Istanbul Medipol University

School of Engineering and Natural Sciences

Graduation Project

2021-2022

PROJECT TITLE
Detecting Trending Topics and Influence on social media
FACULTY ADVISOR
Prof. Reda Alhajj
TEAM MEMBERS
Miray Onaran Ceren Yılmaz Feridun Cemre Gülten

Istanbul Medipol University

School of Engineering and Natural Sciences

Graduation Project

Project Code

Project Title: Detecting Trending Topics and Influence on social media	
Faculty Advisor: Prof. Reda Alhajj	
Project Team Members: Miray Onaran, Ceren Yılmaz, Feridun Cemre Gülten	
Sponsor Company (if any):	

BUDGET (TL)	PROPOSED	CONSENTED
IMU FUNDING	8526 TL	8526 TL
SPONSOR COMPANY FUNDING	17100 TL	17100 TL
TOTAL	25626 TL	25626 TL

PROJECT PLAN	PROPOSED	CONSENTED
PROJECT PLAN		
STARTING DATE		

Istanbul Medipol University

School of Engineering and Natural Sciences

Graduation Project

ADVISOR	**DEPARTMENT CHAIR**
Name: Prof. Reda Alhajj	**Name: Asst. Prof. Mehmet Kemal Özdemir**
Contact Information: Tel : E-mail : ralhajj@medipol.edu.tr	**Contact Information:** Tel : 0216 681 5626 E-mail : mkozdemir@medipol.edu.tr
Signature:	**Signature:**
TEAM MEMBER	**TEAM MEMBER**
Name: Miray Onaran	**Name: Ceren Yılmaz**
Contact Information: Tel : 0535 548 7542 E-mail : miray.onaran@std.medipol.edu.tr	**Contact Information:** Tel : 0545 889 59 91 E-mail : ceren.yilmaz1@std.medipol.edu.tr
Signature:	**Signature:**
TEAM MEMBER	**Sponsor Company**
Name: Feridun Cemre Gülten	**Name:**
Contact Information: Tel : 0553 360 31 35 E-mail : feridun.gulten@std.medipol.edu.tr	**Contact Information:** Tel : E-mail :
Signature:	**Signature:**

Istanbul Medipol University

School of Engineering and Natural Sciences

Graduation Project

Project Title: Detecting Trending Topics and Influence on social media
Faculty Advisor: Prof. Reda Alhajj
Team Members: Miray Onaran, Ceren Yılmaz, Feridun Cemre Gülten
Project Group Title:

ABSTRACT

In today's world, social media platforms have a significant potential impact on individuals in various areas and issues. Twitter is one of these social media platforms with its wide communication network. In addition, Twitter has become a popular social media platform for sharing and communication in the field of health. It has the potential to have a significant impact on individuals in the field of health, as it is a platform used by both the public to obtain information and the scientists to share information. At this point, the analysis of the tweets and tweet owner on Twitter in the field of health for Turkey has been considered as an important topic to be focused on. In light of all this, the main purpose of this study is identifying and predicting trending healthcare topics and influential people in Turkey in a specific time period. Thus, important health issues and problems in Turkey will be determined. Accounts with potential impact in the field of health will be identified. In line with this goal, an efficient model as a method is proposed for determining healthcare trending topics and influential people. The proposed method includes the n-grams algorithm to identify trending topics and social network analysis to identify influencers. In addition, with an interface design, it is aimed to display trend topics and influential people for the desired time interval.

Keywords: Trending Topic, Influential People, Twitter, Social Media, Healthcare.

Istanbul Medipol University

School of Engineering and Natural Sciences

Graduation Project

Project Code

BUDGET PROPOSED– (TL)

	ITEMS				
	PEOPLE (3 students)	**MACHINE-INSTRUMENT**	**MATERIALS**	**SERVICE**	**TRAVEL**
IMU FUND	People Cost (300TL/monthly per student for 7 month)	-	-	Office Programs (106TL/monthly per student for 7 month)	-
SPONSOR COMPANY FUND	-	Laptop (5000TL per student)	-	Turk Telekom Internet (100TL/monthly per student for 7 month)	-
TOTAL	6300 TL	15000 TL	0	5562 TL	

BUDGET APPROVED– (TL)

	ITEMS				
	PEOPLE (3 students)	**MACHINE-INSTRUMENT**	**MATERIALS**	**SERVICE**	**TRAVEL**
IMU FUND	People Cost (300TL/monthly per student for 7 month)	-	-	Office Programs (106TL/monthly per student for 7 month)	-
SPONSOR COMPANY FUND	-	Laptop (5000TL per student)	-	Turk Telekom Internet (100TL/monthly per student for 7 month)	-
TOTAL	6300 TL	15000 TL	0	5562 TL	

*Provide pro forma invoice for machines and materials to be purchased.
*Provide technical specifications for machines and services to be purchased.
*Make a contract for services if necessary

1. OBJECTIVE OF THE PROJECT

The increase in the use of computers and smartphones together with the wide access to the Internet has caused social media sites like Twitter, Facebook, and Instagram to have a very large place in daily life around the world. Social media platforms have become able to affect many areas of our lives. In addition, social media platforms have become important sources in the field of information. These platforms cause significant effects on people in terms of obtaining information, receiving news, sharing ideas and opinions. It is ensured that individuals from various countries, age groups, and occupational groups interact with each other by communicating with each other. Social media platforms are social interaction areas where people generate, share, and exchange ideas on a wide variety of topics. Social media are global platforms where distances are broken, and people can relate to societies. All this communication has a significant impact on people. These platforms can be described as important tools that have an impact in areas such as interpersonal interaction, social behavior of individuals, and social issues. Twitter is a social media platform that is one of the most widely used by people. Twitter is an information network with a large user base, based on the sharing of short messages called tweets limited to 280 characters [1][2]. People can share information and communicate via Twitter on a wide variety of topics such as agenda, history, religion, politics, environment, etc.

Twitter is a social media platform that is widely used for sharing and communicating health-related information. Twitter can be expressed as a data source that is increasing in popularity. Therefore, it can be qualified as a versatile platform for research about health. It is stated that scientists and health professionals commonly prefer Twitter to disseminate and share scientific information. Also, it is emphasized that Twitter has become an accepted resource for the public on health-related issues due to its easy access. In fact, it is stated that Twitter is the most used social media platform professionally by scientists [2]. Twitter has the potential to influence large audiences in the field of health. It is pointed out that influential accounts own the potency to influence health issues such as raising awareness about health problems, advocating health, and causing health-related change at the social level. It has the potency to feed into the development of public health subjects at individual, societal and social degrees. In addition, influential people have significant potential in health issues and interventions, especially in promotion and development [3].

Considering the widespread use and impact potential of social media, as well as Twitter's potential in the field of health, it was deemed necessary to address this area for Turkey as well. Identifying and predicting trending topics in the field of health in Turkey is important in terms of detecting problems in the field of health. In addition, identifying and predicting influential accounts in the field of health is important in terms of mass and social interventions and referrals. Considering all this, the main purpose of this study is identifying and predicting trending healthcare topics and influential people in Turkey in a specific time period.

The purposes of this project can be summarized as follows:

- Identifying and predicting healthcare trending topics in Turkey in a specific time peroid, thus identifying health problems on a country basis or regionally.

- Identifying influential accounts (persons or organizations) in the field of health in Turkey in a specific time period, thus revealing their potential owners in the fields of health-related incentives, development, and intervention.
- Proposing a method model for determining healthcare trending topics and influential people.
- By performing an interface design, a certain time period can be selected, and trend topics and influential people for this time period can be displayed.

The targets of this project can be summarized as follows:

- To draw attention to the impact potential of Twitter in the field of health and to raise awareness.
- Analyzing the status of health problems in Turkey. To draw attention to the solution of health problems that need to be improved and prioritized.
- Encouraging future studies involving healthcare trending topics in Turkey in Twitter and to be able to guide relevant future studies.

