

A scoping review of topic modelling on online data

Mohd Mukhlis Mohd Sharif¹, Ruhaila Maskat¹, Zirawani Baharum², Kamaruzaman Maskat³

¹Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA Shah Alam, Shah Alam, Malaysia

²Technical Foundation, Malaysian Institute of Industrial Technology, Universiti Kuala Lumpur, Kuala Lumpur, Malaysia

³Department of Computer Science, Faculty of Defence Science and Technology, National Defence University of Malaysia, Kuala Lumpur, Malaysia

Article Info

Article history:

Received Jun 18, 2022

Revised May 11, 2023

Accepted May 19, 2023

Keywords:

Natural language processing

Online data

Short text

Systematic review

Topic modelling

ABSTRACT

With the increasing prevalence of unstructured online data generated (e.g., social media, online forums), mining them is important since they provide a genuine viewpoint of the public. Due to this significant advantage, topic modelling has become more important than ever. Topic modelling is a natural language processing (NLP) technique that mainly reveals relevant topics hidden in text corpora. This paper aims to review recent research trends in topic modelling and state-of-the-art techniques used when dealing with online data. Preferred reporting items for systematic reviews and meta-analysis (PRISMA) methodology was used in this scoping review. This study was conducted on recent research works published from 2020 to 2022. We constructed 5 research questions for the interest of many researchers. 36 relevant papers revealed that more work on non-English languages is needed, common pre-processing techniques were applied to all datasets regardless of language e.g., stop word removal; latent dirichlet allocation (LDA) is the most used modelling technique and also one of the best performing; and the produced result is most evaluated using topic coherence. In conclusion, topic modelling has largely benefited from LDA, thus, it is interesting to see if this trend continues in the future across languages.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Zirawani Baharum

Technical Foundation, Malaysian Institute of Industrial Technology, Universiti Kuala Lumpur

Kuala Lumpur, Malaysia

Email: zirawani@unikl.edu.my

1. INTRODUCTION

Unstructured online data has become prevalent. They come in the form of social media, and online reviews. It has been reported that the number of social media users alone in 2021 is 3.78 billion worldwide as compared to only 3 years earlier of 2.86 billion in 2017 [1]. Social media text has been greatly used in sentiment analysis to tap into the emotional opinion polarity of the public on a specific product, issue or service. Once known as opinion mining which utilizes a small sample of people, sentiment analysis differs based on its use of new media i.e., social media e.g., Facebook and Twitter. Reviews are substantially influential in a customer's decision to purchase. Reported that 68% of consumers surveyed will interact with potential businesses due to reading positive reviews [2]. 40% will avoid transactions with businesses that received negative reviews. The remaining 36% is indifferent to any reviews but instead places key importance on location and price. This makes Google the top highly influential online seller worldwide. The underlying premise of this use was triggered by the perspective and later supported by research studies [3]-[7], arguing that from online data, e.g. social media, one can discover genuine thoughts, emotions and sentiments of people. These studies showed that people reveal themselves more openly online due to the absence of the face-to-face element during such communication. Therefore, online data is suitable as a source of genuine viewpoints or stances of many people.

Yet, revealing this viewpoint requires more than just sentiment analysis. This is where topic modelling comes into play.

Topic modelling is a natural language processing (NLP) technique with the primary aim of revealing relevant topics hidden inside a text corpus [8]. It is the process of selecting words from a document that are related to a given topic logically [9]. In general Figure 1, topic modelling starts with the identification of useful sources. Common suitable sources include social media and online forums. A corpus is then formed by engaging either a scraping process or an available application programming interface (API). Then, pre-processing on the corpus is conducted such as tokenization, and stemming. The primary aim is to transform the corpus into a format suitable for the NLP algorithm. Next, the number of topics must be determined before the modelling phase can take place, often manually. Subsequently, machine learning models are employed, e.g., latent dirichlet allocation (LDA), latent semantic analysis (LSA) or normalized mutual information (NMI), which use an automated procedure for determining the relevant topic described in an original text [10]. These models are evaluated before any deployment (e.g., precision, recall). It is challenging to discern topics when dealing with unmoderated, user-generated information on social media networks like Twitter. There are drawbacks to this, such as user confusion over terminology and the length of micro posts, which typically lack context [11]. The purpose of this systematic review is to help fellow academics and practitioners to better understand recent advancements in topic modelling and applied them in their discipline.

The structure of this paper is as follows. In section 2, we explain how the preferred reporting items for systematic reviews and meta-analysis (PRISMA) method was applied to this work and the results are discussed in section 4. We conclude the paper in section 5.

