Sosyal Ağlarda Topluluk ve Konu Tespiti: Bir Sistematik Literatür Taraması

Literatür Makalesi/Review Article

Ömer Ayberk ŞENCAN, İsmail ATACAK, İbrahim Alper DOĞRU

Department of Computer Engineering, Gazi University, Ankara, Turkey
oayberksencan@gazi.edu.tr, iatacak@gazi.edu.tr, iadogru@gazi.edu.tr

(Geliş/Received:21.01.2022; Kabul/Accepted:24.06.2022)
DOI: 10.17671/gazibtd.1061332

Özet— Günümüzde internetin hızlı bir şekilde gelişmesi ve kolay bir şekilde ulaşılır olması; Facebook, Instagram, Twitter ve LinkedIN gibi yaygın kullanılan sosyal iletişim platformlarını büyük veri yığınlarının olduğu ortamlara dönüştürmüştür. Bu durum hem aranan bilgiye kolay bir şekilde ulaşılabilmesi için konu tespiti uygulamalarının, hem de konuyla ilgili paylaşım yapan benzer eğilim ve düşünceye sahip topluluklara toplu hizmet verebilmek için topluluk tespit uygulamalarının bu platformlarda kullanımını zorunlu hale getirmiştir. Bu yüzden araştırmacıların sosyal iletişim ağlarında konu tespiti ve topluluk tespiti alanları üzerine araştırmalar yapması ve problemin çözümü ile ilgili yöntem ve teknikler geliştirmesi bu ortamların etkin kullanımı açısından hayati bir önem arz eder. Bu çalışmada, bu alanlara kapsamlı bir bakış sağlamak için sosyal medya platformlarında konu ve topluluk analizi yapan çalışmalar üzerine sistematik ve derinlemesine bir literatür incelemesi sunulmaktadır. İncelemesi yapılacak çalışmaların çoğu uygulamada başarılı sonuçlar ürettiği bilinen makine öğrenmesi temelli modeller kullanan makalelerden seçilmiştir. Bu çalışmaların incelenmesi neticesinde; topluluk tespiti alanında elde ettiği performans değerleri ile Louvain metodunun öne çıktığı görülürken, performans açısından konu analizi alanında tek bir modelin önerilemeyeceği ve uygun modelin ancak verilen sorunun tüm özellikleri göz önünde bulundurularak, probleme özgü şekilde seçilmesi ya da oluşturulması gerektiği sonucuna varılmıştır.

Anahtar Kelimeler— sosyal ağ, topluluk tespiti, konu tespiti, twitter, doğal dil işleme, makine öğrenmesi

Systematic Literature Review of Detecting Topics and Communities in Social Networks

Abstract— In the recent past and in today's world, the internet is advancing rapidly and is easily accessible; this growth has made the social media platforms such as Facebook, Instagram, Twitter, and LinkedIn widely used which produces big data. This requires both topic Detection applications in order to access the required information, as well as community detection practices in order to provide collective services to communities that can be referred to as individuals with similar interests and opinions over the same subject. Therefore, it is vital for researchers to conduct research on topic detection and community detection research areas in social networks and to develop methods and techniques for problem-solving. In this study, a systematic and in-depth literature review is provided on studies that conduct topic and community analysis on social media platforms to provide a comprehensive overview of the given areas. Most of the studies to be analyzed are selected from articles using machine learning-based models that are known to achieve successful results in practice. As a result of the analysis of these studies; it has been concluded that a single model cannot be proposed in the area of topic detection and that the appropriate model should only be selected or created in a problem-specific way, taking into account all the characteristics of the given problem, while the Louvain method seems to stand out with its results in terms of performance in the area of community detection.

Keywords— social network, community detection, topic detection, twitter, natural language processing, machine learning

It is the extended version of the paper "Systematic Literature Review of Detecting Topics and Communities in Social Networks" accepted at the International Conference on Informatics and Computer Science 2021.

1. INTRODUCTION

Social networks advanced widely over the last decade and become a very important part of our life because they offered a new way to interact and communicate. However, these advances have brought about an exponential increase in the data load on social networks, as it has led to the widespread use of some social communication applications such as Facebook, Instagram, and Twitter. The number of monthly active social media users worldwide is expected to reach 3.43 billion by 2023, about a third of the world's entire population [1].

Twitter among these networks has become one of the most widely used social media apps due to its unique features, including a simple interface and a limit of 140 characters per post. In addition to the growth in the numbers for the usage of social networks due to their nature allow users to create their content in the forms of messages, pictures, etc., and letting them create connections with other users has gained a considerable amount of attention from the researchers. Such interest from the researchers towards the social media has introduced the term Social Media Analysis (SNA) which provides the methods for utilizing the big data received from social networks and allowing the researchers to analyze this data to produce meaningful information.

There are several methods developed by researchers with the goal of analyzing the social media data for a great variety of different purposes such as detection of communities, detection of emerging topics, opinion mining, sentiment analysis, etc. While topic detection is an important tool for providing a comprehensive understanding of the emerging issues and what is the users' perceptions about the given topic, community detection allows us to understand the connections between the users in a much more detailed way. This information can be used for suggestions, marketing reasons, security-based applications visualization, and so on. This information also brought us to the conclusion of Community Detection and Topic Detection concepts can be referred to as linked topics due to the strong connections between their inputs & outputs. Namely, the topics can be extracted from the popular words & word groups in communities and the communities can be extracted by finding the people who are talking about similar topics.

We conduct this SLR (Systematic Literature Review) with an aim to describe the methods and models which are used to solve the above-given problems, compare the performance of these models, and provide a comprehensive view of the overall situation in the fields of topic detection and community detection. Namely, we perform an in-depth systematic search for the existing topic detection and community detection methods. Furthermore, we precisely discuss the performance evaluation of given algorithms. We hope this study will give a broad idea about the current situation in the field to the researchers and will be beneficial for their evaluation of the current state-of-the-art methods in the given domains.

The rest of the paper is organized as follows: Section 2 consists of basic definitions of the research and gives information about the background of this study. The methodology we follow is given in Section 3. The results obtained in this study are presented in Section 4 and Section 5 summarizes and concludes the paper.

2. BACKGROUND

To understand the studies carried out in the field of community detection and topic detection, it is of great importance to first define the terms community and topic as they are used in these studies. In this section, it is aimed to present these two terms and to give a broad idea about the usage of these terms.

