SEBHA UNIVERSITY JOURNAL OF PURE & APPLIED SCIENCES VOL.25 NO. 2 2026
DOI: 10.51984/MWSG6Z34

مجلة جامعة سبها للعلوم البحتة والتطبيقية
Sebha University Journal of Pure & Applied Sciences
Journal homepage: www.sebhau.edu.ly/journal/jopas

A Novel Integration of Exploratory Data Analysis and LDA Text Mining for Deeper Understanding of Libyan-Related Tweets

Yousef Salem[a*], Mansour Essgaer[a], Yahyia Ben Yahmed[b], Rokaya Abdulhamid[c]

[a]Computer Science Department, Faculty of Information Technology, Sebha University, Sebha, Libya.
[b]Department of Computer Science, College of Education, Traghen, University of Fezzan, Murzuq, Libya.
[c]Information System Department, Faculty of Information Technology, Sebha University, Sebha, Libya.

Keywords:
Latent Dirichlet Allocation
Clustering.
Text Mining.
Topic Modeling.
Twitter.

ABSTRACT

Identifying topics across digital sources such as scientific articles, web publications, and social media platforms is essential for summarising large text datasets and understanding emerging trends. Social networking platforms, in particular, generate vast amounts of data, creating an urgent need for topic extraction to uncover underlying themes. This study aims to extract topics from Arabic-language tweets posted by Libyan users on Twitter using a structured multi-stage approach. The study follows a four-stage methodology: (1) Data Collection, involving the construction of a Twitter corpus based on Libyan dialect-specific keywords; (2) Exploratory Analysis, which employs statistical analysis to identify key patterns in the data; (3) Preprocessing, which applies natural language processing techniques to prepare the dataset; and (4) Topic Modelling, where Latent Dirichlet Allocation (LDA) is applied to extract dominant topics. The exploratory analysis revealed behavioural patterns such as a high frequency of positive emoji usage, peak Twitter activity at 9 p.m., and the dominance of Android as the operating system. The topic modelling results identified major themes including Friendship, Travel, Political Issues, and Education. These findings provide insights into the interests and communication patterns of Libyan Twitter users and demonstrate the effectiveness of topic modelling in capturing culturally and regionally relevant discourse in social media data.

دمج مبتكر لتحليل بيانات استكشافي واستخراج المواضيع باستخدام خوارزمية ديريشلي الكامنة لفهم أعمق للتغريدات المتعلقة بليبيا

يوسف سالم[a*]، منصور الصغير[a]، يحي بن يحمد[b]، رقية عبدالحميد[c]

[a]قسم علوم الحاسب، كلية تقنية المعلومات، جامعة سبها، سبها، ليبيا.
[b]قسم علوم الحاسب، كلية التربية، تراغن، جامعة فزان، مرزق، ليبيا.
[c]قسم نظم المعلومات، كلية تقنية المعلومات، جامعة سبها، سبها، ليبيا.

الكلمات المفتاحية:
التجميع باستخدام تخصيص ديريشلي الكامن.
استخراج النصوص.
نمذجة المواضيع.
تويتر.

الملخص

إن تحديد المواضيع عبر مختلف المصادر الرقمية، مثل المقالات العلمية والمنشورات على شبكة الإنترنت ومنصات وسائل التواصل الاجتماعي، أمر ضروري لتلخيص مجموعات البيانات النصية الكبيرة بكفاءة والتنبؤ بالاتجاهات المستقبلية. وتولد مواقع التواصل الاجتماعي، على وجه الخصوص، كميات هائلة من البيانات، مما يخلق حاجة ملحة لاستخراج المواضيع للكشف عن الموضوعات الأساسية. تهدف هذه الدراسة إلى استخراج المواضيع باللغة العربية التي يناقشها المستخدمون الليبيون على تويتر، باستخدام نهج منظم متعدد المراحل لاكتشاف مجالات الاهتمام الرئيسية. تنقسم عملية البحث إلى أربع مراحل: (1) جمع البيانات، والذي يتضمن جمع مجموعة من تويتر باستخدام الكلمات الرئيسية الخاصة

*Corresponding author.
E-mail addresses: you.salem@sebhau.edu.ly, (M. Essgaer) man.essgaer@sebhau.edu.ly, (Y. Yahmed) yah.benyehmed@fezzanu.edu.ly, (R. Abdulhamid) rouk.mehimed@sebhau.edu.ly.
Article History : Received 17 September 25 - Received in revised form 18 June 26 - Accepted 23 June 26

باللهجة الليبية؛ (2) التحليل الاستكشافي، والذي يستفيد من النمذجة الإحصائية لتحديد الأنماط الرئيسية في البيانات؛ (3) المعالجة المسبقة، والتي تطبق تقنيات معالجة اللغة الطبيعية لإعداد البيانات لمزيد من التحليل؛ و(4) نمذجة المواضيع، والتي يتم فيها تطبيق تخصيص ديريتشليت الكامن لاستخراج المواضيع الأكثر انتشارًا. أسفرت المرحلة الاستكشافية عن نتائج قيّمة، مثل شيوع استخدام الرموز التعبيرية التي تُجسّد تعابير السعادة، ووقت ذروة نشاط تويتر الساعة التاسعة مساءً، وسيادة نظام أندرويد كنظام التشغيل المُفضّل. حدّد نموذج المواضيع عدة مواضيع رئيسية، منها الصداقة، والسفر، والقضايا السياسية، والتعليم. تُقدّم هذه النتائج رؤى قيّمة حول اهتمامات مستخدمي تويتر الليبيين وأنماط تواصلهم، وتُظهر إمكانات نمذجة المواضيع في رصد المواضيع ذات الصلة ثقافيًا وإقليميًا في بيانات وسائل التواصل الاجتماعي.

1. Introduction

The growing availability of digital text data, particularly from social networking platforms, has highlighted the importance of topic modelling for extracting meaningful insights. Platforms such as Twitter generate vast volumes of user-generated content daily, making efficient methods for summarisation and analysis essential. Topic modelling, a subset of machine learning, addresses this challenge by identifying latent themes within large text corpora. It has applications in diverse domains, including network security, public governance, financial analysis, and healthcare [1], [2].

Arabic presents particular challenges for natural language processing due to its linguistic complexity and strong dialectal variation. Spoken across 22 countries, Arabic includes diverse regional varieties such as Gulf, Levantine, and North African dialects, all of which significantly influence online communication. These variations create challenges for NLP systems, which must account for non-standardised linguistic structures to ensure accurate analysis [3], [4].

Topic modelling techniques such as Latent Dirichlet Allocation (LDA) have been widely adopted for discovering and clustering latent topics in text datasets [5], [6] . However, traditional approaches often neglect contextual information at the sentence level, focusing primarily on document-level keyword distributions. This limitation can reduce performance in dialect-rich and context-dependent languages such as Arabic, where meaning is often implicit and context-sensitive. In addition, real-world topic structures are often overlapping and non-exclusive, further complicating accurate topic extraction.

This study aims to extract and analyse topics discussed by Libyan Twitter users in Arabic using the LDA algorithm. The methodology involves constructing a Twitter corpus using Libyan dialect-specific keywords, preprocessing the dataset to improve quality, and conducting exploratory analysis to identify preliminary patterns. LDA is then applied to extract dominant themes, revealing user interests and communication behaviours.

The findings are expected to provide insight into cultural, social, and political discourse among Libyan Twitter users. By addressing the challenges of dialectal Arabic and topic ambiguity, this study contributes to the growing field of natural language processing and social media analytics, with applications in trend analysis, policy support, and digital communication strategy.

The remainder of this paper is organised as follows: Section 2 reviews related work on LDA. Section 3 describes the methodology. Section 4 presents experimental results and analysis. Section 5 concludes the paper and outlines future research directions.