2. LITERATURE REVIEW

2.1 Background Information

Social Media

All the visual and auditory tools that convey all kinds of information to individuals and society and have three basic responsibilities such as entertainment, information and education are called media. When the usage of media becomes internet-based it can be called as social media. Social Media is an online network where accounts and share their own opinion or media. Some of the popular platforms are Facebook, Instagram, YouTube, Pinterest, Twitter etc. It has positive effects such as recognizing and learning different ideas, quick access to information, acquiring new business contacts, promoting, and marketing the products for the companies etc. On the other hand, it has some negative aspects such as it may create addiction, users' information may be collected from the people who they do not know, also it may affect the spread of misinformation.

Influencer

People who can reach, influence, and direct the thoughts of many people on social media are called influencers. Influencers have an impact on marketing strategies. They are also effective in the emergence of topics that will create trends in society. Influential users and trending topics can be related to any topic such as technology, fashion, economy, or health etc. For example, the ideas shared by the influential users in the vaccine campaigns developed for the covid-19 virus may have changed the viewpoints of the peers against the vaccine. Therefore, the effects of these influencers and occurred the trending topics on social media can be the subject of research.

Twitter

Twitter has become a popular site in which accounts can share their ideas, media etc. Users can post the tweets which are limited as 280 characters. The tweet may be mentioned to other accounts by including ‘’@’’ and hashtags (a keyword or a phrase used to describe a topic or a theme by placing the symbol “#” in front of them) [14][15]. Retweet and favorite are the features that create an effective interaction between the users of the platform. It can be said that as the number of followers of an account increase, its popularity also increases. However, the account with many followers may not be an influential person. Thus, retweet, mention, favorite and key words of the tweets by taking into account the relationship of the users should be investigated to detect influential users and trending topics.

Twitter API

Twitter permits to get the public tweets and accounts. Twitter API is a tool which provides getting data from the unprotected accounts or users. With the Twitter API, the app coded to query it can share one of the quotes from the pre-made list for a user each day, then search for tweets about a user and save them in a file. Twitter data differs from data shared by many other social platforms in that it reflects information that users choose to share publicly. Twitter API supplies reaching public data that users share with the people in the Twitter [16].

Graph

Graph is a network structure that is used to show many real-life problems by establishing a logical relationship [17].

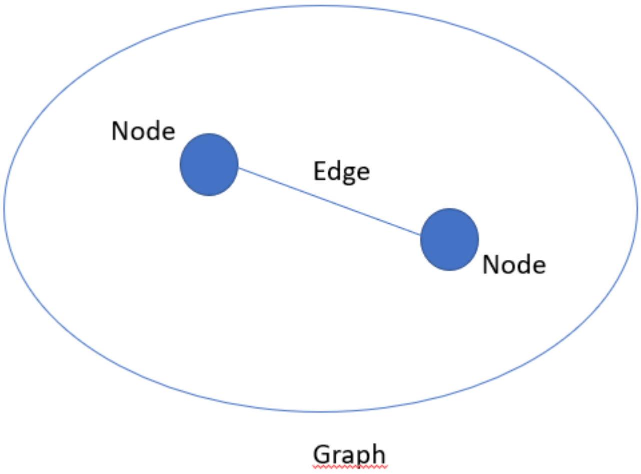


Figure 1:Basic Network Structure.

Node

A node v is an endpoint or an intersection point of a graph [17].

Figure 2: A Single Node.

Edge

A link between two nodes is called an edge [17].

Figure 3: An Edge.

Directed Graph

If the edges in a graph have direction information indicating where the connection starts and ends, such graphs are called directed graphs. Twitter is an example of a directed graph [17].

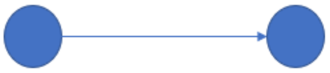


Figure 4: A Graph Representing Two Nodes with a Directed Edge.

Directed Edge

Directed edges, indicated by arrows, are represented by ordered pairs of nodes [17].

Undirected Graph

If the edges in a graph do not have direction information indicating where the connection starts and ends, such graphs are called undirected graphs. Facebook is an example of an undirected graph. [17].

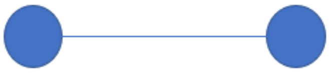


Figure 5: A Graph Representing Two Nodes with an Undirected Edge.

Undirected Edge

Undirected edges are represented by unordered pairs of nodes [17].

Flask

Framework Flask is a web framework from Python language. Flask has a collection of libraries and code that can be used to construct websites. With these features, there is no need to do everything from the beginning [33].

2.2 Scientific Articles

There are various studies about detection and identifying trending topics and influencers. Before starting our study, we investigated some studies related to our project. These related works are summarized and analyzed below.

2.2.1 Towards Identifying Personalized Twitter Trending Topics

Fiaidhi et al. (2012) mentioned some issues about Twitter trending topic in their study. The problem-focused in the study is that although Twitter offers its users a list of the most trending topics, it is hard for understanding of the content of trending topics when it is far from personalization. Based on this problem, Twitter clients has been developed in the study. The developed client can filter personalized tweets and trending topics with the help of algorithms. RSS (Really Simple Syndication) feeds allow for personal preferences to be included. The developed twitter client allows users to group tweets with captions that are explicitly or implicitly defined based on their preferences. Thus, Twitter users will be able to view the elements they are interested in more easily. Two algorithms, which are the Levenshtein Distance algorithm and the LDA (Latent Dirichlet Allocation), were used as a method in the article in order to perform topic clustering. The data part of the study can be summarized as, The United States (New York) and Canada (Toronto) were selected for the location status of the tweets. Geocoding API is used to receive data with this location information. Stream data was collected using the Twitter4j API. Total 2736048 (economy 1795211, education 89455, health 390801, politics 60265, sports 400316) tweets were collected [4].

The positive aspects of the research have been identified as: Firstly, with this study, it will simplify the search and identify tweets of attention by finding personalized trending topics and grouping tweets according to consistently clustered trending topics for more straightened exploration. Thus, users will be able to access tweets in the area they are interested in more easily. They will not get lost in a huge pile of tweets. Secondly, the related research section, where information about topic identification and some methods (TF-IDF and segmentation, latent semantic indexing, latent dirichlet allocation) is presented, is useful because it provides summary information about the relevant techniques. Lastly, the results have been properly presented by means of graphs. The negative aspects of the research have been identified as: the negative side of the article is that it only focuses on shaping tweets based on personal preferences. From the article, we can access information on this subject in general terms.

2.2.2 Influence of Social Media on the Social Behavior of Students as Viewed by Primary School Teachers in Kwara State, Nigeria

Adegboyega (2020) examined the effect of social media in his study. This study was directed to examine the effects of social media and its social behavioral effects on students. It is emphasized as social media has many good as well as many bad sides. Children can be badly influenced by social media when they are not well-watched or monitored. Students and young people can be affected by the negative effects of social media. This study aims to examine and analyze the effect of social media on students' social behavior with the observations of primary school teachers in Ilorin metropolis, Nigeria. Based on the stated problems, a research question was determined about the effects. The descriptive survey method was used in the study as being quantitative research. Surveys were used to collect data within the scope of this method. The purpose of using the surveys is to determine the effect of social media about the social attitude of students. Simple random sampling technique among all 60,054 primary school teachers in Ilorin Metropolis 200 teachers were selected and participated in the survey. Afterward, 3 null hypotheses were assumed and tested in the study. According to the tests conducted, the hypotheses were not rejected, and according to the opinions of the teachers, there was no statistically considerable difference in the effect of social media on social behaviors according to gender, age, and education level [1].

The positive aspects of the research have been identified as: Firstly, in the study, the effect of social media on people and students in today's world was emphasized. The impact and scope of social media are discussed in detail. The positive and negative effects are mentioned. Emphasizing these parts is important and beneficial in order to understand the impacts of social media. Secondly, some statistical methods have been applied. The study includes information on the application and use of statistical methods. The negative aspect of the research has been identified as: Only statistical methods have been used. Apart from that, information in areas such as tweet review, data extraction from tweets, tweet analysis, which are related to our project, were not obtained.