2.1. Community Definition

The issue of community detection has gained an important place on social media platforms. This is mainly because a number of useful information about the users can be extracted and evaluated through the detection of users who has certain features in common.

Community detection aims to identify compatible clusters or groups in real-world graphs, such as social media networks [2]. It also aims to identify the modules of the graph as well [3]. Identifying user communities with similar interests, determining which advertisement to show to which users, and identifying the spam-posting accounts and communities are some of the results that are intended to be obtained [4]. These results are used for many different purposes such as blocking spam-posting accounts [5], showing the related advertisement to the users with similar interests [6], and so on. This also shows the diverse nature of the studies based on community detection.

2.2. Emerging Topics Definition

The emerging topic is considered simultaneously seen in words in word-groups [7]. The detection of topics is defined as the process of obtaining and summarizing trending topics in a form that will contain useful information. Therefore, the emerging topic detection concept can be defined as the detection of emerging issues on social networks. Also known as hot topic detection, these methods aim to group thematically related documents from a temporal set of documents into an unknown number of topics, and then find a series of topics that are frequently seen over a period of time [8].

3. METHODOLOGY AND FINDINGS

In social media analysis, topic detection and community detection are known as some of the most common research areas. Thus, a large number of studies can be expected to exist which is related to the objective of this study. A systematic literature review (SLR) is chosen to be able to identify suitable articles from a large number of publications. Some useful features such as the method being accurate, robust, and transparent are taken into consideration when making this selection [9].

3.1. Review Planning Phase

During the study of a new field of information, researchers usually conduct a bibliographic examination to determine publications on a particular topic. However, such reviews do not use a systematic approach and do not offer any system to prevent factors such as deviation, bias, misapprehension, and tendentiousness that may occur during the selection of publications to be analyzed [10].

While conducting the identification of publications, another way of analyzing primary studies is to use the systematic examination method [11], because it allows gaining clear and unbiased information about the research topic [12]. This systematic review study is conducted to summarize information about approaches involving topic and community detection through social networks and therefore to investigate the advantages and limitations of these studies. The process used in this systematic review is given below.

Table 1. Identification of the purpose of the study using the GQM Method [13]

Analysis	Studies on topic and community detection
Purpose	Characterizing and summarizing the studies
Scope	Originality and performance of the proposed model
Viewpoint	Researcher
Context	Primary studies involving topic and community detection

The following sections describe the basic information about the conduct of this systematic review. A more detailed description can be found in the study of Barceló's & Travasos [14].

As given in Table 2, selected studies are limited to be published after 2016 to focus on relatively state-of-art studies. In addition to that, there are also rules about the selection process given in the 'Context' tab of Table 2, such as 'The studies should be tested with a dataset.' and 'The performance metrics of the study should be specified in the article.'. These rules are made to be able to compare the performance of the studies correctly.

3.1.2 Search Terms

Based on the keywords and research questions, the following search strings are chosen.

- "Topic Detection" OR "Emerging Topic Detection" AND ("Social Network" OR "Social Media" OR "Twitter")
- "Community Detection" OR "Communities Detection" AND ("Social Network" OR "Social Media" OR "Twitter")

The above-given search strings are used at the selected resource portal in order to find the related works.

Table 2. Systematic review protocol master information based on the template of Biolchini et. al. [15]

Research Question Target	Review and evaluation of the studies in the field of community and topic detection.	
Research Questions (RQ)	RQ1	What are the state-of-the-art solutions for detecting topics and communities in social networks?
	RQ2	What are the methods/models used in those solutions?
	RQ3	Is there any specific model or models that are more successful than the others?
Context	(1) The studies should be published after 2016.	
	(2) The studies should be primary.	
	(3) The studies should be tested with a dataset.	
	(4) The performance metrics of the study should be specified in the article.	
	(5) The studies should have the SCI/SCI-e/SSCI index.	
	(6) The studies should be in the English language.	
	(7) The data that used in the studies should be fetched from Twitter or equivalent social platforms.	
Keywords	Community detection, emerging topic detection, topic detection, social networks, twitter, deep learning, machine learning	
	(1) The resource should have a search engine suitable for finding the articles.	

It is the extended version of the paper "Systematic Literature Review of Detecting Topics and Communities in Social Networks" accepted at the International Conference on Informatics and Computer Science 2021.

Resource Selection Criteria	(2) The resource should have official recognition in the academic field.
	(3) The resource should consist of Q1, Q2 or at least Q3 indexed articles
Resource List	Science Direct Portal [16]
Inclusion, Exclusion Criteria	The articles should be accessible and include the topics mentioned in the Question tab.

3.2. Review Phase

There are two steps in the execution of this systematic review. The first of these steps consists of scanning scientific publications related to the above-mentioned topics. This scan is carried out by using search engines of the resource websites given in Table 2 to find the related papers. This search was done with custom search strings specified using specific keywords defined in the protocol. The relevant information is provided in Table 2.

Table 3. Steps of SLR approach used in this study

No	Step
1	Search for the studies using custom search strings specified using specific keywords defined in the protocol and filter the results according to the given date (which should be after 2016 as defined in Table 2), paper type (Journal Paper & Conference Paper), and language which should be English. Then, download the search results to a local drive.
2	Reading the abstract and introduction sections: of the downloaded papers and checking the relevance with the aim of this research.
3	Performing Quality Evaluation using the method proposed by Dyba et. Al. [17].
4	Determined the final set of papers used in the study.

After the identification process, the summary and introduction part of each publication were analyzed and it was determined whether they were selected for a more comprehensive analysis according to the criteria for inclusion or exclusion.

This systematic review is carried out in January 2022. In this review, we used the IEEE Explore and Science Direct databases to access the papers as given above in Table 2. We followed a four-step approach to conduct this systematic literature review, which is given in Table 3, and used the Quality Assessment Method introduced by Dyba & Dingsoyr [17].

3.2.1. Topic Detection

Following the steps given in Table 3, a total of 83 scientific studies were found after the search process and 15 of them were deemed to meet the relevant criteria after the first and second review steps, 20 studies were selected out of 83 total studies found with the given keywords. After step three, 5 more studies were excluded from this research based on the quality evaluation result.

