Ibn Al-Haitham Journal for Pure and Applied Sciences
Journal homepage: jih.uobaghdad.edu.iq
PISSN: 1609-4042, EISSN: 2521-3407
IHJPAS. 2025, 38 (1)

Detection the topics of Facebook posts using text mining with Latent Dirichlet Allocation (LDA) algorithm

Shahlaa Mashhadani[1*]

[1]Computer Science Department, College of Education for Pure Sciences (Ibn Al-Haitham), University of Baghdad, Baghdad, Iraq.
*Corresponding Author.

Received: 8 August 2024 Accepted: 11 November 2024 Published: 20 January 2025
doi.org/10.30526/38.1.4033

Abstract

The development of artificial intelligence technologies has led to their massive integration in various fields, including daily life. Text data plays a pivotal role in the world of artificial intelligence, especially in machine learning, allowing valuable insights to be extracted from massive data sets to help make informed decisions. Latent Dirichlet Allocation (LDA) and digital forensics intersect through analyzing and classifying textual digital evidence in social media, including Facebook, in which text data is the main focus. This technique is particularly a useful topic modeling technique for uncovering hidden patterns in text data, which can be particularly useful in digital forensics taken from Facebook, including text analysis and evidence discovery, where LDA is used to extract large amounts of unstructured text data from meaningful topics, such as emails, documents, or chat logs. Investigators often deal with huge amounts of text-based evidence, so this technique helps them identify topics, such as fraud, especially in relation to text data, which is the core of our research. It not only improves effort and time but also carries a huge potential for security packages. This work presents a method for processing Facebook posts with the help of a Latent Dirichlet Allocation (LDA) ruleset to classify these texts into coherent themes. The significance of the research lies in its ability to discover themes within each post, which is crucial for analyzing user behavior and addressing security concerns. The use of relevant Facebook data enhances the real-world relevance of the results, facilitating targeted analysis based on the language patterns used by users in these posts and thus contributing to the success of security objectives. In evaluating existing methodologies, this study demonstrates improved performance by optimizing the LDA ruleset to more accurately match the unique features of the target statistics. This improvement leads to improved performance and reduced errors. The results of this study

demonstrate the effectiveness of using the LDA approach, as it showed significant improvements over traditional strategies in terms of accuracy and applicability to real-world security situations and digital analytics.
Keywords: Data Mining, Text Mining, Latent Dirichlet Allocation (LDA) algorithm, Digital Forensics , Machine Learning Techniques.

1. Introduction

Recently, the use of publicly available digital data has become more commonplace and at the same time unwanted since it has been increasing at an exponential rate. However, the amount of data that is generated and processed on a daily basis creates a certain complexity with respect to how this data can be processed, analyzed, and used. The ethical concerns regarding big data have also come up as a problem area, as has been noted in the literature (1). These include the ethical issues that must be considered, principles that should be respected, and practices that should be avoided that are likely to cause harm. Addressing the ethical considerations concerning big data analytics requires good data governance policies and practices, which can help ensure that maximization and minimization of big data risks are achieved. In recent years, topic models have been used and applied to derive useful information from large and unstructured textual documents and databases. Past works have indicated notable advances in using topic models aimed at unearthing hidden geometrical order in textual corpus along the edges of the LDA Model (2). The LDA approach has become a popular method that easily identifies unknown subjects in large amounts of text documents, which is relevant to the area of big data where the assessment of texts is not possible in a simple way.

The LDA algorithm is, without a doubt, one of the most critical procedures developed for modeling topics in recent decades. It works under the hypothesis that each document is a distribution of several topics, and each topic is expressed using certain words in a probabilistic manner (3). Thanks to its probabilistic nature, the LDA model is able to reflect hidden structures within text collections, which can extend the analytical parameters beyond traditional keyword analyses. Such algorithms' abilities have also boosted their popularity in text mining, particularly in social media analytics. The social networking sites that include Facebook, Twitter, and YouTube have large volumes of text data worth mining and include user behavior, likes, and opinions, among other aspects. With the use of LDA in social media data, it is possible for researchers to identify the topics of interest and track them across the variances of users (4). This way, besides expanding the knowledge of how social media works, certain areas where this information is applicable, like marketing, security, or public policy, can be utilized efficiently. In **Figure 1**, the process of the interface evolution comprises three stages, with the first one being the allocation of topics to each text/s, the second being the frequency of words belonging to those topics, and the third is the estimation of the likelihood that a particular topic is present in the text (4).

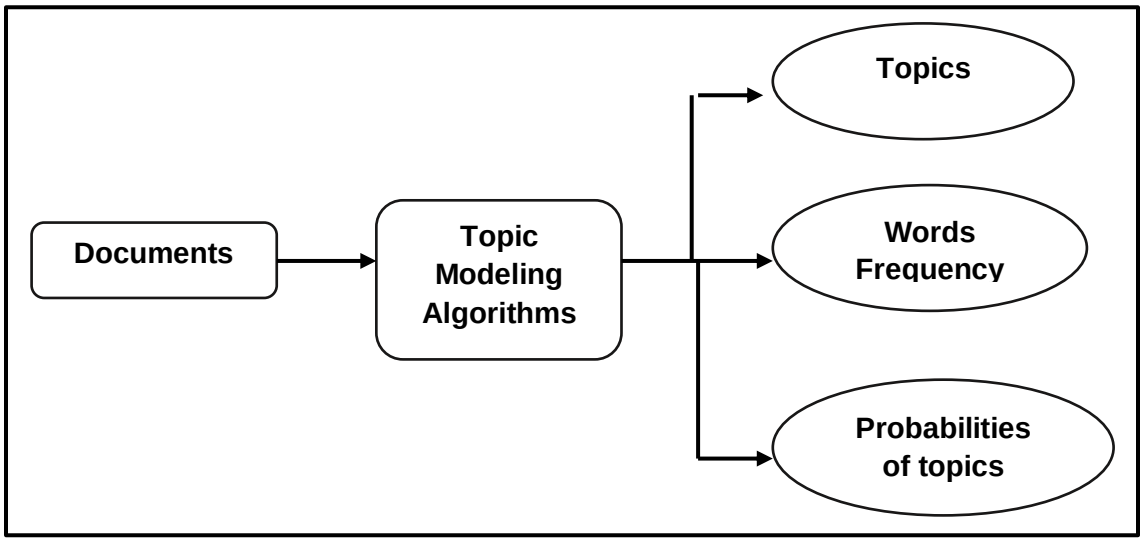

Figure 1. Topic Modeling Stages.

