Bitlis Eren Üniversitesi Fen Bilimleri Dergisi
BİTLİS EREN UNIVERSITY JOURNAL OF SCIENCE
ISSN: 2147-3129/e-ISSN: 2147-3188
VOLUME: 12 NO: 1 PAGE: 272-291 YEAR: 2023
DOI:10.17798/bitlisfen.1238241

Russia-Ukraine Conflict: A Text Mining Approach through Twitter

İbrahim Miraç ELİGÜZEL[1*]

[1] *Gaziantep University, Industrial Engineering, 27310 Gaziantep, Turkey*
(ORCID:0000-0003-3105-9438)

Keywords: Fuzzy c-means, Russia, Ukraine, War, Topic modeling, Twitter

Abstract
The focus of this study is to use social media to investigate the Russia-Ukraine conflict. With the assent of the Russian parliament, Russian President Vladimir Putin proclaimed that they will begin invading Ukraine on February 24, 2022. During the Russia-Ukraine conflict, social media, particularly Twitter, has been heavily used. For that reason, it becomes to strong tool for handling processes during the conflict such as political decision making, organizing humanitarian activities, and proving assistance for victims. As a result, social media becomes the most up-to-date, comprehensive, and large information source for current scenario analysis. A total of 65412 tweets are gathered as a dataset for analysis in the proposed study between February 24 and April 5. Then, for each tweet, a topic modeling method called Latent Dirichlet Allocation (LDA) is used to collect significant topics and their probabilities considering each tweets. Then, using the specified probabilities, Fuzzy c-means is utilized to generate clusters for the entire document. Finally, seven unique clusters have been gathered for processing. N-grams and network analysis are used to examine each resulting cluster for a better understanding. As a result of this study, worldwide public opinion, current situation of civilians, course of the conflict, humanitarian issues during the Russia-Ukraine conflict are extracted.

1. Introduction

Following the resignation of Ukraine's then-President, pro-Russian Viktor Yanukovych, in February 2014, Russia annexed Crimea and increased its support for pro-Russian separatists in the Russian-speaking Donbas region in the country's east [1]. Separatists in Donbas declared Donetsk and Luhansk People's Republics and seized state institutions shortly after. This action triggered an armed conflict between the Ukrainian government and separatists [2]. While Russia is not officially involved in the conflict in Ukraine, experts, international organizations, the media, and the Ukrainian government have all stressed Russia's backing for rebels since 2014. Recently, with Russia's recognition of the independence of separatist forces in eastern Ukraine on February 21, tensions in the region rose and Russia invaded Ukraine on February 24. As the war continues to contribute to the ongoing humanitarian and refugee catastrophe in Ukraine, the usage of social media by both sides of the conflict has produced a variety of responses.

The advancement of Web 2.0 technology has made it possible for individuals without programming skills to create content on the Internet [3]. With the social media applications built around the idea of Web 2.0, it has become very easy to reach social or cultural phenomena instantly. Data has been generated by users and shared as a result of the widespread use of available software and hardware to access social media platforms through the Internet. In this way, users have become accustomed to receiving regular updates on major personal or worldwide events. As a result, social media platforms like Twitter and Facebook have become increasingly popular as communication tools around the world. Twitter is the first platform that springs to mind for disseminating information in real-time crisis situations.

User-generated data collected from social media refers to unstructured data such as text, images, and videos. As a result of social media shares

[*]Corresponding author: *miraceliguzel@gmail.com*
Received: 17.01.2023, Accepted: 03.03.2023

of many users, huge amounts of data emerge in short time periods. To give an example from Twitter, posts, comments, likes constitute big data. Big data collected from social media is useless unless it is used to drive decision making by converting massive amounts of social data into meaningful information [4]. The fact that the data is large and contains many-to-many interactions, unlike traditional media, has significant potential for researchers. With the popularity of data-driven systems, the analysis of social networks gains importance in understanding various social phenomena.

One of the processes converting data into information, topic modeling [5] has been applied in many social media platforms. It is an attractive tool for detecting hidden text patterns in content. It aids in determining the relevance of a topic based on how often it is mentioned and how it is related with other topics. Topic modeling is important, both from the perspective of individuals, as well as, analyzing the public opinion about ongoing conflicts at local and international levels. Without a prepared text data set with established schemes, the topic modeling can help discover and explain broad subjects of interest on social media. The potential of this method to uncover hidden subjects or patterns in text data on its own without supervision is what makes it popular among scholars in a wide range of areas [6-8].

Russia's attempt to invade Ukraine, as in many social events, attracted great attention in the social media in a short time. This study examines Russia's attempt to invade Ukraine, which is one of the current global issues, through public opinions in tweets. In this regard, this study proposes using machine learning techniques on Twitter data containing tweets related to Russia-Ukraine war. Following contributions are made in the current study: i) a twitter dataset is collected containing #UkraineRussiawar hashtag, ii) topic extraction is performed using LDA, iii) extracted topics are clustered with Fuzzy c-means, another machine learning technique, and iv) clusters are deciphered by using n-gram technique.

The contribution of the proposed paper is to provide a general view of society from different perspectives on the war between Ukraine and Russia via social media analysis. In addition, the proposed study is the first one that analyzes the attitude of society through Twitter by utilizing topic modelling and Fuzzy c-means clustering technique with an n-gram analysis approach. Therefore, it aims to provide a basis for decision processes such as political decisions, humanitarian activities, and effective and efficient support for victims.

The remaining part of the study is organized as follows. Section 2 outlines the extant literature on the topic modeling and its applications on social networks. Section 3 explains the integration of the methodologies of topic modeling, Fuzzy c-means and n-grams. Our experiment is conducted and their results are discussed in Section 4. The paper is concluded with Section 5.

2. Literature Review

Several models for interpreting text data are offered by machine learning-based text analysis. Topic modeling can help identify and explain general interest topics in social media without a textual dataset with predefined schemas. Several works use topic modeling methodologies such as LDA, LSA (Latent Semantic Analysis), and their extensions, since topic modeling has the ability to derive key characteristics of particular topics. Among other approaches, LDA has become a standard tool as it is frequently preferred in topic modeling [9, 10].

During the last decade a considerable amount of literature has been published on topic modeling [11-14]. The development of social media platforms and the ease of access to data also accelerated its pace [15, 16]. In particular, what the data say about social phenomena was frequently examined. Vazquez et al. [17] analyzed online news about Venezuela migration for 4 years period. They applied topic modeling to the news that were decided to be related to migration with the binary classifier. They found that the factors that cause migration are unemployment, medicine and food shortages. Tang et al. [18] grouped the construction industry into four clusters and compared how the industry was perceived in terms of workers, companies, unions and the media over 3200 most recent tweets. Using the ability of text mining, especially in social issues, the view of society can also be revealed. For example, Lee and Jang [19] analyzed how the 2021 Atlanta shooting ignited debates. They explored the emergent topics of Twitter from the first 7 days' data about #StopAsianHate.

Because topic detection in large documents is too challenging to do manually, topic modeling methods are frequently combined with clustering algorithms. The goal is to decompose a group of objects in such a way that objects in the same cluster are more similar to each other than objects in other clusters. The center of each cluster is interpreted as topics in topic modeling [20]. K-means is the most common clustering algorithm because its simplicity and efficiency. According to the k-means algorithm, since each object may belong to a cluster, each document also belongs to a subject in topic modeling. However, real-world data may belong to more than one topic. For example, a textual data can be a combination of several topics. Fuzzy c-means (FCM) is a clustering method that groups objects into multiple clusters with a membership degree [21]. Since FCM is more suitable than other clustering algorithms, it has been widely used in literature. Abri and Abri [22] integrated FCM and topic modeling to provide a personalized model for a search engine. FCM results random initializations on each run. To avoid randomness, Alatas et al. [23] proposed non-negative double singular value decomposition (NNDSVD) as the initialization method of FCM in topic modeling. Prakoso et al. [24] used eigenspace-based FCM (EFCM) for conducting the clustering process in low dimensional textual data. Fearing that dimension reduction would reduce accuracy, the authors used a kernel trick to improve accuracy. Parlina et al. [25] also used EFCM to characterize the dimensions of smart sustainable cities over the related literature. Sutrisman and Murfi [26] used NNSVD as the initialization method of EFCM and applied it on Indonesian online news for topic modeling. Trupthi et al. [27] and Mandhula et al. [28] took advantage of possibilistic FCM (PFCM) topic modelling for twitter sentiment analysis and amazon customer's opinion prediction, respectively. Taking it a step further, Kolhe et al. [29] proposed a robust system that integrates LDA and modified grey wolf optimizer. Authors first identified optimal keywords through the superior topic modeling, then clustered them into positive and negative forms through quantum inspired FCM.

Table 1 Comparison of literature papers

Study Reference	*Focus of Study*	*Methods*	*Data*
[17]	Migration crisis in Venezuela	Text classification, word	10000 news articles
[18]	Construction industry	Sentiment analysis, topic	3200 Twitter messages
[19]	Hate crimes, racism, and	Topic modelling	- Twitter messages
[24]	Sensing trending topics	Clustering	- Twitter messages
[25]	Smart sustainable city	Topic detection, topic	Scientific literature
[26]	News analysis	Clustering , Topic detection	Online news
[27]	Twitter sentiment analysis	Topic modeling, sentiment	479 Twitter messages
[28]	Customer's opinion on	Topic modeling, Sentiment	Amazon customer review
Proposed study	Russia-Ukraine Conflict	Topic modeling, clustering,	65412 Twitter messages

The following contributions are made by this study as its shown in Table 1: i) a Twitter dataset including the #UkraineRussiawar hashtag is collected, ii) topics are extracted using LDA, iii) retrieved topics are clustered using Fuzzy c-means, another machine learning technique, and iv) clusters are decoded using the n-gram technique.

3. Material and Method

The core idea of proposed paper is to provide a public opinion against war between Ukraine and Russia. To obtain mentioned aim, LDA topic modeling and Fuzzy c-means clustering technique are integrated. The total of 65412 tweets are gathered between February, 24 and April, 5 by using #UkraineRussiawar hashtag. In order to achieve gathering tweets, Snscrape library in Python software is used. After that, some sequential stages are applied in order to obtain different cluster as seen Figure 1. Each cluster can demonstrate different opinions and views against war. First of all, pre-processing technique is used to provide clean data. The right after, LDA topic modeling is used as feature extraction method to obtain a feature vector. Then, Fuzzy c-means model is applied and various clusters are provided. Lastly, each cluster is analyzed by using n-gram technique.