2. Related Work

The proliferation of online social networks has resulted in a surplus of user-generated content, posing difficulties for individuals in extracting valuable information [7], [8], [9]. In response to this issue, machine learning and NLP algorithms, specifically focusing on topic modeling techniques, have emerged to scrutinize extensive quantities of social media data. Various methods of topic modeling, such as latent semantic analysis, LDA, non-negative matrix factorization, and BERTopic have emerged in literature [10], [11]. These methodologies exhibit potential in identifying crucial topics and deriving insightful conclusions from customer reviews and brief textual information [12].

a. Topic Discovery

Recent research emphasizes the increasing significance of Arabic text mining within the realm of social media analysis. Twitter has emerged as a pivotal data source for Arabic text mining studies, with Support Vector Machine (SVM) and Naïve Bayes (NB) classifiers standing out as the most commonly employed techniques [13]. Researchers have put forth frameworks that combine sentiment and subjective analysis to gauge user interest in specific topics, achieving notable success with an accuracy rate of 89% [14], [15]. However, conducting sentiment analysis on Arabic texts presents challenges stemming from limited resources and the intricacies of the language. To address these obstacles, studies have utilized machine learning methodologies like NB and Decision Tree algorithms for the classification of Arabic tweets, underscoring the necessity for enhanced Arabic text processing tools and lexicon development [16]. These endeavors play a crucial role in advancing Arabic text mining capabilities across diverse domains, encompassing NLP, sentiment analysis, and event detection. In the academic landscape, while the focus on Arabic text mining in social media platforms has gained traction, there remains a notable scarcity of scholarly literature dedicated to exploring Arabic content on these platforms [17]. This gap underscores the need for further research to delve deeper into the nuances of Arabic text analysis within the context of social media. [18] presented a study that aimed to assess the level of sports enthusiasm evident in Twitter posts through the utilization of a computational model for content evaluation. This research sheds light on the potential applications of text mining techniques in understanding user behaviour's and interests within the sports domain on social media platforms.

Similarly, [19] conducted an analytical investigation focusing on establishing theoretical foundations for text mining and sentiment analysis in the social media landscape. Their study not only introduces fundamental concepts integral to this field but also delves into the methodologies and strategies employed in text mining within social media contexts. By bridging these gaps in academic research and building upon existing studies, future endeavors in Arabic text mining can aim to enhance the efficacy and breadth of analysis in social media contexts, ultimately contributing to a more comprehensive understanding of user interactions, sentiments, and trends within Arabic content online. Recent research has highlighted the potential of social networking sites as valuable databases for measuring public opinion. [20] emphasized the importance of text mining, particularly opinion mining, in uncovering public sentiments and attitudes toward specific topics through the vast information available on the internet. This approach enables researchers to gauge public opinion more effectively.

Further exploring public sentiment, a study by [21] utilized machine learning techniques to analyze discussions surrounding the outbreak of the COVID-19 virus on Twitter. The research, which focused on tweets collected between March 1 and May 30, 2020, identified five primary topics of public concern: the healthcare environment, psychological and emotional support, the business economy, social change, and stress-related issues. This categorization underscores the multifaceted nature of public discourse during health crises. In the context of Arabic-language research, there exists a notable scarcity of studies addressing text mining and sentiment analysis. [18] attempted to fill this gap by assessing sports fanaticism in tweets. Their study employed a computer model for content analysis to evaluate public

opinions expressed on Twitter. Similarly, [22] provided a descriptive analytical framework for text mining and sentiment analysis in social media, discussing key concepts and methodologies pertinent to the field.

In addition to sentiment analysis, machine learning applications have been developed for identifying threatening content on social platforms. [23] designed a model using neural networks to classify comments on Instagram as either threatening or non-threatening. This model was trained on a manually categorized dataset, demonstrating the effectiveness of machine learning in content moderation. [24] focused on detecting violence-related discourse in the Jordanian dialect on Twitter and Facebook, further illustrating the versatility of machine learning technologies in analyzing social media content. [25] also contributed to this body of work by proposing a model to identify instances of bullying on social media platforms, while [26] addressed hate speech directed at political figures in the Arab world.

Moreover, [27] utilized a neural network algorithm to identify irony in Arabic social media content, highlighting the challenges of accurately interpreting nuanced language in digital communication. Additionally, [20] conducted a study using text mining techniques to uncover significant topics within Arabic tweets related to COVID-19. This research employed automated text clustering to identify and categorize prevalent themes effectively.

Finally, a study focusing on Twitter discourse from Saudi Arabia during the COVID-19 pandemic examined public concerns regarding government measures and social sustainability from February 1 to June 1, 2020 [28]. This work underscores the importance of contextualizing public sentiment within specific socio-political frameworks.

b. Document Clustering

Several methods for document clustering and topic discovery have been explored using three datasets from Reddit and Twitter [29]. The study employed four different feature representations derived from Term-Frequency Inverse-Document-Frequency (TF-IDF) matrices and word embeddings, combined with four clustering methods, including (LDA) for topic discovery. The findings indicated that clustering methods utilizing neural embedding feature representations yielded the best performance across the datasets, as measured by appropriate extrinsic evaluation metrics. Additionally, the research demonstrated a clustering approach that identified key topics through a top-words methodology, leveraging TF-IDF weights alongside embedding distance measures. [30] investigated the application of clustering techniques to develop a standardized sizing system for Sudanese army officers' uniforms. This research aimed to uncover valuable insights from a large dataset of tailoring information, utilizing the k-means algorithm—a widely recognized clustering method. The study compared the resulting clusters against German sizing standards to evaluate their effectiveness. In a comparative analysis of document clustering versus topic modelling, [31] identified two primary methods for clustering, which divides documents into subsets based on related content. Their experiments revealed a relationship between document groups and sample subjects within a large preliminary text set. Interestingly, while both methods produced coherent results independently, the description of their outcomes showed significant similarities, suggesting a surprising mutual reinforcement that enhances confidence in both approaches as effective tools for illustrating and defining document content.

According to [32] various clustering algorithms were applied to short texts from two health-related datasets, including tweets and emails. The study implemented multiple clustering algorithms using two distinct representations: TF-IDF and Doc2Vec. The researchers employed cluster health indicators to assess the performance of subject modeling and clustering effectiveness without relying on external information, comparing the results to known external class labels. They noted that model performance varied significantly based on the data source and hyperparameters, such as the number of subjects and training iterations. Error analysis using a Hamming loss measure revealed that the GSDMM algorithm achieved the lowest error rate across both datasets.

Recent studies have investigated the use of LDA and Bag-of-Words (BOW) techniques for analyzing Arabic text, particularly in the context of tweets. LDA has proven effective in extracting semantic topics from Arabic tweets for sentiment analysis, yielding better results than conventional approaches [33]. Additionally, LDA has been employed to classify topics related to sustainability in Arabic tweets, aligning them with the United Nations Sustainable Development Goals [34]. In a large-scale study focused on authorship verification of Arabic tweets, researchers utilized BOW and TF-IDF in conjunction with a Naive Bayes classifier within a Hadoop framework, achieving an accuracy rate of 61.6% [35]. Furthermore, another study introduced by [36], proposed a hybrid LDA-SVM approach for classifying Arabic text, leveraging topics generated by LDA as features instead of traditional word vectors. This method yielded impressive results, with a macro-averaged F1 score of 88.1% and a micro-averaged F1 score of 91.4%.

c. Latent Dirichlet Allocation Approaches

These scholarly works delve into methodologies for analyzing Twitter data utilizing Latent Dirichlet Allocation (LDA) and associated strategies [37]. [31] introduce MS-LDA, a model that integrates LDA with social connections and word similarity to extract hierarchical user interests for multidimensional scrutiny. [38] introduce a two-step process utilizing LDA and topic mapping to enhance tweet clustering. [39] employ LDA topic modeling in conjunction with sentiment analysis to efficiently extract insights from Twitter data. These methodologies aim to address the challenges posed by the brevity and informal nature of tweets. Through experiments, the researchers showcase the efficacy of their approaches, demonstrating enhancements in Twitter data analysis and topic extraction. The proposed techniques hold promise for personalized recommendations, opinion mining, and business intelligence applications [31], [37]. Collectively, these studies contribute to the progression of social media data analysis through LDA-based techniques.