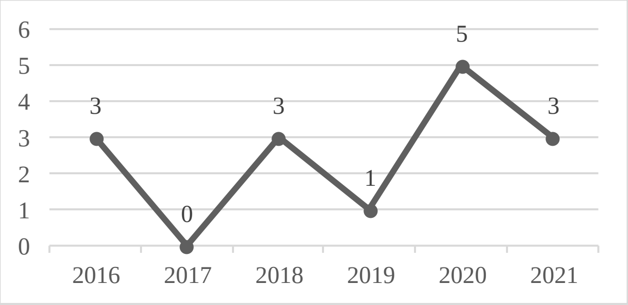

Figure 1.The distribution of the topic detection studies within years

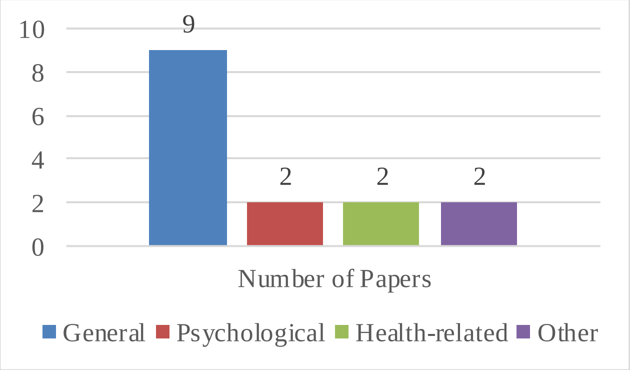

Figure 2. The distribution of the topic detection studies based on the research area

The distribution of the selected papers in the topic detection research area within the years is given in Figure 1. According to this distribution, the highest rate for publications in topic detection has taken place in 2020 on the other hand 2017 and 2019 have the lowest rate with zero and one selected publication to analyze respectively. The distribution also shows that the remaining studies have been distributed equally over 2016, 2018, and 2021 with three publications selected for each year. Figure 2 shows the main research areas for the selected papers.

According to Figure 2, there are different perspectives on the studies about community detection, and the majority (%60) of the selected papers are mainly focused on finding topics in general use whereas some of them aimed to find specific topics within social media content such as psychological, health-related, etcetera.

Table 4. Used algorithms, datasets & performance metrics for selected topic detection papers ('N/A' refers to 'not available')

Study	Used Algorithm / Method	Dataset	Performance Metric			
			Accuracy	Precision	Recall	f-score
[18]	Researcher developed model based on Graphs	3 Twitter datasets collected by Aiello et. al. [19] containing 2.999.870 tweets	N/A	0,615*	N/A	N/A
[20]	LDA & SVM	A user-made dataset containing an unknown number of tweets collected using Twitter API	N/A	0,870*	N/A	N/A
[21]	GloVe & GRU	C-SSRS gold standard dataset containing 500 users' posts	N/A	N/A	N/A	0,243
[22]	ARM	3 subsets of the Aiello dataset [19] containing 4.535.691 tweets in total	N/A	0,709	0,631	0,562
[23]	FCA	The Replab 2013 dataset containing 2200 tweets	N/A	N/A	N/A	0,406
[24]	HUPM	Aiello dataset [19]	N/A	0,386	0,624	0,476
[25]	GSDMM	A user-made dataset containing 6.487.842 tweets in English and Portuguese was collected using Twitter API	N/A	N/A	N/A	0,780
[26]	Fine-tuned BERT	A user-made dataset containing 1.769.384 tweets was collected using Twitter API	N/A	0,96	0,96	0,96
[27]	RF-based two-stage classifier	A user-made dataset containing 5487 tweets was collected using Twitter API	N/A	0,869	N/A	N/A
[28]	LDA	A user-made dataset containing 11.703 tweets was collected using Twitter API	0,830	N/A	N/A	N/A
[29]	A researcher-developed model using k-NN	First Story Detection Dataset containing 51.879.318 tweets in total [30]	N/A	N/A	0,844	N/A
[31]	RNN+LSTM	A user-made dataset containing 234.088 tweets collected using Docteur Tweety	0,670	0,750	0,931	0,831
[32]	MNBMC	Twitter dataset acquired from John Hopkins University (2019-nCOV Data Repository) containing an unknown number of tweets	0,700	N/A	N/A	N/A
[33]	One-Class SVM	A user-made dataset containing 1.427.315 posts collected from Vkontakte, Yandex.local and Spb websites	N/A	0,750	0,820	0,780

LSTM: Long Short-Term Memory
LDA: Latent Dirichlet Allocation
SVM: Support Vector Machine
GRU: Gated Recurrent Unit
ARM: Association Rule Mining
FCA: Formal Concept Analysis
HUPM: High Utility Pattern Mining
GSDMM: Gibbs Sampling Algorithm for the Dirichlet Multinominal Mixture
RNN: Recurrent Neural Network
BERT: Bidirectional Encoder Representations from Transformers
RF: Random Forest
MNBMC: Multinominal Naïve Bayes Multi-Class

3.2.2. Community Detection

Similar to the topic detection, we followed the steps given in Table 3 and after the search process, we found 79 papers in total as they meet the initial criteria. After the detailed examination using the first and second steps given in Table 3, 34 of them were selected. To be specific, a total of 45 studies were excluded, 41 of them were excluded because

It is the extended version of the paper "Systematic Literature Review of Detecting Topics and Communities in Social Networks" accepted at the International Conference on Informatics and Computer Science 2021.

they were not related to the research area of this study and 4 more were excluded because they were secondary research papers. Later on, we performed the quality evaluation as Table 3 suggests, and 16 studies were deemed to meet the relevant criteria for selection. The detailed information about the distribution of the selected papers in the community detection area within the years is given in Figure 3. According to the information from Figure 3, the highest rate for community detection studies has achieved in 2021 with 6 selected papers. Even though 2018 follows 2021 with 5 selected papers, it can be said that 2022 will eventually surpass 2018 because this study takes place at the beginning of 2022 and there are already 3 selected papers from the year 2022. In addition to that, the years 2020, 2019, and 2017 follow the rest with equally selected papers, 2 for each and there were no papers selected from the year 2016.

