International Journal of Innovative Research and Scientific Studies, 8(9) 2025, pages: 152-163

IJIRSS
International Journal of Innovative Research & Scientific Studies

ISSN: 2617-6548

URL: www.ijirss.com

Topic modeling as a tool for analyzing tweets: A case study of the Russia-Ukraine war on Arabic social media

Farah Alshanik[1*], Jumana Khrais[1], Rasha Obeidat[1], Lamees Rababa[1], Saif Ziad Aljunidi[2]

[1]*Department of Computer Science, Faculty of Computer and Information Technology, Jordan University of Science and Technology, Irbid, Jordan.*
[2]*Department of Public Law, Faculty of Law, Yarmouk University, Irbid, Jordan.*

Corresponding author: Farah Alshanik (*Email: fmalshanik@just.edu.jo*)

Abstract

The transformation of information dissemination and public discourse was heavily influenced by the rapid rise of social media platforms, especially Twitter, and this was particularly evident in times of conflict and war. The huge explosion of social media usage has, in turn, created a large amount of data that can be analyzed through different methods such as text mining and natural language processing. In this paper, we employ topic modeling to extract and analyze the topics of discussion around the Russian-Ukraine conflict in the Middle East. The analysis is facilitated by collecting dialectical Arabic tweets specifically containing terms relating to the Ukraine-Russia conflict. Investigation and comprehension of the dominant themes and views carried by the discussion are held through comparative research of two important topic modeling tools: BERTopic and LDA. From our findings, the influence of social media in forming public opinion, the spread of information, and the creation of a discourse regarding the Russian-Ukrainian war in the Middle East becomes visible. The topic modeling in our study presents the broad spectrum of views and themes emerging through social media discourse. This comprehensive perspective assists in the understanding of the intricate complexities surrounding this geopolitical conflict and offers a deep dive into the multifaceted nature of the matter.

Keywords: BERTopic, LDA, Public discussion, Social media, Topic modeling, Twitter, Ukraine-Russia conflict.

DOI: 10.53894/ijirss.v8i9.10644
Funding: The authors would like to acknowledge the deanship of Research at Jordan University of Science and Technology for supporting this research (Grant Number: 20230242).
History: Received: 27 August 2025 / **Revised:** 19 September 2025 / **Accepted:** 23 September 2025 / **Published:** 14 October 2025
Copyright: © 2025 by the authors. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).
Competing Interests: The authors declare that they have no competing interests.
Authors' Contributions: All authors contributed equally to the conception and design of the study. All authors have read and agreed to the published version of the manuscript.
Transparency: The authors confirm that the manuscript is an honest, accurate, and transparent account of the study; that no vital features of the study have been omitted; and that any discrepancies from the study as planned have been explained. This study followed all ethical practices during writing.
Publisher: Innovative Research Publishing

1. Introduction

The Russia-Ukraine war, which began on February 24, 2022, has had a significant impact on global affairs and has emerged as a topic of much debate and study. Like any significant conflict, understanding the various aspects and perspectives surrounding this war is essential for policymakers, analysts, and researchers. Among the wealth of resources available for studying this war, social media platforms such as Twitter stand out as a valuable source of information, where users freely express their opinions, share news updates and engage in discussions directly related to the ongoing conflict.

These platforms play a crucial role as vast repositories of user-generated content, offering valuable insights into public sentiment, emerging trends, and the primary topics of discussion surrounding the Russia-Ukraine conflict. The analysis of these topics and discussions is highly important in understanding the dynamics and implications of this ongoing conflict. In particular, studying the topics expressed in the Arabic region provides invaluable insights into the perceptions, concerns, and opinions of the Arabic-speaking population.

Topic modeling, a powerful technique derived from natural language processing and machine learning, is used to identify and extract latent topics from massive amounts of data. By applying topic modeling to dialectical Arabic tweets about the Russia-Ukraine war, we explore deeply into the main issues, concerns, and viewpoints that the Arabic-speaking masses express. The study of topics within the Arab region not only sheds light on perceptions, concerns, and opinions of the Arabic-speaking community but carries its own weight in the larger context. Considering the immense cultural backdrop, politically diverse societies, and heavyweight positioning of the entire Arab world on the global chessboard, it becomes very crucial to understand their views and concerns regarding this conflict. These results have yet more pertinent information for the Arab community, the direct beneficiaries being the policymakers and diplomats. It serves as a strategic instrument to persuade foreign policy orientations, build diplomatic relations, and structure viable frameworks for international cooperation. By delineating perspectives of Arabic-speaking masses, we get a panorama beyond all-theoretical knowledge. It is imperative to sway regional dynamics and public opinion, and it provides some clear options to navigate through the complex interplay of global happenings and the intertwined cultures in Arabic-speaking societies.

However, while there has been a lot of research done on topic modeling in different languages, the Arabic language still hasn't got the attention it deserves. The processing of Arabic language has its own difficulties which are caused by the non-concatenative nature of Arabic morphology, the lack of spelling rules, and the different dialects [1]. Taking these difficulties into account, our research suggests the use of topic modeling methods on dialectical Arabic tweets that are focused on a specific recent crisis: The Ukraine-Russia conflict. This will not only help in overcoming some of the challenges in Arabic language processing but also in getting a better understanding of the main topics and themes that are there in this situation.

This study aims to identify the potential insights from the Middle Eastern tweets about the Russian-Ukrainian War and thus understand the war's influence on the Arab world. To this end, the paper systematically collects Twitter data for the period specified, February 1, 2022, to December 14, 2022, using carefully chosen keywords. The study employs sophisticated topic modeling techniques like Latent Dirichlet Allocation (LDA) and BERTopic with the aim of uncovering the hidden topics contained in these tweets and thus providing a clear understanding of public discussion. The paper, by unraveling these technical aspects, aspires to bring forth a more accurate and in-depth investigation of the complex interplay of the Russian-Ukrainian War within the Middle Eastern Twitter sphere.

The significant contributions of this paper can be summarized as follows:

- Contribute to the available data resources for the Arabic language by collecting and preparing a dataset consisting of dialectical Arabic tweets specifically related to the Russia-Ukraine War.
- Compare the performance of different topic modeling algorithms on the collected dialectical Arabic dataset.
- Provide a comprehensive textual analysis of the tweets, including:
- Discovering the main topics and themes discussed in the tweets.
- Address questions from decision-makers regarding public concerns and opinions within the Arabian context concerning the Ukraine-Russia conflict.