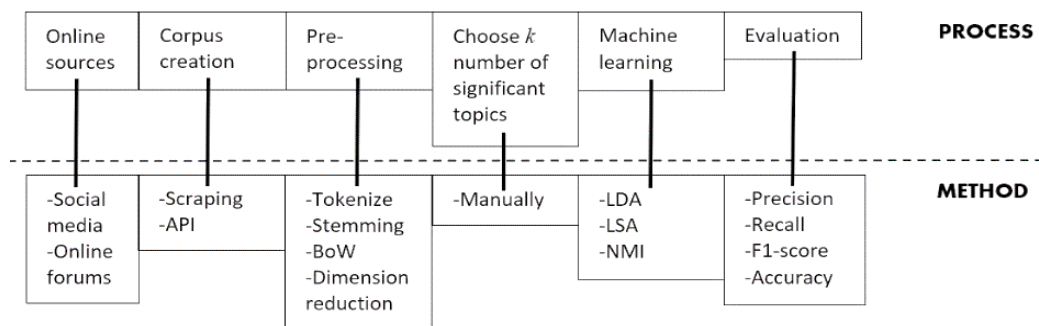


Figure 1. General topic modelling process and method

2. RESEARCH METHOD

Prior to conducting our review, we looked at existing survey papers on topic modelling techniques published within the last five years. Two were published. Jelodar *et al.* [12] focused on LDA-based techniques only. Qiang *et al.* [13] compares the performance of modelling techniques as recent as the year 2020. Our review, in contrast, covers end-to-end processes important for topic modelling. It can be useful for entry-level researchers as well as researchers with interests in methods involved in topic modelling. We focused on four important areas. They are the dataset, pre-processing technique, modelling method and evaluation technique. We conducted our review by employing the PRISMA method. PRISMA is a minimum collection of evidence-based elements that are used to guide the formulation and structure of systematic reviews and other meta-analyses [14]. The review protocol contains the following steps. They are: defining the research questions, identifying online sources, formulating several search queries, determining the inclusion and exclusion criterion specifications and screening relevant papers. The implementation of each step of the PRISMA is elaborated in the following subsections of our research method.

2.1. Research questions definition

The questions formulated in this paper provide an understanding of the background of recent publications on topic modelling (e.g., publication year). Additionally, aspects of its general processes and methods (e.g., the dataset employed, domains involved, pre-processing techniques used, recent methods implemented, and the evaluation techniques implemented). Table 1 lists the composed research questions and their motivations.

Table 1. Research questions and motivations

Research questions	Motivations
RQ1. What is the language, type and source of the datasets employed?	The answer to this question shows the language, type and source of the datasets used in topic modelling.
RQ2. What pre-processing techniques were used?	The answer to this question identifies pre-processing techniques used on datasets of differing languages.
RQ3. Which topic modelling method produced high performance?	The answer to this question reveals the most notable topic modelling methods explored.
RQ4. What evaluation techniques were adopted for topic modelling?	The answer to this question identifies evaluation techniques widely used for topic modelling.

2.2. Search strategy

In this step, we identified several information sources, consisting of digital libraries and search engines. They are shown in Table 2. An overlapping of papers existed from search engines such as Google Scholar with some of the digital libraries, nevertheless, this duplication was removed during the screening process. The finalized total of papers reviewed is thus unique.

To extract papers from these sources, we composed 2 search queries. S1 aims to collect all literature related to topic modelling within the determined year range. S2 queries topic modelling literature relating to social media. Table 3 lists our composed search queries. Each query is run on each of the information sources.

Table 2. Information sources

Source	Type	URL
Science Direct Elsevier	Digital library	http://www.sciencedirect.com/
IEEE Xplore	Digital library	http://ieeexplore.ieee.org/Xplore/home.jsp
Wiley online library	Digital library	https://onlinelibrary.wiley.com/
Google Scholar	Search engine	https://scholar.google.com/
MDPI	Digital library	https://www.mdpi.com/
Springer	Digital library	https://www.springer.com/
Semantic Scholar	Search engine	https://www.semanticscholar.org/

Table 3. Search query

Queries Used
S1 “Topic Modelling”
S2 “Topic Modelling” AND “social media”

2.3. Inclusion and exclusion criteria determination

Table 4 lists both our formulated inclusion and exclusion criteria. The aim here is to narrow down the collection of found papers to include only complete papers that are likely to answer the research questions. Publications written in languages other than English were not included as much literature is in this language.

Table 4. Inclusion and exclusion criteria

Inclusion criteria	Exclusion criteria
Studies should be listed in one of the research databases	Studies that are not written in English
Studies should meet at least one of the search terms	Studies that are missing full text
Studies should be published/in-press at a journal or conference	Studies that do not have DOI
Studies should provide answers to the research questions	Papers that are not relevant to topic modelling

2.4. Screening relevant papers

Based on the inclusion and exclusion criteria previously set, noncompliant studies are removed. Additionally, a screening technique is used to choose articles that are relevant to our goals. Our screening technique is as:

- Duplicate removal.
- Title and abstract screening: we eliminate irrelevant papers after perusing each paper’s title and abstract. Keywords and embedded information are used in our decision-making. Papers having abstracts that met at least 40% of the inclusion criteria are kept for the next screening.
- Full-text screening. After reading the full text of each paper that succeeded in the previous screening process, papers not directly answering the research questions formulated in Table 1 are purged.