By using metadata analysis in digital forensics, investigators gain a powerful tool to extract insights and patterns from text-heavy datasets, facilitating more efficient and comprehensive investigations. These include: (1) Clustering evidence according to what digital forensics topics require to classify documents into relevant topics for investigation, where the role of LDA is to cluster documents to make it easier for forensic analysts to focus. (2) LDA can prioritize investigative leads and help forensic investigators prioritize the leads from documents that are most relevant to a case. (3) In cases involving cybercrimes and uncovering them from text, LDA can analyze chat logs from social media posts to uncover discussions related to malicious activity. Even with LDA's benefits, there are still barriers that interfere with the utilization of this method for text mining, especially the analysis of low and dense classification, such as social media data (5). It is not easy to avoid those problems since it is hard to detect and classify topics with the presence of slang, diagrams or letters, and non-ordinary writing. Moreover, the amount, diversity, and quality of user-generated content differ, and this can compromise the effectiveness of topic modeling algorithms. Still, these problems create a need for faster and better cp-algorithms. To overcome these issues, various scholars have adopted the first, second, and third refinements to LDA. Various research in the literature has shown that topics detected via a combination of LDA with additional techniques are accurate and robust (6). Such improvements come in handy, especially in situations where there is a lot of variation or noise in the text data. The addition of other contextual cues during the topic modeling process has also been effective in increasing the clarity and activity of the topics yielded.

The actual applications of LDA are not merely academic as they have already been put into practical use. For example, LDA has found application in systems designed for automatic lip reading regarding the analysis and understanding of voice even in the presence of background noise, exemplifying how this algorithm can be applied in various forms of textual and linguistics (7). This further solidifies the position of LDA among the popular text-mining tools. Apart from social media, LDA has been put to good use in the field of medicine and has complemented the

improvement of patient data management. The application of topic modeling to multi-channel healthcare data allows researchers to correlate various diseases with outcomes that assist the patients better and formulate more targeted therapies (8). This application emphasizes the general applicability of LDA in multiple areas. In a multilingual setting, the concern of LDA with the handling of documents in various languages is of paramount importance. Papers that pertain to the analysis or modeling of topics across different topical documents written in different languages, including aids for knowledge management retrieval and sharing, aren't uncommon (9). This aspect is important in cross-sectional studies conducted in many countries.

Integrating user-generated content from e-commerce data represents another promising application of LDA. It is used to link customer evaluations with product features. Helping companies Gain insights into consumer needs and improve their product offerings. (10) This method also helps to identify emerging trends based on customer feedback. This creates a competitive advantage in the market. For example, consumer perceptions on social media platforms can be analyzed by examining fashion trends. LDA-based topic modeling is also benefited from. The study explores how LDA can be used to analyze fashion trends without gender and processing unstructured data from social media (11). This research highlights the increasing importance of LDA in understanding and predicting consumer behavior in the dynamic online environment. This advanced fast text retrieval technique has also been developed to complement the LDA algorithm, which improves the speed and accuracy of text analysis. These advances enable efficient processing of large data sets, making LDA a feasible option for real-time applications in many areas, including security, marketing, and public health (12).

The structure of the paper is described as follows: Section 2 analyzes the latest advances in LDA. Next, Section 3 details the research methodology used in this study. Session 4 discusses the findings and implications. Finally, Session 5 concludes this article.

2. Related Studies

In the past few years, literature and scientific analysis have been carried out through a review of the available literature. Articles were collected, categorized, and reviewed to identify important information about various stakeholders in the area of interest that can be used later. Some studies focus on modeling topics and topic frameworks that have various implementations in natural processing languages. We divided the important studies into two parts; the first part shows the general studies using LDA, as shown in **Table 1.** The second focuses on various applications of LDA-based techniques in textual evidence analysis and digital forensics contexts, as demonstrated in **Table 2.**

Table 1. Summary of Related Studies (11-15)

Reference no#	Model Employed	Method Used	Limitations Problem
(11)	The LDA algorithm and linear regression were used to analyze customer promotions,	This research analyzed textual data from blogs and social media from 2018 to	The findings primarily highlighted customer interest in fragrances, fashion, and beauty products.

	opinions, and fashion trends through text mining.	2020 using Python version 3.7.	
(12)	LDA, NGRAM, and LSDM were used to extract information and model comparisons between posts and user accounts.	The research analyzed texts and documents using the LDA algorithm, which allows for information retrieval and document summarization.	LDA has been used in text mining, and it is very suitable in this field, but it has limitations, as it has only been applied to the English language.
(13)	LDA, an unsupervised machine learning technique, was employed to identify topics and uncover patterns by clustering word groups and similar expressions within a document collection.	The study used LDA	The paper's limitations are: a. Limited scope: Only data from one journal was used, potentially affecting the findings' generalizability. b. No direct modeling of topic occurrence correlations. c. Requirement for manual topic labeling. d. Short dataset span: The datasets cover only about nine years, possibly missing long-term trends.
(14)	The proposed technique uses historical COVID-19 data and textual data from news articles for improved feature extraction. Machine learning algorithms then utilize these features to predict COVID-19 trends. This model differs from traditional LDA by incorporating numerical COVID-19 data alongside text from specific time frames.	The method uses an LDA-based feature extraction model that includes COVID-19 data and online news articles. This approach enhances topic identification and feature extraction by linking features to responses to COVID-19's spread. These features are then used as inputs for machine learning algorithms to predict changes in COVID-19 case numbers.	The proposed technique is based on the assumption that news articles and COVID-19 data are reliable sources of information.
(15)	The model used Pandemic-LDA (PAN-LDA), a modified LDA that incorporates daily statistics of new COVID-19 cases and globally published news articles for feature extraction.	The development of a new feature extraction model, PAN-LDA, was employed to enhance results by integrating news articles with data on confirmed COVID-19 cases.	The main limitation of the paper was its failure to investigate the optimal hyperparameter values for machine learning algorithms like XGBoost and LightGBM, and it did not utilize these values.

