Hindawi
Mobile Information Systems
Volume 2022, Article ID 6028739, 12 pages
https://doi.org/10.1155/2022/6028739

Research Article

Rapid Text Retrieval and Analysis Supporting Latent Dirichlet Allocation Based on Probabilistic Models

S. Gnanavel,[1] **Vinodhini Mani**,[2] **M. Sreekrishna**,[2] **R. S. Amshavalli**,[2] **Yomiyu Reta Gashu**,[3] **N. Duraimurugan**,[4] **and Namburi Srinivasa Rao**[5]

[1] *Department of Computing Technologies, School of Computing, SRM Institute of Science and Technology, Kattankulathur 603203, Chengalpattu, Tamil Nadu, India*
[2] *Department of Computer Science and Engineering, Sathyabama Institute of Science and Technology, Chennai, India*
[3] *Mizan Tepi University, Tepitown, Addis Ababa, Ethiopia*
[4] *Department of Computer Science and Engineering, Rajalakshmi Engineering College, Chennai, India*
[5] *IT Department, Bapatla Engineering College, Bapatla, India*

Correspondence should be addressed to Yomiyu Reta Gashu; yomiyu@mtu.edu.et

Received 4 April 2022; Revised 26 April 2022; Accepted 23 June 2022; Published 8 August 2022

Academic Editor: M. Praveen Kumar Reddy

Text mining, also known as text analysis, is the process of converting unstructured text data into meaningful and functional information. Text mining uses different AI technologies to automate data and generate valuable insights, allowing enterprises to make data-based decisions. Text mining enables the user to extract important content from text data sets. Text analysis encourages machine learning ability for research areas such as medical and pharmaceutical innovation fields. Apart from this, text analysis converts inaccessible data into a structured format, which can be used for further analysis. Text analysis emphasizes facts and relationships from large data sets. This information is extracted and converted into structured data for visualization, analysis, and integration as structured data and refines the information using machine-learning methods. Like most things related to Natural Language Processing, text mining can seem like a difficult concept to understand. But the fact is, it does not have to be. This research article will go through the basics of text mining, clarify its different methods and techniques, and make it easier to understand how it works. We implemented Latent Dirichlet Allocation techniques for mining the data from the data set; it works properly and will be in future development data mining techniques.

1. Introduction

Nowadays, there is an increase in public discussions about a product on social media. Analysis of this type of data requires modern technical features of analytics tools and text analytics. Text analysis attempts to understand the meaning of the written word. This is very difficult because it depends on the communication situation of humans. Therefore, the development of social media surrounds a level of communication inability. There is a need to identify the conversations that take place in the community such as on YouTube, Facebook, blog posts, and tweets, where the comments should be in the public mind. Text Analytics is a promising solution and is useful in these times [1, 2].

Text analysis applies to categorizing or mining information from the text about the environment [2]. The purpose of text analysis is to create a semantic structure based on differences in the concept, meaning, and attitude of the context. It starts by answering some questions that cannot be answered in the text to be analyzed. Sentiment analysis is a type of text analysis that identifies the polarity of the content. [3].

Text analysis is a type of data processing that attempts to detect text formats from large unstructured sources. Also known as text data mining, it is the process of retrieving trivial information and knowledge from unstructured text [4].

Patil and Manjrekar explained that text analytics can hide unstructured or semi-structured databases such as fulltext documents, HTML files, emails, blogs, academic papers, and newspaper articles. Finally, they said text analysis is about learning data processing, information extraction, and machine learning. The computer is unaware of the relationship between texts and words. The system must be trained by humans for the relationship between words [5].

Sentiment analysis is the most important task in text analysis. People often have problems recognizing others, even when in direct contact. An error of text is often caused by a lack of facial expressions or voice traces. This legal writing is hard work and tries to avoid mistakes. People working with systems are the solution to this problem. Systems can identify whether a given word or sentence is positive or negative. Individuals with specific domain expertise can analyze low confidence results and teach the machine the way to standardize these low confidence results. Over time, the computer expert will understand this feature and become more accurate and effective. Our proposed model works perfectly for the given data set and it is able to identify parts of speech like nouns, verbs, etc, and also count the number of words, characters, and character spaces.

This article is organized as follows: Section 2 explained Related works, Section 3 Methodology, Section 4 Results and Discussion, and Section 5 conclusions.

2. Related Works

Park et al. proposed improvements in word-embedding help to implement a concept from minimal seed terms. But the primitive use of those techniques creates false positive errors due to the principle of natural language. To solve this problem, a visual analysis system called Feed Vector is provided, which prompts a user to use feedback to generate feedback and analyze documents. A bipolar concept model has been introduced to refer to irrelevant words [6].

Packiametal and Prakash said that the usage of big data increased but it has some real-time issues. In particular, the key part of information retrieval is the best experience with big data. The main purpose is to find or improve the best low-cost and reliable techniques for extracting values from high terabytes and petabytes of available data. Hence big data analysis is important. Conventional analysis concentrates on structured data, but it is not suitable for large amounts of unstructured data. Text analysis is a technique for gaining prominence from unstructured text to identify transformations and patterns [7].

Medoc et al. provide a Visual Analytics tool for research work and status awareness for text streams. To achieve the data model, encoding streaming text is designed in several dynamic frequency measurements. Visualization includes two dynamic theme rivers. These two dynamic theme rivers enable real-time travel of most features derived from texts stored in long-term and short-term buffers [8].

Bradel et al. suggested that semantic communication delivers difficult statistical models and intrinsic communication techniques between human users. Semantic communication concentrates on manipulating direct spatiality by controlling users who handle sample parameters. However, this semantic communication technique cannot be substantially measured for many text documents. To solve this problem, the multimodel semantic communication concept is put forward in which semantic interactions can trigger multiple models at different levels of the data model so that users can manage large data issues [9–13].

Chiranjeevi et al. presented text documents that are a source for storing personal or public information. Currently, text documents are produced at a tremendous speed and the search engine needs to process the data immediately to improve,. providing a computer that improves the process of retrieving information from unstructured data from text documents in the search engine [14].

Park investigated Java-based Visual Analytics tool to read a collection of various text data sources and retrieve relationships, keywords, and events from text data using ontology and language processing methods. Then they implemented Java-based Visual Analytics tool to provide users with an integrated and intuitive search interface to support useful and robust inquiry into large and difficult data sets [15].

Vandierendonck et al. proposed an optimized solution for text mining. The processing of text data has high performance and is remarkable. A library of high-performance text analytics is provided. This library enables programmers to make maps for heavy numeric representations that can effectively manage text data. The library integrates three performance enhancements that are as follows:

(i) Text data's effective memory management

(ii) Parallel computation on relative data structures which map text to values

(iii) Optimization of the relative data structure's type depending on the program context [16]

Heimerl et al. proposed that word clouds have developed a direct and visually appealing visual technique for text. They are used in different characteristics to provide an overview by mining the text for words that are expressed with more events. In particular, it is done consistently with pure text abstractions. For tasks with general text analysis, the use of word clouds is described [17].