3. RESULTS AND DISCUSSION

This section details our review's results and provides answers to our above-mentioned research questions. Figure 2 shows the processes conducted by this review. Each process reduces the size of the paper collection and gets the work closer to the optimal subset of significant papers.

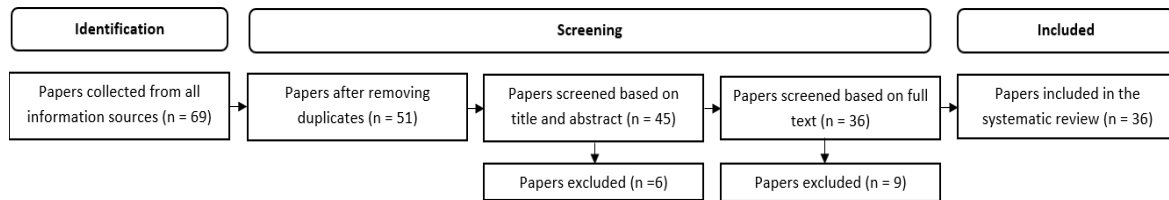


Figure 2. Systematic review process

During the identification stage, we collected 69 papers from all information sources. Next, duplicates were removed to result in 51 papers. This constitutes the screening process. The eligibility of the papers was then determined via two subprocesses: the first is from perusing the title and abstract whereas the second is from the full text. Based on the title and abstract of the remaining papers, 6 were found not eligible and thus excluded. At this point, we had 45 relevant papers. In the second phase of the eligibility determination, 36 papers were found relevant and 9 were not. The selection decision was based on no duplicates, the title, the abstract and full text. Figure 2 visualizes these steps. The resulting 36 papers used in this review are listed in Table 5. The following subsections describe the answers to every research question.

3.1. Answer to RQ1: What is the language, type and source of the datasets employed?

The majority of the dataset used is in English, which comprises 33 publications, followed by African-American (e.g. *wussup n*ggga, what's up bro!*), Chinese, Dutch, and multiple languages (i.e. English, Arabic, Spanish, Italian, German, French and Japanese) each comprising 1 paper. This suggests that more work in non-English languages about dataset curation is needed in topic modelling. This includes slang. From a different perspective, efforts have been made to model topics from other languages. With the increasing awareness of the benefits of topic modelling and the widening use of data analytics, we expect to see exciting developments in these languages as well as low resources languages e.g., Hindi, Malay, and Zulu to list a few.

Textual data can be divided into two types: short text and long text. To date, there is no agreed-upon formal definition of a short text in the academic literature. At present, the categorization of the source medium is determined by the authors of each paper. According to [15], short texts are uniquely characterized as either being 1-4 words or 1-140 characters whereas an earlier publication [16] described the short text as typically having 200 characters except for Microsoft Windows Live Messenger permitting up to 400 characters text. Therefore, in this review, we divide the literature into two groups based on the length of the text. From 36 papers, we identified 19 studies utilizing short text. They consist of AOL query log, skytrax reviews, IMDB, Facebook, clinical notes, accommodation reviews, Reddit posts, Airbnb reviews and most are Tweets. Scopus Databases, Previous Paper Review, Department of Health press releases, web of science (WoS), 20NewsGroups corpus, New York Times corpus, BBC news dataset, religious scriptures, Regulated USA market finance, DBPedia, the corpus of 100 English novels, WiseSearch news articles, Scopus, ScienceDirect, One Million E-Books, Stack Overflow, CSR annual report and Google Scholar provided the remaining 17 long text datasets. One paper [17] were not included to answer this research question since no dataset were used. The nature of this paper was more towards a comparative analysis of earlier works. We discovered that Twitter is the main contributor to 8 datasets. WoS comes second with 2 and the rest recorded one occurrence each. We agree with the common perspective behind this phenomenon due to the easy availability of Twitter API for data extraction whereas other data sources, e.g., Facebook, are less or just not available.

3.2. Answer to RQ2: What pre-processing techniques were used?

Across different papers in our review, several pre-processing techniques can be discovered. Our result discovered that stop word elimination is the most conducted pre-processing technique. The stop words removal process holds the premise that stop words are low-level information found in corpora and should be removed. This is to allow significant information to be included whereas insignificant words such as "and," "or," and "so" are discarded. Removal may rely on some pre-formulated rules and are language-dependent. Another commonly used pre-processing is the conversion of cases into the lower form. The aim is to provide a standard

format where the programming language is case-sensitive. From the literature, punctuations, symbols/special characters and number elimination are also conducted in most of the studies. This indicates that the corpus used was stripped bare to leave only words. Tokenization was also conducted. Two other closely similar pre-processing tasks performed are stemming and lemmatization. Both aim at converting words into their basic forms, however, stemming exclusively reduces a word somewhat mechanically with the absence of any information or context. Other pre-processing techniques employed are term-document matrix (TDM), clustering, bag-of-words (BoW), Word2vec (feature representation), term frequency-inverse document frequency (TF-IDF), HTML and URL removal.