Also, below is a table summarizing the work on the relationship between Latent Dirichlet Allocation (LDA) and digital forensics. Studies have focused on various applications of LDA-based techniques in textual evidence analysis and digital forensics contexts, as shown in **Table 2**.

Table 2. Summary of the relationship between LDA and digital forensics (16-20)

Reference no#	Dataset	When to Use	Importance	Applications	Advantages	Challenges
(16)	Social media chat logs.PDF files	When clustering large unstructured text datasets	Efficient evidence prioritization	Social media analysis, document clustering	Reduces manual effort	High computational requirements
(17)	Forensic Investigation reports	To automate reporting based on case evidence	Simplifies report generation	Case-specific summary generation	Time-saving for investigators	Subjectivity in topic labeling
(18)	Insider threat datasets(email logs)	When focusing on employee communications	Detects patterns related to insider threats	Organizational security	Scalable for large organizational datasets	Privacy concerns with sensitive data
(19)	Corporate datasets, email archives	For proactive identification of risks	Aids in forensic readiness and prevention	Fraud detection, compliance monitoring	Prepare organizations for future issues	Difficulties in dynamic environments
(20)	Cloud storage textual data	When analyzing large-scale cloud data	Ensures cloud- based evidence is processed	Cloud-based forensic analysis	Handles distributed datasets efficiently	Cloud security and data privacy concerns

3. Methodology

In this section, a structure is proposed to determine which topic each message belongs to. It is explained step by step. The structure consists of several main steps, such as feature selection. Deleting unrelated words Using the LDA algorithm and topic model visualization. The method is illustrated in **Figure 2**, where each process leads to the extraction of clear and understandable text from a given overlay. It is derived from documents inserted via Facebook. The result is a presentation of both topics provided in well-structured and consistent content. At last, the results will be an obvious and understandable text extracted from a set of overlapping data for documents entered from Facebook media, which gives clear texts.

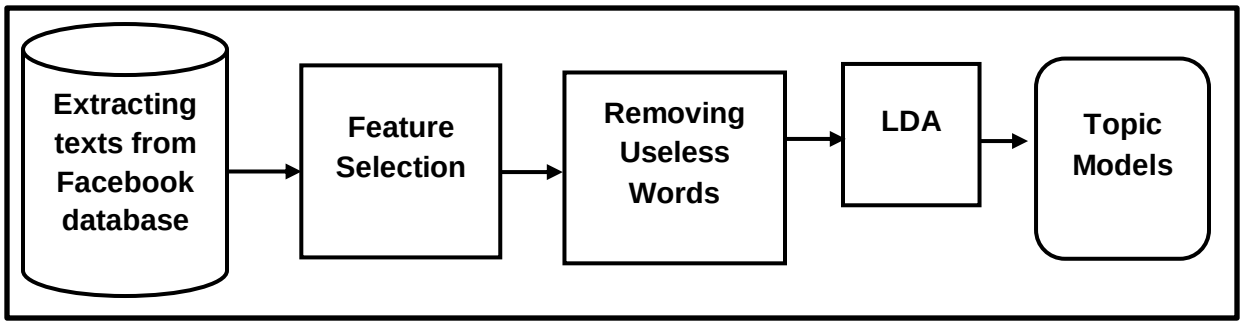

Figure 2. The technique used of the text mining digital forensics**.**

3.1 Dataset

Datasets that is used non-script are called FactoryReports.csv. This appears to be a report related to an incident at the factory. Each item in the dice set has a text description of the event along with category labels. Text descriptions are the main focus of the analysis. When the LDA uses these data to find specific topics or topics hidden in our reports, this dataset usually consists of a collection of descriptions of events in a factory. This can provide details of events that have occurred, such as equipment malfunctioning. Maintenance rate or security incidents by analyzing these text descriptions. The LDA can identify recurring patterns and themes. It helps categorize and understand the types of events that occur in a factory (21).

3.2 Algorithm Parameters

The input parameters guide how the Latent Dirichlet Allocation (LDA) model works to find hidden topics in a set of documents. The first item is the given item. This can be displayed in many different ways. For example, data might consist of individual document words or phrases along with the frequency at which they appear. This is an important starting point because the model relies on word data to identify document patterns and themes. The second input, K, is the number of topics the model expects to find, which represents the number of topics or categories the model should discover within the text collection. Choosing the correct number of topics is essential. Because too few topics may not capture enough details, too much can make the results confusing or less meaningful. Finally, additional options allow users to fine-tune the process by configuring different settings. These options may include precise model accuracy. Or should additional information be required during processing? This flexibility helps adapt the model to different data sets and analysis objectives.

3.3 Text-Preprocessing

The script uses a custom function called (preprocess text) to prepare text data for systematic analysis. The first step of this process is tokenization. Each document is divided into individual tokens or words. Once converted to a token, the next step is to make it a small file to facilitate processing and analysis. This involves discussing the word at its root. This guarantees that different conjugations of the same word (such as "run," "run," and "run") are treated as the same token. Improved consistency between two text dice. By formatting words in this way, the script reduces redundancy and increases the accuracy of the LDA model in identifying topics.

After inserting punctuation marks, the script will remove them. It eliminates symbols such as commas, periods, and question marks. This is because they do not contribute meaningful data to the topic model. This function also removes common, unrelated words such as "and," "of," and "or," which are commonly used words. But there is little meaning in searching for topics. To further refine the dice set, short words (less than three letters) and long words (more than fifteen letters) were filtered. This step ensures that the analysis focuses on words with a higher probability. To be involved: It discards words that are too short or unusually long and might introduce non-model noise. Together, these pre-processing steps create clean, structured inputs for the LDA model, helping us efficiently uncover hidden topics derived from the text.