The paper is structured as follows: Section 2 presents a comprehensive review of topic modeling applications in the existing literature. Section 3 outlines the methodology employed in this study. Section 4 presents the experiment setup in a comprehensive manner. Section 5 presents and discusses the obtained results. Finally, Section 6 concludes the work by summarizing the key findings and providing insights for future research.

2. Related Works

The section is devoted to a detailed review of the literature on the subject of topic modeling applications in both Arabic and English, placing special emphasis on short text (tweets) and long text (documents) analysis. The review outlines any studies which compare the effectiveness of different topic modeling algorithms and those that analyze a particular set of tweets regarding global events, crises, and conflicts through topic modeling. Moreover, some studies consider topic modeling as an essential part of the Natural Language Processing (NLP) application systems being developed.

The authors focus in their paper the deployment of topic modeling in the English language to different kinds of dataset which covered a wide range of topics such as: education, healthcare, environmental issues and international conflicts. .

One of the studies in the education field was performed by Mujahid, et al. [2] who used the Latent Semantic Analysis (LSA) method on English tweets to reveal the major concerns raised during the COVID pandemic regarding the effectiveness of online learning. Additionally,

Zankadi, et al. [3] demonstrated that topic modeling facilitated the recommendation of online courses based on social interactions. The authors of Zankadi, et al. [3] went on to enhance course preferences within Massive Open Online Courses (MOOCs) extracting students' topics of interest from social media content. They employed three widely recognized topic modeling techniques, namely Latent Dirichlet Allocation, Latent Semantic Analysis, and BERTopic. The experimental results demonstrated that BERTopic surpassed the other models in terms of effectively extracting pertinent topics and analyzing textual features. While, Silveira, et al. [4] addressed the necessity for a system that can provide crucial points from legal documents to students, legal scholars, lawyers, and judges on a daily basis. To fulfill this need, a dataset of legal documents was compiled, and a novel topic modeling approach was designed for legal texts. The authors then experimented with LEGAL-BERT1 as the topic modeling algorithm and on further information involving law citations so as to enable the topic model to better determine topics.

A notable example of applying topic modeling on healthcare-related texts is that by Lossio-Ventura, et al. [5]. The authors have comprehensively evaluated some topic modeling algorithms on very short texts related to healthcare. The studies Dahal, et al. [6] and Jin and Spence [7] attempt to understand the topics of social media conversations about environmental events. Dahal, et al. [6] were analyzing tweets about climate change, while Jin and Spence [7] were focusing on tweets about Hurricane Maria.

In contrast, Akpatsa, et al. [8] and Aslan [9] performed analyses of public opinion concerning the conflicts in the US-Afghan war and the Ukraine-Russian wars, respectively, using topic modeling methodologies to shed light on public sentiment and decision-making. The studies by Akpatsa, et al. [8] and Aslan [9] are of utmost importance to our studies, as they implement topic modeling to assess the trends and themes of public opinion surrounding conflicts occurring within their respective countries. Utilizing topic modeling, these studies considerably contribute to the understanding of public sentiment and decision-making processes. Aslan [9] conducted a study using topic modeling to extract crucial issues discussed on Twitter concerning the Ukraine-Russia war. The study specifically focused on English tweets. The authors employed standard LDA and identified seven dominant topics within the dataset. Word cloud visualizations were employed to showcase the most prevalent words associated with each topic. However, it is worth noting that no specific evaluation of the topic quality

As to the deployment of the topic modeling in Arabic language, Alami, et al. [10] and Hammo, et al. [11] utilized topic modeling as a component of the Arabic text summarization system with a role of reducing the dimensionality of the documents from word space into topic space. Alami, et al. [10] introduced an automatic summarization system using LDA and neural networks, addressing domain-dependency issues in Arabic text summarization by assigning hidden topics to groups of similar documents. While in Hammo, et al. [11] topic modeling was used as an identifier to the main thematic structure of the document. The evaluation was done using 1200 human evaluators.

Alhazmi [12] used topic modeling to explore discussions on distance education among Saudi Arabian students using the Biterm Topic Model (BTM) on a dataset of relevant tweets. Alshalan, et al. [13] utilized the NMF topic modeling on the ArCOV-19 dataset to identify hate speech in tweets.

A deep dive into the literature on topic modeling for the Arabic language pinpoints a considerable lack of research on dialectal Arabic. However, Habbat, et al. [14] tried to fill this gap in their study. They gathered a dataset of Moroccan Twitter tweets and focused on sentiment classification, categorizing sentiment within negative, positive, and neutral classes, through topic modeling methods. The authors compared and analyzed topics brought forth by the two well-known algorithms, NMF and LDA, concluding that and gave results stating that LDA exhibited superior performance over NMF on the measure of topic coherence.

In another study, Abuzayed and Al-Khalifa [15] applied pre-trained language models to topic modeling first by comparing BERTopic with LDA and NMF. The authors studied the performance of BERTopic when used alongside word embeddings from various pre-trained language models. They carried out a comparative analysis of BERTopic with the other two models, LDA and NMF. Unlike LDA and NMF, BERTopic does not require the number of topics to be set in advance. The dataset used in the study was drawn from (DataSet for Arabic Classification) and consists of 108,789 documents in MSA. The results showed that LDA performed poorly, NMF performed quite well, and BERTopic coupled with AraVec gave promising results.

The conducted literature review has brought to light the lack of studies applying the existing topic modeling techniques on dialectal Arabic text in the context of conflicts and crises. To address this gap, we have introduced a comprehensive methodology by employing two distinct topic modeling techniques on an Arabic dataset that covers diverse dialects and focuses on discussions related to the Russia-Ukraine Conflict.

3. Methodology

The main aim of this research is to reveal hidden topics and themes in the Twitter data about the Ukraine-Russia conflict through the application of two topic modeling methods which are LDA and BERTopic. The entire study process as shown in figure 1 is described in this part. Our procedure for extracting Twitter data is based on the following three main steps: Data Collection, Data Preprocessing (which involves the quality evaluation of data with the help of BERTopic), and Topic Modeling. The diagram in Figure 1 clearly explains the connection and execution of these steps within our data analysis process. Using these techniques, we are expecting to acquire significant understanding regarding the topics of the conversations of people in the Arab region about the Ukraine-Russia conflict.