3.3. Answer to RQ3: Which topic modelling method produced high performance?

Most of the publications employed LDA-24 papers. LSA 4 papers and structural topic modelling (STM)-3 papers. Other techniques found are bidirectional adversarial topic (BAT) [18], bitern topic model (BTM) [19], correlated topic model (CTM) [20], dirichlet multinomial mixture (DMM) [13], embedded topic modelling (ETM) [21], fuzzy topic model (FTM) [22], gaussian bidirectional adversarial topic (G-BAT) [18], hawkes binomial topic model (HBTM) [23], latent dirichlet allocation-MALLET (LDA-MALLET), latent semantic indexing (LSI), negative sampling and quantization topic model (NQTM), neural variational document model (NVDM), non-negative matrix factorization (NMF), principal component analysis (PCA), probabilistic topic model (PLSA), probabilistic topic modelling (PTM), product of expert LDA (PRODLDA), random projection (RP), word2vec-based latent semantic analysis (W2V-LSA).

Another important aspect in research is which methods recorded best performance. From the literature, the methods are LDA, ETM, FTM, CTM, G-BAT, NQTM, BTM and W2V-LSA. It is exciting to see the usage of these methods in future works on topic modelling, or will we see more variants of LDA dominating the scene, suggesting its continuous effectiveness.

- a) LDA has been acknowledged as the most efficient method for topic modelling to date. Its performance is drawn from its underlying nature of utilizing a dirichlet distribution to identify topics as well as words relating to those topics [19], [21], [24]-[29]. LDA generalizes better than most other methods due to this distribution. Whenever a new document is presented, LDA utilizes the distribution to find samples of the document [9], [30]-[34]. Another way of viewing is LDA using the dataset as training data. Topics in LDA are identified based on a group of words displaying a high probability of describing the same topic [35]-[39]. Although it is regarded as the most effective so far, LDA is not without weakness-all text is assumed to originate from one process, therefore, external variables e.g., author and writing time are not taken into account [40]-[44]. This is where STM or structural topic modelling comes into the scene.
- b) LSA while LDA is known as the most efficient method in topic modelling, LSA is recognized as a foundational method when topic modelling is concerned. The idea of LSA stems from the notion that humans use similar terms to describe a topic. Underlying it are two matrices, the document-topic matrix and the topic-term matrix, formed with the help of TF-IDF [19], [24]. LSA takes advantage of single value decomposition (SVD) to reduce the dimension space in the formed matrices to result in a k -dimensional vector of each document and term. k is the number of interesting topics. Similarity comparison between documents and between words can be easily conducted within the remaining dimensional space using cosine similarity [25], [26].
- c) STM introduces the structure of documents, converted into metadata, to identify topics. Its notion is similar to LDA with the addition of estimating the relationship between topics and document metadata [45]. A topic in STM is identified as a subset of words that are likely to belong to a topic while a document is seen to consist of multiple topics [8]. In STM, a matrix is constructed based on the document metadata to produce topic prevalence (e.g., time, author) and another matrix for topic content (e.g., ideology and geography) [46], [47]. Topic prevalence represents how much a document is related to a topic whereas topic content calculates the probability of words belonging to a topic.

3.4. Answer to RQ4: What evaluation techniques were adopted for topic modelling?

Only 12 of the research papers used evaluation techniques to determine the quality of their models in discovering topics. Precision [19], [24], [25], [27], [35], [48] and recall [19], [24], [25], [27], [35], [48] were the most frequently used, followed by F1-score [19], [25], [27], [35], [48] and accuracy [22], [24], [27] which fell after topic coherence. Considering these four measurements are common to prediction using ML, therefore, we describe other measurements more specific to topic modelling.

- a) Topic coherence [13], [18], [26], [49] examines the collection of words in subjects created by the model and assesses the topic's informativeness. Several metrics compute coherence value in various ways. The examined topic coherence techniques use the top words of the subject as input and compute the total confirmation measure across pairs of words. These approaches have been developed to capture the context between words in a topic.

- b) Cluster evaluation [13]: in topic modelling, cluster analysis refers to the process of identifying groupings of topics that are related to one another but distinct from those in other groups.
- c) NMI [34] is a scaling of the mutual information (MI) score between 0 (no mutual information) and 1 (complete mutual information).
- d) Adjusted mutual information [34]: The MI score is adjusted to account for a change in adjusted mutual information (AMI). it explains why regardless of whether there is more information shared, the MI is typically greater for two clusterings with a larger number of clusters.
- e) Adjusted Rand Index [34]: by evaluating all pairs of samples and counting pairings that are allocated in the same or different clusters in the anticipated and true clustering, the Rand Index computes a similarity measure between two clusterings.