2.2.3 Identifying Twitter influencer profiles for health promotion in Saudi Arabia

Albalawi and Sixsmith (2015) mentioned the effects of social media platforms on public health in their study. The main contribution of this study can be summed up as follows: Contributing to the promotion and development of public health in Saudi Arabia through Twitter, one of the new media platforms. Albalawi and Sixsmith aimed to use ways of applying and comparing multiple impact indicators in the field of public health for the implementation of promotion and development efforts. In line with this, the most influential Twitter accounts have been specified, characteristics (corporate and individual) of accounts have been specified, classification of accounts have been specified in this study. In order to determine the most popular 100 Twitter users 4 different scoring tools, which are primary list development, filtering, classification, and analysis, were used. In the primary list part, a list was obtained using Tweepar and SocialBakers sites that service a list of the top accounts for the country. The final list consists of 182 users. In the filtering stage part, a filtering process was implemented for the primary list. The list is included four several influence-scoring tools. Social Authority by MOZ, PeerIndex, Kred, and Klout was used in order to get a score of 4 influence for each Twitter account. In the classification part, the accounts were first divided into 2 groups, either individually or organizationally. Then they were divided into 2 groups as men and women. Then, 10 groups were classified as religious, sport, media, political, twitter, health, new media, services, royal family, unknown. The final list consists of 99 accounts. In the analysis part, simple calculations, ratios, and percentages were utilized in order to examine data and indicate the best influential accounts [3].

The positive aspects of the research have been identified as: Firstly, the working methodology consists of 4 main parts as Primary list, filtering, classification, and analysis. These 4 stages contain useful perspectives and information. The logic of the filtering and classification stages and the tools used in the primary list and filtering stages include beneficial perspectives and information. Secondly, the study clearly emphasizes the importance and impact of new media platforms and especially Twitter. The negative aspect of the research has been identified as: Too many ready-made tools were used in the method part.

2.2.4 Implications of Twitter in Health-Related Research: A Landscape Analysis of the Scientific Literature

Yeung et al. (2021) worked on Twitter and the healthcare area in their study. The problem can be stated as lack of bibliometric analysis of Twitter usage in terms of research about health, lack of identification of relevant scientific literature, and lack of quantitative analysis. The main virtue of this study can be summed up as follows: the scientific literature on the use of Twitter in the health context was researched and analyzed using a bibliometric approach. Since there were not many studies that performed bibliometric analysis on the use of Twitter in the field of health, Yeung et al. were targeted to determine scientific literature and make quantitative analysis by obtaining a new perspective. In addition, another contribution of the study was to assist researchers who will work in this field in identifying collaborative partners and journals in which research results can be published. The methodology can be summarized as follows: The Web of Science, which is an electronic database, was searched with analyze function in order to find Twitter and articles about health. Basic bibliographic information was acquired by searching. More detailed analysis was performed using VOSviewer, a special bibliometric software. Term map and keyword map synthesis were created. The purpose of this was to visualize repeated words in titles, abstracts, and keywords. A set of 2582 articles was obtained most recently [2].

The positive aspects of the research have been identified as: Firstly, the study draws attention to how social media platforms such as Twitter can be utilized for posts about health. Secondly, a study that analyzes academic articles on Twitter and health can guide people who will work in this field in terms of resources. Lastly, a useful study in terms of bibliographic work in the field of Twitter and health. The negative aspect of the research has been identified as: It was a study that analyzed articles on Twitter and health in the Web of Science database. It is not a detailed study in terms of technique and algorithm. Any other information from this study cannot be accessed.

2.2.5 Automatic Summarization of Tweets in Providing Indonesian Trending Topic Explanation

Winatmoko and Khodra (2013) worked on automatic summarization in their study. The definition of the problem can be summarized as follows: Twitter provides a list of trending topics to its users. Users cannot understand the meaning of the topic because the topics are presented only as a list. In the article, it is aimed to define the characteristics of the trending topic. Studies have been made on automatically explaining the trending topic based on tweet collections. The tweet summarization techniques available do not focus on the trending topic but this article proposes a new method. The method automatically creates a representation for the related trending topic. Winatmoko and Khodra target to identify trending topics that required explanation. They also target the generation of a tweet summary to describe trending topics. The study focused on generating explanations for the trending topic in Indonesia, where no similar research has been found. The trending topic explanation method suggested in the article is carried out in 4 steps: pre-process, topic categorization, sentence extraction, and sentence clustering. For the data, 300 random tweets were collected from Twitter on 1-2 January 2013. 8 trend topics were determined [5].

The positive aspects of the research have been identified as: Firstly, there is a beneficial section called related work is about tweet summarization and how ably they apply. Secondly, a useful article in terms of method and algorithm. The methods and algorithms used, and their places of use are clearly explained. The negative aspect of the research has been identified as: More details could have been given in the 4-step method section.

2.2.6 The Importance of Trending Topics in the Gatekeeping of Social Media News Engagement: A Natural Experiment on Weibo

Yang and Peng (2020) focused on Weibo, which is a social media platform that is widely and actively used in China, in their study. The problem definition can be summarized as determining whether closing the trending topic section influences the post and its interactions. This study is related to digital gatekeepers and trending topics in social media. The user interaction data of 36,239 posts over a 3 week period was analyzed, thus using a natural experiment on Weibo. The method part is divided into two as data and measure. Weibo posts are collected as data. Data were collected from 50 accounts for a period of 4 weeks. A python script was written using the selenium package for post collection. Measure section has been analyzed under 4 headings as presence of trending topics, news engagement, top news items, popularity of news accounts [6].

The positive aspects of the research have been identified as: Firstly, investigating the effects of closing the trending topic is a different matter. Secondly, analyzing over a natural experiment is a positive situation. Lastly, various analyzes and results were used in the results and discussion sections. The negative aspect of the research has been identified as: More detailed information about the method part could have been given.

2.2.7 A Methodology for Identifying Influencers and Their Products Perception on Twitter

The goal of this study [7] is determining influencers and their opinions about the products via Twitter. This analyze give the information about influencers and their information can used on e-commercial and promotions. Business companies are interested in this study because they can use these analyses to increase performance of the commercials [7].

In the study, multilayer network, tensor model and SocialAU algorithm are investigated. Their contribution with this study is using distinct methods and not focused on just one specific topic. They investigated all sectors not just marketing or cosmetics [7].

The methodology of this study is consisted of four main steps; collecting tweets related to decided products set and relevant data which related to influencers, building multiple-layer network and tensor model by using collected data, identifying influencers by using SocialAU algorithm and identifying dominant products and related perceptions. Their perspective is showing the idea of general methodology can applied to different product type. In study, they show in literature review this methodology is used on other studies. However, other studies focus on specific topics. So, this study's main approach gets a result with more various topic [7].

In the other studies, graph theory method is generally used. However, method which is used on this study creating using the relationships by analyzing the information from the tweet's content. In this method, the concept is not analyzing only network topology, also related to opinions expressed on topics of interest [7].

In the article, the steps of methodology explained with details. Twitter API is used while downloading and collecting tweets. Collecting tweets are about targeted products. Each collected tweet is labeled with the mentioned products' name. One of the information get from the tweet is author name. Three layers network is built as a network topology. Each one of the three-layer network's layer include one of these topics, users, products, and keywords. In the study, SocialAU algorithm used to find the most authoritative users. These users sharing tweets regarding determined products. Influencers sorted according to popularity based on selected one specific topic. User's score is used as a mark to determine the popularity of influencers. At the same time, keywords are showing the subject of the shared tweets [7]. Data selection is an important proses for this type of study. Mobile phone manufactures are selected as a topic in the study. 24843 French tweet is collected from May 7th to July 27th, 2015. The number of users is eliminated via some criteria. In the end, they investigated 4953 users [7].