Social networking sites are a common platform for sharing and referral. Posts and comments on someone's wall like Facebook judge people in many situations based on the advice of others. These ideas affect one's thinking when determining because there are so many ideas and opinions. As a result, the user likes a particular post and gives negative feedback for it. To solve the problem, a sentiment analysis approach is provided [18–20].

P. Biosciences proposed an algorithm for calculation based on text analysis, provided for quality evaluation of the e-book. The algorithmic approach evaluates the quality features of an e-book, especially its readability, security, and comprehension. These qualitative aspects are explored by parsing the e-book test and retrieving various information such as the ideas explained in the e-book. Characteristic features are evaluated based on numerical values [21].

Zhang et al. developed a new analysis method, the latent Dirichlet allocation topic model, and used it to carry out

Table 1: Comparison of learning the data leakage in the secrecy policy of different social networks.

Name of the social network	Friends	Streaming	Like	Photo, video and other tag	Comment	Group member
Twitter	Sensitivity data leakage may occur	Sensitive data cannot be leaked	Sensitive data cannot be leaked	Not implemented	Not implemented	Sensitivity data may leak
Facebook	Sensitivity data leakage may occur	Sensitivity data may leak	Sensitivity data may leak	Not implemented	Sensitivity data may leak	Sensitivity data may leak
Google+	Sensitivity data leakage may occur	Sensitive data cannot be leaked	Sensitive data cannot be leaked	Sensitive data cannot be leaked	Sensitivity data may leak	Not implemented
Orkut	Sensitivity data always leaks	Sensitive data cannot be leaked	Sensitivity data may leak	Sensitivity data may leak	Sensitivity data may leak	Sensitivity data may leak
My space	Sensitivity data leakage may occur	Sensitive data cannot be leaked	Sensitivity data may leak	Sensitivity data may leak	Sensitivity data may leak	Not implemented

technology evaluation and road mapping analysis of the Blockchain field based on patent data. Their results describe the current state of technology development and predict future development trends. The new analysis method is based on technology life cycle theory, the latent Dirichlet allocation topic model, and text similarity calculation, enabling a more effective analysis of the status of R&D in the Blockchain field [22].

Alotaibi analyzed the data clustering problem. Symmetry can be considered a precare aspect that can improve shapes and objects, as well as reconstruction and recognition. Symmetric-based clustering methods look for symmetrical clusters with respect to their centers. Furthermore, the K-means algorithm is considered to be one of the most common clustering methods. It is easy to implement and can run quickly in most situations. However, it was initiated by sensitivity and it could easily get caught up in local targets. The Tabu Search algorithm is a standard universal optimization technique, while Adaptive Search Memory is a key component of TS [23].

An efficient model Heap Bucketization-Anonymous has been proposed to balance privacy and usability with multiple sensitivity properties. The Heap Packetization-Anonymous model used anatomy to divide the data vertically into (1) semi-identifier table and (2) sensitivity attribute table. Semi-identifier is anonymized by enabling Q-anonymity and Sensitivity attributes are anonymized by the use of slicing and hip bucketing [24].

Alotaibi et al.'s study provided the characteristics of the recommendations are analyzed and the suggestion mining extraction process for categorizing the recommended sentences from online customer reviews is provided. Classification using the word-embedding approach via the XG Boost classifier is used. The two datasets used in this experiment are related to online hotel reviews andMicrosoft Windows App Studio discussion reviews. F1, accuracy, recall, and accuracy scores are calculated [25].

Rajendran et al. focused on the design of a big data classification model using chaotic pigeon inspired optimization (CPIO)-based feature selection with an optimal deep belief network (DBN) model. This model is executed in the Hadoop MapReduce environment to manage big data. At first, the CPIO algorithm is applied to select a useful subset of features the Harris hawks optimization (HHO)-based DBN model that is derived as a classifier to allocate appropriate class labels. The design of the HHO algorithm to tune the hyper parameters of the DBN model assists in boosting the classification performance [26].

Alsufyani et al. presented an intelligent data management framework for a cyber-physical system (IDMF-CPS) with machine-learning methods. The training approach based on two enhanced training procedures, running concurrently to upgrade the processing and communication strategy and the predictive models, is contained in the suggested reasoning modules [27].

Li et al. contributed to the clustering of natural language and meet the need for dynamic updating of the knowledge system. Its method of extracting dynamic knowledge based on sentence clustering recognition using a neural network-based structure. The process of transitioning from natural language papers to semantic knowledge systems is examined considering the problems related to sentence vectorization. It examines sentence vectorization properties using various basic definitions, judgment theorems, and post-processing components [28, 29].

Table 1 shows the comparison of learning the data leakage in the secrecy policy of different social networks like twitter, Facebook, Google+, Orkut, and My space.

3. Methodology

With the fast development of Internet technology, huge amounts of data can be accessed from online servers. Websites are the major repository of data that often store information in text form. Advances in modern techniques of text analysis have led to the extraction of unstructured data from the online server reaching large data age [30]. To overcome this limit, many researches have been conducted in recent times to do text mining or text analysis from a large database.

Text mining is the technique of retrieving information from an unstructured database. It is linked to extract web content using hypertext from the given documents. To create the data from unstructured data, text was first converted to a manageable form of data. Text representation models such as N-gram and LSTMN are used to sort data into structures.

Text processing provides valuable insights into customer emails, survey comments, social media, call logs, posts, and other sources of text linked data of the method of text processing are homogeneous data processing, in which the text miner focuses on the text in its place of the structured data. Natural language processing (NLP) is a branch of artificial intelligence that enables computers to understand, interpret, and manipulate human language. NLP attracts from many disciplines, including computer science and computational linguistics, with the aim of filling the gap for human communication and computer understanding.

Old-fashioned NLP models were created using rule-based or statistical methods. Studies on the intensity of data analysis in dealing with low-supervised in deep learning models and huge data sets. Neural networks are also used in huge data sets, whose process data are set up functionally. Figure 1 shows the block diagram of mined text data from Facebook accounts.

The text analysis process on Facebook data begins with the extraction of text data. Text analysis is used to control the flow of shared data to users based on the relationship between shared content and the user account received. It extracts text data from the Facebook account to identify the compatibility relationship between the shared post and the user profile and post sharing history. The extracted text data is then analyzed by Latent Dirichlet Allocation (LDA), N-gram algorithms. The group or clusters designed by this allocation mechanism are processed by the Long-Term Memory Network (LSDM) and it sends the shared post to a specific Facebook user to make a decision from the collected database.

3.1. Generative Models for Textlatent Dirichlet Allocation (Lda) Model. In Latent Dirichlet Allocation, K Latent topics correspond to the texts. In vocabulary, every topic is described as a multinomial distribution over the/v/words. The texts are created by sampling a mixture of these latent topics and then sampling words from that sampling mixture [31, 32].

