Zhitomirsky-Geffet, M., Feitelson, D.G., Frachtenberg, E. and Wiseman, Y. (2009) 'A unified strategy for search and result representation for an online bibliographical catalogue', *Online Inf. Rev.*, Vol. 33, No. 3, pp.511–536, DOI: 10.1108/14684520910969934.

Zhou, Y., Sani, N. and Luo, J. (2016) 'Fine-grained mining of illicit drug use patterns using social multimedia data from Instagram', *IEEE Int. Conf. Big Data (Big Data)*, pp.1921–1930.