Recent academic investigations have delved into topic exploration and analysis on Twitter using a variety of text mining methodologies. LDA has been utilized to efficiently pinpoint popular topics within xtensive tweet datasets [37]. The combination of sentiment analysis with LDA can offer insights into the emotions associated with these topics [40]. The TCharM approach integrates cluster analysis and association rule discovery to navigate tweets across content, time, and location dimensions, providing contextually aware trend analysis of topics (Xiao et al., 2017). Topic models have also found application in crisis-related tweets, aiding in the extraction of valuable information during emergencies. In the realm of health-related studies, the Ailment Topic Aspect Model (ATAM) has been utilized to identify health-related topics within Twitter data, showcasing correlations with surveillance data for conditions such as influenza and allergies (Paul &Dredze, 2014). These methodologies illustrate the potential of text mining and exploratory data analysis in revealing insightful findings from Twitter content [41].The following figure 1 shows how the LDA algorithm works.

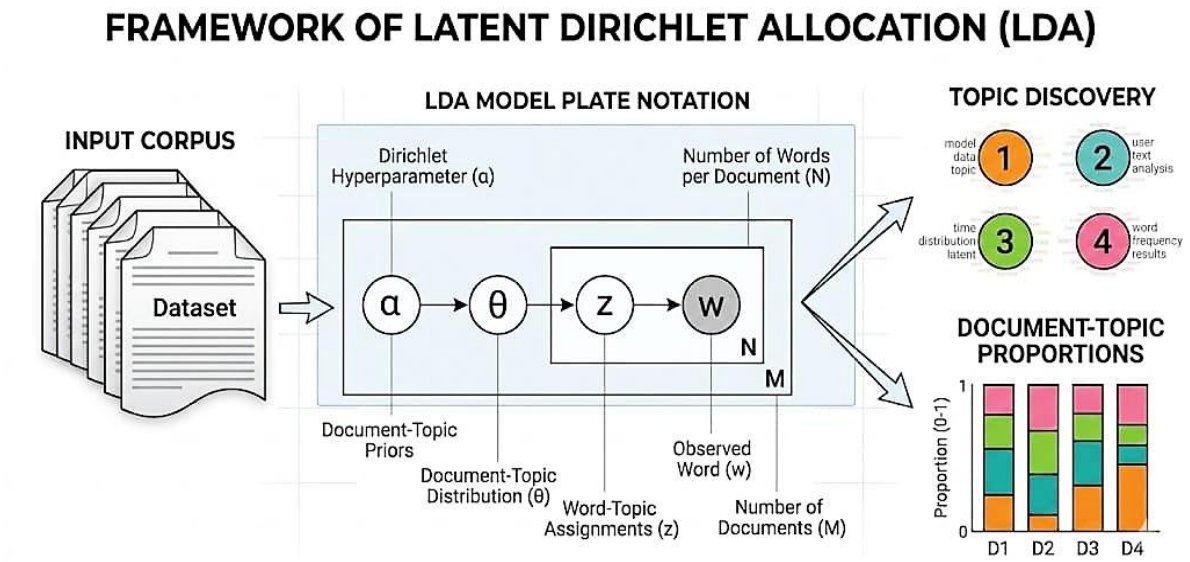

Figure1. The working mechanism of LDA algorithm

In the realm of academic research, a study by [8] delves into the domain of topic modeling, comparing five commonly used methods for modeling topics in short textual social data. The study evaluates the effectiveness of Latent Semantic Analysis and Latent Dirichlet Allocation on two text datasets, assessing their performance based on criteria such as subject quality, recall, accuracy, and topic coherence. Emphasizing the extraction of subjects from text data collected from unstructured and semi-structured sources, the study employs rigorous pre-processing steps to cleanse the data and conducts multiple evaluations with varying numbers of subjects, ultimately presenting the most descriptive subjects obtained compared to alternative techniques.

In a similar vein, [42] address the challenges associated with pre-processing Arabic text in the context of social network data analysis. Focusing on the vast amount of data present on social platforms and the necessity of extracting valuable insights from it, the study proposes a methodology comprising four key stages: data collection, cleaning, enrichment, availability, and the application of pre-processing algorithms for Arabic text normalization and cleaning. Given the proliferation of textual content on social media platforms, particularly in the form of comments and short messages, researchers face difficulties in uncovering pertinent topics. To tackle these issues, the study applies machine learning algorithms, NLP techniques, and topic modeling approaches to unearth significant themes from the data, aiming to avoid the loss of valuable knowledge amidst the deluge of information.

Through a comprehensive review of existing literature, it becomes evident that the extraction of data from social networking sites poses numerous challenges, notably concerning data abundance. The study highlights the necessity of analyzing the copious amounts of data originating from Libyan users on Twitter to unearth hidden insights and prevent the wastage of valuable knowledge. To address these challenges, the study emphasizes the application of topic modeling techniques alongside machine learning algorithms and NLP tools to extract and elucidate important themes from the collected data, facilitating a deeper understanding of the topics discussed on social media platforms.

3. Methodology

The framework followed in this study is shown in figure 2 .

a. Libyan Dialect Data Collection

In this study, data was gathered from the social networking platform Twitter using its Application Programming Interfaces (API). This collection involved utilizing various terms from the Libyan dialect to retrieve the data presented in Table 1.

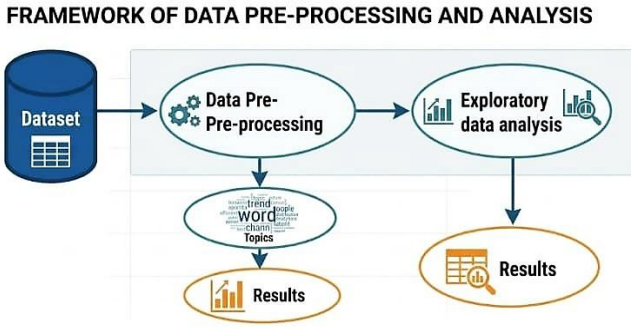

Figure 2. The Experiments of this study

Table 1. The vocabulary of the Libyan dialect that was used to collect the data

مكسد	نعرفش	الفترنيه	حوش	يشبح
هنايا	متاعى	صونية	العشية	لامية
طاسة	مليح	روشن	حقاني	خناب
ماكلة	ضي	بنزينة	هادو	عومان
رقدت	ما عنديش	عزومة	انخنب	اماله
هلبة	تركينة	انخدشت	ماجاش	القوانتي
رقدت	هنايا	كبوطي	متاع	الدخلانى
حوايج	حطيت	تزداني	خبط	زوز
فياق	الكوفينو	اوخيك	معليشي	مصقع
زعمه	قداش	اوني	قعمز	باهي
بكرج	ماظنش	غدوه	نتلاقو	كساد

The Twitter API is widely recognized among web developers as one of the most popular interfaces, and Twitter itself is a significant social platform that promotes user engagement and growth through its API interactions. Access to Twitter data is facilitated by the Twitter Streaming API, which collects real-time tweet data, as well as the Twitter REST APIs, particularly the Twitter Search API. This process enables the collection of historical tweet data, including temporal aspects such as the days of the week and geographic information. Additionally, data concerning followers and friends of specific Twitter users is incorporated. Several processing steps are then applied to this collected data to extract key topics, with the data being gathered from January 2021 to April 2021.

b. Exploratory Phase

The initial exploration phase plays a pivotal role in data analysis as it aims to derive preliminary findings that address the study's inquiries. This phase stands out as a critical and primary step at the onset of data analysis. It serves as an endeavor to comprehend the data stored within the database, unearth concealed insights, determine its characteristics, identify field and record names, ascertain the types of variables present, and assess whether the data exhibits missing, erroneous, or incomplete values that necessitate processing for accurate result

c. Database Description

The consolidated database comprises 13 fields and 153,335 records, as detailed in Table 2. Identify the attributes housing categorical data for utilization in the subsequent topic extraction phase. These fields

encompass the device field (source), documenting the devices utilized in the tweets, the date and time field (date) highlighting the timestamp when the tweet was posted, the text field (Text) encompassing the tweet contents, the username field (user) storing the username of the tweet author, the user's email field (user@) capturing the user's email address, and the dialect field recording the dialect of the tweet author. Table 2 features the categorical fields. Additionally, pinpoint the attributes containing numerical data within the database, as delineated in Table 2.