More exactly, the N words texts $w = \langle w_1, \ldots, w_N \rangle$ are created by the following procedure. Initially, from a Dirichlet distribution, it is sampled. It is sampledmeans it lies in the $(k-1)$-dimensional simplex: θ_{t} tis greater than zero, $\sum_I$ is equal to zero. Then, for each N words, a topic $z_n \varepsilon \{1, \ldots, k\}$ is sampled from Mult (distribution $p(z_n = i \backslash$ At last, in conditioned z_nth topic, each word w_n is sampled from the multinomial distribution $p(w/z_n)$. It is represented as a degree for which the topic is mentioned in the texts. The text probability is defined in the following equation:

$$p(w) = \int \left(\prod_{n=1}^{N} \sum_{zn=1}^{k} p(w_n | z_n; \beta) p(z_n | \theta) p(\theta; \alpha) \right) d\theta, \quad (1)$$

where p is represented as Dirichlet, (z_n/is represented as multinomial parameterized by and p(w_n/z_{n}) is denoted as multinomial over the words. The Latent Dirichlet Allocation (LDA) model is parameterized by the $k \times /v/$matrix, k-dimensional Dirichlet parameters, and these parameters control the k multinomial distributions over words.

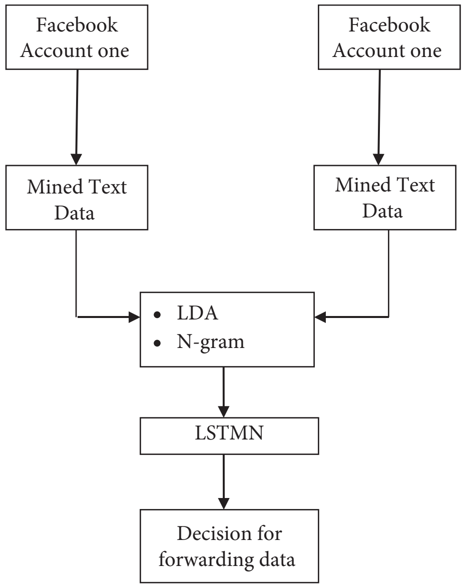


Figure 1: Block Diagram of mined text data from Facebook accounts.

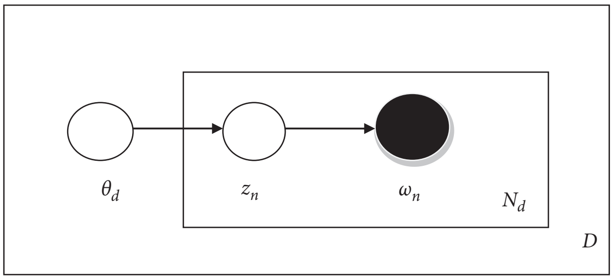


Figure 2: LDA Graphical model representation.

The latent dirichlet allocation models are not as easy as Dirichlet-multinomial clustering model. In LDA model, the innermost plate comprises only w_n as shown in Figure 2. The latent topic node can be sampled only once for every text; and Dirichlet can be sampled only once for the entire collection. In latent dirichlet allocation model, the Dirichlet is sampled for every text, and within the text, the multinomial latent topic node is sampled repeatedly. Compared to prior distribution, the dirichlet is a component in the probability model over the model parameters [33].

The boxes are plates described as replicates. The outer plate described as texts, when the inner plate describes the repeated choice of words and topics within a text.

Equation (1) is a second interpretation of latent dirichlet allocation. From unigram/multinomial model.

$p(w|\theta) = \sum_{z=1}^{k} p(w|z) p(z|\theta)$ Where the unigram models p (the latent dirichlet allocation is the mixture model, and $p()$ provides the mixture weights.

3.2. Related Models. The mixture of unigrams model performed with each text is created by a single arbitrary chosen topic:

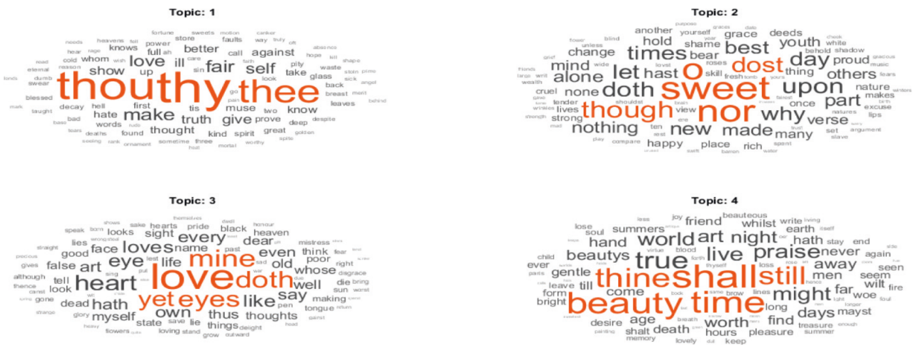


Figure 3: LDA model representation of words used repeatedly in post.

$$p(w) = \sum_{z=1}^{k}\left(\prod_{n=1}^{N}\right)p(w_n|z)p(z). \tag{2}$$

From several topics, the LDA model permits several texts, but sometimes fails to capture the possibility that the text shows several topics. Latent Dirichlet Allocation captures this possibility, which can increase the parameter count otherwise for multinomial p (z), having k-1 parameters over the k topics [34, 35], where k is the free parameter for Dirichlet. Another model is Hofmann's Probabilistic Latent Semantic Indexing (PLSI), which is presented for word 'w' and text label "d." These "d"' and "w" are conditionally independent in the following equation.

$$p(d,w) = \sum_{z=1}^{k} p(w|z)p(z|d)p(d). \tag{3}$$

3.3. Inference and Learning. The inference for the learning problems for Latent dirichlet allocation is described by examining the likelihood contribution created through a single text. Let $z_n^i = 1$ iff z_n is the ith topic and $wnj = 1$, if w_n is the jth word. Let it present the $p(w^3 = 1|z^i = 1), an\ dw = (w_1, \ldots\ldots, wN), z = (z_{1,\ldots\ldots,Z_N})$ modifying the (1), are given by

$$p\left(w;\alpha,\beta+ = \frac{\prod(\sum_i \alpha_i)}{\prod_i \prod(\alpha_i)}\int_\theta^i\left(\prod_{i=1}^{k}\theta_i^{\alpha_{i-1}}\right)\left(\prod_{n=1}^{N}\sum_{i=1}^{k}\prod_{j=1}^{|v|}(\theta_i\beta_{i,j})^{w_n^5}\right)d\theta\right. . \tag{4}$$

The fourth equation is denoted as hypergeometric function that is unfeasible to calculate accurately. For large text collections, the learning algorithms and fast inference are required and have to be used as several techniques to estimate the likelihood in equation (4). Various approximations have to be used for the log-likelihood.