Study is selected specific and distinct subject from similar studies, this gave us advantage to use this methodology on our study. At the same time, each step of the methodology is written detailed. The comparison between different methods gave the advantages of observed and investigated different methods in just one article. As a result, it provides time efficiency and extensive knowledge about methods.

2.2.8 Breaking News Detection and Tracking in Twitter

The defined problem in this study [8] is discovered things requires an effort in Twitter. For this reason, in the study they introduced new solution for these problems which help about collecting, ranking by the number of views, and tracking breaking news in Twitter. They provide webpage which name is Hotstream, to users can easily reach the information. In the study, they examined types of identifying the breaking news which are single message aspect and timeline aspect. Single message aspect investigated in this study. There are two type of single message aspect; one of them is text-based and other one is emotions. In this paper, they are investigated just text-based type single message aspect. The identification can be done by keywords which are nouns and verbs. Some specific nouns and verbs are selected in study, these are names of famous places, people, and events. In the verb part keywords selected as fire, crash, bomb, win etc. They are used story finding method, which is include sampling, indexing, and grouping steps. Most important or read tweets and their topic with hashtags are grouped and ranked. The information, which is gathered from collected tweets, analyzed, and transferred to the application. Users can find and looked trend topics from Hotstream application. Data is collecting with Twitter API. #Breakingnews hashtag is used while data is collecting [8].

Swit Phuvipadawat and Tsuyoshi Murata explained all aspect of the study properly. Organization of paper is neat. The methods and sources which are given in the paper, informed reader. The graph and images of methodology’s steps and distinctions affiliate layout and provided intelligibility.

2.2.9 Identification of Influential Users on Twitter: A Novel Weighted Correlated Influence Measure for Covid-19

This article [9] is focused on propose different approach to identification of influencers on Twitter. The proposed approach is Weighted Correlated Influence (WCI). In the other studies, they used only single parameter method. WCI worked on multiple parameter method. So, this is a new perspective to identification methods. The basic methodology of WCI is calculate the influence score for each chosen user in a network. In WCI graph methodology is used. Graph includes nodes and edges. Nodes and edges are provided relationship between users. Relationship between users represent tweet, retweet, mention, replies, follow-up, and follower. In the study, they use Twitter’s REST API to data collection. At the end of the methodology, algorithms and matrix normalizations applied to the data according to WCI process. Public data is used in the project. These public data collected using #CoronavirusPandemic and #DelhiViolence hashtags. The collected data is store as two separate datasets with 15018 and 18473 number of users [9].

After these articles, we have a more opinion about different methods. The mentioned different methods can be investigated separately. Performance results can extract into our study. Since, in the result part all the varied methods compared with each other, least efficient method can be eliminated.

2.2.10 Sensing Trending Topics in Twitter

In topic detection and tracking subject, there are many different and additional factors to consider. This study [10] is focused on these factors how affect the detection results. In the study, six different detection methods are investigated and explained with details. These methods are Latent Dirichlet Allocation, Document- pivot topic detection, Graph based feature-pivot topic detection, Frequent pattern mining, Soft frequent pattern mining and BN-grams. The collecting data is analyzed with all these methods and results are compared with each other. Data set is consisted of filtered keywords and hashtags. Three different data topic is selected as Duper Tuesday, FA Cup and Election. Each of the data set's collecting time slot is different. For example, Duper Tuesday data collected 1 hour range, FA Cup data collected 1 minute range and Election data collected 10-minute range. The reason of different time slots is each of the datasets change in different time. When it explained detailed, football matched can be changed ant minute, however the election result cannot change every minute. As a result of this study, best method for the noisy data sets is determined which is LDA [10].
In this study, different and various detection methods explained with details. So, six different methods can be observed from the article. Since, there are many numbers of methods in this study comparisons and details can be mixed. So, the reader needs to investigate this study carefully. Three different datasets tested on six methods and in the results part comparison between dataset's results and methods are explained detailed. The comparisons and performance analysis of the methods are beneficial for our study.

2.2.11 Streaming First Story Detection with Application to Twitter

This article [11] introduces different methods to automatically detect outstanding events with a minimal number of insignificances. The main analyze of this method is Story Detection. However, the traditional approach of FSD gives poor performance and high variance with locality sensitive hashing, they modified the FSD method with LSH process. Traditional FSD represent documents as vectors in term of space. After that, each document compared with previous documents and similarity between them is analyzed and finally documents are declared according to this analyze. New modification of FSD virtually eliminates variance and significantly improves performance of the system. In the modified FSD same settings with the UMass system is used. Some steps of this method is to limit the systems to preserving only the top 300 features, set two LSH parameters, the number of hyperplanes k gives a tradeoff between time spent computing hash function. In the result part of the study, the results are compared with manual labeling results. However, the collecting data is nearly 160 million posts. So, some part of the data is manually labeled and compared with the study result. One of the details of the study is # and @ symbols not included to the research. The reason of this researchers doesn't want to some platform specialties doesn't affect the result. As a result of this, this study and method can be used on different platform. Data is collecting for 6 months. The results of this methods are ranking with different threads. These threads are baseline, size of thread, number of users and entropy + users. Size of thread gives the best performance [11].

UMass system features are given in the study. However, the results are not compared with UMass system results. They introduce new method to detect significant events. At the same time, their features not just for Twitter stream. This is an advantage to use this method in multiple streams. This method worked and showed good performance on big datasets. This result is a benefit to who work on big data. However, in the study the modification of FSD needs to be explained more details.

2.2.12 A Structured Mechanism for Identifying Political Influencers on Social Media Platforms: Top 10 Saudi Political Twitter Users

In this study [12], researchers try to determine new criteria to identify social media influencers. They introduce three different features to use in analyzing and determining influencers. These features are the number of followers, social authority, and political hashtags. In this study, researchers focused on Saudi influencers. These selected features matched with three primary metrics. Top 10 political twitter users found by the three metrics by filtering process. The methodology of this study consists of three stages which are gathering data, filtering, and semantic analysis. While data is collecting, users are selected according to primary metrics. In the second filtering process, collected users are filtered with different more personalized criteria. The criteria are, the account needs to be personal account not business or another type of account, the account holder should be Saudi, and it should be an actively used account. The filtering criteria and metric definitions are given in the Table 2. While analyzing the filtered accounts, political hashtags are used. At the end of the study, top 10 Saudi twitter user obtained and analyze with these steps. In this study, public source data is used. The data was not private. As a result of this study, selecting features while filtering can be affecting the results. The filtering features needs to be based on purpose [12].

Table 1: Primary and Secondary Criteria [12].

CRITERIA	CATEGORY	METRIC
Number of followers	Primary	Indegree: A simple importance rank expressed by the number of nodes with a directed edge pointing toward the given node (i.e followers within sampled network)
Social authority	Primary	Interaction: The total number of times that all other users mentioned liked and retweeted the given users within the dataset
Political hashtags	Primary	Knowledge: The number of tweets that a user posts containing context-specific terms divided by the number of tweets in the sample terms derived from a random sample of tweets collected during the sampling period from both networks.

Account age	Secondary
Personal account	Secondary
Sauidi account	Secondary
Active account	Secondary

In the study, the features and filtering process explained properly. It is beneficial for studies working on specialized concepts. Fileting process investigated, it has an advantage while deal with big datasets or concepts. Even though in the paper methods are not written as detailed, determined criteria and process step can help while data collecting [12].

2.2.13 Detecting Influential Users in a Trending Topic Community Using Link Analysis Approach

In this article, the authors proposed a method to detect influential users which is different than the previous research. It is said that influential users are generally detected by considering the retweet, follower, or measurement of centrality. However, this paper underlines that gathering communication relationships among users, users' profile features and link analysis approach are used to find influential users. The compared methods are mentioned as UIRank, FansRank, ToRank, and Retweet Influence. In the compared methods, few features were considered while detecting the influential users in a trending topic. For example, follower amount is considered in the FansRank method. Also, count of retweet of a user's tweet were considered while detecting the influential users. Precision, recall and F1 score were used to compare the proposed method with others. In the paper, the result of each method was constructed as graphs. It was observed that for the Fans Rank method recall, F1 score and the precision at the lowest value. Also, method of Retweet Influence has low value of precision than the proposed method. As a result, it may be concluded that a user's number of followers and tweet counts alone do not indicate whether or not they are an influencer. It may be claimed that in order to determine an influential user using the authors' proposed method, more relationships and traits should be used. [13].