Table 2. Description of the retrieved database

Attribute	Data Type	Description
Id	Numeric	Tweet number
Date	Nominal	Tweet date
Text	Nominal	Tweet text
Favorite	Numeric	Number of favorites for tweet
Source	Nominal	Type of device used to tweet
User	Nominal	User name that changes
@User	Nominal	Name of the account holder in English that does not change
Retweet	Numeric	Retweet
Userfavorite	Numeric	User preferences
Userfollowers	Numeric	Number of the user follows
Userstatuses	Numeric	Number of tweets
Userfollowing	Numeric	Number of people the user is following
Dial	Nominal	The dialect

a. Attributes Description

Through the following table, the description of the fields that contain numerical data within the database has been clarified in terms of their count, the average, the standard deviation, the minimum and maximum values for the data, as well as the quartiles, which are shown in Table 3.

Table 3. shows the description of attributes in database

Attributes	Average	Standard deviation	Minimum	Maximum	25%	50%	75%
Id	1.366	1.208	1.345	1.38	1.34	1.37	1.37
Favorite	8.922	5.136	0.000	1.038	0.00	0.00	1.00
Retweet	5.000	1.5160	0.000	2.84	0.00	0.00	0.00
Usrfavorite	1.666	3.981	0.000	8.71	8.00	4.86	1.63
Usrfollowers	2.018	1.949	0.000	3.05	1.00	4.30	1.44
Usrstatuses	1.927	4.307	1.000	1.50	1.20	5.69	1.98
Usrfollowing	8.143	3.753	0.000	4.78	1.50	3.50	7.50

The (favorite) attribute is the preferred number of tweets, the max value for this attribute is 1.038, and (retweet) attributes shows retweets of the tweets, where the number of retweets that were retweeted is (153335), the max value is 2.84.

d. Description of Attributes

Through the following table, the description of the attributes that contain nominal data within the database has been clarified. Table 4. shows these attributes.

Table 4. shows the description of the attributes that contain nominal data

Attributes	Count	Unique	Maximum	Frequency
date	153335	150158	2021-04-13 15:01:52+00:00	5
text	153335	153335	@NEMO_AL02 المواهب متاعي	1
source	153335	5	Twitter for Android	79397
user	153309	63857	.	389
@User	153335	67303	knowwhathappend	327
dial	153335	1	LY	153335

Through the previous table, we will describe the important attributes from obtained data. These attributes are the (text) attribute contains the most frequently repeated text, this text was (المواهب متاعي@NEMO_AL02), and the number of occurrences of this text was 1. The (Source) attribute contains the source used in tweet, and the number of sources is 153335 sources.The most used was (Twitter for Android) where the number of occurrences for this device was 79397. The user attribute contains 63857 tweets, where the most tweeted person in the collected data was the person using the dot symbol (.), this person is considered who talks the most about Libya and number of his tweets was (153309), the attribute (@user) contains the user's email; The most frequent email was (knowwhathappend) and dial attribute contains the Libyan dialect.

e. Frequent Emoticons

Emojis have played an important role in communication between people on social media, as they are widely used by Internet users to express feelings and moods between the sender and the receiver among themselves, and an application for communication is not devoid of the use of emojis, as almost all social networking sites use them. Then, they were collected from Twitter on emojis, which receive more attention when analyzing the data, as they can express the user's feelings and change the meaning of the text written in the tweets. In this study, the emoji faces were extracted from all the tweets, which were distributed according to Table 5.

Table 5. Shows the most frequent expression of faces

Repetition	55646	12989	7889	6406	3486	3033	2460	2299	2092	2029	1415	1378	1144	1053	981
Emoticons	😂	😭	🤣	❤️	💔	🖤	😊	😅	😍	😁	😔	💙	🤔	🔥	🌹

A summary of the emojis in the entire tweets is shown in Table 6. notes from the table that the total emojis are 154224, which is considered very large compared to the total number of tweets of 153335, as the average of the emojis for all tweets was about 1.0057% of the emojis per post, and that Tweets contain 1,284 unique emojis.

Table 6. Shows a summary of emoticons in tweets

Total Tweets	153335
Total emojis for all Tweets	154224
The rate of emoji faces per tweet	1.0057
Number of unique emoticons	1284

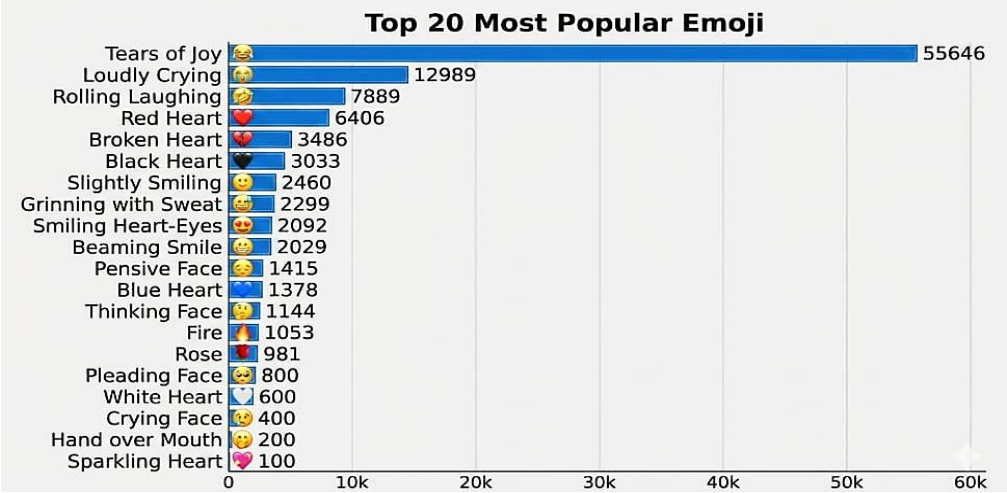

Figure 3. The most frequent expressions

In addition, after knowing the number of tweets that contain emoji shown in the previous figure 3, it will touch on recognizing the most frequent emoji faces, and this leads us to the answer to the question (What are the most frequent emojis in tweets?). Figure 3 shows the most frequent expressions.

Through the figure, we notice that the most frequent expression of faces is the symbol of a face with tears of joy 😂 with a ratio of 55646, followed by the symbol of crying out loud 😭, and this does not indicate that the mood of the Libyans was happy, but if the first five faces all indicate joy, we can say that the mood of the Libyans He was happy.

f. Description of Outliers in the Length Attribute

From the Figure 4 notices that there are abnormal values that need to be addressed.