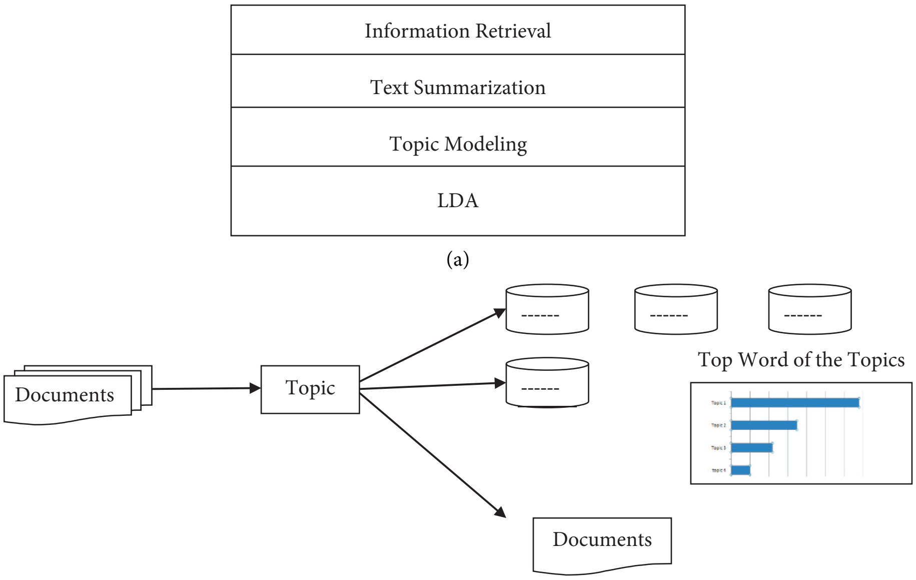
$$\log p(w;\alpha,\beta) = \log\int_\theta^\alpha \sum_z p(w|z;\beta)p(z|\theta)p(\theta;\alpha)$$
$$\cdot\frac{q(\theta,z;\gamma,\varphi)}{q(\theta,z;\gamma,\varphi)}d\theta \geq E_q \log p(w|z;\beta) + \log p(z|\theta)$$
$$+\log p(\theta;\alpha) - \log q(\theta,z;\gamma,\varnothing)]. \tag{5}$$

3.4. LDA Linear Discriminant Analysis of Text in Social Media. Figure 3 shows LDA analysis for 4posts entered by four different users on social media. The topics are as follows: "Truth and loyalty," "Describe a person characteristics," "why I love myself?" and "The Night life of a Man." In the first topic the words "tho," "thy" and "thee" were used repeatedly to emphasize on a subject. The other words which were used to describe the subject were "love," "fair," "self," "make," and "truth."

The second topic describes the person with words such as "sweet," "nor," "dost," which deals with other person and his attributes. The third topic shows the words used to describe oneself. The words such as "love," "mine," "doth," and "eyes" were used repeatedly. Similarly, the words majorly used to describe "The Night life of a Man" includes "thine," "shall," "still," "beauty," and "time." The words which describe person on second scale includes "live," "praise," "age," "hours" and "death." The LDA model provides a comprehensive words used in a large volume of data. In addition, the most used words are highlighted and the corresponding secondary words are shown with smaller fonts. However, the LDA model performs well when the database is small.

4. Result and Discussion

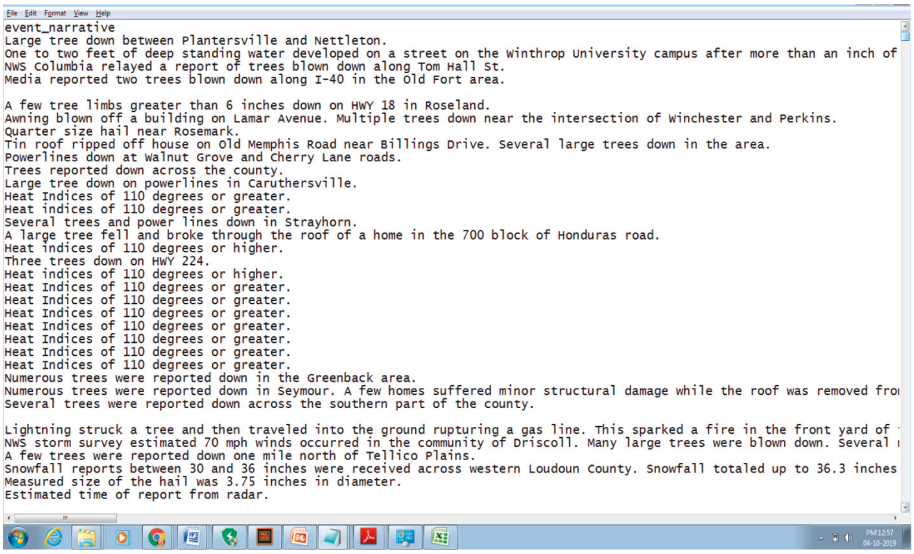
In the present period, information is an important asset of an individual, depending on the condition in which the individual lives, and the information that is shared with others in the same way. When a person shares his information on a



(a)

(b)

Figure 4: (a) Mapping for data analytics based on LDA. (b) Describe an implicit analysis on LDA.



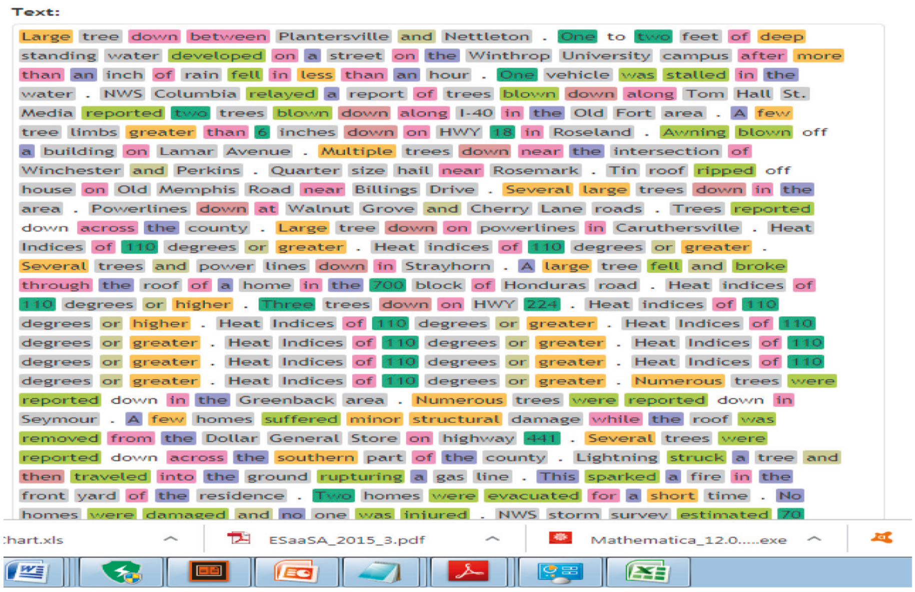
```
event_narrative
Large tree down between Plantersville and Nettleton.
One to two feet of deep standing water developed on a street on the Winthrop University campus after more than an inch of
NWS Columbia relayed a report of trees blown down along Tom Hall St.
Media reported two trees blown down along I-40 in the Old Fort area.

A few tree limbs greater than 6 inches down on HWY 18 in Roseland.
Awning blown off a building on Lamar Avenue. Multiple trees down near the intersection of Winchester and Perkins.
Quarter size hail near Rosemark.
Tin roof ripped off house on Old Memphis Road near Billings Drive. Several large trees down in the area.
Powerlines down at Walnut Grove and Cherry Lane roads.
Trees reported down across the county.
Large tree down on powerlines in Caruthersville.
Heat Indices of 110 degrees or greater.
Heat indices of 110 degrees or greater.
Several trees and power lines down in Strayhorn.
A large tree fell and broke through the roof of a home in the 700 block of Honduras road.
Heat indices of 110 degrees or higher.
Three trees down on HWY 224.
Heat indices of 110 degrees or higher.
Heat Indices of 110 degrees or greater.
Heat Indices of 110 degrees or greater.
Heat Indices of 110 degrees or greater.
Heat Indices of 110 degrees or greater.
Heat Indices of 110 degrees or greater.
Heat Indices of 110 degrees or greater.
Heat Indices of 110 degrees or greater.
Numerous trees were reported down in the Greenback area.
Numerous trees were reported down in Seymour. A few homes suffered minor structural damage while the roof was removed fro
Several trees were reported down across the southern part of the county.

Lightning struck a tree and then traveled into the ground rupturing a gas line. This sparked a fire in the front yard of
NWS storm survey estimated 70 mph winds occurred in the community of Driscoll. Many large trees were blown down. Several
A few trees were reported down one mile north of Tellico Plains.
Snowfall reports between 30 and 36 inches were received across western Loudoun County. Snowfall totaled up to 36.3 inches
Measured size of the hail was 3.75 inches in diameter.
Estimated time of report from radar.
```