3.4 Feature Selection

Here is what happens after you upload the data set and step 1; you need to extract text information in this step under the field description. That includes the bootstrapping and configuration of data prior to analysis. In this stage, we create a bag of words template that flags the unique words present in the documents. Feature selection is a key step that extracts the most informative and relevant words from your dataset before modeling them on topic. The goal is to make the labeled tokens just enough to represent the main topics in the dataset and improve topic extraction (22–24).

3.5 Removing Useless Words

After the feature selection step is finished, we begin removing words that may be irrelevant or do not have any significance for the next step. This part is a very important step in making the information safer, more stable, and clearer after it develops. Text preprocessing converts text data into a more digestible format that could be easily processed for further data mining and machine learning-based ML tasks. Preprocessing is often used during the beginning AI building stages to make sure that the only data that gets processed is what requires getting through as a clean feed. For this task, several steps are performed sequentially, such as tokenization, normalization, and removal of the punctuations and stop words. Here is the preprocess function developed with MATLAB and performed in the given order (25, 26):

i. **Tokenization**: The text is converted into tokens using the `tokenizedDocument` function.
ii. **Normalization**: Spoken words are standardized using the `NormalizeWords` function.
iii. **Punctuation Removal**: Punctuation marks are erased using the `clear punctuation` function.
iv. **Stop Word Removal**: Common words such as "and", "of", and "the" are removed using the `RemoveStopWords` function.
v. **Short Words Removal**: Words with two or fewer characters are discarded using the `RemoveShortWords` function.
vi. **Long Words Removal**: Words longer than 15 characters are removed using the `RemoveLongWords` function.

Preprocessing is used with text tagging where a bag of words is made removing words which do not appear once or more across the dataset. Further, any documents that do not have any searchable words after this process are evacuated from the examination, which guarantees that these staying on-message segments will be both applicable and still reasonable (27).

3.6 LDA and Topic Modeling

LDA is an unsupervised ML algorithm used to find latent topics of which documents are a mixture. The basic assumption of LDA is that each document in a corpus comprises an unobserved probabilistic distribution of topics. Likewise, each topic is a distribution over words so that some words are more probable than others in each topic. LDA is also a probabilistic generative model that tries to explain how documents are created by picking the topic for each word in the document from a discussion of topics for that document and then choosing a word based on that selected topic. The core of the framework is the application of the LDA algorithm. This probabilistic model is designed to classify the documents based on their word distributions, helping to identify key

topics within the dataset. The LDA algorithm works by extracting the relevant words from each group of texts and classifying them according to predefined target topics, which the designer specifies. These topics are stored within the layers of the LDA model, ensuring that the extracted topics are aligned with the intended research goals. The following configurations are made to implement LDA in MATLAB (28-31).

The 'Solver' option is set to savb, controlling the algorithm efficiency-accuracy tradeoff. It is important to choose the right solver, as it directly affects the quality of topics generated and the overall performance of the LDA model. To prevent verbose output during execution, often with large datasets, the Verbose command is set to 0, which simply keeps the output clean and free from unnecessary details so that we can focus on what matters most. It also approximates the LDA model, which means that computational efficiency may be improved for large data sets by using fewer passes on the data and still getting high-quality topics. Few steps can be taken to overcome the challenges in DETROIT, which are specific to social media data since it is often noisy, short, and informal language and improves the performance of the LDA Algorithm on Facebook posts. The steps for enhancing the LDA and implementing it in a Facebook Post are outlined in Table 1 with references (32, 33):

i. *Data Preprocessing and Cleaning*: Handling informal language in Facebook posts is crucial, as they often include slang, abbreviations, and emojis. Preprocessing addresses this by converting informal language into standardized terms, such as changing "u" to "you" and removing unnecessary elements like emojis or special characters. In addition, posts may sometimes include useless or irrelevant data such as hashtags, links, or repeated phrases and sentences, which eliminates this clutter, guaranteeing that the LDA model concentrates on content that is important. Another major step is to remove stop words that include everyday terms like (the, is, and at), which don't aid in topic comprehension. By removing those words, the model is able to identify more insightful patterns. To end this, lemmatization or stemming reduces words to their root forms, such as "running" to "run." Also, ensuring that similar words are handled on a basis, improving the topics that are produced by the LDA algorithm's overall coherence.

ii. Modifying LDA for short texts: LDA often has a difficult time with short documents like Facebook posts, for example, and this is due to the limited context that each document provides. To improve topic coherence in these cases, one approach is to aggregate posts by combining multiple posts from the same user or related to the same topic into a single document. This provides more context for LDA to extract meaningful topics. Another strategy involves integrating word embeddings, such as Word2Vec or GloVe, to enhance the model's understanding of word relationships. Word embeddings capture semantic similarities between words, improving the quality and coherence of the topics inferred by LDA.

iii. *Post-Processing to Improve Topic Coherence*: Filtering low-frequency topics is an important step after running LDA, as some topics may be generated by rare words or outliers that do not contribute meaningfully to the overall analysis. Removing these topics helps to improve the coherence of the remaining topics. Additionally, LDA can sometimes produce highly similar

topics, which may overlap significantly. Clustering and merging these similar topics can create more distinct and meaningful categories, further refining the quality of the topic model and enhancing interpretability.

iv. *Application to Facebook Posts*: Once the LDA model is optimized, it can be applied to a large set of Facebook posts, where each post is analyzed and assigned a distribution of topics. Since the topics generated by LDA are usually unlabeled, manual or semi-automated labeling is necessary to interpret the results. For example, a topic containing words like "vacation," "beach," and "travel" could be labeled as "Travel & Leisure." Additionally, applying LDA to posts over time allows the model to detect trends in user discussions, identifying how certain topics become more or less prevalent. LDA can also be used for user segmentation by grouping users based on the most frequently discussed topics. It enables Facebook or other analysts to create user segments for personalized content or targeted marketing.