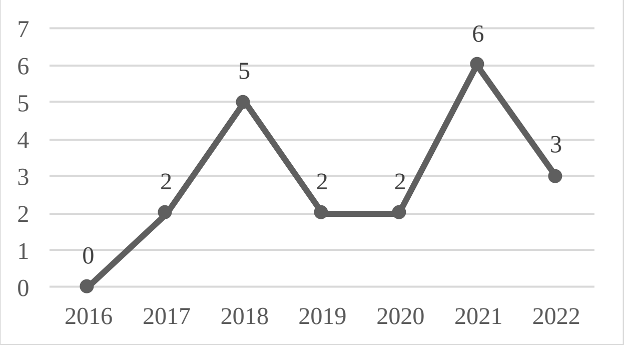

Figure 3. The distribution of the community detection studies within years

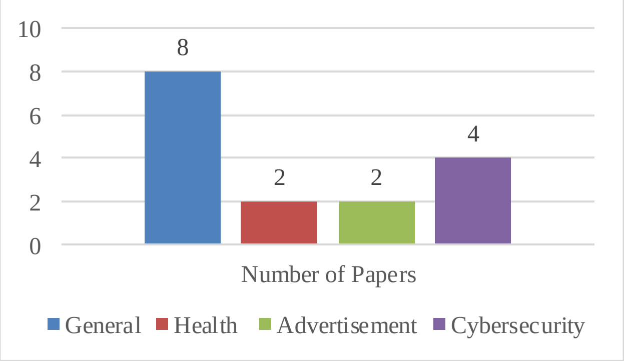

Figure 4. The distribution of the community detection studies based on the research area

Like the topic detection, there are different perspectives on the studies about the community detection research area as well. These perspectives are given in Figure 4 above. According to the data provided in Figure 4, the majority (%53,33) of the papers are focused on finding the communities in social media for general purposes. On the other hand, there are also studies for finding the communities for health-related, adversarial, and cybersecurity concerns.

Health-related papers mostly focus on the effects of the COVID-19 pandemic. While both of the health-related papers are about the COVID-19 pandemic, one of these studies focuses on social media users' views on the COVID-19 vaccine and the other has the goal of visualizing the spread of the virus using social media content.

In addition to that, it is seen that one study focuses on making appropriate friend suggestions on social media while the other aims to increase the success in the field of advertising by identifying the users with the same interests as communities.

Finally, in cyber security-based studies, it has been observed that the aims of the studies vary such as finding and blocking spam sharing communities, finding, and blocking the users who have multiple fake accounts, and preventing hate-speech by finding the communities who suffer from hate-speech.

The metrics used to evaluate the performance of studies in the field of community detection may differ from the performance evaluation metrics used in the topic detection. Because the performance of the community detection methods can also be calculated using the modularity metric, which expresses the strength of the bonds between the nodes in the community. Therefore, although there are some studies using metrics that are also used in the field of topic detection such as Accuracy and F1-Score, it is not possible to evaluate all studies in the field of community detection using those metrics only. On account of that, in this section, the related studies are evaluated by considering the modularity metric as well.

The modularity metric was introduced by Newman [34], [35] and is one of the most common measures to calculate the strength of the connection between the nodes. Namely, the higher the modularity score, the denser the connections in the community and it can take values between -1 and 1. Additionally, modularity scores higher than 0 means that there is a possible presence of a community.

Table 5. Used algorithms, datasets & performance metrics for selected community detection papers ('N/A refers to 'not available')

Study	Used Algorithm / Method	Dataset	Performance Metric				
			Accuracy	Precision	Recall	Modularity	F-score
[36]	A researcher developed two-stage framework based on [37] and LDA	Zachary's karate club [38], dolphin social network [39], political blog dataset [40]	0,600	N/A	N/A	0,915	N/A
[41]	Louvain [42]	A user-made dataset containing an unknown number of tweets	N/A	N/A	N/A	0,870	N/A
[43]	Louvain [42] and its implementation [44]	A user-made dataset containing 83.894.772 tweets	N/A	N/A	N/A	0,988	N/A
[45]	Researcher developed method based on the work of Freeman [46]	ISIS Twitter Network dataset [47]	N/A	N/A	N/A	0,860	N/A
[48]	Researcher developed method based on trust-based modeled weighted directed graphs	FilmTrust [49] & CiaoDVD [50]	0,900*	N/A	N/A	N/A	N/A
[51]	Spectral Clustering [52] & Node2Vec [53]	[38], [39], [54], [55], [56], [57], [58]	N/A	N/A	N/A	0,596	N/A
[59]	Louvain [42]	A user-made dataset containing 6.7 million tweets	0,755	N/A	N/A	N/A	N/A
[60]	Extended version of Louvain [42]	[61]	N/A	0,280	N/A	N/A	0,342
[62]	Louvain [42], LPA [63], walktrap [64], Infomap [65]	HCR [66], OMD [67]	0,770	N/A	N/A	N/A	N/A
[68]	Extended version of Louvain called NI-Louvain	[69], [70]	N/A	N/A	N/A	0,558	N/A
[71]	Initial community set expansion and optimization by Xie et. al. [72]	[73], [74]	N/A	0,926	0,968	0,492	0,930
[75]	Researcher developed method based on CPM	YelpChi & [76]	N/A	0,300	0,100	N/A	0,150
[77]	Researcher developed method based on Fast Greedy	A user-made dataset from Wikipedia	0,948	0,950	0,950	N/A	0,950
[78]	TTSLPA	[79]	N/A	0,850	0,830	N/A	0,850
[80]	The method by Cheng et. al. [81]	A user-made dataset from Reddit containing 257 subreddits	0,970	N/A	N/A	N/A	0,937
[82]	RNN-LSTM & RNN-GRU	Extended version of [83]	0,925	N/A	N/A	N/A	N/A

LPA: Label Propagation Algorithm
CPM: Clique Percolation Method
TTSLPA: Text and Time-series based on Speaker-Listener Label Propagation Algorithm
RNN-LSTM: Recurrent Neural Network – Long Short-Term Memory
RNN-GRU: Recurrent Neural Network – Gated Recurrent Unit

3.2.3. Connection Between Community Detection and Topic Detection

After a detailed examination of selected studies in both Community Detection and Topic Detection fields, as foreseen before conducting this research, a connection has been found between the given two areas. The most noticeable feature of this connection is that it affects both sides. This connection can be inspected in two aspects;

1) Finding Topics based on Communities
2) Finding Communities based on Topics

Based on the explanation of Wu et. al.[2], it is safe to say that networks can be seen as real-world graphs, in which, the nodes are connected to each other and the communities can be found inspecting these connections' density. After the extraction of communities using these graphs, topics can also be extracted by identifying the semantics of the given groups [41].

From the other perspective, similarly to the first argument, community data can be obtained by inspecting the already-found topic data because talking about the same topic causes the users to interact with each other on social networks and therefore, the links between these users' nodes become denser.