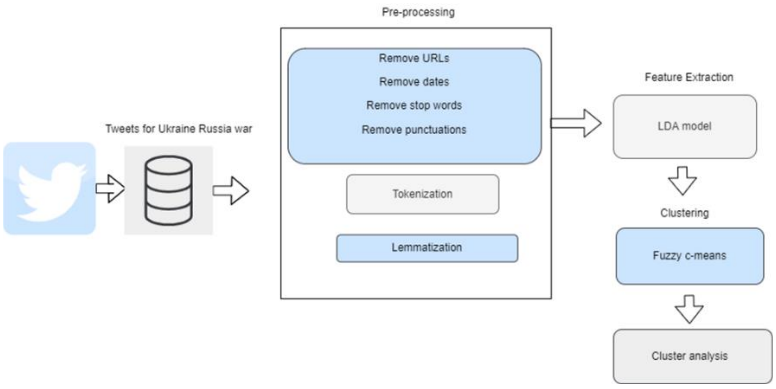

Figure 1. Stages of the proposed study

The gathered tweets are subjected to pre-processing stage. In this level, URLs, dates, stop words, punctuations, and key hashtag (UkraineRussiawar) are removed. After that, tokenization process is implemented. Tokenization is used to split the strings into pieces such as words, symbols, phrases. These pieces are called as token. Tokenization is required before topic process. Lastly, lemmatization is applied in pre-processing stage. Lemmatization is the process of putting together the inflected elements of a word such that they can be identified as a single element, known as the lemma or vocabulary form of the word [30]. After all pre-processing stages are completed, LDA topic modeling is conducted.

3.1. LDA topic modeling

It is called as a generative probabilistic model of a corpus [9]. LDA is utilized to identify underlying topics in a collection of documents and to calculate the probabilities of words in those topics [31]. The key principle is that documents are displayed as random mixtures of latent topics, each of which is described by a word distribution. The structure of LDA model is given in Figure 2, and equation of the probability of a corpus based on this structure is given in Equation (1).

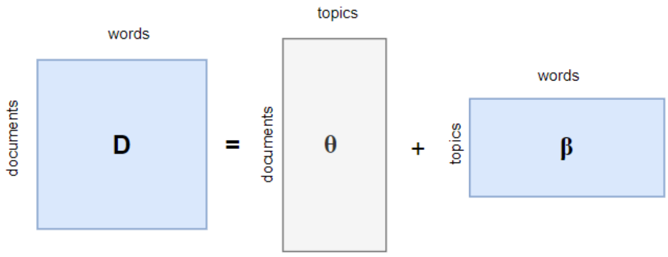

Figure 2. Structure of the LDA model

D refers the document to word matrix. Each document's topic distribution is demonstrated with θ, and each topic-word distribution is demonstrated with β.

$$p(D|\alpha,\beta) = \prod_{d=1}^{M} \int P(\theta_d\,|\alpha)\left(\prod_{n=1}^{N_d} \sum_{z_{dn}} p(z_{dn}\,|\theta_d) p(w_{dn}|z_{dn},\beta)\right) d\theta_d \quad (1)$$

M is the number of documents to analyze, D is the corpus of collection M documents, N is the number of words in the document, α refers Dirichlet-previous concentration parameter of each document topic distribution, β refers corpus level parameter, θ_d refers the document-level variable, z_{dn} refers the topic assignment for w_{dn}, w_{dn} refers the n^{th} word in the d^{th} document. To implement LDA process, MATLAB software is used. The number of documents is takes as 65412 and number of topic is determined according to the validation perplexity as seen in Figure 3.

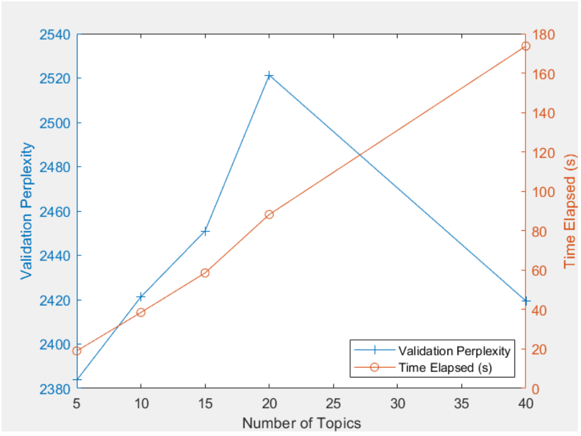

Figure 3. The number of topics for the LDA model

Considering both Figure 3 and the number of topics for text documents in the literature [32], the number of topic for LDA model is considered as 7. Namely, at the end of the LDA process, 65412x7 feature vector is provided. Fuzzy c-means is applied on this vector.

3.2 Fuzzy c-means

Fuzzy c-means clustering technique was first reported by Joe Dunn in 1973 [33], and it was extended by Bezdek [21] in 1984. Fuzzy c-means technique is a popular unsupervised clustering technique. In this technique, objects on the boundaries of many classes are not obliged to fully belong to one of them, but are instead assigned membership degrees ranging from 0 to 1, signifying their partial membership [33]. Fuzzy c-means utilizes fuzzy portioning. The algorithm is demonstrated as follows [33]:

---

1. ***Initialize*** $U=[u_{ij}]$ matrix, $U^{(0)}$
2. At t step: compute the centers vectors $C^{(t)}=[c_j]$ with $U^{(t)}$

$$C_i = \frac{\sum_{j=1}^{n} u_{ij}^{m} xj}{\sum_{j=1}^{n} u_{ij}^{m}}$$

3. Update $U^{(t)}$, $U^{(t+1)}$
4.

$$d_{ij} = \sqrt{\sum_{i=1}^{n}(x_i} - c_i)$$

$$u_{ij} = \frac{1}{\sum_{t=1}^{c}(\frac{d_{ij}}{d_{tj}})^{2/(m-1)}}$$

5. İf $\| U^{(t+1)}-U^{(t)} \| < \varepsilon$ ***then Stop; otherwise return to step 2.***

According to the Fuzzy c-means algorithm; m is a constant real number called as the fuzzifier, u_{ij} refers the membership degree of x_i in cluster j, x_i is the i^{th} of d-dimensional measured data, c_j is the d-dimension center of the cluster. The algorithm considers the distance between cluster centers and data point and memberships are assigned to related each data point. The number of clusters for Fuzzy c-means algorithm is found by using silhouette score. The clustering quality of each data point is measured by constructing a silhouette for that point [34]. The number of the clusters is found as 7. After the clustering process, 7 clusters are analyzed and interpreted. N-gram analysis is used for each cluster.

3.3 N-gram analysis

An "n-gram" is referred as sequence of n words [35]: a 2-gram is called a bigram which is sequence of two words such as "Ukrainian war", "stop war". In the proposed study, after the application of Fuzzy c-means clustering, each cluster is evaluated by utilizing bigram analysis.

4. Results and Discussions

Following the application of the LDA model, seven topics are obtained, as shown in Figure 4. Each topic is made up of six words and represents a theme related to the war. Each topic is given a name based on the words that make up the topic.

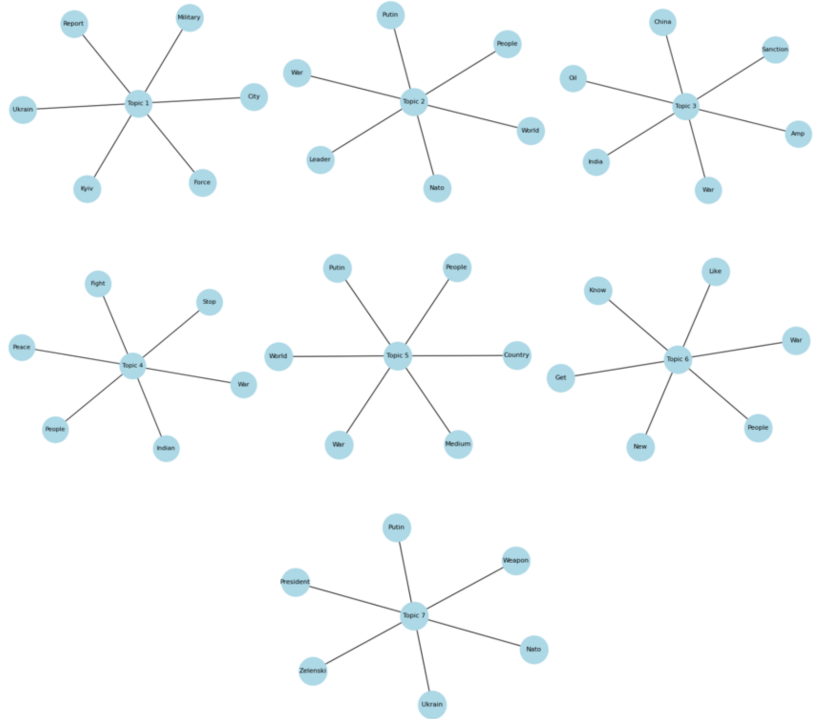

Figure 4. Network visualization of topics

The words Ukraine, city, Kyiv, report, military, and force are all included in the initial topic model. The title of the topic is "Ukraine military force." Words like Putin, leader, conflict, people, world, and NATO appear in the second topic. "Leader Putin" is the title of the second topic. Because of the words China, sanction, India, oil, and war, "Sanction from the World to Russia" can be used as a topic name for the third topic. The most common words in the fourth cluster are peace, people, right, Indian, stop, and conflict. As a result, the fourth topic can be referred to as "end the war." "People in a war" is the title of the fifth topic. The terms people, know, get, news, war, and like appear in Topic 6. The name of topic 6 is "receiving news about the war," as these words illustrate. Finally, by including the words president, Putin, NATO, Ukraine, and Zelensky, topic 7 is referred to as "Putin vs. Zelensky." A matrix of 65412x7 dimensions is generated by examining the seven topics mentioned above. The Fuzzy c-means clustering method is used with the generated matrix. A total of seven clusters is obtained. There are 189, 25, 38471, 2490, 18781, 4803, and 653 tweets in each cluster, correspondingly. The third cluster has the most tweets, with 18781 tweets, and the fifth cluster comes in second with 18781 tweets. There are far fewer tweets in the first and second clusters. The VOSviewer tool is used to evaluate each cluster in the next step. VOSviewer tool demonstrates the connection and number of studies for each selected subject. After that, word frequency and n-gram (bigram) analysis are applied to each cluster. The rest of the discussion section lists all of these applications as well as comments on each cluster.