Figure 5: A collection of information that is most shared on the social website.

social website, his followers also benefit from the information, which is an important issue on the social website. Facebook and Twitter are currently sharing more information about two of the most admired social areas in the world. The code that is used to convey differing information on that social website is a picture, video, audio, animation, information, program, and emotion that can be written in a variety of languages. All types of information are shared based on 10 types of ideas such as Economy, National, Sports, Arts, Politics, Opinion, Local News, Technology, Lifestyle, Entertainment, Social Learning, Team, Travel, and Custom. Through the website, various social websites create a variety of groups that narrow. Bay and Sell, Close Friends, Club, Event and Plans, Family, Video Game, Neighbors, Parents, Project, School or College, Study. They exchange information between these groups and different members of the group, and they are a way of providing a wide variety of information.

4.1. Data Analytics Based on LDA.

Then the information they share is of importance. For example, if there are 2 lakh members in a group, the information shared between them is positive, and as a result, they create an appearance of good, and a negative one in the same group. When information is shared, it creates a social evil. This will provide a good result for analyzing information and whether they should be analyzed or if any of them are harmful or should be shared. Similarly, if a third of people on the website break their account and share a false message with others, they will have



Text:

Large tree down between Plantersville and Nettleton . One to two feet of deep
standing water developed on a street on the Winthrop University campus after more
than an inch of rain fell in less than an hour . One vehicle was stalled in the
water . NWS Columbia relayed a report of trees blown down along Tom Hall St.
Media reported two trees blown down along I-40 in the Old Fort area . A few
tree limbs greater than 6 inches down on HWY 18 in Roseland . Awning blown off
a building on Lamar Avenue . Multiple trees down near the intersection of
Winchester and Perkins . Quarter size hail near Rosemark . Tin roof ripped off
house on Old Memphis Road near Billings Drive . Several large trees down in the
area . Powerlines down at Walnut Grove and Cherry Lane roads . Trees reported
down across the county . Large tree down on powerlines in Caruthersville . Heat
Indices of 110 degrees or greater . Heat indices of 110 degrees or greater .
Several trees and power lines down in Strayhorn . A large tree fell and broke
through the roof of a home in the 700 block of Honduras road . Heat indices of
110 degrees or higher . Three trees down on HWY 224 . Heat indices of 110
degrees or higher . Heat Indices of 110 degrees or greater . Heat Indices of 110
degrees or greater . Heat Indices of 110 degrees or greater . Heat Indices of 110
degrees or greater . Heat Indices of 110 degrees or greater . Heat Indices of 110
degrees or greater . Heat Indices of 110 degrees or greater . Numerous trees were
reported down in the Greenback area . Numerous trees were reported down in
Seymour . A few homes suffered minor structural damage while the roof was
removed from the Dollar General Store on highway 441 . Several trees were
reported down across the southern part of the county . Lightning struck a tree and
then traveled into the ground rupturing a gas line . This sparked a fire in the
front yard of the residence . Two homes were evacuated for a short time . No
homes were damaged and no one was injured . NWS storm survey estimated 70

Figure 6: Analyze a collection of shared posted data about the weather on the social website.

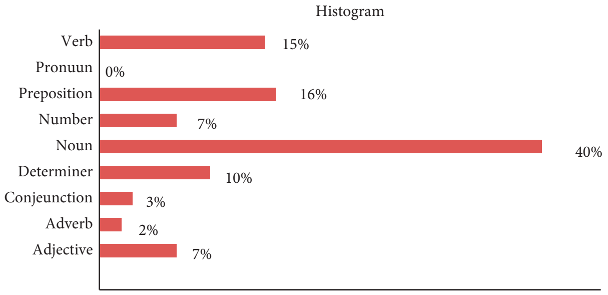
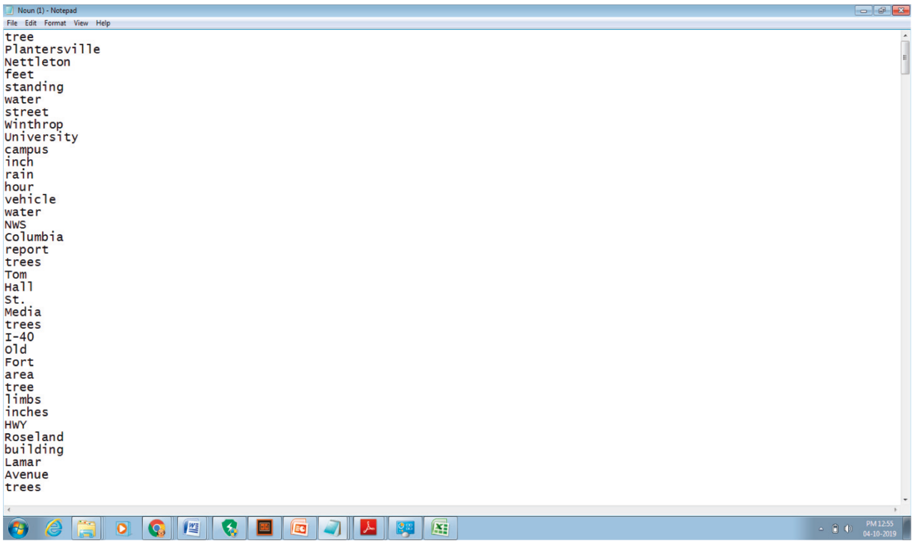
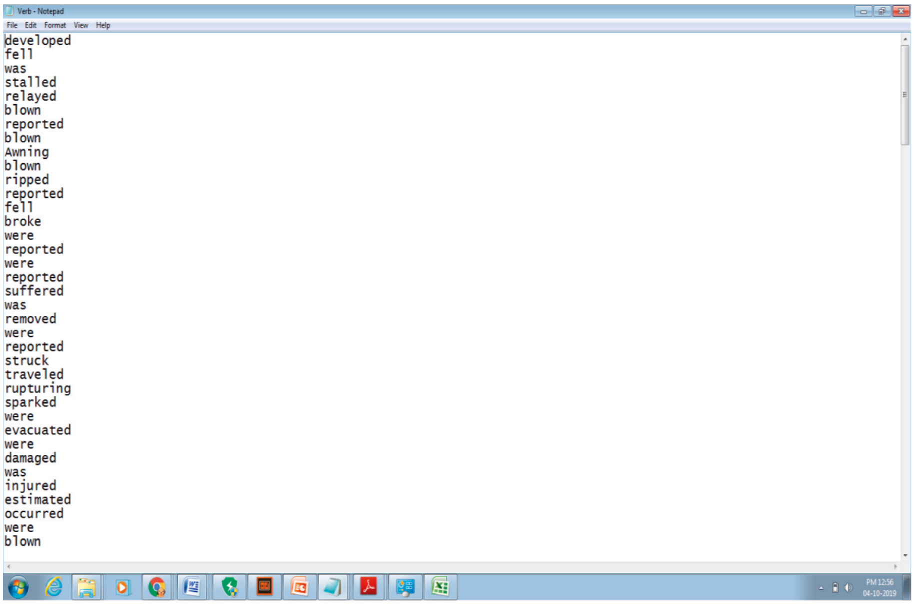