In addition to that, it can also be observed that even if the users talk about different subjects, if the given subjects are somehow correlated, the users will be more likely to interact with each other as a result and this will cause their nodes to be connected to each other stronger to form a community.

In the inspection of the graphs and networks formed by the users in social networks, the nodes represent the users and while the collection of connected nodes forms the community, the edges between these nodes are formed by the topics which cause the nodes to interact and that interaction connects these nodes to each other forming the community.

4. EXPERIMENTAL RESULTS

The results obtained in this systematic literature review which examines the studies in the literature for the topic detection and community detection research areas are discussed in this section. Furthermore, an in-depth overview of the common methods used for detecting topics and communities, performances, and rates of usage of those methods and the superior methods in terms of performance in the given research areas is given in this section.

RQ1 – What are the state-of-the-art solutions for detecting topics and communities on social networks

Although the quantity of studies on social media still rising day by day, a lot of work already has been done in this area. Based on the reviewed papers, it has been seen that a great variety of different models and methods has been used in order to detect topics from the social media content. We further divide the methods and models which are used in these studies into 5 subgroups as follows:

- Supervised Learning
- Unsupervised Learning
- Hybrid Models
- Optimization-Based Models
- Others

The distribution of the usage of these models in selected papers is given in Figure 5.

4.1.1. Supervised Learning

Supervised learning is one of the main subsections of machine learning and it is commonly used for classifying the data [84]. Supervised learning-based methods use pre-labeled data as a baseline to feed the model and predict further unlabeled data. Methods such as Support Vector Machines, Naïve Bayes, and Decision Trees fall in this category.

4.1.2. Unsupervised Learning

Unlike supervised learning, unsupervised learning methods don't need labeled data to be effective [85]. On the contrary, this type of model focuses on finding the patterns or trends in the given data using the attributes of the dataset and is mostly used for clustering [86]. Namely, methods like K-means and Fuzzy C-means fall in this category.

4.1.3. Hybrid Models

Hybrid models are the models that have emerged multiple machine learning models working together in a single frame. Hybrid models also include the models that consists of machine learning methods and graph-based method in the same work [87]. Presently, it is seen that better performance results are achieved as a result of the usage of a combination of two or more of the above-mentioned models in new studies. To exemplify, after the detailed examination of the papers selected for this research, studies that use hybrid-based models and are relatively newer have shown great promise going onwards. For example, [82] achieved 0,925 Accuracy score which is the third-best result in the examined articles. In addition to that, there are also some hybrid models which include machine learning-based algorithms and optimization-based algorithms in the same model to work together [72]. This way, researchers aim to improve the performance of the overall model by taking advantage of both of the given approaches and trying to optimize the selected method by using optimization methods [71].

4.1.4. Others

Although machine learning-based methods are widely used in studies involving sentiment analysis, researchers have also been found to benefit from different-based methods by using some of the attributes that social media networks have offered. Methods such as graph-based models [78] and mathematical-based models such as random walks [65] can be cited as examples in this field.

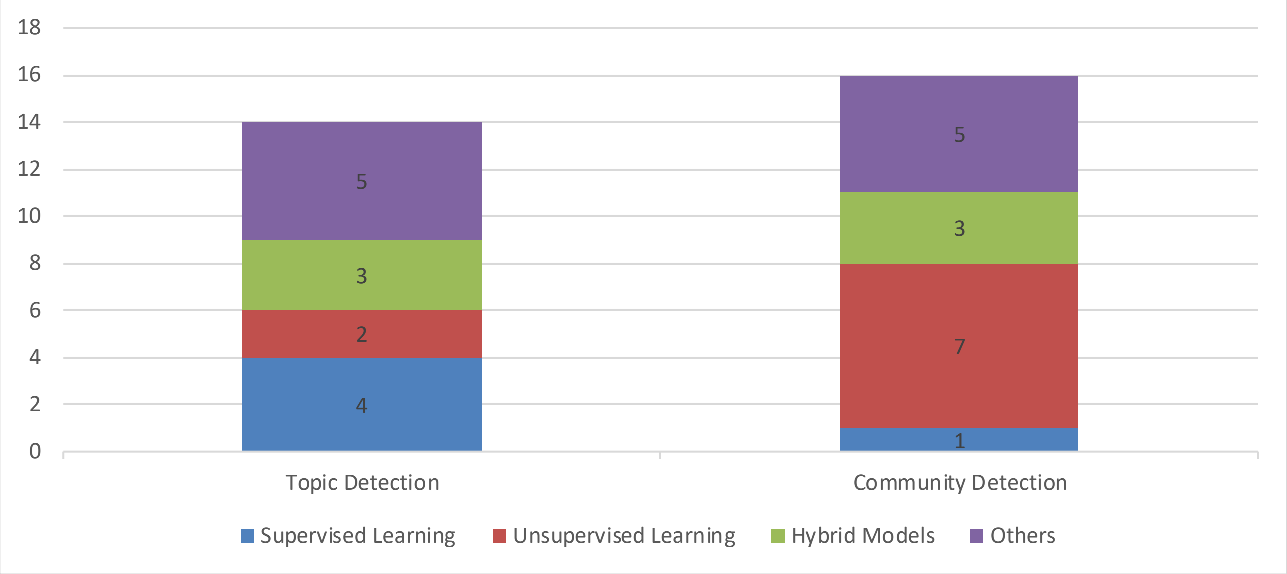

Figure 5. The distribution of the methods based on the models

RQ2 – What are the state-of-the-art solutions for detecting topics and communities on social networks

For the topic detection, it has been observed that there are a great variety of models that are used for this purpose. As seen in Figure 5, four studies benefited from supervised learning-based methods, while two studies used an unsupervised learning-based model and three studies used hybrid models. However, in five studies, it has been seen that they used problem-specific models that can fall into the Others category.

Considering performance metrics and usage rate, the following methods can be shown as state-of-the-art methods in this field:

- BERT – Other
- RNN + LSTM – Hybrid
- RF-based two-stage classifier – Supervised Learning

On the other hand, for the community detection, it has been observed that the Louvain [42] method is frequently selected as the model to use for this purpose by the researchers. This has also led to a significant dominance by the unsupervised learning-based methods in this field in terms of usage rate. Therefore, it was observed that unsupervised learning-based methods were used in 43,75% of the studies examined in the field of community detection. Other methods ranked second with a frequency of 31,25% while hybrid methods and supervised learning-based methods were the least used models with three and one studies respectively. In terms of performance metrics and usage rate, the Louvain method has shown great dominance over others and can be stated as the state-of-the-art method for community detection.