Table 5. Relevant papers

No.	Paper	Year
1	Chen <i>et al.</i> [33]	2020
2	Curiskis <i>et al.</i> [34]	2020
3	Dutta <i>et al.</i> [35]	2020
4	Kaila and Prasad [36]	2020
5	Sutherland <i>et al.</i> [37]	2020
6	Pröllochs and Feuerriegel [38]	2020
7	Xu and Xiong [40]	2020
8	Qundus <i>et al.</i> [41]	2020
9	Sbalchiero and Eder [42]	2020
10	Sutherland and Kiatkawsin [43]	2020
11	Liu <i>et al.</i> [44]	2020
12	Albalawi <i>et al.</i> [24]	2020
13	Asghari <i>et al.</i> [19]	2020
14	Dieng <i>et al.</i> [21]	2020
15	Kalepalli <i>et al.</i> [25]	2020
16	Chen <i>et al.</i> [10]	2020
17	Davidson and Bhattacharya [46]	2020
18	Kim <i>et al.</i> [47]	2020
19	Bagheri <i>et al.</i> [48]	2020
20	Hu <i>et al.</i> [18]	2020
21	Sha <i>et al.</i> [23]	2020
22	Wu <i>et al.</i> [49]	2020
23	Kim <i>et al.</i> [50]	2020
24	Mohammed and Al-Augby [26]	2020
25	Lee <i>et al.</i> [30]	2021
26	Kwon <i>et al.</i> [31]	2021
27	Cuaton <i>et al.</i> [9]	2021
28	Mustak <i>et al.</i> [32]	2021
29	Amara <i>et al.</i> [39]	2021
30	Abri and Abri [22]	2021
31	Isoaho <i>et al.</i> [17]	2021
32	Ryoo <i>et al.</i> [20]	2021
33	Issam <i>et al.</i> [27]	2022
34	Mangsor <i>et al.</i> [28]	2022
35	Khalid <i>et al.</i> [51]	2022
36	Islam <i>et al.</i> [29]	2022

4. CONCLUSION

In this timely paper, we have reviewed topic modelling literature from 2020 to 2022. We employed the PRISMA method and the result is a total of 36 works found relevant to the formulated research questions. Based on the literature review result, we can conclude that there is a need for topic modelling research in non-English languages, especially the low-resources ones. Typical pre-processing techniques, e.g., stop word removal, are applicable to the multi-language literature set. However, this result is skewed as the number of works for non-English language in the literature set is too few to be considered representative of the entire vocabulary of that language. As more research is performed on newer datasets, more specific techniques may become necessary. LDA revealed to be the most frequently used in topic modelling besides LSA and STM. Within a narrow period between 2020 to 2022, 24 published papers used LDA and it recorded as one of the best performing techniques, suggesting LDA is still relevant to topic modelling to this day. Topic coherence is the most preferred evaluation method. More topic modelling research on other languages, especially low resourced, is necessary to understand if pre-processing techniques, modelling techniques and evaluation methods are universally useful.

ACKNOWLEDGEMENTS

Our acknowledgements go to the Malaysian Institute of Industrial Technology, Universiti Kuala Lumpur for sponsoring this paper. Our acknowledgements also go to the research and community service of the College of Computing, Informatics, and Media, Universiti Teknologi MARA, Shah Alam, Selangor and the Department of Computer Science, Faculty of Defense Science and Technology, National Defense University of Malaysia.