The following are the applications to the first cluster and analysis: In the first cluster 189 tweets are evaluated. For the evaluation, VOSviewer word frequency and n-gram techniques are utilized. In Figure 5, sub-clusters of cluster 1 are presented through network visualization.

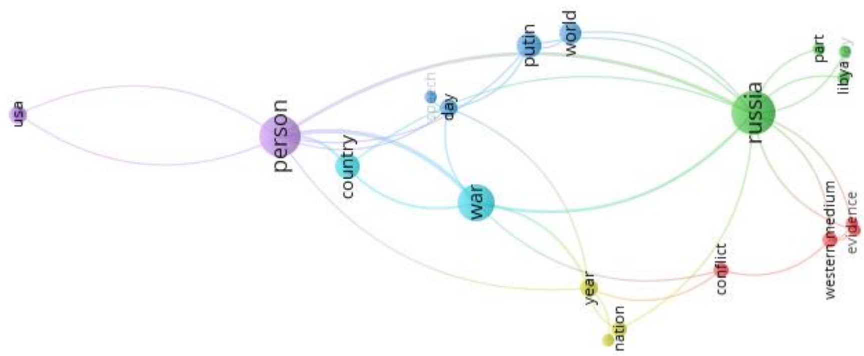

Figure 5. Network visualization of cluster 1 (min. number co-occurrences of a term is 2)

The first cluster is divided into 6 sub-clusters. The red area includes words such as conflict, western media, and evidence. Therefore, it's possible to conclude that there's a lot of uncertainty about the conflict in the Western media. The green area covers words Russia and Libya. Therefore, there can be some political factors between Russia and Libya under the influence of the war. While light blue cluster comprehends words such as country and war, dark blue cluster includes Putin, world, and day words. These areas focus on Russia and president of Russia who is Putin. Lastly, purple cluster includes USA and person words. USA is connected to Putin and Russia. Also, Putin and Russia are connected to Libya. This situation demonstrates that Libya, USA and Russia are focus of first cluster.

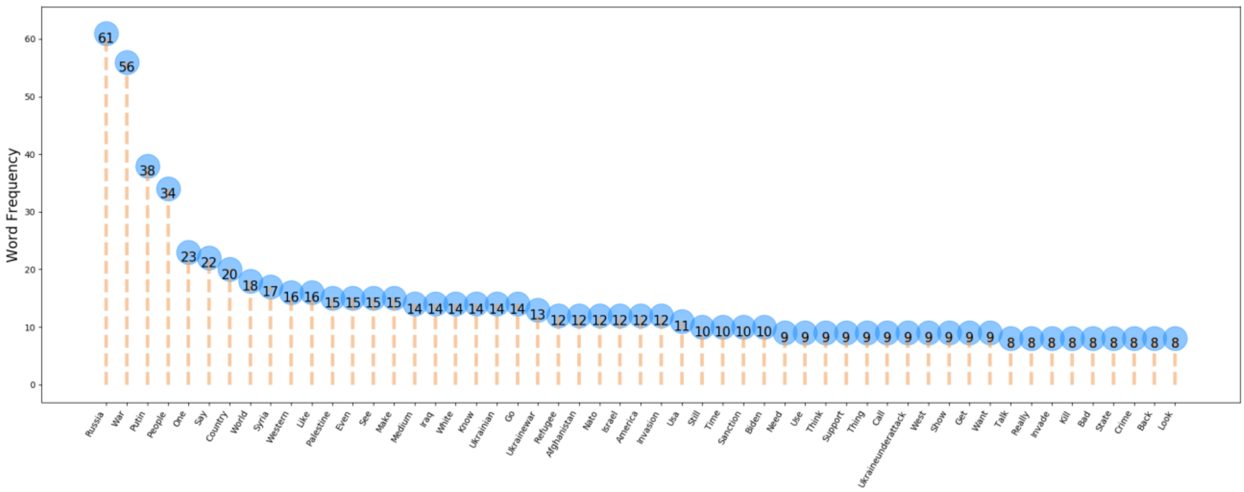

Figure 6. Network Word frequency analysis for cluster 1

As shown in Figure 6, the word Russia is mentioned more frequently on Twitter than the word Ukraine. Some topics such as sanction, invading, crime, and refugee have been debated as a result of this war. Furthermore, discussions concerning NATO are quite common. Syria, Palestine, Iraq, Libya, Yemen, and Afghanistan are among the war-torn countries mentioned. The refugee crisis and an act of war condemnation can both be observed in this cluster.

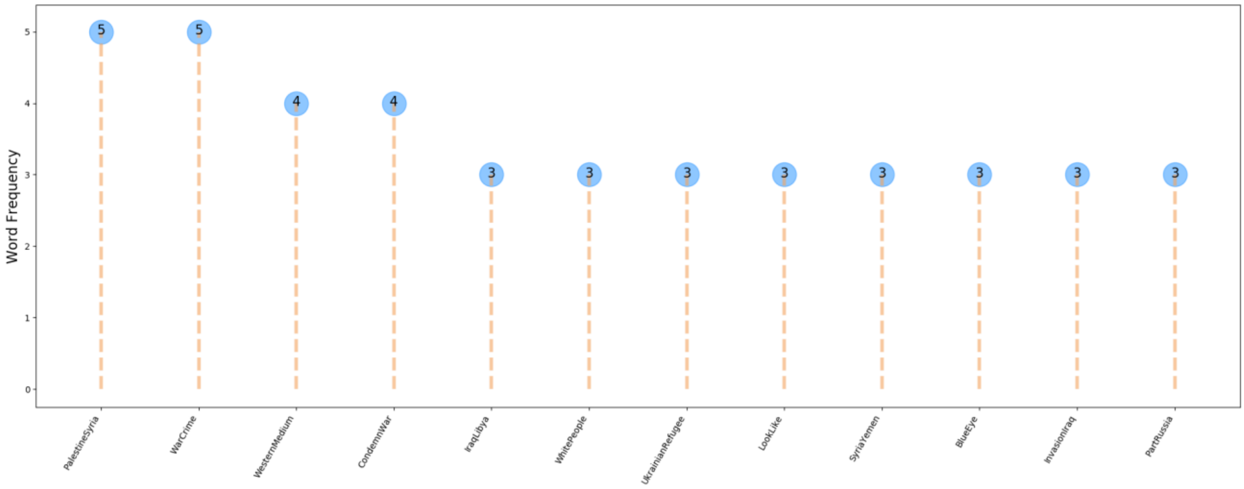

Figure 7. N-gram (bigram) analysis for cluster 1

When n-gram (bigram) analysis is used to look at sequential words, the most regularly used terms are "Palestine Syria" and "War Crime." In the first cluster document, they appear five times. Following that are "Western Media" and "Condemn War." The rest of the sequential terms in the first cluster, such as "Iraq Libya," "White People," "Ukranian Refugee," "Syria Yemen," and "Invasion Iraq," show that the first cluster stresses "war" by considering other nations that have suffered conflict.

After the first cluster analysis, the second cluster is evaluated with the same process as the first cluster. Cluster 2 sub-clusters are seen using network visualization in Figure 8.

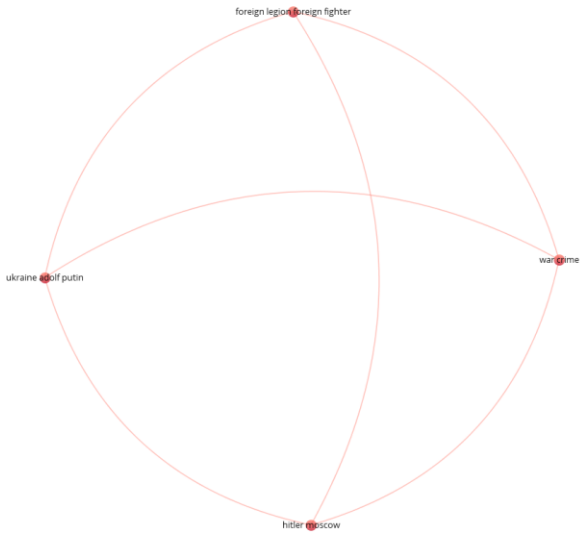

Figure 8. Network visualization of cluster 2 (minimum number co-occurrences of a term is 1)

Twenty-five tweets are analyzed in the second cluster. There are the fewest tweets in this cluster. As a result, there isn't a sub-cluster. In the second cluster, Putin is compared to Adolf Hitler, as shown in Figure 8. Furthermore, war crimes are considered. In Figure 9, word frequency analysis for cluster 2 is demonstrated.

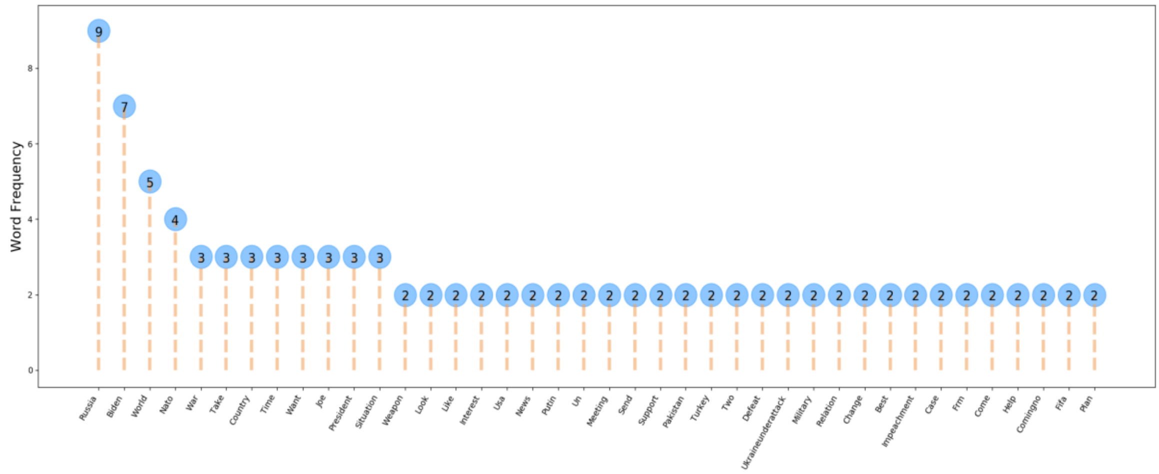

Figure 9. Word frequency analysis for cluster 2

In this cluster, mostly the USA and its president are included. In addition, NATO and weapon can be seen. Besides these points, there are also other words such as "weapon", "military", "send", "help", "meeting", and "support". From the aforementioned words, it can be concluded that tweets under this cluster include the description of the current situation at war in the manner of politics and the army. Figure 10 shows N-gram analysis for cluster 2.