While detecting the influential users, gathering, and processing the data, the methods and tools used in the constructed system are PyQuery, Twitter API, Analytic Hierarchy Process, and PageRank. It is indicated that the algorithm of PageRank was referred for the detection process. Also, PyQuery was used to send queries on the document of XML because it was stated that gathering data dynamically is hard to obtain. It was said that considering a trending topic called "#rp19", 22037 tweets were collected in May 2019 [13].

In this paper, different approaches tried to be used compared to the general research literature that were examined in the paper. The authors score the influential users considering not only for the follower and tweet amount but also the communication relation. Also, the using PR is a useful and clever approach. There is not a lot of negative aspects but it can be said that too much data set cannot be obtained because of cost of the time and extra workload. The features that are focused on by the authors seems few but the techniques that are used provides with successful analysis. It is attractive to add the structure of the proposed system which allows to understand what’s happening in the background. Using AHP and PR is also another interesting method to score and making decisions on the collected data. The PR (link analysis) helps to score influential person which can be used in our project. Additionally, considering the features of a user’s profile and their relationship with the others may help to detect influential more appropriately in our project [13].

2.2.14 Data Science: Identifying influencers in Social Network

In this article, the authors point out that the influential people have a lot of impact on social media users. The effects may change the marketing strategies of a company or industry. Thus, the article focused on to detect influential users about fashion technology topic using the centrality measures such as eigen and degree centrality. Performing social network analysis techniques, the influential users are scored. The centrality is used for understanding the network. Graph theory is used to compute the essentiality of a node in the network. In the paper considering the number of followers and friends, users and their relationships were analyzed. 1000 topmost tweets that are including the fashion technology mentions were collected and 80 users were selected considering the retweet information. It was determined that, based on the information acquired, 90% of the accounts were influenced by fashion technology. The authors mostly focused on the text of the tweets. However, the amount of mentions and retweets was evaluated [14].

To reduce the interaction interfacing, in the article the authors used API call load data to retrieve all the tweet IDs of the given list of 100 tweets based on the fashion industry and stored them in a variable called (data frame). The dataset that is used has a good number of features, but the authors mostly focused on tweet_id and tweet_text. Thus, the number of focused features can be increased [14].

The features that the author focused on seem to be few but the techniques that are used provide a successful analysis. In the article the user with more score is considered as a better influencer in the network about a particular topic or field. In our project the influential people and trending topics can be scored considering some features input as in this article [14].

2.2.15 Agenda Setters of Social Media: A Social Network Analysis on the Agenda Setting Process of Twitter

In this study, Demir and Ayhan (2020) examine the relationship between the policy agenda and Twitter's public agenda using the social network analysis method. The place and importance of the concept of network in today's world has been mentioned. Agenda setting theory and social media issues are mentioned and a comparison of traditional agenda setting model and network agenda setting model is performed. If the method of the study is briefly explained: In this study, the #ayasofyacamii tag created by Twitter users was examined by social network analysis method. Regarding the decision to make Hagia Sophia a mosque again, a hashtag was created on Twitter and approximately 800 thousand tweets were sent to the hashtag #ayasofyacamii between 10 July and 11 July 2020. In this study, after analyzing the dates of 12 and 13 July, the first day of discussion on Twitter, the tweets sent at intervals on 17 July and 19 July were examined in order to determine whether different agenda setters were effective on the public agenda. In addition, on July 24, 2020, the hashtag came to the fore again and approximately 500 thousand tweets were sent. Therefore, it was evaluated on this date. The data of the research were collected through the NodeXL software, which functions as an add-on to the Microsoft Excel program and has an interface that helps to collect the tags or words used by users on Twitter. In the findings and interpretation part of the study, the general view of the networks formed regarding the #ayasofyacamii hashtag, the findings including the relationship between the public agenda and the policy agenda, the analysis results and interpretation of the actors that dominate the network and who are the agenda setters are presented. In order to determine the dominant actors of the agenda, that is, to determine how powerful the actors in the network are on the information flow, they took into account the values of betweenness centrality. The eigenvector centrality method was used to evaluate the agenda setters [18].

2.2.16 A Social Network Analysis Perspective on Twitter Usage Characteristics of Art Schools

In this study, Bakan (2020) examined the corporate Twitter accounts of the forty best art schools in the world between 2018-2019 according to the social network analysis method. In the study, today's internet and social media, Twitter, social network analysis, centrality measures, sample social network analysis studies are mentioned. In this study, Twitter accounts were analyzed according to the social network analysis method, while tweets, hashtags, number of followers, followed, shared photos and information in the videos were taken into account in the analysis. Frequency and mean statistical processes were applied in the analysis of the data. Sentiment analysis, which is one of the content analysis methods, has been used to examine the views on the agenda in art schools as positive, negative or neutral from a class perspective. For the sample of the study, 40 schools determined according to academic criteria were taken into consideration. Information such as the content, followers, likes and comments in the corporate Twitter accounts of these schools were systematically collected for a one-year period (2018-2019). NodeXL program was used for numerical analysis and visualization of the network structure of this study. Within the scope of the research, research questions were determined. With these research questions, the results of measuring density, degree of centrality, betweenness and closeness centrality, and the roles of actors in the network were analyzed. In the study, the distribution of interaction networks on twitter, centrality measurements of universities' twitter networks, ıntuition analysis results of universities' twitter contents are presented and explained [19].

3. ORIGINALITY

When the literature was examined, it was observed that there are several methods and studies while detecting trending topics and influencers on social media. However, the features that are considered for the data are limited. Thus, efficient methods will be proposed in the project by increasing the number of features while investigating the tweets of users. Furthermore, it was discovered that influencers were identified by ranking them based on their number of followers and mentions. In this project, more features will be used while scoring and detecting the influencer users such as retweet, reply, and favorite. Additionally, within our knowledge, considering the literature such a project about detecting influential people and a topic on the health subject has not been observed in Turkey.

4. SCOPE OF THE PROJECT AND EXPERIMENTS/METHODS

As stated in the objective part, the project has two main purposes as to identify trending topics and influencers. In order to achieve these two main objectives, literature research was conducted, and various methods were investigated. Based on the literature reviews and research, some methods have been determined to obtain preliminary results and to make applications. The n-grams algorithm was chosen as the method to determine the trending topics and an application was carried out to get results. Social network analysis was determined as a method to identify influential people and an application was carried out to get results.

4.1 N-grams Algorithm

N-gram is algorithm that used in machine learning. It mostly in used text data in Natural Language Processing. N-grams are basically continuous sequences words, letters or symbols. The operation of the n-gram algorithm is get the data from a document or where data is stored, and separate word groups according to the type of the n-gram. After, it separates the word groups, it lists and stored each word groups separately. The mean of the n-gram's type is value of the n. It can be unigram (n=1), bigram (n=2), trigram (n=3) etc. For each n value, algorithm get that number of words for separation. To explain more clearly, the example can be observed. The sentence is assuming as a example is "The Omicron variant spreads more easily than the original virus that causes #COVID19." [20].

For unigram case (n=1), the sentence separated and stored as below.
"The", "Omicron"," variant", "spreads", "more", "easily", "than" "the", "original", "virus", "that", "causes", "COVID19".
For bigram case (n=2), the sentence separated and stored as below.
"The Omicron", "Omicron variant", "variant spreads", "spreads more", "more easily", "easily than", "than the", "the original", "original virus", "virus that", "that causes", "causes COVID19".
For trigram case (n=3), the sentence separated and stored as below.
"The Omicron variant", "variant spreads more", "spreads more easily", "more easily than", "easily than the", "than the original", "the original virus", "original virus that", "virus that causes", "that causes COVID19".