REFERENCES

- [1] "Global Social Media Users," *Oberlo*. [Online]. Available: <https://www.oberlo.com/statistics/how-many-people-use-social-media>.
- [2] J. Anthony, "62 customer reviews statistics you must learn: 2020/2021 market share analysis and data," *FinancesOnline.com*, 2020. [Online]. Available: <https://financesonline.com/customer-reviews-statistics/>
- [3] S. N. Cassab and M.-B. Kurdy, "Ontology-based emotion detection in arabic social media," *International Journal of Engineering Research & Technology (IJERT)*, vol. 9, no. 08, pp. 923–928, 2020.
- [4] P. Thakor and S. Sasi, "Ontology-based sentiment analysis process for social media content," *Procedia Computer Science*, vol. 53, no. 1, pp. 199–207, 2015, doi: 10.1016/j.procs.2015.07.295.
- [5] U. Bretschneider and R. Peters, "Detecting offensive statements towards foreigners in social media," *Proceedings of the Annual Hawaii International Conference on System Sciences*, vol. 2017-Janua, 2017, pp. 2213–2222, doi: 10.24251/hicss.2017.268.
- [6] E. A. Rissola, D. E. Losada, and F. Crestani, "A survey of computational methods for online mental state assessment on social media," *ACM Transactions on Computing for Healthcare*, vol. 2, no. 2, pp. 1–31, 2021, doi: 10.1145/3437259.
- [7] S. Chancellor and M. D. Choudhury, "Methods in predictive techniques for mental health status on social media: a critical review," *npj Digital Medicine*, vol. 3, no. 1, 2020, doi: 10.1038/s41746-020-0233-7.
- [8] A. Mishler, E. S. Crabb, S. Paletz, B. Hefright, and E. Golonka, "Using structural topic modeling to detect events and cluster twitter users in the Ukrainian crisis," *Communications in Computer and Information Science*, vol. 528, pp. 639–644, 2015, doi: 10.1007/978-3-319-21380-4_108.
- [9] G. P. Cuaton, L. J. B. Caluza, and J. F. V. Neo, "A topic modeling analysis on the early phase of COVID-19 response in the Philippines," *International Journal of Disaster Risk Reduction*, vol. 61, p. 102367, 2021, doi: 10.1016/j.ijdrr.2021.102367.
- [10] X. Chen, D. Zou, G. Cheng, and H. Xie, "Detecting latent topics and trends in educational technologies over four decades using structural topic modeling: a retrospective of all volumes of computers and education," *Computers and Education*, vol. 151, 2020, doi: 10.1016/j.compedu.2020.103855.
- [11] B. Wang, M. Liakata, A. Zubiaga, and R. Procter, "A hierarchical topic modelling approach for tweet clustering," *In Social Informatics: 9th International Conference, SocInfo 2017, Oxford, UK*, pp. 378–390, 2017, doi: 10.1007/978-3-319-67256-4_30.
- [12] H. Jelodar *et al.*, "Latent dirichlet allocation (LDA) and topic modeling: models, applications, a survey," *Multimedia Tools and Applications*, vol. 78, no. 11, pp. 15169–15211, 2019, doi: 10.1007/s11042-018-6894-4.
- [13] J. Qiang, Z. Qian, Y. Li, Y. Yuan, and X. Wu, "Short text topic modeling techniques, applications, and performance: a survey," *IEEE Transactions on Knowledge and Data Engineering*, vol. 14, no. 8, pp. 1–1, 2020, doi: 10.1109/tkde.2020.2992485.
- [14] M. J. Page *et al.*, "The PRISMA 2020 statement: an updated guideline for reporting systematic reviews," *The BMJ*, vol. 372, 2021, doi: 10.1136/bmj.n71.
- [15] H. Wang, "Understanding short texts," *In APWeb 2013: Web Technologies and Applications*, 2013, pp. 1–1, doi: 10.1007/978-3-642-37401-2_1.
- [16] G. Song, Y. Ye, X. Du, X. Huang, and S. Bie, "Short text classification: a survey," *Journal of Multimedia*, vol. 9, no. 5, pp. 635–643, 2014, doi: 10.4304/jmm.9.5.635-643.
- [17] K. Isoaho, D. Gritsenko, and E. Mäkelä, "Topic modeling and text analysis for qualitative policy research," *Policy Studies Journal*, vol. 49, no. 1, pp. 300–324, 2021, doi: 10.1111/psj.12343.
- [18] X. Hu, R. Wang, D. Zhou, and Y. Xiong, "Neural topic modeling with cycle-consistent adversarial training," *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 9018–9030, doi: 10.18653/v1/2020.emnlp-main.725.
- [19] M. Asghari, D. Sierra-Sosa, and A. S. Elmaghraby, "A topic modeling framework for spatio-temporal information management," *Information Processing and Management*, vol. 57, no. 6, p. 102340, 2020, doi: 10.1016/j.ipm.2020.102340.
- [20] J. H. Ryoo, X. Wang, and S. Lu, "Do spoilers really spoil? using topic modeling to measure the effect of spoiler reviews on box office revenue," *Journal of Marketing*, vol. 85, no. 2, pp. 70–88, 2021, doi: 10.1177/0022242920937703.
- [21] A. B. Dieng, F. J. R. Ruiz, and D. M. Blei, "Topic modeling in embedding spaces," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 439–453, 2020, doi: 10.1162/tacla00325.
- [22] S. Abri and R. Abri, "Providing a personalization model based on fuzzy topic modeling," *Arabian Journal for Science and Engineering*, vol. 46, no. 4, pp. 3079–3086, 2021, doi: 10.1007/s13369-020-05048-7.
- [23] H. Sha, M. A. Hasan, G. Mohler, and P. J. Brantingham, "Dynamic topic modeling of the COVID-19 Twitter narrative among U.S. governors and cabinet executives," *arXiv preprint arXiv:2004.11692*, no. 2, pp. 2–7, 2020.
- [24] R. Albalawi, T. H. Yeap, and M. Benyoucef, "Using topic modeling methods for short-text data: a comparative analysis," *Frontiers in Artificial Intelligence*, vol. 3, no. July, pp. 1–14, 2020, doi: 10.3389/frai.2020.00042.
- [25] Y. Kalepalli, S. Tasneem, P. D. P. Teja, and S. Manne, "Effective comparison of LDA with LSA for topic modelling," *Proceedings of the International Conference on Intelligent Computing and Control Systems, ICICCS 2020*, pp. 1245–1250, 2020, doi: 10.1109/ICICCS48265.2020.9120888.
- [26] S. H. Mohammed and S. Al-Augby, "LSA and LDA topic modeling classification: comparison study on E-books," *Indonesian Journal of Electrical Engineering and Computer Science (IJECS)*, vol. 19, no. 1, pp. 353–362, Jul. 2020, doi: 10.11591/ijeecs.v19.i1.pp353-362.
- [27] A. Issam, A. K. Mounir, E. M. Saida, and E. M. Fatma, "Financial sentiment analysis of tweets based on deep learning approach," *Indonesian Journal of Electrical Engineering and Computer Science (IJECS)*, vol. 25, no. 3, pp. 1759–1770, 2022, doi: 10.11591/ijeecs.v25.i3.pp1759-1770.
- [28] N. S. M. N. Mangsor, S. A. M. Nasir, W. F. W. Yaacob, Z. Ismail, and S. A. Rahman, "Analysing corporate social responsibility reports using document clustering and topic modeling techniques," *Indonesian Journal of Electrical Engineering and Computer Science (IJECS)*, vol. 26, no. 3, pp. 1546–1555, 2022, doi: 10.11591/ijeecs.v26.i3.pp1546-1555.