RQ3 – Is there any specific model/models that are more successful than others?

The review process has shown that there is a wide spectrum of choices available for the task of topic detection. Although the methods used can be sorted according to the performance metrics that they achieved, it has also been seen as a result of the same process that selecting or constructing these models specifically for the problem in hand leads to more successful results in terms of performance. Therefore, when the usage rates are taken into account as well, it is concluded that there is not a single algorithm, method, or model in the field that can outperform the others.

It is the extended version of the paper "Systematic Literature Review of Detecting Topics and Communities in Social Networks" accepted at the International Conference on Informatics and Computer Science 2021.

On the other hand, although it has been observed that different methods are used in the field of community detection, it was seen that the Louvain method outperformed other methods in terms of both frequency of use and performance metrics. Therefore, it is possible to mention that there is a method in the field of community detection that can outpace other models and methods in terms of performance.

5. CONCLUSIONS

In this systematic literature review, we identified and discussed the models that are used in state-of-the-art topic detection and community detection studies. The systematic literature review included a total of 30 papers, 14 for topic detection and 16 for community detection which has been published between 2016 and 2022. After the examination of the selected papers, it can be said that there is a constant and growing interest in data mining studies from social networks such as topic detection and community detection. For these purposes, there is a wide spectrum of choices in terms of methods to be used are available for the researchers. Although there was no single method found that is more successful in the field of topic detection than other methods, it is not possible to make the same deduction in the field of community detection. The results of this SLR indicated the Louvain [42] method has outperformed the other methods for detecting the communities both in terms of usage and performance. In addition to that, there were also some methods for the above-given purposes found that are constructed for problem-specific reasons, it has been found the usage of these models is strictly limited based on the problem.

To summarize, this study offers a comprehensive and detailed review study about the studies in the fields of topic detection and community detection to our knowledge. We provide, an up-to-date view of the studies in the above-given fields.

REFERENCES (KAYNAKLAR)

[1] Internet: Social Media - Statistics & Facts, https://www.statista.com/topics/1164/social-networks/#dossierKeyfigures, 05 January 2022.

[2] H.-J. Choi and C. H. Park, "Emerging topic detection in twitter stream based on high utility pattern mining", *Expert Systems With Applications,* 27-36, 2018.

[3] W. Wu, J. Zhao, C. Zhang, F. Meng, Z. Zhang, Y. Zhang & Q. Sun, "Improving performance of tensor-based context-aware reccomenders using bias tensor factorization with context feature auto-encoding", *Knowledge Based Systems,* 71-77, 2017.

[4] S. Fortunato, "Community detection in graphs", *Physics Reports ,* 1(486), 75-174, 2010.

[5] X. Yao, Y. Zou, Z. Chen, M. Zhao & Q. Liu, "Topic-Based Rank Search with Varifiable Social Data Outscoring", Journal of Parallel and Distributed Computing, 1-12, 2019.

[6] H. Byun, S. Jeong & C.-K. Kim, "SC-Com: Spotting Collusive Community in Opinion Spam Detection", Information Processing & Management, 58(4), 2021.

[7] J. W. Kim, K. M. Lee, M. J. Shaw, H.-L. Change, M. Nelson & R. M. Easley, "A Preference Scoring Technique for Personalized Advertisements on Internet Storefronts", Mathematical and Computer Modelling, 44(1-2), 3-15, 2006.

[8] H. Liu, Y. Ge, Q. Zheng, R. Lin & H. Li, "Detecting global and local topics via mining twitter data", Neurocomputing, 120-132, 2017.

[9] W. Ai, K. Li & K. Li, "An effective hot topic detection method for microblog on spark", Applied Soft Computing, 1010-1023, 2017.

[10] M. K. Linnenluecke, M. Marrone & A. K. Singh, "Conducting systematic literature reviews and bibliometric analyses", Australian Journal of Management, 45(2), 175-194, 2020.

[11] B. Kitchenham, "Procedures for Performing Systematic Reviews", Computer Science, 2004.

[12] B. Kitchenham & S. Charters, Guidelines for performing systematic literature reviews in software engineering, Keele University, United Kingdom, 2007.

[13] H. G. Gürbüz & B. Tekinerdoğan, "Model-based testing for software safety: a systematic mapping study", Software Quality Journal, 26(4), 1327-1372, 2018.

[14] V. Basili, G. Caldieira & H. Rombach, "Goal Question Metrics Paradigm", Encyclopedia of Software Engineering, 1994.

[15] R. Barcelos & G. Travasos, "Software Architecture: Identifying the approaches that evaluate its quality", 2004.

[16] J. Biolchini, P. G. Mian, A. C. C. Natali & G. H. Travassos, Systematic Rewiew in Software Engineering, Systems Engineering and Computer Science Department, Rio de Janeiro, 2005.

[17] Internet: Science Direct, https://www.sciencedirect.com/, 5 January 2022.

[18] T. Dyba, T. Dingsoyr & G. K. Hanssen, "Applying Systematic Reviews to Diverse Study Types: An Experience Report", First International Symposium on Empirical Software Engineering and Measurement (ESEM 2007), Madrid, Spain, 2007.

[19] Z. Saeed, R. A. Abbasi, I. Razzak, O. Maqbool & A. Sadaf, "Enhanced Hearthbeat Graph for Emerging Event Detection on Twitter Using Time Series Networks", Expert Systems with Applications, 115-132, 2019.

[20] L. M. Aiello, G. Petkos, C. Martin, D. Corney, S. Papadopoulos, R. Skraba, A. Göker, I. Kompatsiaris & A. Jaimes, "Sensing Trending Topics in Twitter", IEEE Transactions on Multimedia, 15(6), 1268-1282, 2013.

[21] F. Figueiredo & A. Jorge, "Identifying topic relevant hashtags in Twitter streams", Information Sciences, 505, 65-83, 2019.

[22] A. Kumar, T. E. Trueman & A. K. Abinesh, "Suicidal risk identification in social media", 5th International Conference on AI in Computational Linguistics, Bordeaux, France, 2021.

[23] M. Adedoyin-Olowe, M. M. Gaber, C. M. Dancause, F. Stahl & J. B. Gomes, "A rule dynamics approach to event detecion in Twitter with its application to sports and politics", Expert Systems with Applications, 55, 351-360, 2016.