The LDA model's effectiveness depends on its ability to generate relevant and clear known topics from the document collections. With what we have, the model is applied to a varying number of documents (Ns: 5, 7, 10, and 20); this ensures that the topic extraction is tough across different sample sizes (34).

3.7 Visualization of Topics:

After the LDA model has been applied and used, the next step forward is to visualize the results by using MATLAB's word cloud tool, which provides a visual representation of the words with the highest percentage in each topic. Therefore, the word cloud for the first four topics is generated, allowing for a built-in understanding of the well-known terms that are associated with each topic. Additionally, the mix of topics for each document can be visualized by converting the documents into vectors. This enables the analysis of the percentages of each matter that is present and found in documents, revealing the complexity of topics across the dataset. The topic mixture is often visualized for the first Ns number of documents, which helps to identify patterns and the overall thematic structure of the data. The visualization of these topic mixes not only enhances the interpretability of the model results but also facilitates the understanding and knowledge of how topics are distributed within the dataset, allowing further actionable insights (35-40).

4. Result and Discussion

The given and shown method for the text mining was implemented using MATLAB version 2020b. Within this model, the LDA algorithm serves as a topic mapper to detect key major topics in a document corpus and infer word probabilities within these topics. This initial step involved the process of uploading project data, which consisted of a factory report. CSV file that contains text descriptions and categorical labels for events within a plant. For example, data collected from Facebook was used as factory-related information to test the model, identifying four fundamental topics for mining. There was flexibility in choosing the number of subjects because the LDA model was developed using a range of document counts (Ns = 5, 7, 10, 20). To ensure detail, punctuation and small and large words were eliminated during the text mining process. The results represent the effectiveness of this LDA-based model in successfully isolating the topics from the input texts

while also removing fixed determinants; along with that, they are visualized through MATLAB's word cloud tool, providing insights into the division of topics while showing how the proposed algorithm distributes topic probabilities within the dataset. This success is clearly illustrated in **Figures 3, 4**, and **5**.

Figure 3. The presentation of the cloud word using text mining for Ns=5,7,10, and 20 documents.

In **Figure 3**, we see how the visualization of topics evolves with different sample sizes. As the number of documents (Ns) increases, the clarity of topics improves, and more granular and distinct topics emerge. For example, at Ns = 5, the word clouds for each topic appear less defined, whereas at Ns = 20, the topics are more well-structured and characterized by more specific keywords. This indicates that the LDA model is more effective in capturing refined topics when applied to a larger set of documents, allowing for a comprehensive analysis of word distributions and co-occurrences. **Figure 4** illustrates the topic probabilities across various N values. The histograms represent the likelihood of specific topics being associated with each document. The topic probability distribution becomes more spread out as Ns increases, indicating that a larger and more diverse range of topics is detected across the documents. At the same time, smaller Ns values (5 and 7) show concentrated topic distributions, meaning that fewer topics dominate the documents. For instance, at Ns=5, the documents mainly align with two topics (one and four), whereas at Ns=20,

the documents are more evenly and equally spread across a wider set of topics. However, as the document count increases to Ns = 20, the topic mixture becomes more balanced, reflecting the increased complexity and diversity of the data. This illustrates the importance of sample size in LDA applications, as larger datasets enable more detailed and nuanced topic detection. As a result, we can see the model's enhanced sensitivity in capturing thematic variations as the dataset grows. Furthermore, topic mixture probabilities for individual documents are explained in **Figure 5**.