- [29] S. Islam, Y. S. Nugroho, and M. Javed Hossain, "What network simulator questions do users ask? A large-scale study of stack overflow posts," *Indonesian Journal of Electrical Engineering and Computer Science (IJECS)*, vol. 21, no. 3, pp. 1622–1633, 2021, doi: 10.11591/ijeecs.v21.i3.pp1622-1633.
- [30] J. H. Lee, J. Wood, and J. Kim, "Tracing the trends in sustainability and social media research using topic modeling," *Sustainability (Switzerland)*, vol. 13, no. 3, pp. 1–23, 2021, doi: 10.3390/su13031269.
- [31] H. J. Kwon, H. J. Ban, J. K. Jun, and H. S. Kim, "Topic modeling and sentiment analysis of online review for airlines," *Information (Switzerland)*, vol. 12, no. 2, pp. 1–14, 2021, doi: 10.3390/info12020078.
- [32] M. Mustak, J. Salminen, L. Plé, and J. Wirtz, "Artificial intelligence in marketing: Topic modeling, scientometric analysis, and research agenda," *Journal of Business Research*, vol. 124, no. January, pp. 389–404, 2021, doi: 10.1016/j.jbusres.2020.10.044.
- [33] X. Chen, D. Zou, and H. Xie, "Fifty years of british journal of educational technology: a topic modeling based bibliometric perspective," *British Journal of Educational Technology*, vol. 51, no. 3, pp. 692–708, 2020, doi: 10.1111/bjet.12907.
- [34] S. A. Curiskis, B. Drake, T. R. Osborn, and P. J. Kennedy, "An evaluation of document clustering and topic modelling in two online social networks: Twitter and Reddit," *Information Processing and Management*, vol. 57, no. 2, pp. 1–21, 2020, doi: 10.1016/j.ipm.2019.04.002.
- [35] H. S. Dutta, V. R. Dutta, A. Adhikary, and T. Chakraborty, "HawkesEye: detecting fake retweeters using hawkes process and topic modeling," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 2667–2678, 2020, doi: 10.1109/TIFS.2020.2970601.
- [36] A. V. K. Kaila and R. P. Prasad, "Informational flow on twitter-corona virus outbreak-topic," *International Journal of Advanced Research in Engineering and Technology (IJARET)*, vol. 11, no. 3, pp. 128–134, 2020.
- [37] I. Sutherland, Y. Sim, S. K. Lee, J. Byun, and K. Kiatkawsin, "Topic modeling of online accommodation reviews via latent dirichlet allocation," *Sustainability (Switzerland)*, vol. 12, no. 5, pp. 1–15, 2020, doi: 10.3390/su12051821.
- [38] N. Pröllochs and S. Feuerriegel, "Business analytics for strategic management: Identifying and assessing corporate challenges via topic modeling," *Information and Management*, vol. 57, no. 1, 2020, doi: 10.1016/j.im.2018.05.003.
- [39] A. Amara, M. A. H. Taieb, and M. B. Aouicha, "Multilingual topic modeling for tracking COVID-19 trends based on Facebook data analysis," *Applied Intelligence*, vol. 51, no. 5, pp. 3052–3073, 2021, doi: 10.1007/s10489-020-02033-3.
- [40] S. Xu and Y. Xiong, "Setting socially mediated engagement parameters: a topic modeling and text analytic approach to examining polarized discourses on Gillette's campaign," *Public Relations Review*, vol. 46, no. 5, p. 101959, 2020, doi: 10.1016/j.pubrev.2020.101959.
- [41] J. A. Qundus, S. Peikert, and A. Paschke, "AI supported topic modeling using KNIME-Workflows," *CEUR Workshop Proceedings*, 2020, vol. 2535, pp. 1–7.
- [42] S. Sbalchiero and M. Eder, "Topic modeling, long texts and the best number of topics. Some problems and solutions," *Quality and Quantity*, vol. 54, no. 4, pp. 1095–1108, 2020, doi: 10.1007/s11135-020-00976-w.
- [43] I. Sutherland and K. Kiatkawsin, "Determinants of guest experience in Airbnb: a topic modeling approach using LDA," *Sustainability (Switzerland)*, vol. 12, no. 8, 2020, doi: 10.3390/SU12083402.
- [44] Q. Liu *et al.*, "Health communication through news media during the early stage of the covid-19 outbreak in China: digital topic modeling approach," *Journal of Medical Internet Research*, vol. 22, no. 4, 2020, doi: 10.2196/19118.
- [45] M. E. Roberts, B. M. Stewart, and D. Tingley, "stm: R package for structural topic models," *Journal of Statistical Software*, vol. 91, no. 1, pp. 1–40, 2019, doi: 10.18637/jss.v000.i00.
- [46] T. Davidson and D. Bhattacharya, "Examining racial bias in an online abuse corpus with structural topic modeling," *arXiv preprint arXiv:2005.13041*, pp. 2–5, 2020.
- [47] S. H. Kim, N. Lee, and P. E. King, "Dimensions of religion and spirituality: a longitudinal topic modeling approach," *Journal for the Scientific Study of Religion*, vol. 59, no. 1, pp. 62–83, 2020, doi: 10.1111/jssr.12639.
- [48] A. Bagheri, A. Sammani, P. G. M. van der Heijden, F. W. Asselbergs, and D. L. Oberski, "ETM: Enrichment by topic modeling for automated clinical sentence classification to detect patients' disease history," *Journal of Intelligent Information Systems*, vol. 55, no. 2, pp. 329–349, 2020, doi: 10.1007/s10844-020-00605-w.
- [49] X. Wu, C. Li, Y. Zhu, and Y. Miao, "Short text topic modeling with topic distribution quantization and negative sampling decoder," *EMNLP 2020 - 2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, 2020, pp. 1772–1782, doi: 10.18653/v1/2020.emnlp-main.138.
- [50] S. Kim, H. Park, and J. Lee, "Word2vec-based latent semantic analysis (W2V-LSA) for topic modeling: a study on blockchain technology trend analysis," *Expert Systems with Applications*, vol. 152, p. 113401, 2020, doi: 10.1016/j.eswa.2020.113401.
- [51] E. T. Khalid, E. B. Talal, M. K. Faraj, and A. A. Yassin, "Sentiment analysis system for COVID-19 vaccinations using data of Twitter," *Indonesian Journal of Electrical Engineering and Computer Science (IJECS)*, vol. 26, no. 2, pp. 1156–1164, 2022, doi: 10.11591/ijeecs.v26.i2.pp1156-1164.