When n-gram algorithm used, mostly punctuations are removed from the sentences firstly. Because of the punctuations can be affect the analyze. So, the reason of '#' not included into the example is this.
There are various usage areas of n-grams. It can be said most used areas are spam filtering, auto completion of sentences, auto spell check and certain extent. In all of this usage, it learns from the test examples and after that for example in auto completion of sentences purpose, it predicts and suggest some words from previous examples. The type of the n-gram is select according to the project purpose. It can be more efficient to use trigram and fourgram on the spam filtering. However, it can be different for other usage purposes [21].
In this project, n-gram algorithm is used as determine trending topics. The data analysis with n-gram algorithm, after the word groups separated and stored the used number of each word

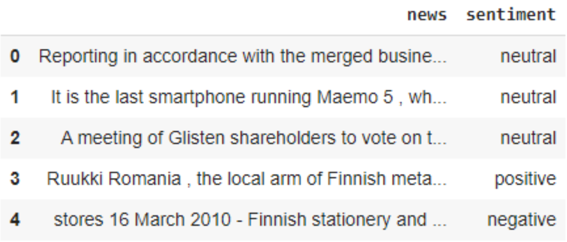
groups is counted. After the counting, the word groups sorted highest number to lower number. As a result of this process, trending topics can be determined from the word group. Because to one topic become trend, one word or word group had to used more than other word/word groups. However, there is a conflict about this part which is if all the words in the dataset will analyze, there will be many stop words, conjunction, articles etc. For this reason, in this project stop words are filtered with the special python library function which is “stop_words ()”. This library includes various words. However, if extra words needed to add to remove, it can be added with code as a extra line. Since the purpose of the project determining trend topics, unigram and bigram gave more proper results for this project.

4.2 Sentiment Analysis

Sentiment analysis is identifying and analyze the text to determine if text is positive, negative or neutral. It used in Natural Language Processing. In nowadays, it can be used for different purposes. Companies use for customer feedback. The big usage can be determined as social media. The comments on the social media can be detected and filtering as positive, negative and neutral with sentiment analysis [22].

In sentiment analysis some steps are followed to get proper result. The data used as a input and after that stop words and some punctuations are filtering. In the next steps, negation handling, stemming and classifications are processed. However, python has a library and specified function for this analysis. So, when the people want to apply sentiment analysis to their own data, they can used directly that library. If there are extra preferences, the classification can be done by that person. Naive Bayes classification, k-NN, decision tree can be used for sentiment analysis. [23]

In this project, NLTK library is used for this project to process the sentiment analysis. NLTK library contains various utilities to analyze specific linguistic data. Since, in the project the data is already split to train and test, pretrained sentiment analyzer which is VADER (Valance Aware Dictionary and Sentiment Reasoner, is used. In the algorithm with the functions, the data is separated as positive, negative and neutral. These results are stored to use on the project and combine with the n-gram part [24].



	news	sentiment
0	Reporting in accordance with the merged busine...	neutral
1	It is the last smartphone running Maemo 5 , wh...	neutral
2	A meeting of Glisten shareholders to vote on t...	neutral
3	Ruukki Romania , the local arm of Finnish meta...	positive
4	stores 16 March 2010 - Finnish stationery and ...	negative

Figure 6: Example of The Sentiment Analysis[23].

4.2.1 Word Cloud

The world cloud is a method to visualize text data analysis. It is based on the repeat number of the world or importance of the world. The importance or the repeat number is increased the size of the word is increased. This visualization method is used to express the general word usage and sorted according to the number or importance. It gives so much advantage. At the same time, since it looks proper and interesting while show the results, it is used when word determination is important in a project. It appeals to the user with those colors and all of words with different sizes. It can implement and process on python directly with the “matplotlib” and “wordcloud” library [25].

In this project, this visualization method is used for as an addition of the n-gram’s result. The reason of this, to observe general written words in the dataset, and attract to users with more colored and different way.

4.3 Social Network Analysis

Social network analysis will be mentioned in general terms and features.

Social network analysis is used widely in areas such as academic studies, business, education, economy, social fields, etc. With the simplest definition, it can be expressed as "the connection between entities". Network is a concept based on graph theory in mathematics. Networks are made up of nodes and connections between nodes. Nodes represented by dots can be composed of various objects such as people, organizations, units. Links between nodes are called edges. Edges are represented by lines or arrows [26].

There are several types of social relationship data in social network analysis. The types relationship can be handled in 4 ways as directed/undirected, weighted/unweighted-binary, two-part network, temporal data set [28].

With another definition, the concept of network is explained as the representation of objects that connect them with links. The three basic elements that make up a network are the actors, the relations of the actors with each other, and the structure that emerges from the different combinations of these relations [26].

Also, social network analysis can be defined as a research approach and method used for the analysis of complex interaction patterns. Individuals, groups, organizations, communities or countries can be considered as units of analysis. Network analysis enables the comparison and analysis of actors at different units and levels [26].

With social network analysis, questions such as who is the most influential person in a network, what is the role of people in the network, how people participate in the social network, how information is spread in the network, how people behave in the social environment can be answered.

Network analysis is defined as a set of approaches that examine the entities that make up the network, the relationships between entities, the relationship models, and the interaction between relationships. Relationship has been defined as “the bond between social entities”. It is accepted as one of the basic features that make up a network of relationships. Entities consisting of various node sets at all levels, such as individuals, groups, organizations, societies, cities, countries, computers, organs, can be specified as vertex/edge in network analysis. In social networks, relationships established between nodes based on one or more types of loyalty are defined as edges or links [29].

There are some concepts that are commonly used in social network analysis. Centrality in these concepts is widely used. Centrality has been defined as a concept that describes the importance of nodes in a network. This concept is a measure of how a node is connected to other nodes, or in other words, the influence a node has on other nodes. Strategic locations in the network usually have the most important or best-known nodes. Different measures of centrality have been proposed to measure this relative importance. The concept of centrality has various measurement methods. Some of these can be listed as degree centrality, closeness centrality, betweenness centrality, eigenvector, pagerank. Centrality metrics enable to identify the most important person or the most central person in the network [29].

4.3.1 Degree centrality

Nodes with more edges according to degree centrality criterion are effective. Degree Centrality is one of the criteria that indicates the extent to which a node is connected to its immediate surroundings and neighbors. Degree centrality is defined as the number of edges connecting a particular node to other nodes. Mathematically, it can be expressed as the sum of each row in the neighborhood matrix representing the network. In terms of social networks, those who communicate with more people achieve greater centrality value. Nodes with a high degree of centrality are recognized by other network members as an important node in a central location in the network [29]. Mathematically, as an example, the degree centrality of an node x, can be expressed as:

$$C_D(x) = \frac{\sum_{y=1}^{N} a_{xy}}{N-1}$$

Quantity of the nodes in the graph is N and *the value of a can be 0 or 1*, considering whether an edge is shared between the nodes x and y.

4.3.2 Closeness centrality

According to the closeness centrality criterion, nodes that have the ability to spread the information in the shortest time are effective. Closeness centrality is based on the concept of distance. Closeness centrality focuses on how close a node is to other nodes in the network. Closeness centrality measures independence or effectiveness. A node that is close to many nodes may have difficulty moving independently without others knowing [29]. It can be expressed mathematically as;

$$C(x) = \frac{N-1}{\sum_{y} d(y,x)}$$

The N represents the quantity of the nodes and d(y, x) is the distance of path between vertices x and y.