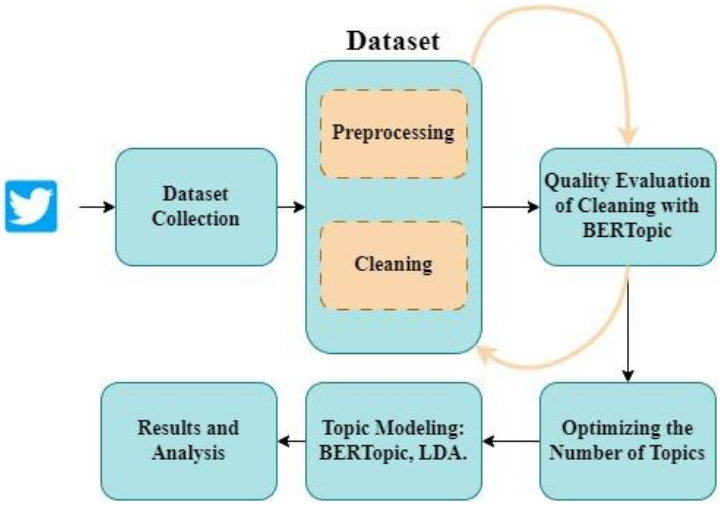

Figure 1.
The Workflow.

3.1. Data Collection

Twitter is a well-known site for microblogging and was chosen as the main source for data extraction. We employed the "snscrape" library in Python to obtain 45,486 tweets in Arabic regarding the war between Ukraine and Russia that were available to the public. All these tweets were made on Twitter from the 1st of February, 2022, to the 14th of December, 2022. For gathering the dataset, the widely accepted Twitter keywords related to the war, which were: war, Russia, and Ukraine (initially written in Arabic but now translated here into English), were utilized. To get the tweets containing all the keywords, these keywords were linked with the help of the "and" operator.

3.2. Data Preprocessing

Data preprocessing is a critical step in text mining and it works as a preliminary stage with two purposes. First, it increases the performance of the prediction algorithms by removing the words that might be a hindrance to the process. Second, it helps to reduce the requirements for storage space, which in turn leads to better computational performance, as per [16]. Data preprocessing is an integral part of our analysis because of the peculiarities involved in the mining of dialectical Arabic texts. It is, therefore, crucial to undertake comprehensive preprocessing so that the dataset can meet our stringent standards. Data preprocessing went through a number of major steps to prepare the data for analysis. Duplicate tweets were first dropped. Next, hyperlinks, usernames, redundant spaces, and emojis were stripped from the tweet content; after that, English letters, numbers, some kinds of punctuation, and some special characters (like "$" and "%") were stripped away. Then, processing steps specific to Arabic were implemented: removing diacritics, resolving tatweel or letter elongation, and normalizing a certain set of letters. They also removed keywords for scraping tweets and their morphological variants: War, Russian, Russia, Ukrainian, and Ukraine. Also removed were words found in the stop-words list given by the Python NLTK library [17] such as: Indeed, He, How, Where, Except, and that which, as they usually do not carry much meaning. The term Urgent was also removed because that word appeared too often in tweets that either contained news or were retweets of news.

Stemming plays a cardinal role in text preprocessing, whose operationalization is to reduce words to their roots, and the Farasa stemmer [18] which has proven to be highly precise, was utilized. Finally, after applying the topic modeling to the datasets resulting from the complete preprocessing, the domain-specific stop words, i.e., those words that frequently occur, don't convey much meaning, and would adversely affect the performance of the topic-modeling algorithms, were detected. It is worth noting that these words vary from one dialect of Arabic to another. To resolve this issue and improve the quality of the dataset, we opted to use BERTopic, as it determines its topics without a fixed number. The method obtained 300 topics; they were reviewed in detail by two domain experts who selected the domain-specific words. The datasets were subsequently cleaned of these words. Examples of the words removed are: This, That, To, Not, Now, Like this, This, What, if, That, Now. After the cleaning stages, the datasets were reduced to 44,583 tweets prepared for analytical processes.

3.3. Topic Modeling

Topic modeling has become an effective method for revealing underlying themes in large amount of text, presenting a comprehensive view of data by probabilistically discovering latent topics that are distributed in words. In this section, we present the two topic modeling methods used in our research:

3.3.1. Latent Dirichlet Allocation

Among various topic modeling techniques, Latent Dirichlet Allocation (LDA) is possibly the most widely known and widely used model. Proposed by Blei, et al. [19] in 2003, LDA is a generative probabilistic model which can infer semantic or hidden (latent) features from a collection of text. Latent Dirichlet Allocation (LDA) is one of the popular and widely used methods for topic modeling. LDA operates on the foundation that a document can be viewed as a bag of words, with

words originating from various topics [19]. Consequently, the model samples the probability distribution over words for each topic and the probability distribution over topics for each document. LDA offers control over these two probabilities via two parameters, denoted as **β** and **α,** respectively [20]. Importantly, LDA is categorized as an unsupervised learning algorithm, meaning that it does not rely on predefined words. Once the number of topics is determined, each topic is assigned a distinct class label.

3.3.2. BERTopic

Recently, a relatively new topic modeling algorithm has been gaining attention. Unlike traditional topic modeling methods such as LDA, BERTopic employs pre-trained language models, especially BERT (Bidirectional Encoder Representations from Transformers) [21] to extract topics from text data. As an advanced language model, BERT has been trained on huge amount of textual data and has shown exemplary performance considering several natural language processes. One prominent advantage of BERTopic is maintaining contextual information and semantic meaning of words in a document. By using contextual embeddings provided by BERT, BERTopic can identify the topics in a much better and effective way [22]. Additionally, since the number of topics is not specified beforehand, BERTopic presents much more flexibility as compared to traditional approaches in topic modeling.

4. Experimental Setup

In this section, we explore the settings and configurations of the topic modeling algorithms we employed, along with our methodology for determining the optimal number of topics.

4.1. Parameter Settings

In Table 1 we provide an overview of the parameter settings for the two algorithms utilized. As previously discussed, the BERTopic model operates without predefining the number of topics. However, not specifying the number of topics led to a high volume of topics, exceeding 300, which exhibited substantial similarity. In our research, we implemented the BERTopic model with word embeddings sourced from AraBERTv0.2. Notably, AraBERTv0.2 is a pretrained transformer model designed for Arabic dialects and tweets [23].