[24] J. Cigarrán, Á. Castellanos & A. García-Serrano, "A step forward for Topic Detection in Twitter: An FCA-based approach", Expert Systems with Applications, 57, 21-36, 2016.

[25] H.-J. Choi & C. H. Park, "Emerging topic detection in twitter stream based on high utility pattern mining", Expert Systems with Applications, 115, 27-36, 2019.

[26] K. Garcia & L. Berton, "Topic detection and sentiment analysis in Twitter content related to COVID-19 from Brazil and the USA", Applied Soft Computing Journal, 101, 2020.

[27] T. Edwards, C. B. Jones & P. Corcoran, "Identifying wildlife observations on twitter", Ecological Informatics, 67, 2022.

[28] S. M. Sarsam, H. Al-Sammaraie, A. I. Alzahrani, W. Alnumay & A. P. Smith, "A lexicon-based approach to detecting suicide-related messages on Twitter", Biomedical Signal Processing and Control, 65, 2021.

[29] H. G. Yoon, H. Kim, C. O. Kim & M. Song, "Opinion polarity detection in Twitter data combining shrinkage regression and topic modeling", Journal of Informetrics, 10, 634-644, 2016.

[30] M. Garg & M. Kumar, "TWCM: Twitter Word Co-occurance Model for Event Detection", 8th International Conference on Advances in Computing and Communication (ICACC-2018), Kochi, India, 2018.

[31] S. Petrovic, M. Osborne & V. Lavrenko, "Using paraphrases for improving first story detection in news and Twitter", 2012 Conference of North American Chapter of the Association for Computational Linguistics: Human Language Technologies , Montreal, Canada, 2012.

[32] G. R, K. S, P. N & P. V, "Tweedle: Sensitivity Check in Health-related Social Short Texts based on Regret Theory", International Conference on Recent Trends in Advanced Computing 2019 (ICRTAC 2019), Chennai, India, 2019.

[33] Ș. Boghiu & D. Gifu, "A Spatial-Temporal Model for Event Detection in Social Media", Procedia Computer Science, 176, 541-550, 2020.

[34] A. Zamiralov, M. Khodorchenko & D. Nasonov, "Detection of housing and utility problems in districts through social media texts", 9th International Young Scientist Conference on Computational Science (YSC 2020), Crete, Greece, 2020.

[35] M. E. J. Newman, "Finding community structure in networks using the eigenvectors of matrices", Physical Review E, 3(74), 2006.

[36] M. E. J. Newman, "Modularity and community structure in networks", Proccedings of the National Academy of Sciences of the United States of America, 103(23), 8577-8582, 2006.

[37] I. Inuwa-Dutse, M. Liptrott & I. Korkontzelos, "A multilevel clustering technique for community detection", Neurocomputing, 441, 64-78, 2021.

[38] I. Inuwa-Dutse, M. Liptrott & Y. Korkontzelos, "Analysis and Prediction of Dyads in Twitter", International Conference on Applications of Natural Language to Information Systems, Saarbrücken, Germany, 2019.

[39] W. W. Zachary, "An information flow model for conflict and fission in small groups", Journal of Anthropogical Research, 4(33), 452-473, 1977.

[40] D. Lusseau, K. Schneider, O. J. Boisseau, P. Haase & S. M. Dawson, "The bottlenose dolpgin community of Doubtful Sound features a large proportion of long-lasting associations", Behavioral Ecology and Sociobiology, 54, 396-405, 2003.

[41] L. A. Adamic & N. Glance, "The political blogosphere and the 2004 U.S. election: divided they blog", The 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Chicago, Ollinois, United States, 2005.

[42] S. Andreadis, G. Antzoulatos, T. Mavropoulos, P. Giannakeris, G. Tzionis, N. Pantelidis, K. Ioannidis, A. Karakostas, I. Gialampoukidis, S. Vrochidis & I. Kopatsiaris, "A social media anlyrics platform visualising the spread of COVID-19 in Italy via expıitation of automatically geotagged tweets", Online Social Networks and Media, 23, 2021.

[43] V. D. Blondel, J.-L. Hoillaume, R. Lambiotte & E. Lefebvre, "Fast unfolding of communitiers in large networks", Journal of Statistical Mechanics: Theory and Experiment, 10, 2008.

[44] T. Hachaj & M. R. Ogiela, "Clustering of trending topics in microblogging posts: A graph-based approach", Future Generation Computer Systems, 67, 297-304, 2017.

[45] M. Jacomy, T. Venturini, S. Heymann & M. Bastien, "ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software", PLOS ONE, 6(9), 2014.

[46] M. Alassad, B. Spann & N. Agarwal, "Combining advanced computational social science and graph theoretic techniques to reveal adversarial information operations", Information Processing and Management, 58, 2021.

[47] L. C. Freeman, "A Set of Measures of Centrality Based on Betweenness", Sociometry, 1(40), 35-41, 1977.

[48] S. Al-khateeb & N. Agarwal, "Deviance in Social Media and Social Cyber Forensics: Uncovering Hidden Relations Using Open Source Information (OSINF)", Springer, 2019.

[49] F. Ullah & S. Lee, "Community clustering based on trust modeling weighted by user interests in online social networks", Chaos, Solutions and Fractals", 103, 194-204, 2017.

[50] G. Guo, J. Zhang & N. Yorke-Smith, "A novel bayesian similarity measure for recommender systems", Twenty-third international joint conference on artificial intelligence (IJCAI), Beijing, China, 2013.

[51] G. Guo, J. Zhang, D. Thalmann & N. Yorke-Smith, "ETAF: An extended trust antecedents framework for trust prediction", IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014), Beijing, China, 2014.

[52] F. Hu, J. Liu, L. Li & J. Liang, "Community detection in complex networks using Node2vec with spectral clustering", Physica A, 545, 2020.

[53] M. Fieler, "Alhebraic connectivity of graphs", Czechoslovak Mathematical Journal, 2(23), 298-305, 1973.

[54] A. Grover & J. Keskovec, "node2vec: Scalable Feature Learning for Networks", Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, California, United States, 2016.

[55] D. E. Knuth, The Stanford GraphBase: a platform for combinatorial computing (Vol. 1.), New York: AcM Press, 1996.

[56] M. Girvan & E. J. Newman, "Community Structure in social and biological networks", Proceedings of the National Academy of Sciences, 12(99), 7821-7826, 2002.