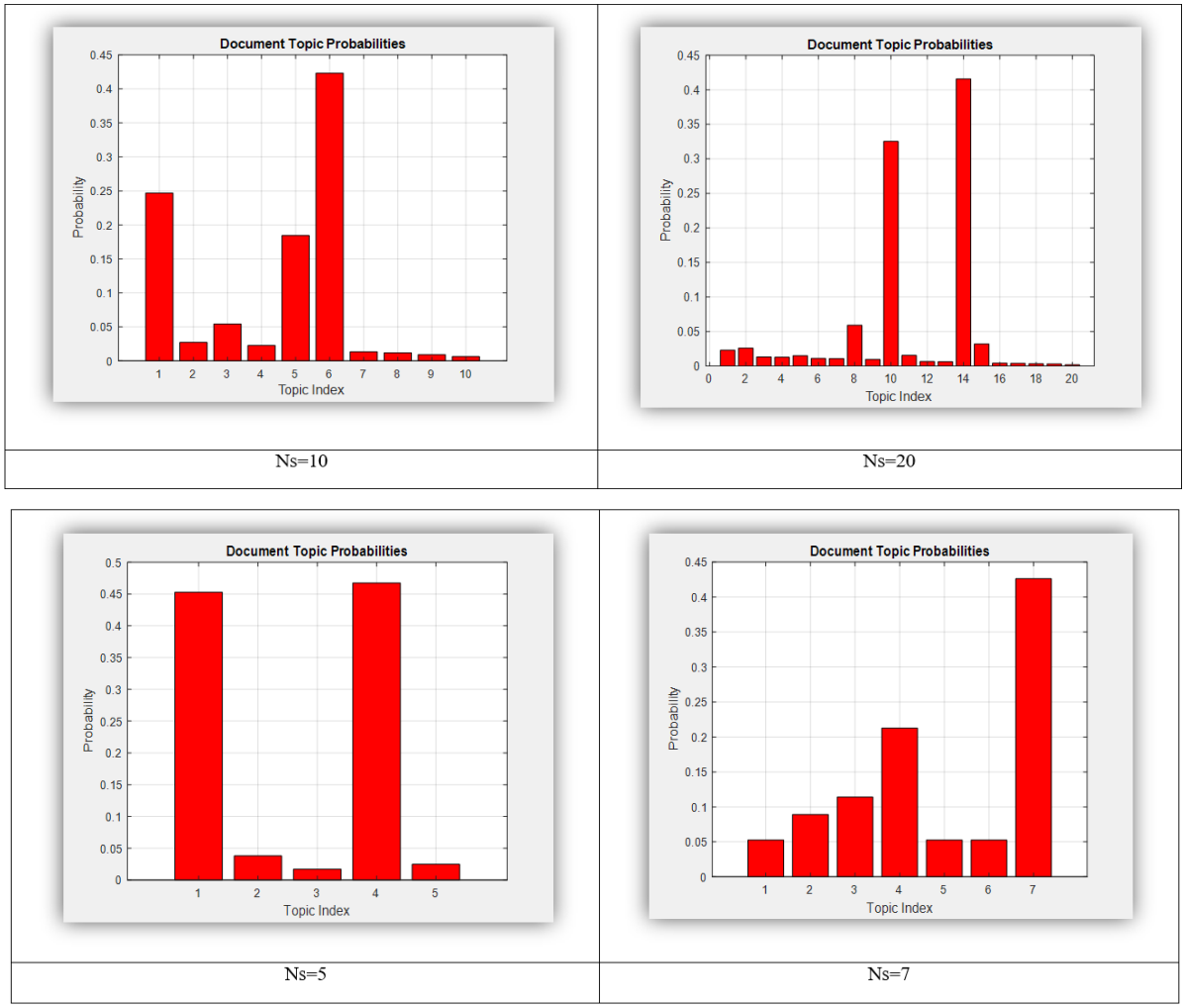

Figure 4. The probabilities achieved for the documented topic with Ns=5,7,10 and 20 documents.

Overall, as demonstrated by the progression of results from Ns = 5 to Ns = 20, the represented LDA model performs effectively across a variety of document sizes. This proves the model's ability to adapt to varying data sizes and produce detailed topic analysis, making it suitable for mining large text datasets and improving the efficiency of topic detection in real-world applications and uses. Thus, our suggested technique will use the LDA algorithm for text mining and apply particular limitations on analyzed and examined text units that meet our design objectives of producing correct and clear topics suitable for safety and privacy applications.

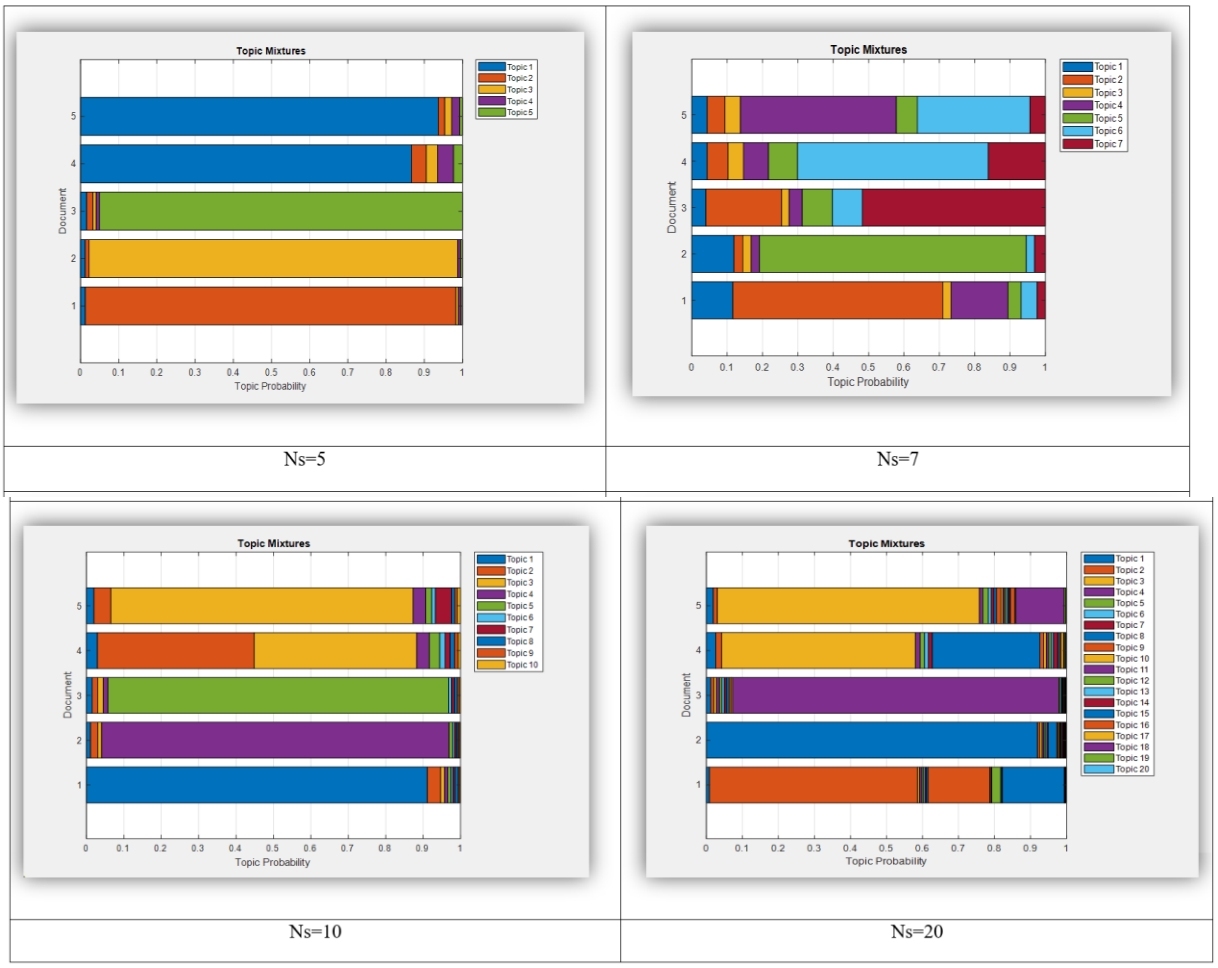

Figure 5. The probabilities of each documented topic mixture with Ns=5,7,10 and 20 documents.