Regarding the LDA method, we used the "LatentDirichletAllocation" model from the Python gensim library in this study. The LDA model is well-established as a robust approach to topic modeling, but needs careful consideration of several important parameters to optimize performance. One of the most important factors of LDA is the number of topics to extract from the dataset (K). Röder, et al. [24] proposed a measure of topic coherence that focuses on identifying salient semantic associations between high-probability words within a topic to provide a more meaningful evaluation of topics. In addition to determining the number of topics (K), the "passes" parameter specifies how many passes through the corpus to take during the training procedure and was set to 10 in this study. This is consistent with an iterative process and was based on consideration of fine-tuning the hyperparameters, which was necessary to induce effective convergence in the LDA model. A value less than 10 created potentials for insufficient convergence and thus impact the performance of the LDA model. In addition, the "iteration" parameter in the LDA model was also important to the resultant output of the topics. The "iterations" parameter specifies how many maximum iterations through the corpus when inferring the distribution of topics. Changing this parameter may affect the model's accuracy and rate of convergence. For two other important parameters, β and α, we chose to keep the values set to their defaults. The defaults represent a symmetric distribution and imply that all topics are equally important.

Table 1.
Training Parameter Settings.

Model	Parameters
BERTOPIC	- Language: "Arabic" - Embedding model: AraBERTv0.2 - Number of topics: 22
LDA	- Corpus: bag of word corpus - Number of topics: 22 - id2word: implemented dictionary - Passes: 10 - Iterations: 100 - Alpha and beta: default symmetric distribution

5. Results

5.1. Descriptive Analysis

After preprocessing the raw data, our final dataset consisted of a total of 43,819 tweets. We focused on identifying the most frequently tweeted bigrams related to the Russia–Ukraine war. Bigrams comprise two consecutive words, irrespective of their grammar structure or semantic meaning, and may not be self-explanatory. The applicability of this technique, which has been previously used in Xue, et al. [25] and Xue, et al. [26] to identify the leading topics and trends in Twitter conversations has been proven successful through the use of bigrams. Some of the identified bigrams include NATO, European Union, United States, foreign minister, Soviet Union, European country, America-Europe, nuclear weapons, impose sanction, and America West. Additionally, popular unigrams (single words) related to the conflict include America,

country, West, global, Europe, NATO, said, Putin, economy, and power. An overview of the most frequently occurring unigrams and bigrams associated with the Ukraine–Russia conflict is provided in Table 2, and these insights are further visually represented through word clouds in Figures 2 and 3.

Table 2.
Top 20 unigrams and bigrams and their distributions.

Unigram	Percentage (%)	Bigram	Percentage (%)
America	1.887122	NATO	0.172612
country	1.241229	European Union	0.122856
West	1.069818	United States	0.100179
global	0.894672	foreign minister	0.085772
Europe	0.826775	Soviet Union	0.076035
NATO	0.672038	Europe country	0.074567
said	0.519836	America Europe	0.072433
Putin	0.515834	Europe America	0.067097
economy	0.485820	nuclear weapon	0.063229
power	0.485553	impose punishment	0.058427
income	0.476749	America West	0.056292
China	0.457274	Middle East	0.055359
day	0.457007	West country	0.054158
weapon	0.447269	national security	0.046955
president	0.440600	Arab country	0.046021
intend	0.410053	West America	0.046021
military	0.392178	military operation	0.045221
people	0.388443	China Taiwan	0.045087
Arabs	0.370301	prime minister	0.044420
desire	0.327482	Security Council	0.044020

Figure 2.
The word cloud of the most popular unigram.

Figure 3.
The word cloud of the most popular bigram.

5.2. Russia-Ukraine War–Related Topics

5.2.1. Russia-Ukraine War–Related Topics Using LDA

As previously mentioned, we utilized the coherence metric as our guiding principle to determine the optimal number of topics for LDA and BERTopic. This metric provides understanding not only about the interpretability of the topics but also about the words' closeness in each topic's top N words [27]. The C_npmi measure was used in our research for coherence. The C_npmi measure is based on sliding windows of words and simultaneously computes mutual information between all possible word pairs but only considers the most frequent ones [28]. To what extent coherence and number of topics were interconnected, we generated the coherence plots for the entire range of topic numbers from 2 to 50 and found the one with the highest coherence score. The plots showing the development of the coherence values of both LDA and BERTopic are shown in Figures 4 and 5 respectively. According to the data presented in Figures 4 and 5, the number of topics, giving the highest score in terms of coherence differed for the two models. The BERT model determined the number to be 22, while the LDA model confirmed it to be 10. To address this difference, we applied two different metrics from the R package, CaoJuan2009 and Deveaud2014, to make our decision for the number of topics [29]. We set our goal to minimize the CaoJuan2009 measures and on the other hand, maximize the Deveaud2014 measures.

It should be mentioned that the CaoJuan2009 measure normally declines with the rise in number of topics while the Deveaud2014 measure does the opposite and increases with the number of topics going up. Thus, we were looking for a point when the metrics became stable which would indicate that there was no change in the outcome. After the scores were evaluated, a very interesting pattern was observed. The CaoJuan2009 score reached its lowest point with 14 topics, whereas at the same time the Deveaud2014 score was at its highest point. The agreement of the two measures was so strong that we were able to conclude without any doubt that the best number of topics, referred to as 'k', is actually 14. The clarity of this discovery is artistically displayed in Figure 6, which illustrates the results obtained from both measures.

Since initially LDA and BERTopic produced different numbers of topics based on coherence metric and CaoJuan2009 and Deveaud2014 metrics, we opted for empirical experimentation to determine the ideal number of topics. In that process, we threw three numbers 10, 14, and 22 as possible numbers of topics to human evaluation about topic interpretability.

In the end, after thorough considerations, we found the number 22 as the best choice. These findings are indicative of the excellent performance of BERTopic. In the coherence metric, the highest coherence score for BERTopic was recorded at 22 topics, which was more than LDA's score. This means that at the chosen optimal number, BERTopic produced topics that were more coherent and interpretable, thus, validating its efficiency in our analysis. To provide strong evidence for the selection of 22 as the optimal number of topics in the LDA model, we carried out an additional analysis. We computed the topic distance and plotted the intertopic distance in a 2D plane, as stipulated in Chuang, et al. [30]. Figures 6 and 7 exhibit the results, where each circle on the plot represents a certain topic, either Topic 1 to Topic 14 or 22 in this research. The positions of these circles were determined based on the calculated distances between the topics. An important observation is that in the visualization, when employing 22 topics, the circles did not overlap, signifying the emergence of new topics. This absence of overlap served as a validation of the appropriateness of the 22-topic configuration.