Figure 7: A Histogram based on the data analyzed in parts of the shared sentence.



```
tree
Plantersville
Nettleton
feet
standing
water
street
Winthrop
University
campus
inch
rain
hour
vehicle
water
NWS
Columbia
report
trees
Tom
Hall
St.
Media
trees
I-40
Old
Fort
area
tree
limbs
inches
HWY
Roseland
building
Lamar
Avenue
trees
```

Figure 8: Nouns obtained from the analyzed data.



```
developed
fell
was
stalled
relayed
blown
reported
blown
Awning
blown
ripped
reported
fell
broke
were
reported
were
reported
suffered
was
removed
were
reported
struck
traveled
rupturing
sparked
were
evacuated
were
damaged
was
injured
estimated
occurred
were
blown
```

Figure 9: Verbs obtained from the analyzed data.

Table 2: Number of words/character involved in text feeds on social websites.

Symbols involved in text feeds	NO. Of words/character involved in text feeds
Words	1290
Characters (including spaces)	7265
Characters (without spaces)	5965

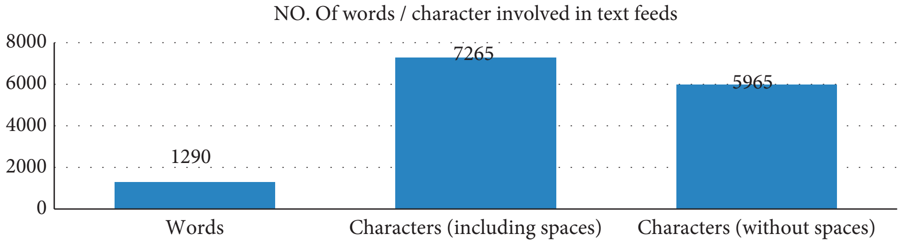


Figure 10: Number of words statistics involved in text feeds on social websites.

Table 3: Analysis of data according to English grammar based on LDA.

Word statistics	Word count	Cumulative	Percentage of cumulative
Syllables	2045	2045	51.58
Sentences	126	2171	3.18
Unique words	371	2542	9.36
Average word length (char)	4.6	2546.6	0.12
Average sentence length (word)	10.2	2556.8	0.26
Monosyllabic words (1 syllable)	749	3305.8	18.83
Polysyllabic words (≥3 syllables)	179	3484.8	4.52
Syllables per word	1.6	3486.4	0.04
Paragraph	79	3565.4	1.99
Difficult Words	399	3964.4	10.07

Table 4: Finding the most used word in a document.

Single keyword words	Frequency	Percentage of single words (%)
Down	35	2.7
Reported	26	2
Trees	24	1.9
Tree	19	1.5
Across	18	1.4
County	17	1.3
Blown	16	1.2
mph	16	1.2
Wind	16	1.2
Inches	14	1.1

Table 5: Finding the most used double words in a document.

Double keyword words	Frequency	Percentage of double words (%)
(1) Trees were	14	1.1
(2) Were reported	14	1.1
(3) Blown down	13	1
(4) Reported down	13	1
(5) Down in	11	0.9
(6) Heat indices	11	0.9
(7) Indices of	11	0.9
(8) Of 110	11	0.9
(9) 110 degrees	11	0.9
(10) Degrees or	11	0.9

Table 6: Finding the most used triple words in a document.

Triple keyword words	Frequency	Percentage of triple keywords (%)
(1) Heat indices of	11	0.9
(2) Indices of 110	11	0.9
(3) Of 110 degrees	11	0.9
(4) 110 degrees or	11	0.9
(5) Trees were reported	10	0.8
(6) Were reported down	10	0.8
(7) Degrees or greater	9	0.7
(8) Or greater heat	7	0.5
(9) Greater heat indices	7	0.5
(10) Wind gusts of	7	0.5

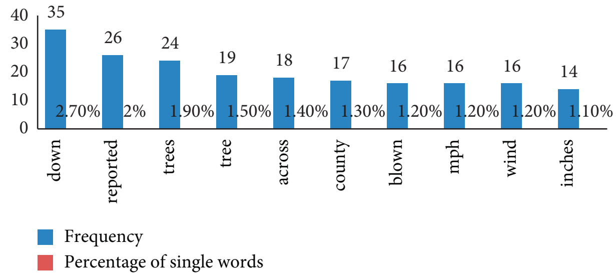


Figure 11: Percentage of the most used word in a document.

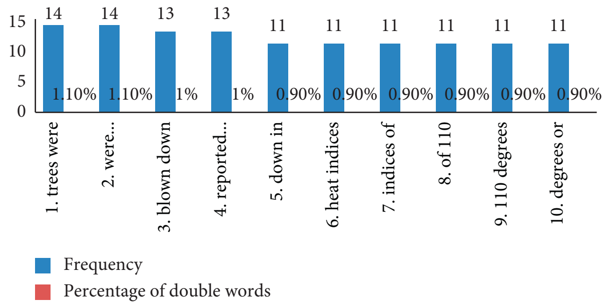


Figure 12: Percentages of most used double words in a document.

a huge impact on the community. This has resulted in the social website being damaged and an economic loss to the community, which we have proposed to avoid. We consider a frame work to be a good solution.