Table 3 compares the proposed method's results with recent work and shows its effectiveness.

Table 3. A comparison with recent work.

No	Recent Work	Method Used
1	It is based on a relatively small sample of texts that cannot be validated using big data.	It is implemented on a large amount of data and can be applied to big data in the future.
2	The images in the tweet can be considered when implementing the model.	The images in the tweet can be considered when implementing the model.
3	The performance can be affected by more enormous datasets or images.	The performance will not be that much affected in the case of more enormous datasets or images.

5. Conclusion

This paper presents a strongly built approach to text mining using the Latent Dirichlet Allocation (LDA) algorithm, specifically applied to social media data. This method proves its effectiveness in identifying key topics within large and complex datasets while providing a more structured understanding of unorganized text data. This is done by improving the efficiency and accuracy of topic detection. The proposed approach addresses the challenges and difficulties of processing vast amounts of data, such as those found on platforms like Facebook. It utilized real-world social

media data to enhance the relevance of the findings for practical applications, including sentiment analysis, trend prediction, and security monitoring. This method improved existing LDA-based techniques and reduced errors and performance by tuning the preprocessing and text classification steps. These advancements make it more suitable for handling the noisy, diverse, and unstructured data that social media platforms generate. Moreover, the Latent Dirichlet Allocation (LDA) algorithm has proven to us that it is a valuable tool in the field of digital forensics involving textual data analysis and topic modeling. It has been shown to be able to uncover topics in unstructured textual data, which also makes it useful in analyzing digital evidence, such as emails and documents, and efficiently summarize data that might otherwise remain hidden. Despite its advantages, the algorithm has challenges, such as its reliance on parameter tuning for large data sets and its inability to interpret message patterns directly. In conclusion, linear data analysis technology contributes significantly to the analytical toolkit used in digital forensic social media investigations such as Facebook. Integrating this approach into security and surveillance applications opens new possibilities for real-time monitoring and threat detection using user-generated content. Advances in linear data analysis technology, combined with its integration into other algorithms, could further enhance its effectiveness and strengthen its role in managing the growing complexity of digital forensic investigations.

Acknowledgment

Our researcher extends his Sincere thanks to the editor and members of the Ibn AL-Haitham Journal of Pure and Applied Sciences preparatory committee.

Conflict of Interest

There are no conflicts of interest.

Funding

There is no funding for the article

References

1. Bormida MD. The Big Data World: Benefits, Threats and Ethical Challenges. Ethical Issues in Covert, Security and Surveillance Research. Advances in Research Ethics and Integrity. Leeds: Emerald Publishing Limited; 2021. p. 71–91. Available from: https://doi.org/10.1108/S2398-601820210000008007
2. Jelodar H, Wang Y, Yuan C, Feng X, Jiang X, Li Y, Zhao L. Latent Dirichlet allocation (LDA) and topic modeling: Models, applications, a survey. Multimedia Tools Appl. 2019;78:15169–211. Available from: https://doi.org/10.1007/s11042-018-6894-4
3. Chauhan U, Shah A. Topic modeling using latent Dirichlet allocation: A survey. ACM Comput Surv. 2021;54(7):1–35.
4. Amenah HA, Bara'a AA, Rashid AN, Al-Ani M. A new evolutionary algorithm with locally assisted heuristic for complex detection in protein interaction networks. Appl Soft Comput. 2018;73:1004–25. Available from: https://doi.org/10.1016/j.asoc.2018.09.031

5. Hashim AR. Pupil detection based on color difference and circular Hough transform. Int J Electr Comput Eng. 2018;8:3278–84. Available from: https://doi.org/10.11591/ijece.v8i5.pp3278-3284
6. Abdulsalam WH, Alhamdani RS, Abdullah MN. Emotion recognition system based on hybrid techniques. Int J Mach Learn Comput. 2019;9(4). Available from: https://doi.org/10.18178/ijmlc.2019.9.4.831
7. Rajab MA, Hashim KM. An automatic lip reading for short sentences using deep learning nets. Int J Adv Intell Inform. 2023;9(1). Available from: https://doi.org/10.26555/ijain.v9i1.920
8. Lu HM, Wei CP, Hsiao FY. Modeling healthcare data using multiple-channel latent Dirichlet allocation. J Biomed Inform. 2016;60:210–23. Available from: https://doi.org/10.1016/j.jbi.2016.02.003
9. Lui M, Lau JH, Baldwin T. Automatic detection and language identification of multilingual documents. Trans Assoc Comput Linguist. 2014;2:27–40. Available from: https://doi.org/10.1162/tacl_a_00163
10. Zoghbi S, Vulić I, Moens MF. Latent Dirichlet allocation for linking user-generated content and e-commerce data. Inf Sci. 2016;367:573–99. Available from: https://doi.org/10.1016/j.ins.2016.05.047
11. Kim H, Cho I, Park M. Analyzing genderless fashion trends of consumers' perceptions on social media: Using unstructured big data analysis through Latent Dirichlet Allocation-based topic modeling. Fashion Text. 2022;9(1):1–21.
12. Gnanavel S, Mani V, Sreekrishna M, Amshavalli RS, Gashu YR, Duraimurugan N, et al. Rapid Text Retrieval and Analysis Supporting Latent Dirichlet Allocation Based on Probabilistic Models. Mob Inf Syst. 2022. Available from: https://doi.org/10.1155/2022/6028739
13. Yadav K, Kumar N, Maddikunta PKR, Gadekallu TR. A comprehensive survey on aspect-based sentiment analysis. Int J Eng Syst Model Simul. 2021;12(4):279–90. Available from: https://doi.org/10.1504/IJESMS.2021.119892
14. Zhang Y, Chen M, Huang D, Wu D, Li Y. iDoctor: Personalized and professionalized medical recommendations based on hybrid matrix factorization. Future Gener Comput Syst. 2017;66:30–5. Available from: https://doi.org/10.1016/j.future.2015.12.001
15. Sun S, Luo C, Chen J. A review of natural language processing techniques for opinion mining systems. Inf Fusion. 2017;36:10–25. Available from: https://doi.org/10.1016/j.inffus.2016.10.004
16. Anwar W, Bajwa IS, Choudhary MA, Ramzan S. An empirical study on forensic analysis of Urdu text using LDA-based authorship attribution. IEEE Access. 2019;7:3224-3234. doi: https://doi.org/10.1109/ACCESS.2018.2885011
17. Qadir AM, Varol A. The Role of Machine Learning in Digital Forensics. In: 8th Int. Symp. Digit. Forensics Secur. ISDFS 2020; 2020. https://doi.org/10.1109/ISDFS49300.2020.9116298
18. Bin Sarhan B, Altwaijry N. Insider Threat Detection Using Machine Learning Approach. Appl Sci. 2023;13(1). doi: https://doi.org/10.3390/app13010259
19. Knn RU, Regression L. Credit Card Fraud Detection: An Improved Strategy for High. 2023.
20. Lu Y, Wang J. Constructing a Digital Capability Evaluation Framework for Manufacturing Enterprises in the Context of Digital Economy: Based on LDA, Entropy Weight and TOPSIS Model. 2024. doi: https://doi.org/10.4108/eai.23-2-2024.2345917
21. Xu Z, Liu Y, Xuan J, Chen H, Mei L. Crowdsourcing based social media data analysis of urban emergency events. Multimedia Tools Appl. 2017;76:11567-11584. doi: https://doi.org/10.1007/s11042-015-2731-1
22. Chen T-H, Thomas SW, Hassan AE. A survey on the use of topic models when mining software repositories. Empir Softw Eng. 2016;21(5):1843-1919. doi: https://doi.org/10.1007/s10664-015-9402-8