[57] P. Gleiser & L. Danon, "Community structure in jaxx", Complex Systems, 4(6), 5656-573, 2002.

[58] R. Guimera, L. Danon, A. Diaz-Guilera, F. Giralt & A. Arenas, "The real communication network behind the formal chart: Community structure in organizations", Journal of Economic Behavior & Organization, 4(61), 653-667, 2006.

[59] L. Salwinski, C. S. Miller , A. J. Smith, F. K. Pettit, J. U. Bowie & D. Eisenberg, "The Database of Interacting Proteins: 2004 update", Nucleic Acids Research, 32, 449-451, 2004.

[60] J.-H. Park & H.-Y. Kwon, "Cyberattack detection model using community detection and text analysis on social media", ICT Express, 2021.

[61] M. Huang, Q. Jiang , Q. Qu, L. Chen & H. Chen, "Information fusion oriented heterogenous social network for friend recommendation via community detection", Applied Soft Computing, 114, 2022.

[62] S. Kwon, M. Cha & K. Jung, "Rumor Detection over Varying Time Windows", PLOS ONE, 1(12), 2017.

[63] Z. Xiaomei, Y. Jing, Z. Jianpei & H. Hongyu, "Microblog sentiment analysis with wead dependency connections", Knowledge-Based Systems, 142, 170-180, 2018.

[64] N. R. Usha, A. Réka & S. Kumara, "Near linear time algorithm to detect community structures in large-scale networks", Physical Review E, 3(76), 36-106, 2007.

[65] P. Pons & M. Latapy, "Computing Communities in Large Networks Using Random Walks", International Symposium on Computer and Information Sciences (ISCIS 2005), Istanbul, Turkey, 2005.

[66] M. Rosvall & C. T. Bergstrom, "Maps of random walks on complex networks reveal community structure,» Proceedings of the National Academy of Sciences of the United States of America", 4(105), 1118-1123, 2008.

[67] M. Speriosu, N. Sudan, S. Upadhyay & J. Baldridge, "Twitter polarity classification with label propagation over lexical links and the follower graph", Proceedings of the First Workshop on Unsupervised Learning in NLP, Edinburgh, Scotland, 2011.

[68] D. A. Shamma, L. Kennedy & E. F. Churchill, "Tweet the debates: understanding community annotation of uncollected sources", Proceedings of the first SIGMM workshop on Social media, 3-10, 2009.

[69] D. Singh & R. Garg, "NI-Louvain: A novel algorithm to detect overlapping communities with influence analysis", Journal of King Saud University - Computer and Information Sciences, 2021.

[70] U. Brandes & J. Hildebrand, "Smallest graphs with distinct singleton centers", Network Science, 3(2), 416-418, 2014.

[71] J. Qiu, Q. Chen, Y. Dong, J. Zhang, H. Yang, M. Ding, K. Wang & J. Tang, "GCC: Graph Contrastive Coding for Graph Neural Network Pre-Training", Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, California, United States, 2020.

[72] M. Huang, G. Zou, B. Zhang, L. Yue , G. Yajun & K. Jiang, "Overlapping community detection in heterogenous social networks via the user model" Information Sciences, 432, 146-184, 2018.

[73] J. Xie , S. Kelley & B. K. Szymanski, "Overlapping community detection in networks: the state-of-the-art and comperative study" ACM Computing Surveys (csur), 4(45), 43, 2013.

[74] A. Arenas, A. Diaz-Guillera & C. J. Perez-Vicente, "Synchronization processes in complex networks", Physica D: Nonlinear Phenomena, 21-2(224), 27-34, 2006.

[75] A. Arenas, A. Diaz-Guilera & C. J. Perez-Vicente, "Synchronization Reveals Topological Scales in Complex Networks", Physical Review Letters, 11(96), 102-114, 2006.

[76] G. Xu, M. Hu & C. Ma, "Secure and smart autonomous multi-robot systems for opinion spammer detection", Information Sciences, 576, pp. 681-693, 2021.

[77] S. Rayana & L. Akoglu, "Collective Opinion Spam Detection: Bridging Review Networks and Metadata", Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data, New York, United States of America.

[78] Z. Yamak, J. Saunier & L. Vercouter, "SocksCatch: Automatic detection and grouping of sockpuppets in social media", Knowledge-Based Systems, 149, 124-142, 2018.

[79] W. Jia, R. Ma, L. Yan, W. Niu & Z. Ma, "TT-graph: A new model for building social network graphs from texts with time series", Expert Systems With Applications, 192, 2022.

[80] C. Tu, H. Liu, Liu Zhiyuan & M. Sun, "CANE: Context-aware network embedding for relation modeling", Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Vancouver, Canada, 2017.

[81] J. R. Ashford, L. D. Turner, R. M. Whitaker, A. Preece & D. Felmlee, "Understanding the characteristics of COVID-19 misinformation communities through graphlet analysis", Online Social Networks and Media, 27, 2022.

[82] B. B. Y. Cheng, B. Ryan, D. A. Copland & S. J. Wallace, "Prognostication in post-stroke aphasia: speech pathologists' clinical insights on formulation and delivering information about recovery", Disability and Rehabilitation, 1-14, 2020.

[83] Z. Mossie & J.-H. Wang, "Vulnerable community identificaiton using hate speech detection on social media", Information Processing and Management, 3(57), 87-102, 2020.

[84] M. R. M. Talabis, R. McPherson, I. Miyamoto, J. L. Martin & D. Kaye, "Chapter 1 - Analytics Defined", Information Security Analysis, Boston, Syngress, 1-12, 2015.

[85] Y. Chen, R. Kong & L. Kong, "14 - Applications of artificial intelligence in astronomical big data" Big Data in Astronomy, Elsevier, 347-375, 2020.

[86] C. McCue, "Chapter 7 - Indentification, Characterization, and Modeling", Data Mining and Predictive Analysis (Second Edition), Boston, Butterworth-Heinemann, 137-155, 2015.

[87] N. Tuna, A. Sebatlı Sağlam, F. Çavdur, "Covid-19 Salgını ile İlgili Paylaşımlar Üzerinde Veri Analizi", Journal of Information Technologies, 15(1), 13-23, 2022.

It is the extended version of the paper "Systematic Literature Review of Detecting Topics and Communities in Social Networks" accepted at the International Conference on Informatics and Computer Science 2021.