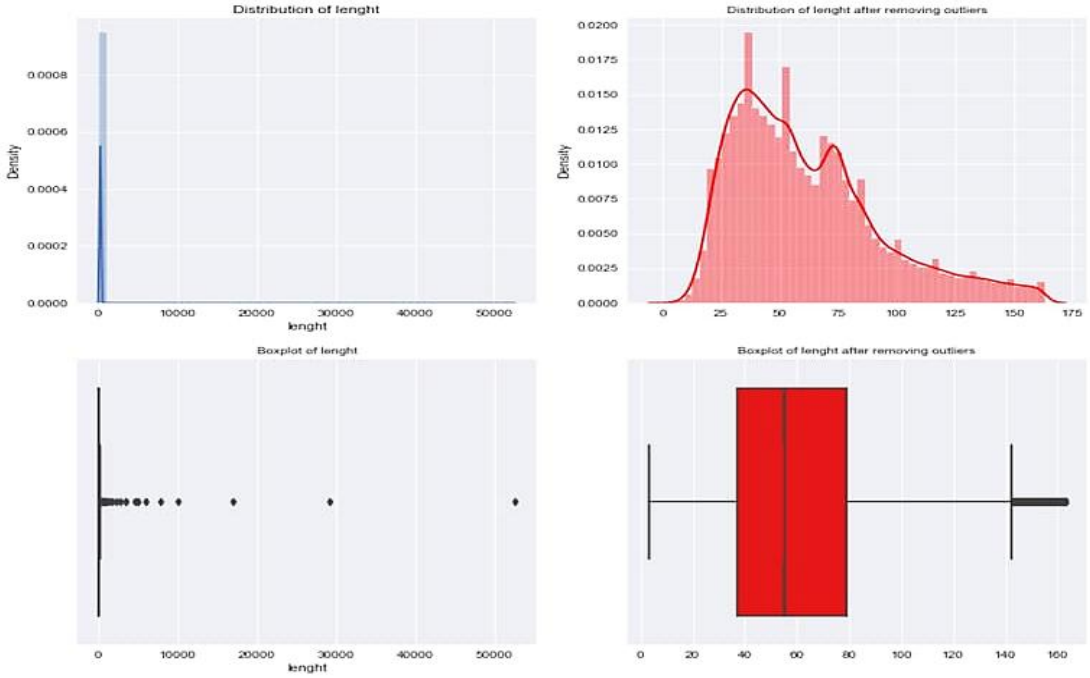

Figure 4. The abnormal values in the length attribute

g. Average Length of Tweets After Removing Outliers

The following figure 5 shows the average tweet length after deleting the outliers. If the result ranged from 5 to 70, this indicates that most Libyans had a general pattern of writing their tweets short.

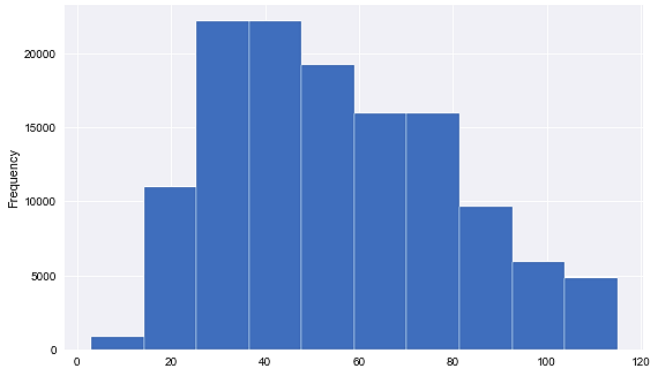

Figure 5. The average lengths of Tweets after deleting outliers

We note from the previous figure that the average length of Tweets is about 200,000 or more, and this indicates that the writing style of Libyan tweets is not short.

b. Preferred Number of Hours

Through Figure 6, it was found that the most preferred number of hours is 9 pm, and this indicates that most Libyans who used tension in the evening.

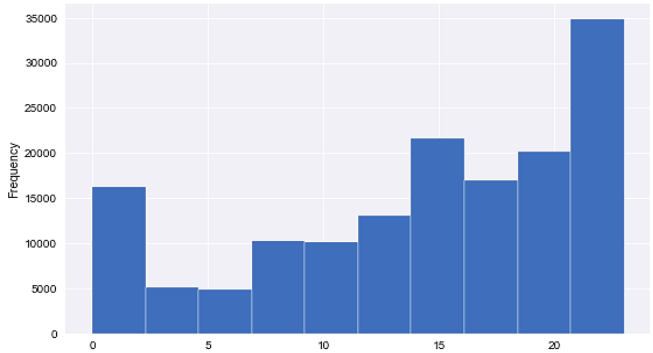

Figure 6. The most preferred number of hours

c. Most Preferred Months

Figure 7 shows the most preferred months, and the month (January) was the most preferred month, and this may be related to the presence of a specific event in that month.

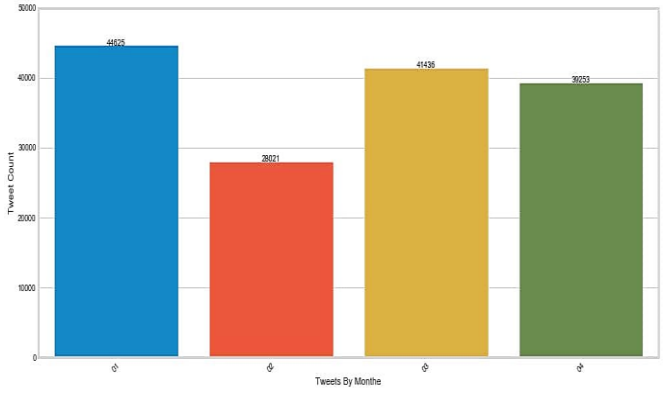

Figure 7. The most favourable months

h. Devices Source Used in The Tweet

The source field shows the source of devices that the user used to make the tweets. The following table shows the type of device, the number and percentage that were made using the application used in the tweet, and the cumulative percentage. Table 7. shows the sources of devices used in the tweet

Table 7. Shows the source of the devices used in tweets

Device's source	Number	Percentages	Cumulative percentage
Twitter for Android	79397	51.780089	51.780089
Twitter for iPhone	53989	35.209835	86.989924
Twitter Web App	19327	12.604428	99.594352
Twitter for iPad	492	0.320866	99.915218
Tweet bot for iOS	130	0.084782	100.00000

We note from the previous table that the data entry sources for Twitter were not all from devices, the source (Twitter Web App and Tweet bot for iOS) was software used in software and advertisements, so it will be deleted from other sources; The most used device is (Twitter for Android), followed by (Twitter for iPhone), which indicates that most Libyans use their phones to express their opinions.

i. Data Pre-Processing Phase

The task of extracting important topics from Twitter is a challenge in

itself, and it is becoming more and more difficult to analyze the retrieved data, as more errors are published by social media users; Including typos, spelling errors, sarcasm, and the use of proverbs and emoticons. The most prominent of these obstacles that constitute difficulty in extracting important topics from the data collected in the Libyan dialect are the similarity of some Libyan words in the tweets collected with dialects from other countries, such as the word (باهي، شوية). These words are used in other countries such as the state of Tunisia, and the other obstacle is sarcasm, which is represented in the fact that users of communication sites use a meaning or feeling different from the feeling to be expressed, which makes it more difficult to know the true feeling carried by the text. These difficulties require some special treatment when Extracting topics from these texts, including:

j. Remove Hyperlinks and Hashtags

Data collected from social networking sites is expected to contain some link tags or hashtags. hyperlinks and hashtags are common in social media communication and are widely used. So they should be removed from the text, as well as usernames, which are names preceded by the letter "@ or #" that are used to refer to another person within the tweet. An example, "@Konorkonor_94@fathi_bashagha باهي شن يعني بتقعد ديما حرب لبلاد و لا ###باهي شن يعني بتقعد ديما حرب لبلاد و لا " convert to "كيف"

k. Remove Stop Words from Indexing

There are a number of unnecessary Arabic words that do not carry any meaning, but their presence affects the validity of the results, such as (من، في،عن،كان،هذا،الذي) which must be removed from the text of the tweets. In this study, a list of words was used, including (اه،اها،اى،انت،انتن) that were removed, and this list was updated in the study as a result of the presence of words from the Libyan dialect that were not included in the mentioned list.

l. De-Diacritical Characters

This step is to reduce some of the similarity in the data by removing diacritics in the text, which are the symbols located above or below the letters of the Arabic text. These symbols are fatha, tanween with fatha, dammah, tanween with dammah, kasra, and tanween with kasrah and sukoon, so they are removed from the text to unify the forms of words. The removal of diacritics from letters such as " يا دائِمَ الفَضلِ عَلى البَرِيَّة، يا باسِطَ اليَدينِ بِالعَطِيَّة" will clean to " يا دائم الفضل علي البرية، يا باسط اليدين بالعطية".

m. Eliminate Repeated Characters

Repeated letters are one of the most popular techniques used on social media platforms. It is an attribute that is often used for confirmation or exclamations such as (رااااااائع) (المفضضضل), so these repeated letters must be removed.

n. Removing Characters

Twitter users use characters to express their tweets, Characters do not give much importance to get topics, so symbols in all their forms are removed from the text. These characters may be symbols, icons, numbers, spaces or English characters such as: ! "$ %9&\'() *+,-. ;<>=?@ 5 []^_`{=}— "•‹|~«°"]+.

o. Processing of Emoji's & Emoticons

After removing non-Arabic letters and numbers, and due to the fact that expressive images are English letters and numbers arranged in a specific format and used to represent a symbol for the meaning of a positive, negative, or neutral feeling. Exploratory and delete these expressive faces from the texts because they are not important in discovering topics. Table 8. shows the text before and after removing the emoticons.