4.3.3 Betweenness Centrality

According to the betweenness centrality criterion, nodes that act as bridges in information transfer are effective. Betweenness centrality is based on determining the extent to which a particular node is in between other nodes in the network. Betweenness centrality is the measure of nodes that act as a bridge between two or more sets of nodes that cannot communicate with each other. Mathematically, this criterion is calculated by finding the shortest paths between all pairs of nodes in the network, then proportioning how many of these paths are involved in that node [29].

$$C_B(x) = \sum_{u \neq v \neq x} \frac{\sigma_{uv}(x)}{\sigma_{uv}}$$

The number of shortest paths between vertices u and v is represented with denominator; number of shortest paths between vertices u and v that pass through vertex x represented with numerator.

4.3.4 Eigenvector Centrality

According to the eigenvector centrality criterion, nodes with more edges and nodes with a common edge are effective. Eigen vector centrality is a centrality criterion that calculates not only the centrality of a node, but also the centrality of other nodes connected to that node, showing the effect values of strategically connected nodes. Eigenvector centrality is an expression of the importance and influence of the node in the network. This criterion shows that the edges that are taken into account when calculating the centrality and reflect the relationship of the node with other nodes are not of equal importance. The more centralized nodes it is connected to, the more centralized that node will be. When calculating this criterion, the sum of the centrality degrees of the neighbors is taken into account [29]. In the project, the network(Twitter) that is considered is a directed network. The eigenvector centrality measures mainly work for undirected networks. However, knowing this centrality measure type, gives opinion to understand the networks.

The eigenvector centrality calculation is more complex compared to the others'. The Mathematically the eigenvector centrality is calculated with the equation;

$$C_E(x) = \frac{1}{\lambda} \sum_{y \in M(x)} C_E(y) = \frac{1}{\lambda} \sum_{y \in G} a_{x,y} C_E(y)$$

4.3.5 PageRank

One of the most important applications of the Eigenvector approach is PageRank technology. It was created by Google founders to calculate the essence of webpages from the hyperlink network structure. The method considers the each nodes' score, then find the importance of them [29].

This measure is similar to the eigenvector approach, but it additionally includes the direction of the links between nodes and the importance of the links, that can provide identify the influencer people. Another difference from the eigenvector measure is, the pagerank mainly works for directed networks such as Twitter which is the network studied in the project.

5. PROJECT TARGETS AND SUCCESS CRITERIA

In Detecting Trending Topics and Influence on Social Media project, there are four important success criteria; determine top influencers and trend topics, proposing an efficient model for detection, collecting tweets, create a user interface. Each one of them has a significant impact to project process and result. When each criterion is examined separately, each of them effects can be observed.

Determine top influencers and trend topics is the main goal of the project. If any of the trending influencers cannot be determined in a specific subject, the project will fail. So, at least the top 5 influencers need to be determined to get proper results. When results are compared with real-life trends and influencers, we want 80% of the result matched with the real-life trends and influencers.

Different methods were observed during the literature review. The originality of the project is basically using a distinct method compared to the other project methods. For this reason, there are three choices to extract methods into the project: combining two different methods and obtaining a unique method, choose more than one method and try each of them and select most efficient one as a project method and finally get a hybrid model from existing data and methods. For this reason, finding a unique and efficient method has effect on the originality of the project and result of the project.

Other criteria are collecting tweets as data. The goal of the project is to detect trending topics and influencers. So, gathered data is used for analyzing process. In the project, the number of collected data is important to get more efficient and accurate results. The cost will increase as the amount of data is increased. It is, nonetheless, beneficial to the precision of the result. As a result, the diversity of data will be expanded. Creating a user interface is one of our study’s success criteria. The difference between these criteria and other is there is no significant effect into the study. It does, however, make the observing process easier. Users can view the results by selecting the desired date interval. It is a useful tool for determining the study's impact.

Table 2: Success Criteria and Percentage of Effect.

Success Criteria	Effect to Project (%)
Determine Top Influencers and Trend Topics with 80% accuracy	35%
Proposing an efficient model	30%
Collecting Tweets as data	25%
Creating user interface	10%

6. RISKS AND B PLANS

RISK	B PLAN
Existing tools and techniques may not directly fit the target of the project. Combining them may not be possible.	We may try to find alternative techniques and tune up the techniques to make them together.
The initial design of the proposed model may not work well for the domain.	We may adjust the design by integrating some components that will enrich can the design to make it fit better.
For the interface, classical design tools may not satisfy our target well.	We may try to use new techniques and tools.
Failure to achieve the 80% accuracy rate specified in the success criteria.	Ensuring the most possible optimum rate of accuracy. We may try to do up the system to increase the accuracy by adjusting components.
Failure to collect the expected number of tweets as data for a large period.	We will try to collect concise but descriptive data.

7. WORK TIME PLAN OF THE PROJECT

The work time plan of the project is presented in the tables at the end of the report templates. There are 3 main tables as project activities and work plan, list of work packages, and work package distribution.

8. FINANCIAL EVALUATION

Financial Evaluation is presented in the tables at the top of the report template. There are tables named budget, budget proposed and budget approved.

9. RESULTS

PRELIMINARY RESULTS

9.1 N-grams Algorithm with Open-Source Dataset

In this project one of the determined algorithms is n-grams algorithm. The details of n-grams algorithm are explained in methods section. As a preliminary result, the open-source dataset is tried on unigram and bigram. Since the words used in the tweets can be different structure such as, single word, compound word or two words can make sense when put together. So, unigram and bigram results are analyzed jointly. The open-source dataset gets from Kaggle. The dataset is about the perspective of retail investors. Even dataset is not collected dataset or not related to health, it contains sentences. So, sentimental analyze can be done on this dataset and trend words can be determined. It gave proper results as a preliminary result for this project.

In algorithm there are simply steps.

- Extracting and Reading the dataset
- Train-Test Split for Analyze
- Removing Punctuation and Stop-Words
- Generating n-grams
- Separating words according to Sentimental
- Visualize Result

1.Extracting and Reading the dataset:

In python with a help of pandas library, the csv/excel/text files can be read directly with ‘*read_csv*’ function. Open-source dataset save as csv file. So, it can directly read with specified function.

2. Train-Test Split for Analyze:

Splitting dataset as a train and test part, is for sentimental analysis. The ratio of the train-test split selected as %60 for train and %40 for test. This is a most common splitting ratio in machine learning. However, the ratio can select by trying most optimal ratio.

3. Removing Punctuation and Stop-Words:

Removing punctuation is a critical point of the analysis. Since most of the sentences can include period, comma, question mark, exclamation point, quotation mark etc. If these punctuations are not removed from the dataset, it can affect results. Stop words are defined by python. It includes so many words such as of, why, between, and etc. However, if there are extra words to need to add the stop words to remove from the dataset, it can be added to stop words section with stop_words(‘english’) + [‘the’, ‘food’, ’covid’]. These stop words are used in almost each sentence and not specify the topic of the sentences. In python, there are libraries for both punctuation and stop words. So, this step can be completed easily. However, in stop word case, there are few available languages. Turkish is one of the available languages[27]. Since, in this project Turkish tweets will be collected, available languages are crucial for this part. The other important point is stop words sensitive for upper and lower case. When additional work added to the list such as “the”, if the sentence includes “The”, it won’t be remove from the dataset. Since specified word begins with lower case however dataset includes upper case version. For this problem, beginning of the algorithm, all dataset letters change to the lower case with ‘*str. lower ()*’ function.

4. Generating n-grams:

There is algorithm for generating n-grams. In this code part, when select specified n-gram value. The algorithm works for that n-grams version. In this project, unigrams and bigrams are used. For unigram case, n-grams defined as 1 and for bigram n-grams defined as 2. In unigrams, algorithm separate sentences to one-by-one word. In bigrams, algorithm separate sentences two words each time. The advantages of the various type of n-grams are, it can be used for specified purpose. In this project, since trend topics needs to be determined, unigram and bigram are most suitable options.

5. Separating words according to Sentimental:
With this algorithm, sentimental analysis is done. The sentences can separate as positive, negative and neutral. There is library in python for this purpose which is *defaultdict from collections*. Dataset analysis and for each sentences separate to determined store (positive, negative, neutral) with this algorithm. After this separation, each store sorted to find top 10 most used words. These words stored with two features; one of them is the word itself and how many times this word is used.