23. Debortoli S, Müller O, Junglas I, Vom Brocke J. Text mining for information systems researchers: An annotated topic modeling tutorial. Commun Assoc Inf Syst (CAIS). 2016;39(1):7.
24. Debortoli S, et al. Text mining for information systems researchers: an annotated topic modeling tutorial. CAIS. 2016;39:7.
25. Sun X, et al. Exploring topic models in software engineering data analysis: A survey. In: 2016 17th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD); 2016. IEEE.
26. Pejić Bach M, Krstić Ž, Seljan S, Turulja L. Text mining for big data analysis in financial sector: A literature review. Sustainability. 2019;11(5):1277. doi: https://doi.org/10.3390/su11051277
27. Amado A, Cortez P, Rita P, Moro S. Research trends on big data in marketing: A text mining and topic modeling based literature analysis. Eur Res Manag Bus Econ. 2018;24(1):1-7. doi: https://doi.org/10.1016/j.iedeen.2017.06.002
28. Wang Y-C, Burke M, Kraut RE. Gender, topic, and audience response: an analysis of user-generated content on Facebook. In: Proceedings of the SIGCHI conference on human factors in computing systems; 2013. ACM.
29. Alashri S, Kandala SS, Bajaj V, Ravi R, Smith KL, Desouza KC. An analysis of sentiments on Facebook during the 2016 US presidential election. In: 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM); 2016. IEEE; 2016:795-802. doi: https://doi.org/10.1109/ASONAM.2016.7752329
30. Bastani K, Namavari H, Shaffer J. Latent Dirichlet allocation (LDA) for topic modeling of the CFPB consumer complaints. Expert Syst Appl. 2019;127:256-271. doi: https://doi.org/10.1016/j.eswa.2019.03.001.
31. Li H, Lu W. Learning latent sentiment scopes for entity-level sentiment analysis. In: Proceedings of the AAAI Conference on Artificial Intelligence. 2017;31(1). Available from: https://doi.org/10.1609/aaai.v31i1.11016
32. Mouhssine E, Khalid C. Social big data mining framework for extremist content detection in social networks. In: 2018 International Symposium on Advanced Electrical and Communication Technologies (ISAECT). IEEE; 2018 Nov. p. 1–5. Available from: https://doi.org/10.1109/ISAECT.2018.8618726
33. Quan X, Kit C, Ge Y, Pan SJ. Short and sparse text topic modeling via self-aggregation. In: Yang Q, Wooldridge M, editors. Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI). 2015. p. 2270–6. Available from: https://www.ijcai.org/Abstract/15/321
34. Laureate CDP, Buntine W, Linger H. A systematic review of the use of topic models for short text social media analysis. Artif Intell Rev. 2023;56(12):14223–55. Available from: https://doi.org/10.1007/s10462-023-10471-x
35. Mhamdi C, Al-Emran M, Salloum SA. Text mining and analytics: A case study from news channels posts on Facebook. In: Intelligent Natural Language Processing: Trends and Applications. 2018. p. 399–415. Available from: https://doi.org/10.1007/978-3-319-67056-0_19
36. Li H, Lu W. Learning latent sentiment scopes for entity-level sentiment analysis. In: Proceedings of the AAAI Conference on Artificial Intelligence. 2017;31. Available from: https://doi.org/10.1609/aaai.v31i1.11016
37. Mouhssine E, Khalid C. Social big data mining framework for extremist content detection in social networks. 2018. Available from: https://doi.org/10.1109/ISAECT.2018.8618726
38. Niu Y, Zhang H, Li J. A Pitman-Yor process self-aggregated topic model for short texts of social media. IEEE Access. 2021;9:129011–21. Available from: https://doi.org/10.1109/ACCESS.2021.3113320

39. Laureate CDP, Buntine W, Linger H. A systematic review of the use of topic models for short text social media analysis. Artif Intell Rev. 2023. Available from: https://doi.org/10.1007/s10462-023-10471-x
40. Mhamdi C, Al-Emran M, Salloum SA. Text mining and analytics: A case study from news channels posts on Facebook. In: Intelligent Natural Language Processing: Trends and Applications. 2017. p. 399–415. Available from: https://doi.org/10.1007/978-3-319-67056-0_19