BIOGRAPHIES OF AUTHORS






Mohd Mukhlis Mohd Sharif    is a Ph.D. candidate with over 19 years of experience in the IT industries. He recently completed his master's degree in data science from the School of Computing Sciences, College of Computing, Informatics and Media, Universiti Teknologi MARA Shah Alam, Malaysia. He is passionate about data analytics and machine learning and has strong problem-solving and communication abilities, along with a history of completing data projects on time and within budget. Currently, he is completing a doctorate in computer science at UiTM Shah Alam with research focusing on knowledge graphs and ontologies using social media corpora. Mukhlis can be reach at mohdmukhlis@gmail.com.






Ruhaila Maskat    is a senior lecturer at the School of Computing Sciences, College of Computing, Informatics and Media, Universiti Teknologi MARA Shah Alam, Malaysia. In 2016, she was awarded a Ph.D. in Computer Science from the University of Manchester, United Kingdom. Her research interest then was in Pay-As-You-Go dataspace which later evolved to data science where she is now an EMC Dell Data Associate as well as holding four other professional certifications from RapidMiner in the areas of machine learning and data engineering. Recently, she was awarded with the Kaggle BIPOC grant. Her current research grant with the Malaysian government involves conducting analytics on social media text to detect mental illness. She can be contacted at: ruhaila256@uitm.edu.my.



Zirawani Baharum    finished her doctorates degree for Doctor of Philosophy in Computer Science in 2017. She received her B.Sc. in Computer Science majoring in Modelling and Industrial Computing from Universiti Teknologi Malaysia (UTM), in 2003. Then, she received M.Sc. in Information Technology from UTM in 2005. She is currently a senior lecturer in Technical Foundation section, Universiti Kuala Lumpur, Malaysian Institute of Industrial Technology (UniKL MITEC). Her research interests are in the computer modelling and simulation, integrated model development, computer science and information and communication technology (ICT). She can be contacted at: zirawani@unikl.edu.my.



Kamaruzaman Maskat    has a Masters in Information Security from the University of Technology Malaysia (UTM). Currently working as an academician in National Defense University of Malaysia under Computer Science Department. He had two grants of RACE and RAGS. Obtained two professional certificates in EC-Council Network Security Administrator (ENSA) and Certified Ethical Hacker (CEH). Research interests are in information security, cybersecurity, machine learning, IOT and network. Had publication of about over 20 papers. He could be contacted at: kamaruzaman@upnm.edu.my.