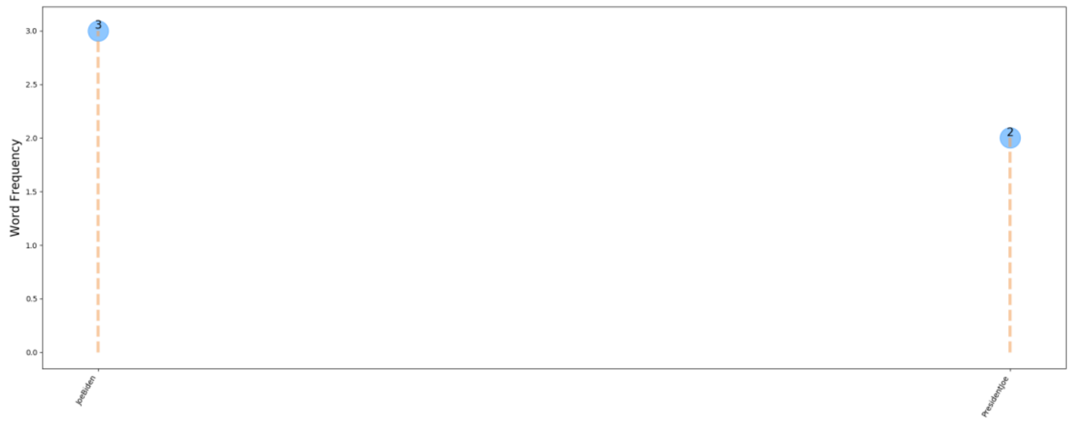

Figure 10. N-gram (bigram) analysis for cluster 2

The second cluster is related to USA president Joe Biden. Since the number of tweets in this cluster is low, mostly Joe Biden is mentioned. All in all, word frequency analysis emphasizes NATO, USA president Joe Biden, weapon and some countries such as Turkey and Pakistan in the second cluster. Figure 10 demonstrates that “Joe Biden” and “President Joe” (USA president) are the most frequently seen sequential words in the cluster. Therefore, it can be said that this cluster includes the tweets related to the USA president. A total of 38471 tweets are analyzed when the third cluster is considered. As a result, this cluster contains the most tweets. The result of network analysis using VOSviewer is shown in Figure 11.

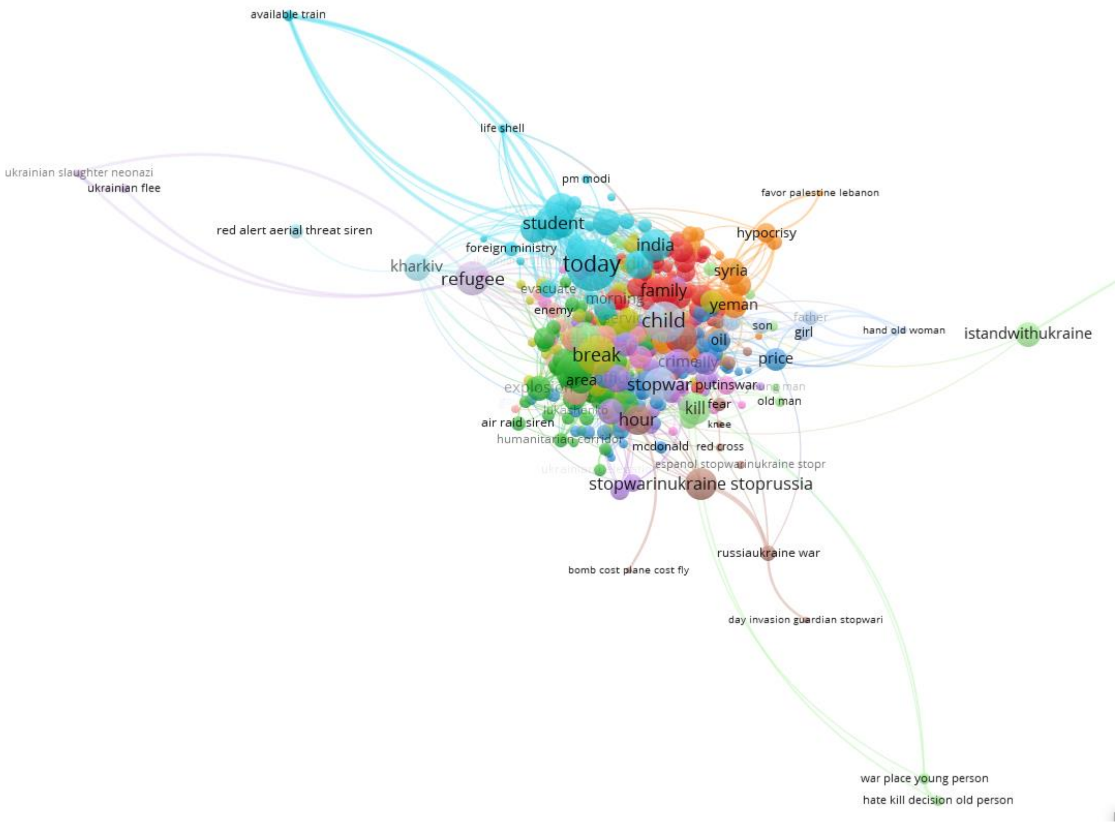

Figure 11. Network visualization of cluster 3 (minimum number co-occurrences of a term is 10)

There are too many sub-clusters in the third cluster since it has the most tweets. The cluster is divided into more than eight sub-clusters. The purple area focuses on the refugee crisis. The light blue area covers Indian students. The red area is related to the family. The orange area includes Syria and Yemen. The light green area focuses on Istanbul and is related to negotiations. The brown area emphasizes stopping the war. In Figure 12, Word frequency analysis for cluster 3 is given.

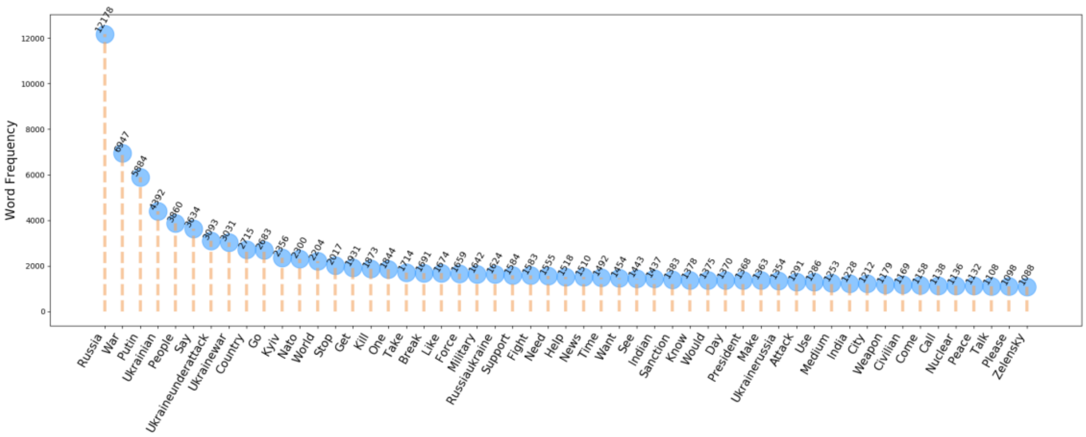

Figure 12. Word frequency analysis for cluster 3

It is true that most people refer to Putin (the Russian President) rather than Zelensky (the President of Ukraine). This indicates that the attacking side is mentioned more than the defending side. As in the other clusters, NATO is included in this one as well. N-gram analysis for cluster 3 is demonstrated in Figure 13.

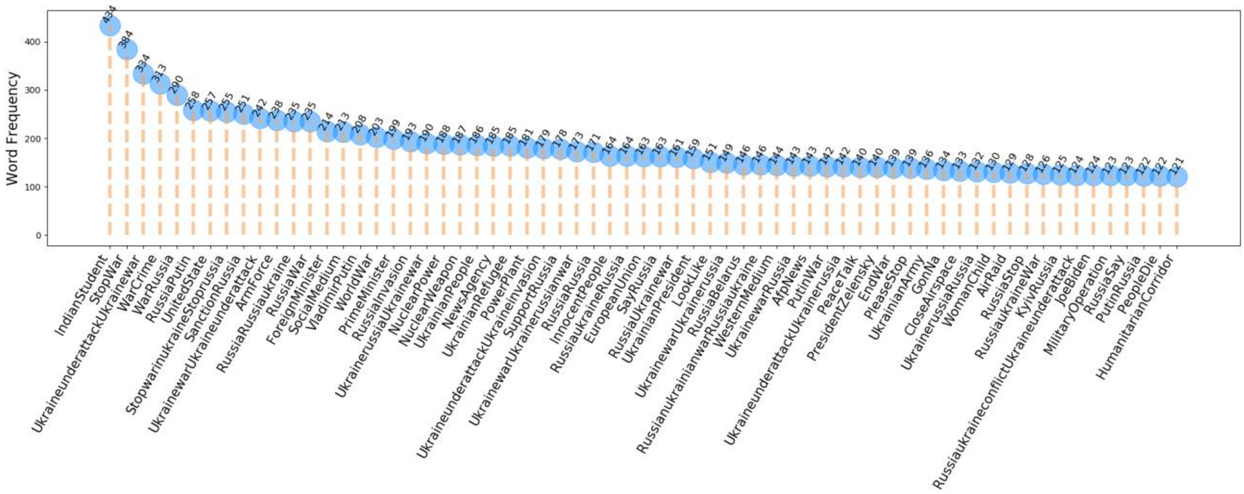

Figure 13. N-gram (bigram) analysis for cluster 3

"Indian students" is the most commonly used term in the third cluster. This demonstrates that Indian students are having difficulties throughout the war. The strikes by Russia, the use of nuclear weapons, the necessity to end the war, and innocent people, women, and children are all discussed in this cluster. To summarize, word frequency analysis reveals that the most usually seen word in terms of being under attack is Ukrainian. "Sanctions on Russia," "Indian students," "war crimes," "power plants," "nuclear weapons," and "United States" are all recognized as consecutive words. The third cluster contains certain sanctions imposed by other countries as well as the weapons employed by Russia in its attacks on Ukraine.