Table 8. Shows the removal of emoticons from the text

Text	Cleaning text
@Rayed0006 شويه رحمه= ضياع بطولات في المتناول☺	شويه رحمه ضياع بطولات في المتناول
@faoma13 😇😇بالعكس اطلعي 😇لا والله ادهوري وأمشي وين يحب قلبك 😍😍 البيت ممل هلبة	لا واله بالعكس اطلعي ادهوري و امشي وين يحب قلبك البيت ملهلبه

p. Tokenization

In this step, the string of text is cut into other meaningful words. The aim of the text slicing is to explore the words in the sentence. Table 9 shows the cutting of the text

Table 9. Shows text segmentation

معك	فزعه	تلفزيون	كرته	الجدول
رقابه	النوم	ماضيك	خيبه	والمسلمات
نعطي	مخطوف	جطلاوي	عماتي	اتابع
هذيكا	إبراهيم	وزعلك	يحتاجك	والكافي
لمستشفي	السواد	طبيعتك	مر	وبنختار
هلنا	حضرته	والفعل	ونكران	مناعتك
كباراتو	يصفقلهم	نتعارك	الخدين	والحريات

q. Normalization of Writing

It is the process of replacing some letters with other letters in order to unify the shapes of the letters used in writing, such as replacing "ة " instead of "هـ", as well as replacing "ي" instead of "ى", and replacing "ء" instead of "ؤ". and " إ أ آ " instead of " ا ", as most social media users have trouble using the correct letters when writing, which makes it more difficult to extract topics. Table 10 shows the text before and after the normalization of the writing method.

Table 10. Shows standardization of writing method

Text	Cleaning text
@AtooGadema96 @Mohamedelmonfy باهي شن ذنبنا نحن اللي في بنغازي قاعدين وساكتين عندنا صبايا وهل نخافو عليهم تدوي من هنا يخرطوك من هنا وربي يفرج علينا واسمي مجهول الهوية مدايره فوق جو الايميل بس يعلم ربي ماشورنا حاجه 😔	باهي شن ذنبنا نحن اللي في بنغازي قاعدين وساكتين عندنا صبايا وهل نخافو عليهم تدوي من هنا يخرطوك من هنا وربي يفرج علينا واسمي مجهول الهوية مدايره فوق جو الايميل بس يعلم ربي ماشورنا حاجة
@LnAFOBW0AYrgeOR هذا شء طبيعي يجيك شويه صداع يخوف شويه كتير	هذا شئ طبيعي يجيك شويه صداع يخوف شوية كتير

r. Converting Text to Features

many studies examine various text representation and classification techniques for Arabic social media content. Bag-of-Words (BOW) and Term Frequency-Inverse Document Frequency (TF-IDF) are commonly used for feature extraction in Arabic sentiment analysis and authorship authentication (Kasri, Birjali et al. 2019) ; (Albadarneh, Talafha et al. 2015). However, distributed representations like word embeddings can outperform BOW in opinion mining tasks (El-Halees 2017). For detecting false information in Arabic tweets, a hybrid approach combining Harris Hawks Optimizer for feature selection with machine learning models, particularly Logistic Regression, has shown promising results (Thaher, Saheb et al. 2021). While these methods have demonstrated effectiveness, challenges remain due to the complexity of the Arabic language and the unique characteristics of social media text. The studies highlight the importance of appropriate feature extraction and selection techniques in improving the performance of various NLP tasks for Arabic content.

The process of converting text into structured properties is a crucial stage essential for the effective analysis of natural languages. Given that machine learning algorithms operate on digital attributes, this conversion stage, also known as text tokenization, is imperative. Text data, being inherently unstructured, requires preprocessing to be effectively utilized by algorithms, necessitating the transformation of this data into structured formats. In this context, a robust model is applied to represent textual data, enabling the exploration of meaningful topics. Various techniques exist for converting text into features; one fundamental model in NLP is the Bag of Words technique, as highlighted in the study by Zhang (Zhang, Jin et al. 2010). This approach converts text into a digital format that highlights the most frequent words in the dataset, facilitating the extraction of important topics prevalent among Libyan Twitter users from this structured representation.

4. Result and Discussion

Several experiments conducted to adjust the parameters of LDA algorithm, noting that two algorithms used to extract the important topics raised by Libyans on the social networking site (Twitter), namely the original algorithm (LdaMulticore), and the modified algorithm (lda_model).

a. Parameters of LdaMulticore Algorithm

Table 11. The parameters used in LdaMulticore algorithm

Parameter name	Description	Value
num_topics	Number of topics to be extracted	2 5
Threads	Number of Threads	1

b. Parameters of Lda_Model Algorithm

Some Parameters of the algorithm are different from LdaMulticore algorithm as shown in Table 12.

Table 12. The parameters used in Lda_model algorithm

Parameter name	Description	Value
eta	Preconceived notion about topic word distribution	Symmetric Auto
alpha	Preconceived notion about topic distribution	Symmetric Auto Asymmetric
num_topics	Number of topics to be extracted	2 5

The eta operator does not accept an asymmetric value when applied.10 topics were tested based on the study of (Yang and Zhang 2018). The topic (2, 5) was chosen because it shows explanatory topics. On the basis of the above parameters, the results will be divided according to the experiments as follows:

- Two experiments when using the original algorithm (LdaMulticore) by fixing all parameters except for the number of threads, which resulted in:
  1. Experience when the topics were 2.
  2. Another experiment when the topics were 5.
- Six experiments when changing the parameters of the modified algorithm were as follows:
  1. An experiment when the parameters for eta and alpha were symmetric and the number of topics is 2
  2. An experiment when the parameters for eta and alpha were symmetric and the number of topics is 5
  3. An experiment when the parameters for eta are equal to symmetric and alpha are equal to asymmetric and the number of topics is 2
  4. An experiment when the parameters for eta were equal to symmetric and alpha were equal to asymmetric and the number of topics is 5
  5. An experiment when the parameters for eta and alpha were Auto and the number of topics is 2
  6. An experiment when the parameters for eta and alpha were Auto and the number of topics is 5

c. Results of LdaMulticore Algorithm

Two experiments show in Table 13 noted the most words were supposed to be explanatory topics for LdaMulticore algorithm were not explained, all the seven topics did not have any significance or connection for the words extracted for each topic, therefore we were not able to form a specific topic, as the words most likely include several.