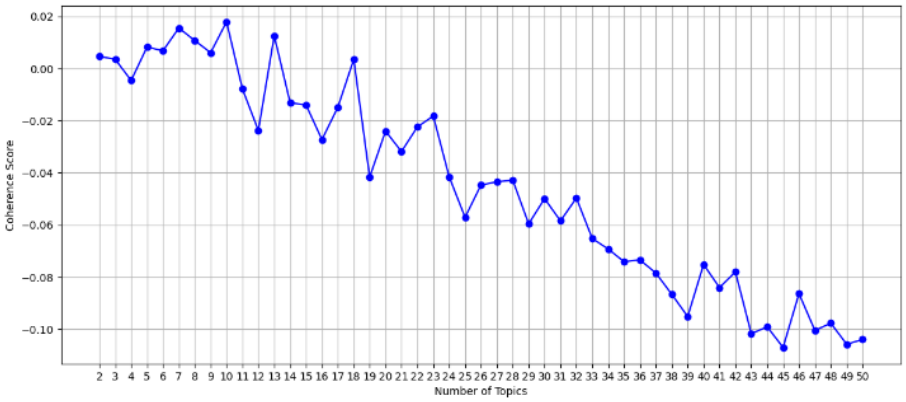

Figure 4.
LDA – coherence.

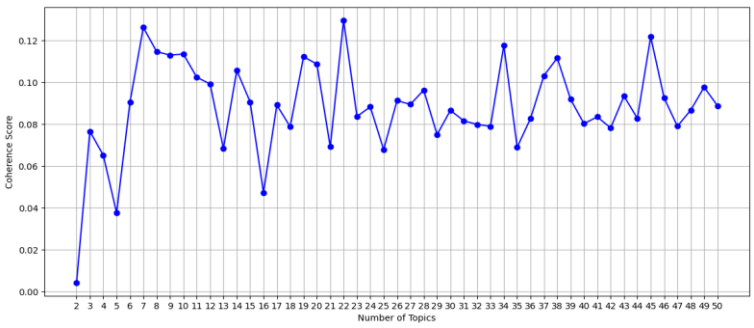

Figure 5.
BERT – coherence.

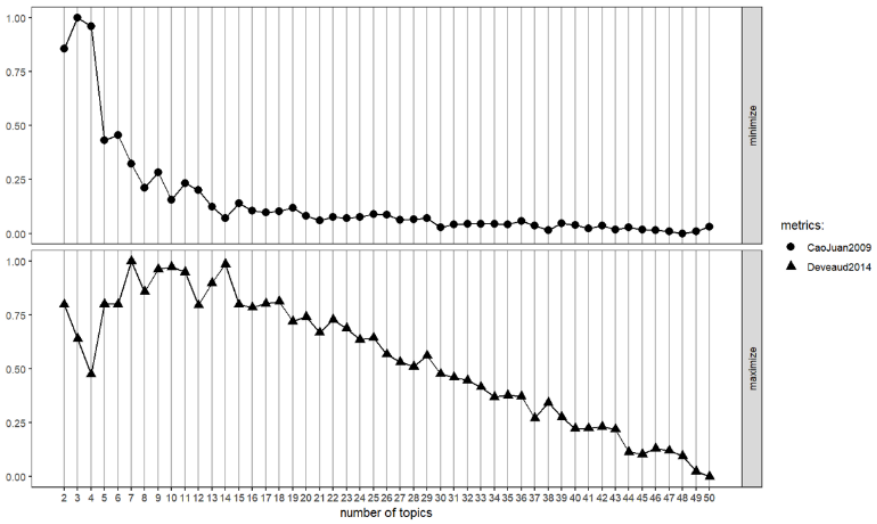

Figure 6.
Metrics for estimation of the best fitting number of topics for 5 to 50 topics.

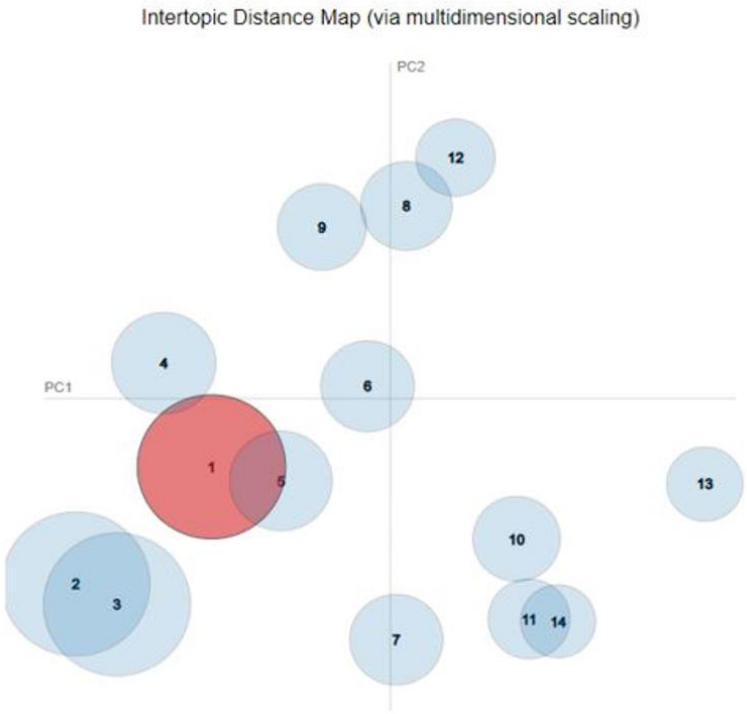

Figure 7.
LDA - Intertopic Distance for k=14.