After the third cluster analysis, the standard procedure (same techniques as in the other clusters) is applied to the fourth cluster. A total of 2490 tweets are included in this cluster. Cluster 4 sub-clusters are seen using network visualization in Figure 14.

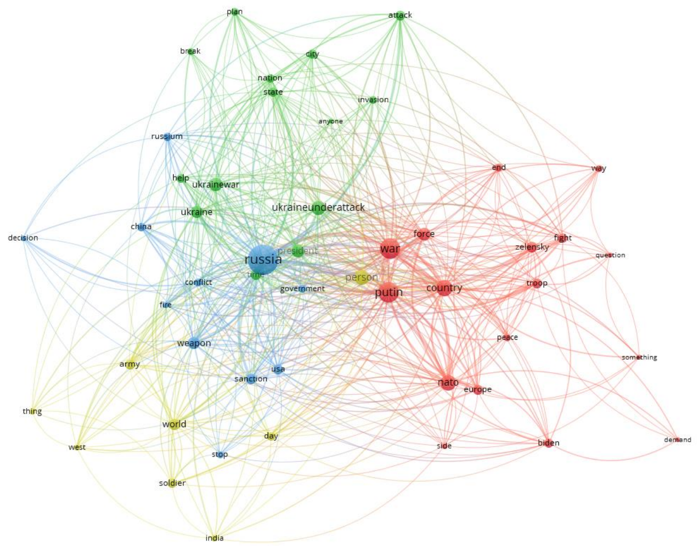

Figure 14. Network visualization of cluster 4 (min. number co-occurrences of a term is 7)

The fourth cluster is divided into four sub-clusters. Also, it can be seen that the network between sub-clusters is more intense compared to the aforementioned clusters. The blue area focuses on sanctions imposed on Russia by other countries such as the United States and China. When it comes to red area, tweets include the efforts of NATO and Europe to provide peace between two sides by referring to the presidents of both sides. The tweets in the green area focus on the invasion and attack of Russia on Ukraine. Lastly, the yellow area is mostly concerned with the military. Figure 15 demonstrates the word frequency analysis for cluster 4.

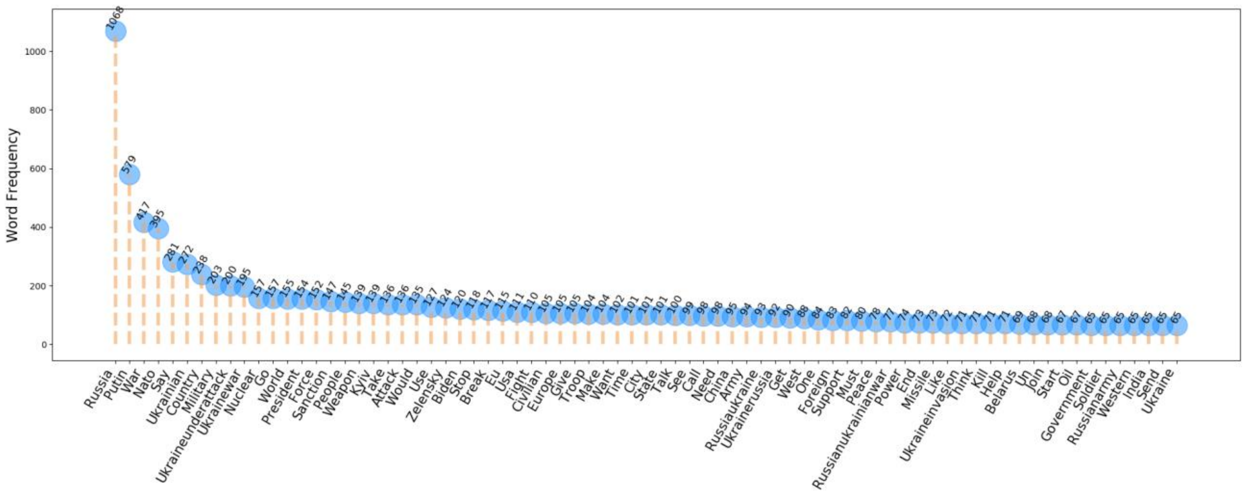

Figure 15. Word frequency analysis for cluster 4

According to the word frequencies indicated in Figure 15, there are several other countries involved in the war from a political standpoint, such as China, Belarus, and the United States. Apart from Russia, Putin, and war, NATO is the most often highlighted word in this cluster, according to the data. Therefore , it can be inferred that NATO's engagement is desirable in order to bring the war to an end as soon as possible. Overall, word frequencies show that nuclear, attack, sanction, and NATO are the most often used terms, as shown in Figure a. As a result, it can be determined that tweets cover issues such as nuclear attack and NATO engagement in the conflict. Figure 16 gives N-gram analysis of cluster 4.

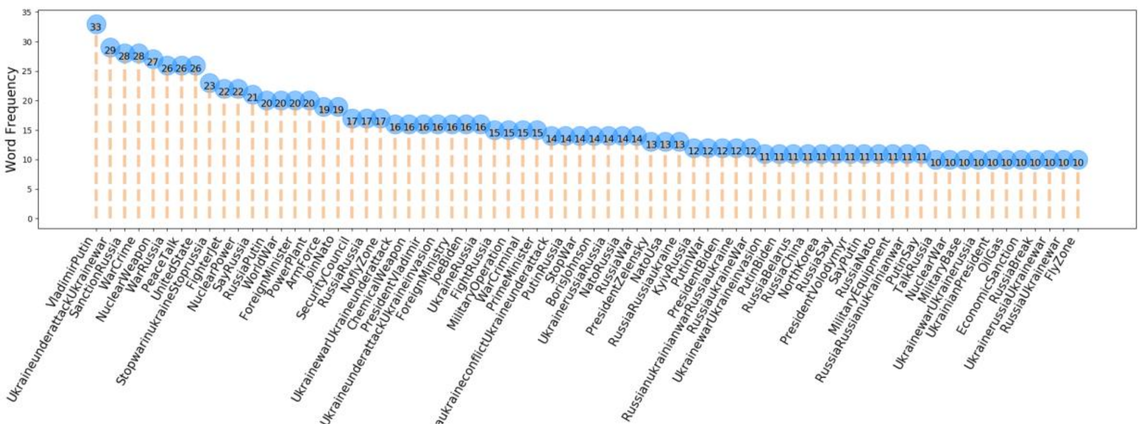

Figure 16. N-gram (bigram) analysis for cluster 4

The most frequently occurring sequence terms, according to bigram analysis, are "Vladimir Putin," "Ukraine under attack," "sanction Russia," "war crime," "nuclear weapon," "war Russia," and "peace talk." The conclusion drawn from the analysis is that there is a widespread desire to avoid conflict, avoid nuclear weapons, and need the participation of foreign ministers.

Cluster 5 has 18781 tweets, making it the second largest cluster out of a total of seven. The majority of tweets in this cluster are about war crimes, Vladimir Putin, and a common desire to end the conflict. Using network visualization, Cluster 5 sub-clusters can be visualized in Figure 17.

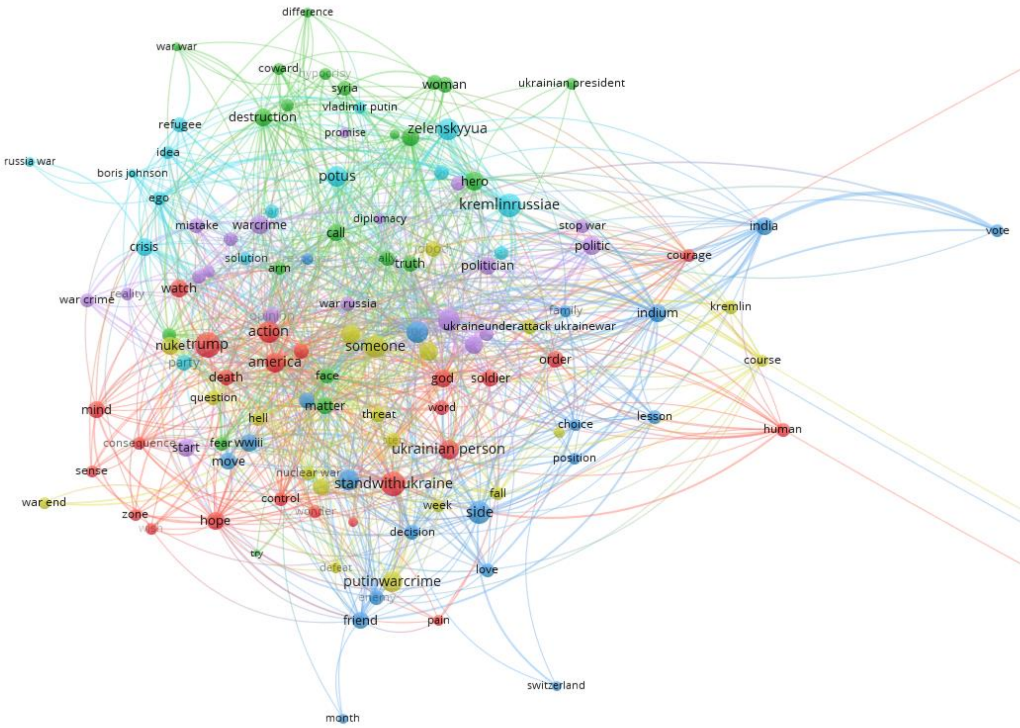

Figure 17. Network visualization of cluster 5 (min. number co-occurrences of a term is 15)

The fifth cluster is divided into 6 sub-clusters indicated by different colors, which are green, light blue, dark blue, purple, red, and yellow. In the sub-cluster referred to by green, tweets related to destruction, women, heroes, and Zelensky. The status of women in the war, the destruction wrought by conflict, and society's perception of Zelensky as a hero may all be inferred from these remarks. The words encountered in the light-blue sub-cluster are Kremlin, Potus, refuge, crisis, and solution. Therefore, it can be concluded that tweets are related to the discourse of America to the Kremlin about the solution to the refugee crisis. The dark-blue sub-cluster consists of words like India, indium, lesson, side, decision, position, and choose. These words demonstrate that decision about the situation of Indian students who are suffering from the ongoing war. Ukraine under attack, war crimes, politicians, politics, and mistakes are the words that are mostly underlined in the purple sub-cluster. These words indicate that the main theme of the sub-cluster is about wrong political decisions and the occurrence of war crimes. The Red sub-cluster includes words that stand with Ukraine, Trump, America, and action. These words demonstrate that the sub-cluster is related to American support for Ukraine by taking some actions such as military support. The last sub-cluster is illustrated by the yellow color, which includes threat, someone, nuclear war, Putin's war crimes, and ending war words. As a result, it can be determined that this sub-cluster is about Russia's threat of using nuclear weapons, which is considered a war crime.