Table 13. The results of the experiments of original algorithm (LdaMulticore)

Experiment	Number of topics	Result	Topic
First	2 topics	وعقل، ماتبيش، عارفك، اداب، تقابلت، يكمل، اجرني، لكبيره، مقياسك.	No topic explained
		يكسرو، درتهن، وراءه، ويحاول، ربوع، تاليفون، مخطوف، فنفسك، وبعقله، ماكنتش.	No topic explained
Second	5 topics	تعبني، يكسرو، أفتر، لفرح، تقابلت، بطبيعتي، والمحافظين، غفرتهالنا، بتجاوز، درتهن	No topic explained
		الطمع، محرمته، احباب، نجي، لامحاله، شلتك، حاقده، سادسه، بالتساوي، الدهيماء.	No topic explained
		الاغفرته، عطاك، الاجازه، فانيه، امو، اداب، يكلمنيش، "نجوم، اعملوا، يهون	No topic explained
		مختص، نهنيك، تغذيه، وخبره، بكامل، تصنيف، حيكون، التربيزه، صادقي، وحناك	No topic explained
		ولاحطيت، غاليات، بحلم، روعه، الاسباب، ولابد، نايا، اسلام، حارمها، هعيش	No topic explained

Separate indications and were not related to one subject, from that it is possible to say that the original algorithm did not provide explanatory results on the data of the Libyan dialect when changing the parameter of subjects from 2 to 5.

d. Results of Lda_Model Algorithm

First experience: After applying Lda_model algorithm and changing the values of the parameters, the results of the first experiment, when the number of topics was two, the words of the first topic did not show an explanatory topic, and the words of the second topic indicated friendship.

Second experiment: Shows five topics based on Lda_model algorithm, increased the interpreted results for words that do not indicate a clearly explained topic, as the correctly interpreted topics indicated (study, recess, sick leave).

Third experiment: We came to one explanatory topic out of two topics; this topic is about (respect and love).

Fourth experiment: Most of the words were linked, and explained topics showed, these topics are about (disease, the political parliament, military forces for a political figure and a political figure).

Fifth experiment: The results of Lda_model algorithm show some topics that did not show a correctly interpreted topic.

Sixth experiment: Five topics extraction based on Lda_model algorithm, but the interpreted results for patterns do not specify an obviously explained topic.

It is noted from the experiments illustrates in Table 14, that most of words are repeated and did not show a correctly interpreted topic, and there is no correlation between the topics, so we will not be able to extract explanatory topics for those words.

Table 14. The results of the experiments of modified

Experiment	Number of topics	Result	Topic
First	2 topics	بادي، الاعمار، لاكتر، لاخر، مشبيه، احباب، عسانا، هيبه، ماهمني، وطبيبه	No topic explained
		ماهمني، هيبه، عسانا، احباب، مشبيه، لاخر، الاعمار، بادي، نلقاها	Friendship
Second	5 topics	الاعمار، احباب، هيبه، لاخر، نلقاها، وتي، الروجمتصله مدايره، تلبسي	Travel
		بادي، تعبني، زدت، أفتر، ريم، الذهبي، بغير، اجرني، مصيبتي	No topic explained
		لاكتر، وطبيبه، نجي، وهبني، توب، اضاعت، بايدن، خيانته، أحاول، واكتشفت	No topic explained
		عسانا، دواه، ماخاب، ومروح، يكلمنيش، داكن، المركب، الاجازه، ويبخصني	Sick leave
		ماهمني، مشبيه، واتيه، روج، سكول، نقري، مستخدمين، نقله ويدري، واضطر	Study
Third	2 topics	لاخر، مشبيه، احباب، الاعمار، ماهمني، بادي لاكتر، عسانا، هيبه، وهبني	Respect and Love
		هيبه، عسانا، لاكتر، بادي، ماهمني، الاعمار، احباب، مشبيه، لاخر، وطبيبه	No topic explained
Fourth	5 topics	ماهمني، احباب، لاكتر، دواه، وتي، واتيه، وطبيبه نلقاها زدت، داكن	Illness
		لاخر، هيبه، الاعمار، الروج، متصله، أفتر، ولايقرا، الذهبي، ريم، مصيبتي	No topic explained
		نجي، وهبني، توب، أحاول، بايدن، امارته، وعاصمة، ومجده، نعاودوا، ماعشناش	Political figure
		عسانا، ماخاب، ومروح، ويبخصني، قواته، وخرج، يشفيك، لهيبه، وبعقله، ولصغيره	Military forces for a political figure
		مشبيه، نقله، وزال، والنواب، وتفهم، صحيحه، كثيرا، مشروعه ملباك، ماكنش	Political Parliament
Fifth	topics	الاعمار، بادي، احباب، عسانا، ماخاب لاكتر، مشبيه، لاخر، هيبه، ماهمني	No topic explained
		بادي، لاخر، هيبه، ماهمني، مشبيه، الاعمار، احباب، عسانا، لاكتر،ماخاب	No topic explained
Sixth	5 topics	مشبيه، لاخر، هيبه، ماهمني، لاكتر الاعمار، بادي، احباب، عسانا، ماخاب	No topic explained
		مشبيه، لاخر، هيبه، ماهمني، لاكتر الاعمار، بادي، احباب، عسانا، ماخاب	No topic explained
		مشبيه، لاخر، هيبه، ماهمني، لاكتر الاعمار، بادي، احباب، عسانا، ماخاب	No topic explained

Nevertheless, upon identifying the algorithm's parameters and their application to the gathered data, an issue emerged concerning the extraction of results, despite modifications to the parameter values in an attempt to improve the topics. This challenge is illustrated in previous trials detailed in Table 14, where certain topics lacked clarity and failed to provide coherent explanations, despite the correct implementation of processing steps such as numeral removal, standardization of letter formats, elimination of English characters, hyperlinks, and emoticons from the dataset. The topics that emerged included themes like friendship, respect, love, illness, leisure, sick

leave, political figures, military associations of political figures, parliamentary affairs, and education. Most of the topics derived from these experiments pertained to politics; however, some results revealed topics that proved challenging to decipher. This difficulty arises from the intricacies of tweets composed in the Libyan dialect, posing obstacles to accurately interpreting the extracted topics.

5. Limitation

This study has several limitations that should be acknowledged. First, the use of Latent Dirichlet Allocation (LDA) on short-text Twitter data may reduce topic coherence and interpretability due to the sparse nature of tweets. Second, newer short-text topic modelling approaches such as BER Topic and Top2Vec were not implemented for comparison. Additionally, topic coherence metrics such as CV, UMass, and C_npmi were not used to optimize the number of topics. The preprocessing pipeline was extensive; however, the quantitative effect of each cleaning stage was not reported. Finally, the dataset reflects a specific time period and may not fully represent evolving Twitter discussions over time.

6. Conclusion

Based on findings from exploratory data analysis and Latent Dirichlet Allocation (LDA) applied to Twitter data written in the Libyan dialect, several key patterns were identified.

First, emoji usage analysis revealed that 91,701 tweets did not include emojis, while 15,759 tweets contained two emojis. This suggests that a substantial proportion of users employ emojis to express sentiment and meaning in their communication. Among emoji types, the "face with tears of joy" 😂 was the most frequently used (55,646 occurrences), followed by the loudly crying face 😭. While the dominance of positive emojis may indicate expressive communication practices, it cannot be directly interpreted as a definitive measure of overall user sentiment.

Temporal analysis showed that tweet activity peaked at approximately 9 p.m., suggesting that users are more active during evening hours, likely reflecting leisure time. Monthly trends indicated higher activity in January, while weekly patterns showed increased engagement on Mondays and Sundays, with lower activity on Tuesdays and Fridays. These patterns may reflect social routines and work–leisure cycles.

The topic modelling results identified several dominant themes, including friendship, respect, love, illness, leisure, sick leave, political figures, military discourse, parliamentary affairs, and education. Political topics appeared particularly prominent, suggesting heightened political engagement during the data collection period.

Despite the overall consistency of preprocessing and modelling steps, some topics remained less coherent, highlighting the sensitivity of LDA results to data preparation quality. This underscores the importance of rigorous preprocessing in ensuring reliable topic extraction.

In conclusion, the integration of exploratory data analysis and LDA-based topic modelling provides valuable insights into behavioural patterns, temporal dynamics, and thematic structures in Libyan Twitter discourse. The findings contribute to a deeper understanding of how users express opinions and interact within a dialect-rich social media environment.

7. References

[1] Q. Shen, 'Topic Discovery and Future Trend Prediction In Scholarly Networks'. 2016.

[2] T. Porturas and R. A. Taylor, 'Forty years of emergency medicine research: Uncovering research themes and trends through topic modeling', *Am. J. Emerg. Med.*, vol. 45, pp. 213–220, 2021.