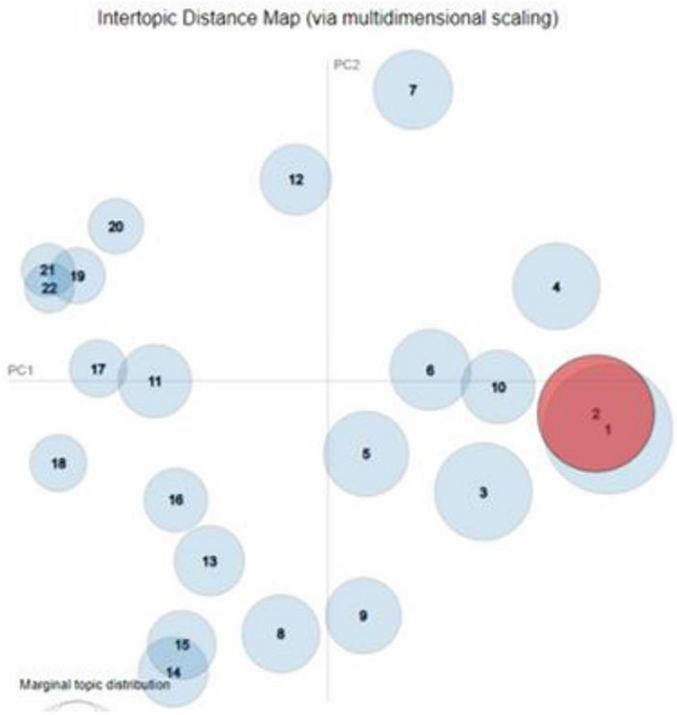

Figure 8.
LDA - Intertopic Distance for k=22.

5.2.2. BERTopic

The 22 topics generated by the BERTopic are presented in the Table 3. The top ten words per topic are presented. Generally speaking, the topics are very coherent, and they relate to a specific subject matter. Good coherence within the topic words ensures that the algorithm works well. This can be due to the fact that the used word embedding takes into account the dialects and nuances of the Arabic language and the nature of tweets. For example, the eighth topic concerns the oil and energy market. This is consistent with the fact that wars can disrupt the production, transportation, and distribution of oil and other energy sources, leading to a decrease in supply and an increase in prices. In topic 12, words are related to the wheat market. The Russian-Ukraine war can disrupt the global wheat market in the region due to the fact that Ukraine is one of the major wheat-producing countries in the world. Topic 2 related to the complex political situation in Syria, in which Russia has played a significant role.

Table 3.
BERT- Top words for each topic - Translated into English.

Topic Number	Topic Words in English
-1	America, country, west, global, Europe, NATO, economy, power, weapon, said
0	Global, knew, upset, solution, you, doesn't know, cup, said, club, entered
1	America, Iraq, country, Yemen, west, China, Turkey, Arab, Europe, power
2	Putin, Biden, president, NATO, America, Soviet, west, union, Nuclear, Europe
3	Syria, Lebanon, Bashar, will, criminal, Assad, people, Syria, Israel, Putin
4	Crime, NATO, minister, president, public, foreign affairs, France, Europe, alliance, committed
5	Iran, march, foreign, use, response, Tehran, confirm, minister, utilized, weapon
6	Attack, Libya, said, Haftar, now, Tripoli, Libyan, wants, like this, sheep
7	Power, series, part, army, NATO, Blinken, used, weapon, beginning, nuclear
8	Oil, price, company, OPEC, market, share, barrel, increase, decrease, production
9	Kuwait, Saddam, Kuwaiti, parliament, National, Assembly, Iraq, government, what said, session
10	Morocco, of, Algeria, to, go, will, Qais, country, Tunisia, Happy
11	Remained, want, global, another, people, know, more, said, a lot, subject
12	Price, Egypt, wheat, America, destroyed, happened, life, global, know, country
13	Avoidance, ambassador, London, joining, resignation, NATO, alliance, request, Chinese, transformation
14	Yedioth Ahronoth, Israeli, taking, Hebrew, neutral, pressure, leadership, stance, government
15	Church, Catholic, Orthodox, religious, patriarch, cross, Christian, Orthodox, denomination
16	Salman, king, new, event, son, crown prince, Khalid, video, prince
17	Arrested, released, Saudi, mediation, Reuters, prisoner, covenant, crown prince, captive, Salman
18	Prisoner, exchange, battalion, Azov Battalion, commander, announced, fighter, military, sacrifice, information

19	Invitation, lesson, Muslim, officer, Fajr (dawn), volunteered, championship, champion, foreign, terrorist
20	British, anchored, port, ship, transport, requested, prevention, minister, debate, eastern

Table 4.
LDA- Top words for each topic - Translated into English.

Topic	Top Words in English
0	International, decision, policy, law, tension, sons, issue, imminent, community, series, Ukrainian, word, legal, happening, period, relationship, relation, tomorrow
1	Will, man, said, news, people, studied, played, blood, situation, benefited, laughed, mind, understood, sure, account, good, women, knew, hate, million
2	President, minister, launch, foreign affairs, America, defense, name, council, invasion, said, attack, day, official, statement, security, crime, British, Europe, accused, information
3	Entered, global, said, day, occupied, nothing, enabled, invaded, want, year, words, Iraq, Kuwait, Black, Sea, subject, time, sympathy, month, will not go
4	Gas, crisis, price, country, oil, wheat, economy, global, Europe, market, increase, raised, source, largest, increased, energy, impact, oil, prices, significant
5	Weapon, army, nuclear, soldier, force, missile, country, military, strike, oppressor, used, entered, aircraft, land, sent, strongest, use, mobilization, destroyed, defense
6	Yemen, citizen, homeland, stance, country, order, day, departure, remained, Saudi Arabia, people, Houthi, came, message, government, defense, number, greater, reached, aggression
7	NATO, alliance, country, entered, Putin, joining, NATO, global, annexation, intervention, entry, Crimea, military, requested, non, status, direct, general, joined, threat
8	Global, began, new, nuclear, expected, event, outbreak, ended, system, day, situation, erupted, end, matter, beginning, possibility, subsequent, change, situation, current
9	News, based, happened, country, work, house, life, escape, tried, like this, political, followed, concerned, security, his work, tried, altercation, beating, system
10	West, media, son, herding, revealed, event, lie, image, truth, channel, spoke, guardian, racism, right, means, freedom, appeared, dissemination, the world
11	Egypt, people, said, issue, know, a little, coronavirus, must, happened, work, become, together, want, work, original, enough, become, day, where, finished
12	China, America, global, Taiwan, ally, Korea, Chinese, alliance, entered, West, northern, power, withdrew, country, front, historical, broke out, pole, new
13	Syria, Iraq, criminal, bombardment, people, city, killed, crime, Putin, Libya, Lebanon, Yemen, child, invasion, destroyed, victim, Afghanistan, committed, terrorism, right
14	America, Europe, country, union, will, west, entered, interest, Soviet, president, ally, defend, Europe, people, support, fought, entry, largest, Afghanistan, force
15	Arabs, Muslims, West, country, Israel, Palestine, land, Jewish, people, Zionist, peace, evil, hope, entered, Islam, occupation, people, stand, Islamic, innocent
16	Separation, act, exposure, witnessed, condemned, once, invaded, caller, lacking, possibilities, stopped, what, immediately, appeared, edge, apparent, I don't know (peace), percentage, amateur, consideration
17	West, economy, sanction, America, military, will, Putin, attrition, goal, imposed, force, believed, country, loss, political, managed, endure, long, support, act
18	East, known, Biden, solution, Middle East, region, retreat, hour, horror, background, south, during, start, you, Donetsk, upset, barrel, will be, obstacle
19	Security, global, third, national, victory, project, viewed, party, Cuba, named, invasion, group, produced, hero, discourse, nation, face, theoretical, declaration, council
20	Turkey, gang, street, fighting, army, mercenary, encountered, imagination, training, I can, intention, force, aggression, spirit, Sudan, Syria, great, alive, fighter, story
21	United, Europe, Germany, state, France, president, military, force, announcement, America, aerial, operation, Washington, announced, house, ambassador, white, ban, Britain, situation