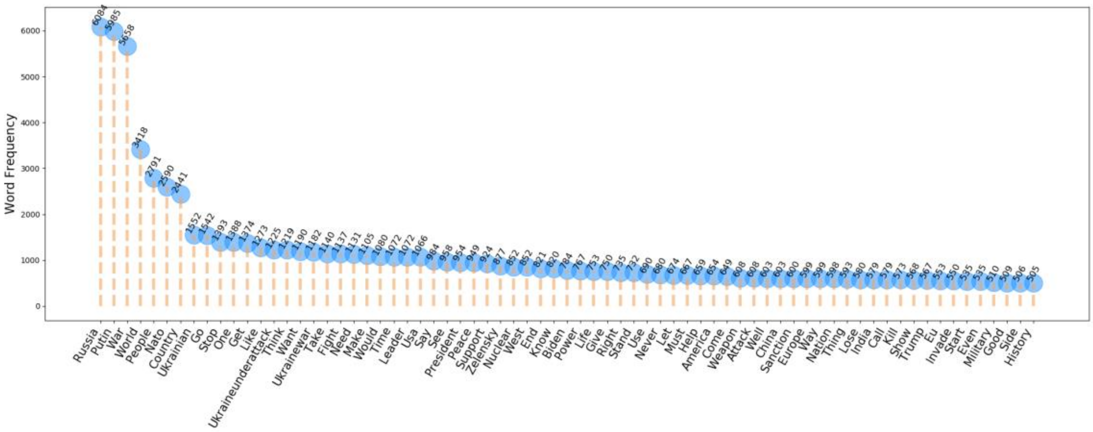

Figure 18. Word frequency analysis for cluster 5

The word frequency analysis of the fifth cluster demonstrates Russia's dominance over Ukraine (Figure 18). People's desires, anti-war sentiments, and peace terms are all part of this cluster. Russia's attacks and the world's reactions to Russia can be seen in this cluster. Figure 19 gives N-gram analysis of cluster 5.

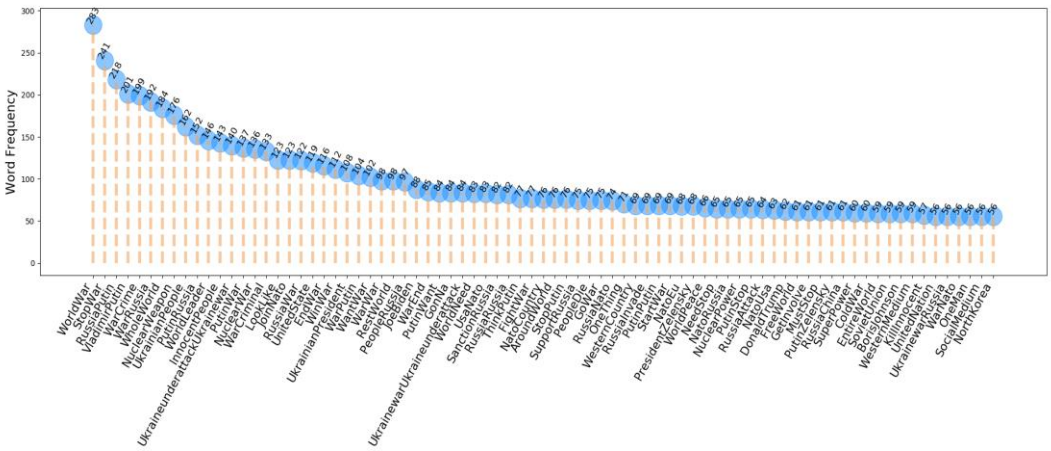

Figure 19. N-gram (bigram) analysis for cluster

The most often occurring sequential term in cluster 5 is "World War", which is followed by the "Stop War" word in the sequence. Therefore, it can be concluded that the war is at an important point for the whole world to overcome, and the general tendency in social media is in favor of ending the war. Also, from the point of war victims, tweets focused on "war crimes" and "innocent people" are observed in this cluster. From the analysis of this cluster, it can be deduced that the other focus of tweets is "nuclear weapons". As a result, bigram analysis depicts Putin's war crimes as well as the rest of the world's judgment on the war.

The sixth cluster includes a total of 4803 tweets. This is the third big cluster. The network visualization of cluster 6 is given in Figure 20.

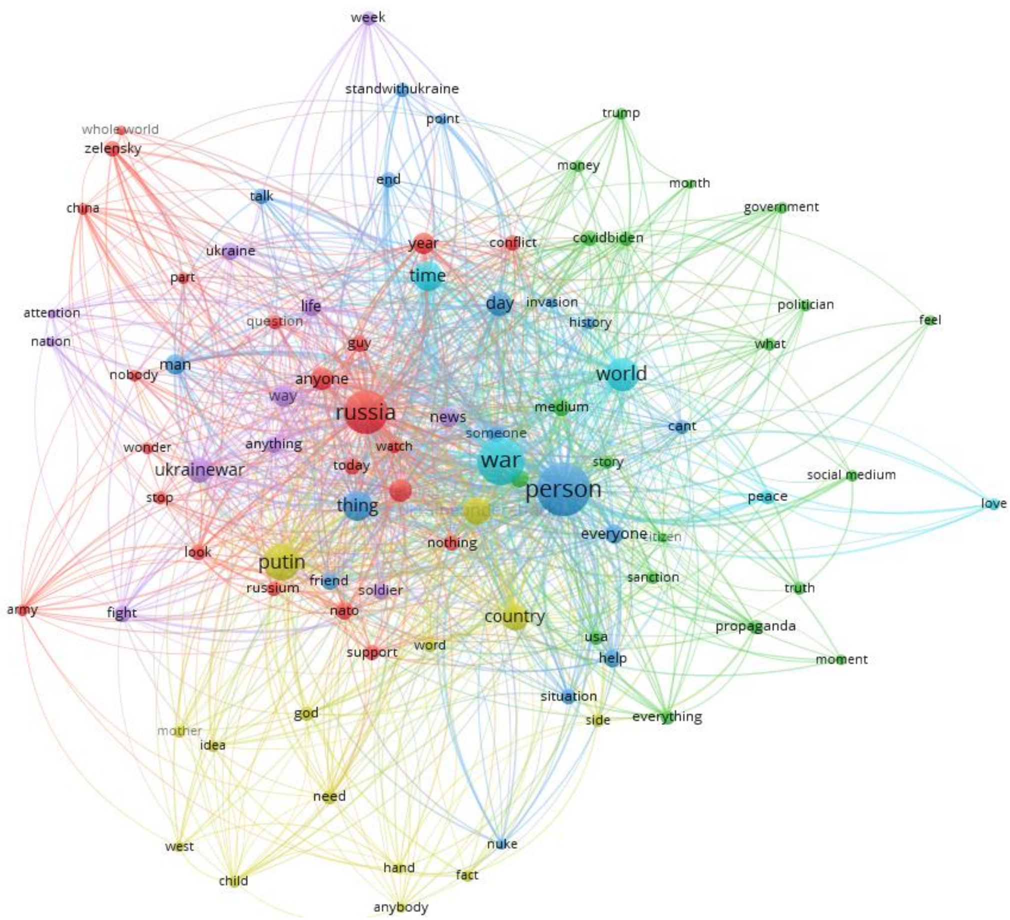

Figure 20. Network visualization of cluster 6 (min. number co-occurrences of a term is 10)

6 sub-clusters are shown in Figure 20 for the cluster 6. Russia, person, war, and Putin are the focus points of this cluster. The green area focuses on social media and propaganda while taking into account political considerations such as the United States and its former and current presidents. The blue area emphasizes people in distress and asks for help. The light blue area reflects the global community's attitude on war and yearning for peace. The red area denotes a conflict, and the war must come to an end. The other area is referred by purple color, that includes military and media terms such as soldier and news with the main focus on Ukraine. Finally, spiritual and emotional concepts such as mother, child, and god are represented in the yellow area. Word frequency analysis for sixth cluster is given in Figure 21.

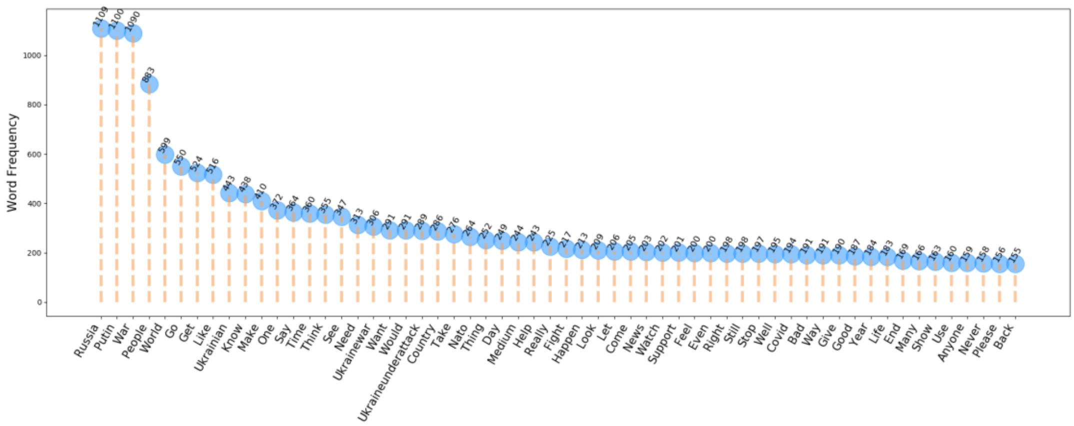

Figure 21. Word frequency analysis for cluster 6

Russia, Putin, and war are seen as the top three words in terms of word frequency analysis. Figure 21 shows that Ukraine is being attacked, and the world talks about it. Therefore, this analysis gives the result that almost same with aforementioned results. Next Figure is Figure 22 which demonstrates the bigram analysis of sixth cluster.