[3] R. Abdelhamed, M. Essgaer, A. Agaal, and A. Shibani, 'Enhancing Topic Modeling in Scientific Literature: A Comparative Study of LDA with Word2Vec, Doc2Vec, and SciBERT Embeddings', in *Selected Papers from the International Conference on Artificial Intelligence*, A. O. Albaji, Ed., Cham: Springer Nature Switzerland, 2026, pp. 634–648.

[4] R. Abdelhamed and M. Essgaer, 'Exploring Knowledge Landscapes: Clustering and Topic Modeling of Sebha University Scientific Publications', in *2024 IEEE 4th International Maghreb Meeting of the Conference on Sciences and Techniques of Automatic Control and Computer Engineering (MI-STA)*, May 2024, pp. 712–717. doi: 10.1109/MI-STA61267.2024.10599683.

[5] M. B. Mutanga and A. Abayomi, 'Tweeting on COVID-19 pandemic in South Africa: LDA-based topic modelling approach', *Afr. J. Sci. Technol. Innov. Dev.*, vol. 14, no. 1, pp. 163–172, 2022.

[6] Y. Chen and L. Liu, 'Development and research of topic detection and tracking', in *2016 7th IEEE International Conference on Software Engineering and Service Science (ICSESS), IEEE*, 2016.

[7] A. Krishnan, 'Exploring the Power of Topic Modeling Techniques in Analyzing Customer Reviews: A Comparative Analysis'. 2023.

[8] et al Albalawi R., 'Using topic modeling methods for short-text data: A comparative analysis', *Front. Artif. Intell.*, vol. 3, p. 42, 2020.

[9] R. Egger, 'Topic Modelling', *Appl. Data Sci. Tour.*, 2022.

[10] S. Kumar and V. Bhatnagar, 'A review of regression models in machine learning', *J. Intell. Syst. Comput.*, vol. 3, no. 1, pp. 40–47, 2022.

[11] et al Bohra N., 'Popularity Prediction of Social Media Post Using Tensor Factorization', *Intell. Autom. Soft Comput.*, vol. 36, no. 1, 2023.

[12] A. andL. B. R. Meddeb, 'Using Topic Modeling and Word Embedding for Topic Extraction in Twitter', in *International Conference on Knowledge-Based Intelligent Information & Engineering Systems*, 2022.

[13] et al Alothman M., 'Review of Researches on Arabic Social Media Text Mining'. 2021.

[14] N. F. Bin Hathlian and A. M. Hafezs, 'Sentiment - subjective analysis framework for arabic social media posts', in *2016 4th Saudi International Conference on Information Technology (Big Data Analysis) (KACSTIT): 1-6*, 2016.

[15] N. F. B. Hathlian and A. M. Hafez, 'Subjective Text Mining for Arabic Social Media', *Int J Semantic Web Inf Syst*, vol. 13, pp. 1–13, 2017.

[16] L. Al-Horaibi and M. B. Khan, 'Sentiment analysis of Arabic tweets using text mining techniques', in *International Workshop on Pattern Recognition*, 2016.

[17] et al Ghani N. A., 'Social media big data analytics: A survey', *Comput. Hum. Behav.*, vol. 101, pp. 417–428, 2019.

[18] et al Alqmase M., 'Sports-fanaticism formalism for sentiment analysis in Arabic text', *Soc. Netw. Anal. Min.*, vol. 11, no. 1, pp. 1–24, 2021.

[19] et al Asif M., 'Sentiment analysis of extremism in social media from textual information', *Telemat. Inform.*, vol. 48, p. 101345, 2020.

[20] S. C. McGregor, 'Social media as public opinion: How journalists use social media to represent public opinion', *Journalism*, vol. 20, no. 8, pp. 1070–1086, 2019.

[21] et al Hung M., 'Social network analysis of COVID-19 sentiments: Application of artificial intelligence', *J. Med. Internet Res.*, vol. 22, no. 8, 2020.

[22] et al Ragini J. R., 'Big data analytics for disaster response and recovery through sentiment analysis', *Int. J. Inf. Manag.*, vol. 42, pp. 13–24, 2018.

[23] S. A. AlAjlan and A. K. J. Saudagar, 'Machine learning approach for threat detection on social media posts containing Arabic text', *Evol. Intell.*, vol. 14, no. 2, pp. 811–822, 2021.

[24] et al Khalafat M., 'Violence Detection over Online Social Networks: An Arabic Sentiment Analysis Approach', *iJIM*, vol. 15, no. 14, p. 91, 2021.

[25] et al Kanan T., 'Cyber-bullying and cyber-harassment detection using supervised machine learning techniques in Arabic social media contents', *J. Internet Technol.*, vol. 21, no. 5, pp. 1409–1421, 2020.

[26] et al Guellil I., 'Detecting hate speech against politicians in Arabic community on social media', *Int. J. Web Inf. Syst.*, 2020.

[27] etal Allaith A., 'Neural Network Approach for Irony Detection from Arabic Text on Social Media', in *FIRE (Working Notes)*, 2019.

[28] T. Al-Khalifi, 'Social media data mining and its applications in media research: Sentiment analysis as a model', *J. Media Res. Stud.*, vol. 8, no. 8, pp. 1–73, 2019.

[29] et al Curiskis S. A., 'An evaluation of document clustering and topic modelling in two online social networks: Twitter and Reddit', *Inf. Process. Manag.*, vol. 57, no. 2, p. 102034, 2020.

[30] E. F. E. Elnour, 'Using Data Mining Techniques to Establish Standard Sizing System for Sudanese Army Officers Poshirt'. 2018.

[31] et al Yuan M., 'Document Clustering vs Topic Models: A Case Study', in *Proceedings of the 25th Australasian Document Computing Symposium*, 2021.

[32] et al Lossio-Ventura J. A., 'Evaluation of clustering and topic modeling methods over health-related tweets and emails', *Artif. Intell. Med.*, vol. 117, p. 102096, 2021.

[33] M. H. Beseiso, 'New Sentiment Analysis Model Using LDA for Arabic Tweets', in *Proceedings of the 3rd International Conference on Advances in Artificial Intelligence*, 2019.

[34] et al Al-Qudah I., 'Applying Latent Dirichlet Allocation Technique to Classify Topics on Sustainability Using Arabic Text', *Sai*, 2022.

[35] et al Albadarneh J., 'Using Big Data Analytics for Authorship Authentication of Arabic Tweets', in *2015 IEEE/ACM 8th International Conference on Utility and Cloud Computing (UCC): 448-452*, 2015.

[36] et al Zrigui M., 'Arabic Text Classification Framework Based on Latent Dirichlet Allocation', *J Comput Inf Technol*, vol. 20, pp. 125–140, 2012.

[37] et al Wang H., 'Exploring the Chinese public's perception of omicron variantson social media: lda-based topic modeling and sentiment analysis', *Int. J. Environ. Res. Public. Health*, vol. 19, no. 14, p. 8377, 2022.

[38] L. Zou and W. W. Song, 'LDA-TM: A two-step approach to Twitter topic data clustering', in *2016 IEEE International Conference on Cloud Computing and Big Data Analysis (ICCCBDA): 342-347*, 2016.

[39] S. Yang and H. Zhang, 'Text Mining of Twitter Data Using a Latent Dirichlet Allocation Topic Modeland Sentiment Analysis'. 2018.

[40] A. Omar, M. Essgaer, and K. M. S. Ahmed, 'Using Machine Learning Model To Predict Libyan Telecom Company Customer Satisfaction', in *2022 International Conference on Engineering & MIS (ICEMIS)*, Jul. 2022, pp. 1–6. doi: 10.1109/ICEMIS56295.2022.9914055.

[41] et al Oshikawa R., 'A survey on natural language processing for fake news detection'. 2018.

[42] et al Hegazi M. O., 'Preprocessing Arabic text on social media', *Heliyon*, vol. 7, no. 2, 2021.