5.2.3. LDA

In Table 4, we outline the results of the 22 LDA topics, illustrating the most frequently appearing words within each topic.

The LDA and BERTopic algorithms both generated similar topics, such as those related to Food, Oil, Egypt, and Syria. However, the topics produced by BERTopic had better coherence. This can be due to the fact that the word embeddings used by BERTopic are considered the dialect of the Arabic language.

5.3. Russia-Ukraine War–9–Related Themes

To strengthen the reliability of our findings obtained through the BERTopic and LDA model, we integrated a qualitative method focused on gaining a more profound insight into the identified themes. In particular, we followed the

established six-step thematic analysis framework outlined by Braun and Clarke [31]. This thematic methodology centered on human interpretation, recognizing that it could be influenced by individual comprehension of the themes and potential biases.

Using thematic analysis, we organized the identified topics into separate themes. The 22 topics have been grouped into nine main themes, including:

- International Relations and Diplomacy ("International Involvement"): This theme encompasses topics related to several key players and regions in the context of international relations. Notably, it includes the involvement of the United States, Iraq, Yemen, China, Turkey, the Arab world, Europe, and the NATO alliance. Discussions center around a spectrum of issues, such as China's growing global influence, the intricacies of power dynamics and alliances among these nations, and the pivotal role of NATO in global security. Additionally, concerns surrounding Russian actions and their implications for international relations are a significant aspect of these discussions.
- Comparing Past Conflicts in the Middle East: The Russian-Ukrainian conflict has been compared to previous conflicts in the Middle East, such as the Syrian Civil War and the Gulf War. Commonly, words such as "Iraq," "Kuwait," "Saddam," "Syria," and "Bashar" emerge frequently in discussions related to this theme.
- Conflict and Military Operations: This theme focuses on various military strategies, maneuvers, and operations carried out by the involved parties. For instance, this theme discussed the Battle of Donetsk Airport.
- Economic Consequences: this theme focuses on the financial and economic impacts caused by the conflict. It examines the repercussions on the economies of the involved nations, as well as the global economic landscape. Notably, certain countries heavily rely on purchasing large amounts of wheat and oil from Ukraine and Russia.
- Nuclear Warfare Concerns: This theme encompasses discussions regarding the potential of nuclear war as a consequence of the Russian-Ukrainian conflict and people expressing worries about the prospect of another war.
- Media and Information Warfare: This theme revolves around the aspects of media engagement and information warfare in the context of the Russia-Ukraine war, highlighting the significance of propaganda and information dissemination and their influence on public opinion and perceptions.
- Humanitarian Impact and Displacement: this theme, as identified within the LDA analysis, focuses on evaluating the impact of the Russia-Ukraine conflict on humanitarian conditions, including issues of displacement, refugee crises, and associated challenges.
- Faith Amidst Conflict: this theme explores religious practices and beliefs during times of war. The theme also includes expressions of gratitude and good wishes for nations and people.
- Geopolitical Influences on International Sports "The Russia-Ukraine Conflict and FIFA's World Cup Dilemma": this theme revolves around the examination of how the Russia-Ukraine conflict influences international sports, particularly within the context of FIFA's World Cup.

In Table 5, the themes and the corresponding topics of each model are presented. Based on the table, the dominant themes are International Relations and Diplomacy, Comparing Past Conflicts in the Middle East, Literature Conflict and Military Operations and Economic Consequences.

Table 5.
Theme and Topic Numbers.

Theme	Topic Number from BERT	Topic Number from LDA
International Relations and Diplomacy	-1, 1, 2, 4, 13	0, 2, 7, 12, 14, 21
Comparing Past Conflicts in the Middle East	3, 9	3, 6, 15
Literature Conflict and Military Operations	6, 7, 18	5, 16, 18, 20
Economic Consequences	8, 12, 20	4, 17
Nuclear Warfare Concerns	11	8, 19
Media and Information Warfare	5, 14, 16, 17	1, 9, 10
Humanitarian Impact and Displacement	-	11, 13
Faith Amidst Conflict	15, 19	-
Geopolitical Influences on International Sports	0, 10	-

6. Conclusions

In this paper, we have applied LDA and BERTopic, which are two advanced topic modeling methods, to examine the perspectives of Arab native speakers regarding the Russian-Ukrainian war. The viewpoints of the Arab-speaking community were obtained from a Twitter dataset that was created using relevant keywords. Different metrics helped to fix the number of topics for both LDA and BERTopic, thus ensuring that our topic extraction was relevant and interpretable. The application of this methodology made it possible for us to uncover the topics and themes in the data that were not only insightful but also well-defined, thus providing a thorough analysis of the conflict from the perspective of the Arab world. The analysis revealed a complex set of opinions and debates, thus highlighting the diversity and complexity of the views held by Arab-speaking populations. One of the major themes was the international involvement, which dealt with the discussion of the main global players and the Arab world's place in the geopolitical scene. The reference to the past Middle East conflicts, on the other hand, sheds more light on the understanding of the regional conflict. The economic

consequences, especially the dependence on Ukrainian and Russian resources and the fear of nuclear warfare are other findings of great significance.

References

[1] A. Farghaly and K. Shaalan, "Arabic natural language processing: Challenges and solutions," *ACM Transactions on Asian Language Information Processing (TALIP)*, vol. 8, no. 4, pp. 1-22, 2009.