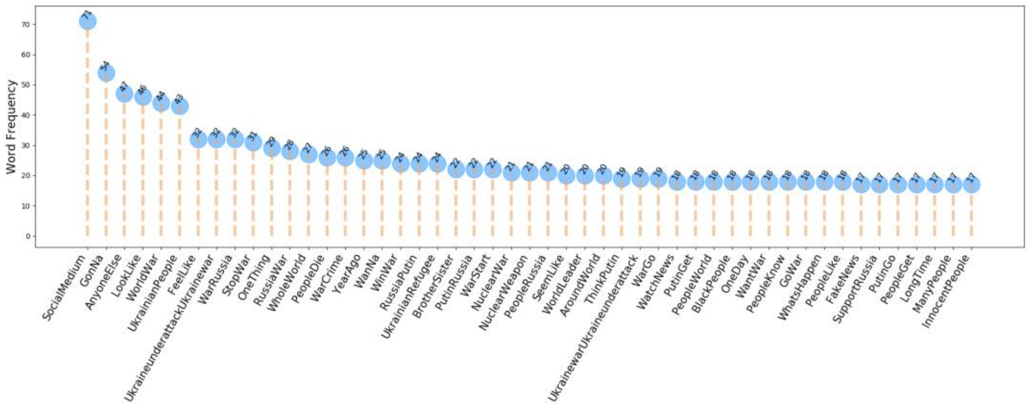

Figure 22. N-gram (bigram) analysis for cluster 6

The first consecutive term in this cluster is “social media”. From the perspective of real-world, this cluster demonstrates the relevance of social media in a war. This analysis demonstrates that the war affects the entire world. Feelings and emotions are included in this cluster. Furthermore, the terms "nuclear war" and "Ukrainian refugee" are significant consecutive words in this cluster. The final cluster contains a total of 653 tweets. Figure 23 depicts the network visualization of cluster 7. Although there are a limited number of tweets in this cluster, the number of sub-clusters is high. However, sub-clusters are not represented in as many tweets.

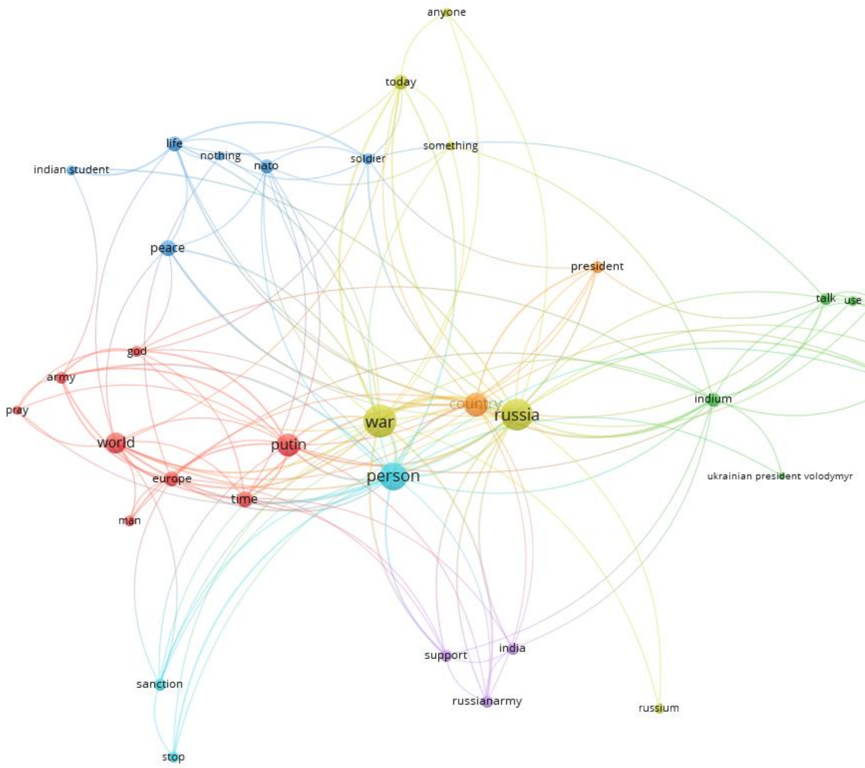

Figure 23. Network visualization of cluster 7 (min. number co-occurrences of a term is 3)

7 sub-clusters are seen in cluster 7. Indian students have an impact across several sub-clusters. The green area represents questions about Indian students. Also, the blue are focuses on Indian students together with peace. The purple area emphasizes Russian army and India. Therefore, it can be concluded that the problem related to the Indian students is discussed in the several sub-clusters. The words encountered in the red sub-cluster are world, Europe, army, god, and time. It illustrates that war is not viewed from afar by the rest of the world or Europe. The light blue area consists of sanctions on Russia. The yellow and orange areas are intersected and focus on Russia and its president. In Figure 24 word frequency analysis of cluster 7 is demonstrated.

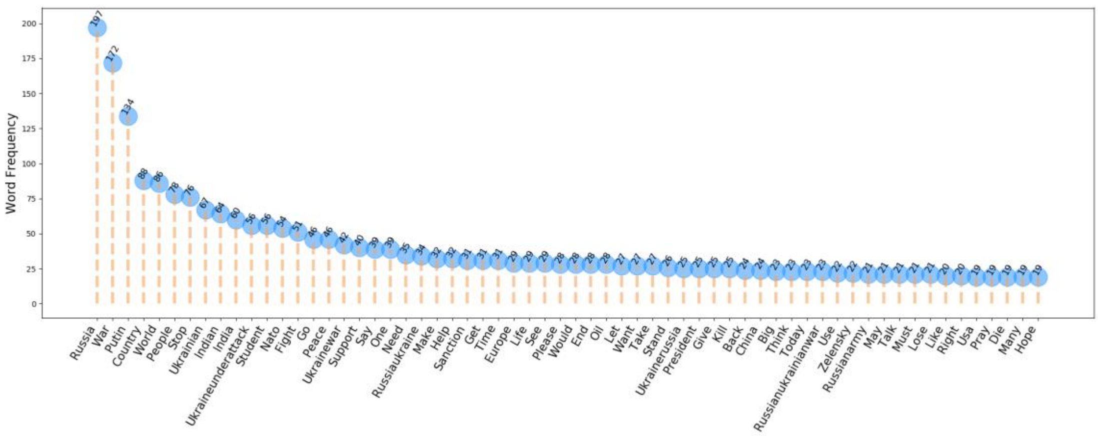

Figure 24. Word frequency analysis for cluster 7

After the top three words, which are "Russia", "war", and "Putin", it is seen that "people", "Ukrainian", "Indian", and "student" are the mostly underlined words from Figure 24. Therefore, the interpretation of this cluster can be expressed as considered tweets mostly focused on war victims and their issues. On the other hand, words like "NATO", "peace", "support", "please", and "end" lead the interpretation to the concept of ending war, waiting for international support, and seeking peace. Bigram analysis of cluster 7 is given Figure 25.

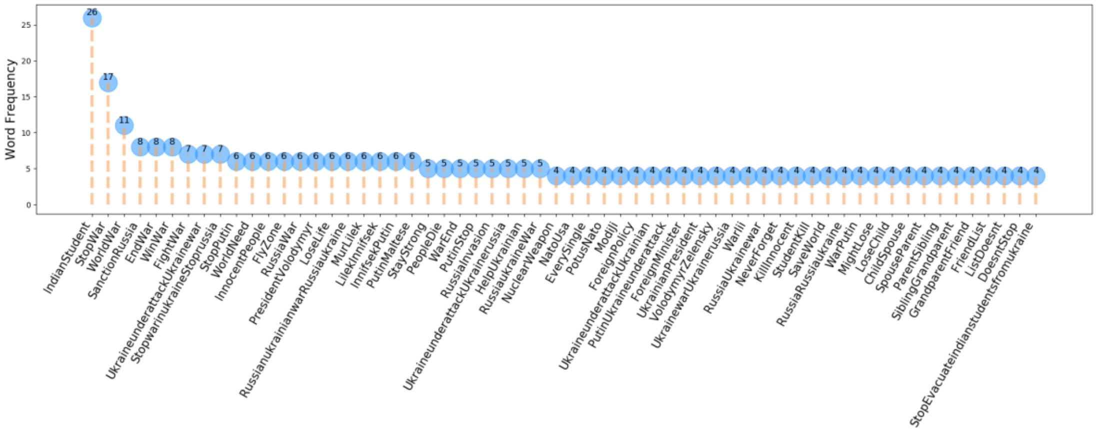

Figure 25. N-gram (bigram) analysis for cluster 7

As given in network visualization of cluster 7, "Indian students" is the most discussed consecutive word. In addition, "stop war", "world war", "sanction Russia", and "end war" are the other popular consecutive words in this cluster. It can be concluded that other nations living in Ukraine are experiencing problems as a result of the war.

4. Conclusion and Suggestions

In a nutshell, the aim of the conducted study is to reveal the objectives of social media during the war. As a result, it can provide a better understanding of society and their general perspective on the issues that have arisen. In that way, some of the decision and assistance processes can be provided with a more accurate and efficient approach. Some of the aforementioned processes can be summarized as: political decisions can be taken with a deeper understanding of the current situation; humanitarian activities can be conducted in accordance with detected needs; and support can be provided to victims in an effective and efficient way. Analysis in the proposed study consists of two main parts. The first part is LDA application and Fuzzy c-means clustering. In detail, LDA is applied to extract topic probabilities in order to define each tweet in accordance. After that, these results are utilized as an input for Fuzzy c-means to gather clusters of retrieved tweets for further analysis. Furthermore, 7 distinct clusters are gathered to be processed. To provide a better understanding, the second part of the analysis evaluates each obtained cluster using n-grams and network analysis. Lastly, the implications of the proposed study are given as follows:

- Putin is discussed more than Zelensky as a leader, and every action taken by Putin has become a social media trend.
- During the war Ukraine has been under attacked heavily and the global public opinion is in the favor of ending war.
- America, China, Europe, and the Western media are the main actors who comment on war across the world.
- Sanctions have been imposed on Russia by other countries.
- Problems related to foreign students are discussed throughout the Indian students.
- The refugee crisis has arisen as a result of people leaving Ukraine.
- Other war-torn nations (such as Syria, Palestine, Iraq, Yemen, and Libya) are referenced in the context of the Ukraine-Russia conflict.
- Situation of women and children who experienced war is underlined.
- According to both cluster analysis and LDA application, the oil problem has arisen as a result of war.
- Assistance from the NATO and other countries are demanded to Ukraine by Ukraine and other people.