Figure 4(a) illiterates the mapping for data analytics based on LDA algorithm. The first layer is information retrieval this layer retrieve the data from social media. The second layer is text summarization, and it is summarizes the

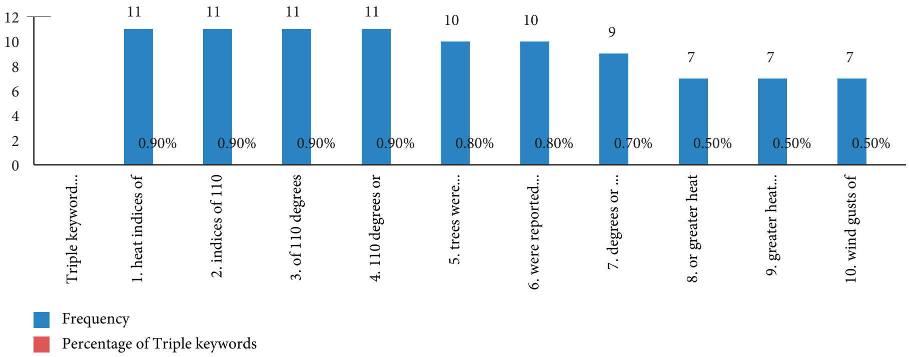


Figure 13: Percentages of most used triple words in a document.

text based on predefined rule, the next thread layer is topic modeling, and the fourth one is LDA algorithms.

Information obtained from social websites is stored in a dataset as a weather dataset, which is used to easily analyze information and extract it from words. We break down the words into different groups and explain how many times each word has been used in those groups with an analysis and description of maximum usage which shown in Figure 4(b).

4.2. Data Analyzed In Parts of the Shared Sentence. By using this method, I have proposed to LDA analyze framework shared information through others; the words in the information are grouped based on the English grammar and in order to understand the simple. In this, we have shown in percentage that the number of words in each sentence is used as nouns or verbs, following the English grammar. This nouns and verbs are a source of the wrong way to convey an opinion or to inform others in the right way. When a user is expressing his opinion to others, some of them have evil intent and are analyzed before he can inform others which are shown in Figures 5–7. Additionally, the information is represented in Figures 8 and 9.

Figure 8 shows that nouns are obtained from the given data set. The most used nouns in the top of the figure and the least used noun in the bottom of the figure.

Figure 9 shows that verbs are obtained from the given data set.The most used verbs in top of the figure and least used verbs in bottom of the figure.

4.3. Number of Words and Character Count. Table 2 shows the number of words or characters involved in text feeds on social websites and count symbols involved in text feeds, and Figure 10 is a graphical representation of the number of words or character involved in text feeds.

Table 3demonstrates an analysis of data according to English grammar based on LDA algorithms, Table 4 displays the most used single word in a document, Table 5 shows the most used double words in a document, and Table 6 shows the most used triple words in a document.

4.4. Most Used Single, Double, and Triple Word. From Figure 11, we can note that the word "down" is highly used whereas the word "'inches"is the least used in the given data set.

From Figure 12, we can note that the words "trees were" is highly used double words whereas the words "degrees or" are the least used in the given data set.

From Figure 13, we can note that the words "heat indices of" is highly used triple words whereas the words "greater heat indices" are the least used in the given data set.

The model works perfectly for the given data set and is able to identify parts of speech like nouns, verbs, etc, and also count the number of words, characters, and character spaces.

5. Conclusion

Text mining is achieved by two methods that are linguistic rules and machine learning systems. Linguistic rules operate based on a rule-based pattern-matching model based on simple Boolean keywords. It is also performed by creating a complex model developed by field experts. Linguistic rules are used for quick analysis. The machine learning approach uses forms of text database. Statistical methods are used to compare documents with each other and generate the most important text information from a large text corpus or database. ML approaches range from simple to complex processes that collect valuable information and unique forms from a provided database. The dataset collected from social media such as Facebook was used to analyze the text to avoid anonymous users unnecessarily sharing the public post. The N-Gram LSDM and LDA algorithms are used for mining the information and generating patterns that compare the connection between the post and the user account. Finally, the LDA algorithm works properly, and we conclude that it is suitable for text mining. It is limited to the English language and it has not been tried with others languages. In the future, we plan to implement this algorithm in other languages and use the Latent Semantic Analysis (LSA) algorithm and make comparisons between LDA and LSA.

Data Availability

Data are available based on request

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] N. Ahmad and J. Siddique, "Personality assessment using twitter tweets," *Procedia Computer Science*, vol. 112, pp. 1964–1973, 2017.

[2] E. H. Ko and D. Klabjan, "Semantic properties of customer sentiment in tweets," *Proc. - 2014 IEEE 28th Int. Conf. Adv. Inf. Netw. Appl. Work. IEEE WAINA*, vol. 4, no. 5, pp. 657–663, 2014.

[3] H. Ming, C. Rohrdantz, H. Janetzko, U. Dayal, and A. Daniel, "Keim, lars-erik haug and mei-chun hsu, "visual sentiment analysis on twitter data streams," in *Proceedings of the VAST 2011 IEEE Conference on Visual Analytics Science and Technology (VAST)*, vol. 3, no. 2, pp. 277-278, Providence, RI, USA, October 2011.

[4] R. C. Basole, C. D. Seuss, and W. B. Rouse, "IT innovation adoption by enterprises: knowledge discovery through text analytics," *Decision Support Systems*, vol. 54, no. 2, pp. 1044–1054, 2013.

[5] A. R. Patil and A. Manjrekar, "An innovative approach to classify and retrieve text documents using feature extraction and Hierarchical clustering based on ontology," in *Proceedings of the Int. Conf. Comput. Anal. Secur. Trends*, vol. 3, no. 2, pp. 371–376, Pune, India, December 2016.

[6] D. Park, S. Kim, J. Lee, J. Choo, N. Diakopoulos, and N. Elmqvist, "ConceptVector: text visual analytics via interactive lexicon building using word embedding," *IEEE Transactions on Visualization and Computer Graphics*, vol. 24, no. 1, pp. 361–370, 2018.

[7] R. M. Packiam and V. S. J. Prakash, "An empirical study on text analytics in big data," in *Proceedings of the 2015 IEEE Int. Conf. Comput. Intell. Comput. Res. ICCIC*, vol. 3, no. 4, pp. 1–4, Madurai, India, December 2015.

[8] N. Medoc, M. Stefas, M. Ghoniem, and M. Nadif, "Visual analytics of text streams through multiple dynamic frequency matrices," in *Proceedings of the IEEE Conf. Vis. Anal. Sci. Technol. VAST 2014 - Proc*, vol. 3, no. 4, pp. 381-382, Paris, France, October 2014.