[2] M. Mujahid *et al.*, "Sentiment analysis and topic modeling on tweets about online education during COVID-19," *Applied Sciences*, vol. 11, no. 18, p. 8438, 2021. https://doi.org/10.3390/app11188438

[3] H. Zankadi, A. Idrissi, N. Daoudi, and I. Hilal, "Identifying learners' topical interests from social media content to enrich their course preferences in MOOCs using topic modeling and NLP techniques," *Education and Information Technologies*, vol. 28, no. 5, pp. 5567-5584, 2023. https://doi.org/10.1007/s10639-022-11373-1

[4] R. Silveira, C. G. Fernandes, J. Araujo Monteiro Neto, V. Furtado, and J. E. Pimentel Filho, "Topic modelling of legal documents via legal-bert," *Topic Modelling of Legal Documents via LEGAL-BERT*, vol. 1613, p. 73, 2021.

[5] J. A. Lossio-Ventura, S. Gonzales, J. Morzan, H. Alatrista-Salas, T. Hernandez-Boussard, and J. Bian, "Evaluation of clustering and topic modeling methods over health-related tweets and emails," *Artificial Intelligence in Medicine*, vol. 117, p. 102096, 2021. https://doi.org/10.1016/j.artmed.2021.102096

[6] B. Dahal, S. A. P. Kumar, and Z. Li, "Topic modeling and sentiment analysis of global climate change tweets," *Social Network Analysis and Mining*, vol. 9, no. 1, p. 24, 2019. https://doi.org/10.1007/s13278-019-0568-8

[7] X. Jin and P. R. Spence, "Understanding crisis communication on social media with CERC: topic model analysis of tweets about Hurricane Maria," *Journal of Risk Research*, vol. 24, no. 10, pp. 1266-1287, 2021. https://doi.org/10.1080/13669877.2020.1848901

[8] S. K. Akpatsa *et al.*, "Sentiment analysis and topic modeling of Twitter data: A text mining approach to the US-Afghan war crisis," *Available at SSRN 4064560*, 2022.

[9] S. Aslan, "MF-cnn-bilstm: A deep learning-based sentiment analysis approach and topic modeling of tweets related to the ukraine-russia conflict," *Available at SSRN 4218398*, 2022.

[10] N. Alami, M. Meknassi, N. En-nahnahi, Y. El Adlouni, and O. Ammor, "Unsupervised neural networks for automatic Arabic text summarization using document clustering and topic modeling," *Expert Systems with Applications*, vol. 172, p. 114652, 2021. https://doi.org/10.1016/j.eswa.2021.114652

[11] B. H. Hammo, H. Abu-Salem, and M. W. Evens, "A hybrid arabic text summarization technique based on text structure and topic identification," *International Journal of Computer Processing of Languages*, vol. 23, no. 01, pp. 39-65, 2011.

[12] H. Alhazmi, "Detection of students' problems in distance education using topic modeling and machine learning," *Future Internet*, vol. 14, no. 6, p. 170, 2022. https://doi.org/10.3390/fi14060170

[13] R. Alshalan, H. Al-Khalifa, D. Alsaeed, H. Al-Baity, and S. Alshalan, "Detection of hate speech in covid-19–related tweets in the arab region: Deep learning and topic modeling approach," *Journal of Medical Internet Research*, vol. 22, no. 12, p. e22609, 2020.

[14] N. Habbat, H. Anoun, and L. Hassouni, *Topic modeling and sentiment analysis with LDA and NMF on moroccan tweets*. Cham: Springer International Publishing, 2021, pp. 147-161.

[15] A. Abuzayed and H. Al-Khalifa, "BERT for Arabic topic modeling: An experimental study on BERTopic technique," *Procedia Computer Science*, vol. 189, pp. 191-194, 2021.

[16] F. Alshanik, A. Apon, A. Herzog, I. Safro, and J. Sybrandt, "Accelerating text mining using domain-specific stop word lists," presented at the IEEE International Conference on Big Data (big data), 2020.

[17] S. Bird, E. Klein, and E. Loper, *Natural language processing with Python: analyzing text with the natural language toolkit*. Sebastopol, CA: O'Reilly Media, Inc, 2009.

[18] A. Abdelali, K. Darwish, N. Durrani, and H. Mubarak, "Farasa: A fast and furious segmenter for arabic," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, 2016, pp. 11–16.

[19] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, no. Jan, pp. 993-1022, 2003.

[20] U. Chauhan and A. Shah, "Topic modeling using latent Dirichlet allocation: A survey," *ACM Computing Surveys*, vol. 54, no. 7, pp. 1-35, 2021.

[21] J. D. M.-W. C. Kenton and L. K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of naacL-HLT*, 2019, vol. 1, no. 2.

[22] M. Grootendorst, "BERTopic: Neural topic modeling with a class-based TF-IDF procedure," *arXiv preprint arXiv:2203.05794*, 2022.

[23] W. Antoun, F. Baly, and H. Hajj, "AraBERT: Transformer-based model for arabic language understanding," presented at the LREC 2020 Workshop Language Resources and Evaluation Conference 11–16 May 2020, p. 9, 2020.

[24] M. Röder, A. Both, and A. Hinneburg, "Exploring the space of topic coherence measures," in *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, 2015, pp. 399-408.

[25] J. Xue, J. Chen, and R. Gelles, "Using data mining techniques to examine domestic violence topics on Twitter," *Violence and Gender*, vol. 6, no. 2, pp. 105-114, 2019.

[26] J. Xue *et al.*, "Twitter discussions and emotions about the COVID-19 pandemic: Machine learning approach," *Journal of Medical Internet Research*, vol. 22, no. 11, p. e20550, 2020.

[27] E. Zvornicanin, *What is yolo algorithm?| baeldung on computer science*. London, UK: Baeldung, 2022.

[28] S. Kapadia, "Evaluate topic models: Latent dirichlet allocation (LDA)," *Towards Data Science*, 2019.

[29] M. Nikita and M. M. Nikita, "ldatuning: Tuning of the Latent dirichlet allocation models parameters," 2016.

[30] J. Chuang, D. Ramage, C. Manning, and J. Heer, "Interpretation and trust: Designing model-driven visualizations for text analysis," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2012, pp. 443-452.

[31] V. Braun and V. Clarke, "Using thematic analysis in psychology," *Qualitative Research in Psychology*, vol. 3, no. 2, pp. 77-101, 2006. https://doi.org/10.1191/1478088706qp063oa