Limitations of the proposed study can be underlined as: because it's difficult to analyze tweets in many languages at the same time, only those posted in English are evaluated. The

"#UkraineRussiawar " term and hashtag were used to find tweets. It's possible, however, that there are additional tweets that don't use the hashtag but are nonetheless related to the topic. As a result, the outcomes may be more representative of English-speaking communities. For the future study, the system can be generated to demonstrate the interpretations of tweets on the chosen hashtags. Different text mining approaches can be applied. Integration of the languages rather than English can be investigated and implemented in the analysis.

Conflict of Interest Statement

There is no conflict of interest between the authors.

References

[1] F. Bordignon, I. Diamanti, and F. Turato, "Rally'round the Ukrainian flag. The Russian attack and the (temporary?) suspension of geopolitical polarization in Italy," *Contemporary Italian Politics*, pp. 1-17, 2022.

[2] L. Eras, "War, Identity Politics, and Attitudes toward a Linguistic Minority: Prejudice against Russian-Speaking Ukrainians in Ukraine between 1995 and 2018," *Nationalities Papers*, pp. 1-22, 2022.

[3] N. A. Ghani, S. Hamid, I. A. Targio Hashem, and E. Ahmed, "Social media big data analytics: A survey," *Computers in Human Behavior*, vol. 101, pp. 417-428, 2019/12/01/ 2019.

[4] A. Gandomi and M. Haider, "Beyond the hype: Big data concepts, methods, and analytics," *International Journal of Information Management*, vol. 35, no. 2, pp. 137-144, 2015/04/01/ 2015.

[5] D. M. Blei and J. D. Lafferty, "*Topic models*," in Text mining: Chapman and Hall/CRC, 2009, pp. 101-124.

[6] S. A. Curiskis, B. Drake, T. R. Osborn, and P. J. Kennedy, "An evaluation of document clustering and topic modelling in two online social networks: Twitter and Reddit," *Information Processing and Management*, Article vol. 57, no. 2, 2020, Art no. 102034, doi: 10.1016/j.ipm.2019.04.002.

[7] M. E. Roberts et al., "Structural topic models for open-ended survey responses," *American Journal of Political Science*, Article vol. 58, no. 4, pp. 1064-1082, 2014, doi: 10.1111/ajps.12103.

[8] H. Yuan, R. Y. K. Lau, and W. Xu, "The determinants of crowdfunding success: A semantic text analytics approach," *Decision Support Systems*, Article vol. 91, pp. 67-76, 2016, doi: 10.1016/j.dss.2016.08.001.

[9] 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.

[10] Y. Yang, J. H. Hsu, K. Löfgren, and W. Cho, "Cross-platform comparison of framed topics in Twitter and Weibo: machine learning approaches to social media text mining," *Social Network Analysis and Mining, Article* vol. 11, no. 1, 2021, Art no. 75, doi: 10.1007/s13278-021-00772-w.

[11] P. Hu, W. Liu, W. Jiang, and Z. Yang, "Latent topic model for audio retrieval," *Pattern Recognition, Conference Paper* vol. 47, no. 3, pp. 1138-1143, 2014, doi: 10.1016/j.patcog.2013.06.010.

[12] L. Yao et al., "Concept over time: The combination of probabilistic topic model with wikipedia knowledge," *Expert Systems with Applications*, Article vol. 60, pp. 27-38, 2016, doi: 10.1016/j.eswa.2016.04.014.

[13] M. Pavlinek and V. Podgorelec, "Text classification method based on self-training and LDA topic models," *Expert Systems with Applications*, Article vol. 80, pp. 83-93, 2017, doi: 10.1016/j.eswa.2017.03.020.

[14] S. Xiong, K. Wang, D. Ji, and B. Wang, "A short text sentiment-topic model for product reviews," *Neurocomputing*, Article vol. 297, pp. 94-102, 2018, doi: 10.1016/j.neucom.2018.02.034.

[15] L. Hong and B. D. Davison, "Empirical study of topic modeling in twitter," in Proceedings of the first workshop on social media analytics, 2010, pp. 80-88.

[16] E. Lee, F. Rustam, I. Ashraf, P. B. Washington, M. Narra, and R. Shafique, "Inquest of Current Situation in Afghanistan Under Taliban Rule Using Sentiment Analysis and Volume Analysis," *IEEE Access*, Article vol. 10, pp. 10333-10348, 2022, doi: 10.1109/ACCESS.2022.3144659.

[17] P. Vazquez, J. C. Garcia, M. J. Luna, and C. Vaca, "Temporal topics in online news articles: Migration crisis in Venezuela," in 2020 *7th International Conference on eDemocracy and eGovernment, ICEDEG* 2020, 2020, pp. 106-113, doi: 10.1109/ICEDEG48599.2020.9096804.

[18] L. Tang, Y. Zhang, F. Dai, Y. Yoon, Y. Song, and R. S. Sharma, "Social Media Data Analytics for the U.S. Construction Industry: Preliminary Study on Twitter," *Journal of Management in Engineering, Article* vol. 33, no. 6, 2017, Art no. 04017038, doi: 10.1061/(ASCE)ME.1943-5479.0000554.
[19] C. S. Lee and A. Jang, "Questing for Justice on Twitter: Topic Modeling of #StopAsianHate Discourses in the Wake of Atlanta Shooting," *Crime and Delinquency*, Article 2021, doi: 10.1177/00111287211057855.
[20] J. Allan, *Topic detection and tracking: event-based information organization*. Springer Science & Business Media, 2012.
[21] J. C. Bezdek, R. Ehrlich, and W. Full, "FCM: The fuzzy c-means clustering algorithm," *Computers & Geosciences*, vol. 10, no. 2, pp. 191-203, 1984/01/01/ 1984.
[22] S. Abri and R. Abri, "Providing a Personalization Model Based on Fuzzy Topic Modeling," *Arabian Journal for Science and Engineering*, Article vol. 46, no. 4, pp. 3079-3086, 2021, doi: 10.1007/s13369-020-05048-7.
[23] H. Alatas, H. Murfi, and A. Bustamam, "Topic Detection using fuzzy c-means with nonnegative double singular value decomposition initialization," *International Journal of Advances in Soft Computing and its Applications*, Article vol. 10, no. 2, pp. 206-222, 2018.
[24] Y. Prakoso, H. Murfi, and A. Wibowo, "Kernelized Eigenspace based fuzzy C-means for sensing trending topics on twitter," in *ACM International Conference Proceeding Series*, 2018, pp. 6-10, doi: 10.1145/3239283.3239297.
[25] A. Parlina, K. Ramli, and H. Murfi, "Exposing emerging trends in smart sustainable city research using deep autoencoders-based fuzzy c-means," *Sustainability (Switzerland)*, Article vol. 13, no. 5, pp. 1-28, 2021, Art no. 2876, doi: 10.3390/su13052876.
[26] R. T. Sutrisman and H. Murfi, "Analysis of non-negative double singular value decomposition initialization method on eigenspace-based fuzzy C-Means algorithm for Indonesian online news topic detection," in 2018 *6th International Conference on Information and Communication Technology, ICoICT* 2018, 2018, pp. 55-60, doi: 10.1109/ICoICT.2018.8528791.
[27] M. Trupthi, S. Pabboju, and G. Narsimha, "Possibilistic fuzzy C-means topic modelling for twitter sentiment analysis," *International Journal of Intelligent Engineering and Systems*, Article vol. 11, no. 3, pp. 100-108, 2018, doi: 10.22266/IJIES2018.0630.11.
[28] T. Mandhula, S. Pabboju, and N. Gugulotu, "Predicting the customer's opinion on amazon products using selective memory architecture-based convolutional neural network," *Journal of Supercomputing*, Article vol. 76, no. 8, pp. 5923-5947, 2020, doi: 10.1007/s11227-019-03081-4.
[29] L. Kolhe, A. K. Jetawat, and V. Khairnar, "Robust product recommendation system using modified grey wolf optimizer and quantum inspired possibilistic fuzzy C-means," *Cluster Computing*, Article vol. 24, no. 2, pp. 953-968, 2021, doi: 10.1007/s10586-020-03171-6.
[30] D. Khyani, B. Siddhartha, N. Niveditha, and B. Divya, "An Interpretation of Lemmatization and Stemming in Natural Language Processing," Shanghai Ligong Daxue Xuebao/*Journal of University of Shanghai for Science and Technology*, vol. 22, pp. 350-357, 2020.
[31] N. Eligüzel, C. Çetinkaya, and T. Dereli, "Comparison of different machine learning techniques on location extraction by utilizing geo-tagged tweets: A case study," *Advanced Engineering Informatics*, vol. 46, p. 101151, 2020.
[32] K. Crockett, D. Mclean, A. Latham, and N. Alnajran, "Cluster analysis of twitter data: A review of algorithms," in *Proceedings of the 9th International Conference on Agents and Artificial Intelligence*, 2017, vol. 2: Science and Technology Publications (SCITEPRESS)/Springer Books, pp. 239-249.
[33] J. C. Dunn, "A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters," *Journal of Cybernetics*, vol. 3, no. 3, pp. 32-57, 1973/01/01 1973, doi: 10.1080/01969727308546046.
[34] M. Rawashdeh and A. L. Ralescu, "Fuzzy Cluster Validity with Generalized Silhouettes," in MAICS, 2012.
[35] D. Jurafsky and J. H. Martin, "*N-gram Language Models*," in Speech and Language Processing, 3rd ed. draft ed., 2021.