[9] L. Bradel, C. North, L. House, and S. Leman, "Multi-model semantic interaction for text analytics," in *Proceedings of the IEEE Conf. Vis. Anal. Sci. Technol. VAST 2014 - Proc*, vol. 4, no. 3, pp. 163–172, Paris, France, October 2014.

[10] K. Yadav, N. Kumar, P. K. R. Maddikunta, and T. R. Gadekallu, "A comprehensive survey on aspect-based sentiment analysis," *International Journal of Engineering Systems Modelling and Simulation*, vol. 12, no. 4, p. 279, 2021.

[11] M. Ahmed, "Elmogy , usman tariq , atef ibrahim and ammar mohammed, "fake reviews detection using supervised machine learning"," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. No. 1, 2021.

[12] P. P, G. G, G. Srivastava, P. K. R. Maddikunta, and T. R. Gadekallu, "A two-stage text feature selection algorithm for improving text classification," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 20, no. 3, pp. 1–19, 20Issue 3May 2021.

[13] T. R. Gadekallu, M. Alazab, R. Kaluri et al., "Hand gesture classification using a novel CNN-crow search algorithm," *Complex & Intelligent Systems*, vol. 7, pp. 1855–1868, 2021.

[14] H. S. Chiranjeevi, K. ManjulaShenoy, S. Prabhu, and S. Sundhar, "DSSM with text hashing technique for text document retrieval in next-generation search engine for big data and data analytics," in *Proceedings of the 2nd IEEE Int. Conf. Eng. Technol. ICETECH*, vol. 3, no. 5, pp. 395–399, Coimbatore, India, March 2016.

[15] J. Park, "Integrated visual analytics tool for heterogeneous text data," *IEEE Conf. Vis. Anal. Sci. Technol. VAST 2014 - Proc*, vol. 3, no. 2, pp. 325-326, 2014.

[16] H. Vandierendonck, K. Murphy, M. Arif, and D. S. Nikolopoulos, "Hpta: high-performance text analytics," in *Proceedings of the IEEE Int. Conf. Big Data*, vol. 4, no. 3, pp. 416–423, Washington, DC, USA, December 2016.

[17] F. Heimerl, S. Lohmann, S. Lange, and T. Ertl, "Word cloud explorer: text analytics based on word clouds," in *Proceedings of the Annu. Hawaii Int. Conf. Syst. Sci*, vol. 4, no. 3, pp. 1833–1842, Waikoloa, HI, USA, January 2014.

[18] R. Singh, R. Bagla, and H. Kaur, "Text analytics of web posts' comments using sentiment analysis," in *Proceedings of the Int. Conf. Work. Comput. Commun*, vol. 4, no. 3, pp. 1–5, Vancouver, BC, Canada, October 2015.

[19] S. Ramakrishnan and S. Ramakrishnan, "HD video transmission on UWB networks using H.265 encoder and ANFIS rate controller," *Cluster Computing*, vol. 21, no. 1, pp. 251–263, 2018.

[20] S. Gnanavel and S. Ramakrishnan, "N Mohankumar "Wireless video transmission over UWB channel using fuzzy based rate control technique"," *Journal of Theoretical and Applied Information Technology 28th February*, vol. 60, no. 3, pp. 491–503, 2014.

[21] S. Khurana, M. Relan, and V. K. Singh, "A text analytics-based approach to compute coverage, readability and comprehensibility of eBooks," *International Conference on Contemporary Computing (IC3)*, vol. 1, pp. 1–9, 2013 Sixth.

[22] H. Zhang, T. Daim, and Y. P. Zhang, "Integrating patent analysis into technology roadmapping: a latent dirichlet allocation based technology assessment and roadmapping in the field of Blockchain," *Technological Forecasting and Social Change*, vol. 167, Article ID 120729, 2021.

[23] Y. Alotaibi, "A new meta-heuristics data clustering algorithm based on tabu search and adaptive search memory," *Symmetry*, vol. 14, no. 3, p. 623, 2022.

[24] J. Jayapradha, M. Prakash, Y. Alotaibi, O. I. Alghamdi, and S. A. Alghamdi, "Heap bucketization anonymity-an efficient privacy-preserving data publishing model for multiple sensitive attributes," *IEEE Access*, vol. 10, pp. 28773–28791, 2022.

[25] Y. Alotaibi, M. Noman MalikNoman Malik, H. Hayat KhanHayat Khan et al., "Suggestion mining from opinionated text of big social media data," *Computers, Materials & Continua*, vol. 68, no. 3, pp. 3323–3338, 2021.

[26] S. Rajendran, O. I. Khalaf, Y. Alotaibi, and S. Alghamdi, "MapReduce-based big data classification model using feature subset selection and hyperparameter tuned deep belief network," *Scientific Reports*, vol. 11, no. 1, p. 24138, 2021.

[27] A. Alsufyani, Y. Alotaibi, A. O. Almagrabi, S. A. Alghamdi, and N. Alsufyani, "Optimized intelligent data management framework for a cyber-physical system for computational applications," *Complex & Intelligent Systems*, vol. 1-13, 2021.

[28] G. Li, F. Liu, A. Sharma et al., "Research on the natural language recognition method based on cluster analysis using

neural network," *Mathematical Problems in Engineering*, vol. 2021, pp. 1–13, 2021.

[29] K. Kaluri, R. Gundluru, N. Alzamil, and Z. S. Rajput, "review on deep learning techniques for IoT data," *Article in Electronics*, vol. 11, no. 10, 2021.

[30] G. T. Reddy, M. P. K. Reddy, K. Lakshmanna et al., "Analysis of dimensionality reduction techniques on big data," *IEEE Access*, vol. 8, pp. 54776–54788, 2020.

[31] R. Nallapati and W. W. Cohen, "Link-PLSA-LDA: a new unsupervised model for topics and influence of blogs," *Second International AAAI Conference on Weblogs and Social Media*, vol. 2, no. 1, 2008.

[32] Y. Lu, Q. Mei, and C. Zhai, "Investigating task performance of probabilistic topic models: an empirical study of PLSA and LDA," *Information Retrieval*, vol. 14, no. 2, pp. 178–203, 2011.

[33] V. Rajyalakshmi and K. Lakshmanna, "A review on smart city - IoT and deep learning algorithms, challenges," *International Journal of Engineering Systems Modelling and Simulation*, vol. 13, no. 1, Article ID 122733, 3 pages, 2022.

[34] N. Gundluru, D. S. Rajput, K. Lakshmanna et al., "Enhancement of detection of diabetic retinopathy using Harris hawks optimization with deep learning model," *Computational Intelligence and Neuroscience*, vol. 2022, no. 24, pp. 1–13, 2022.

[35] K. Lakshmanna, R. Kavitha, B. T. Geetha, A. K. Nanda, A. Radhakrishnan, and R. Kohar, "Deep Learning-Based Privacy-Preserving Data Transmission Scheme for Clustered IIoT Environment," *Nature-Inspired Computing for Web Intelligence*, vol. 2022, no. 6, pp. 1–11, 2022.