6. Visualize Result:
In visualization part, *matplotlib* library help to create bar charts for results. The results are separated two main parts: unigram results and bigram results. In these two parts, it separated three figures: positive, negative and neutral. It can be observed when unigram and bigram results are compared, the bigram shows more reasonable topics. Because in the unigram, the words can be apart from the context even filtering is used. However, it can be change for purpose and dataset.

- Unigram

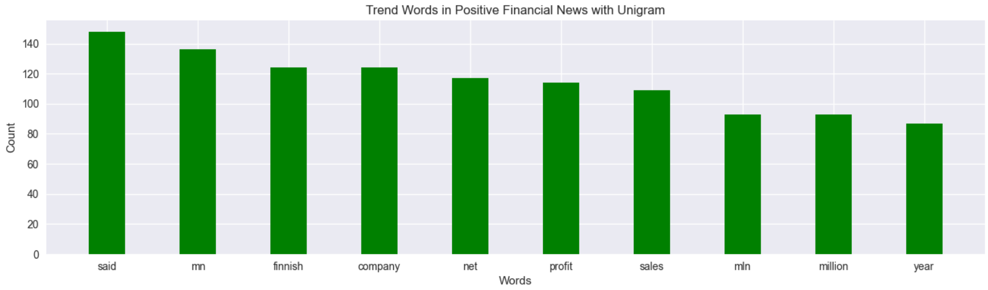


Figure 7: Top 10 Trend Words in Positive Financial News with Unigram (ngrams=1)

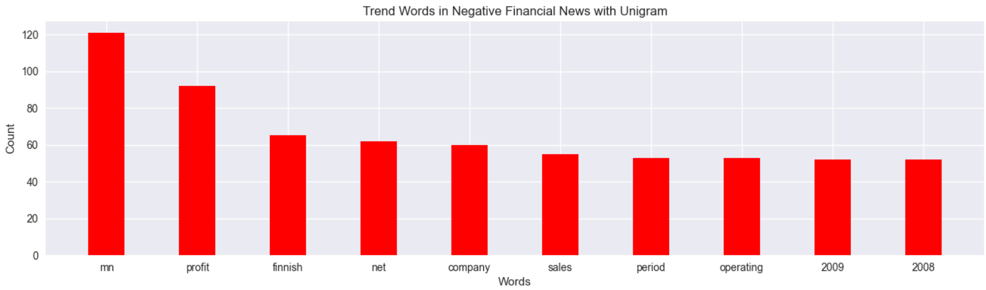


Figure 8: Top 10 Trend Words in Negative Financial News with Unigram (ngrams=1)

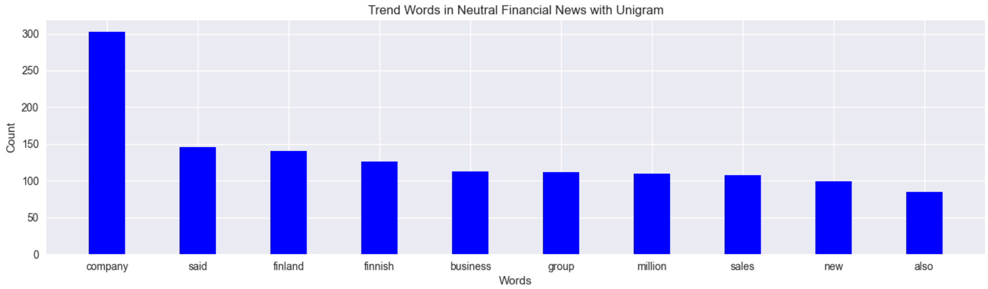


Figure 9: Top 10 Trend Words in Neural Financial News with Unigram (ngrams=1)

- Bigram

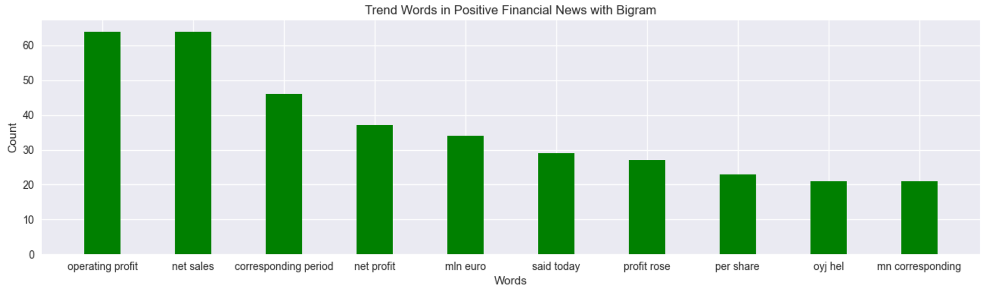


Figure 10: Top 10 Trend Words in Positive Financial News with Bigram (ngrams=2)

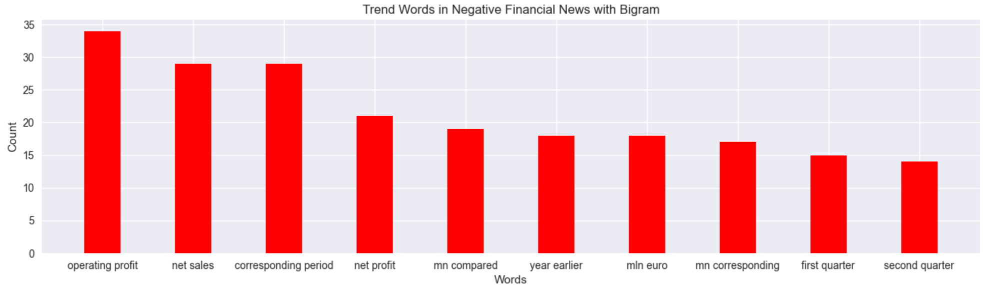


Figure 11: Top 10 Trend Words in Negative Financial News with Bigram (ngrams=2)

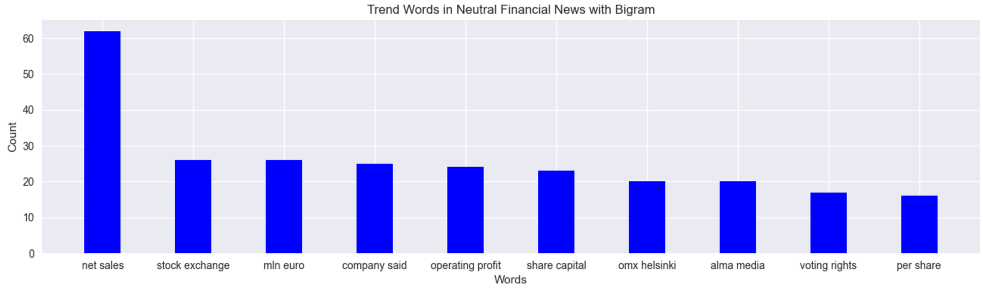


Figure 12: Top 10 Trend Words in Neutral Financial News with Bigram (ngrams=2)

These preliminary results show, it can work proper on sentences. It is not determined just the trend words, at the same time it find and specify if the word used as positive way, negative way or neutral way. When, the tweets will be collected in this project, tweets will be saved as csv file. After, saving the original dataset as csv file, this code can be used directly with minor changes such as name of the columns etc. As a result of this, n-grams will be helpful in the project's process and development with other algorithms. If there is no problem, It can gave the trend topics directly.

9.2 Word Cloud

To get familiar with sentiment analysis which is relevant with detecting trending topics part of the project, a word cloud was created using Python by importing the TextBlob and WordCloud libraries. It is basically a visualization method to get most frequent words in texts. A Covid-19 mock data which includes 1000 tweets from 2020 to 2021[32]. was used to create a wordcloud. To obtain words that emerged during the pandemic and show the impacts of covid-19 on the social media, some stop words were used such as covid, corona, coronavirus, virus etc.

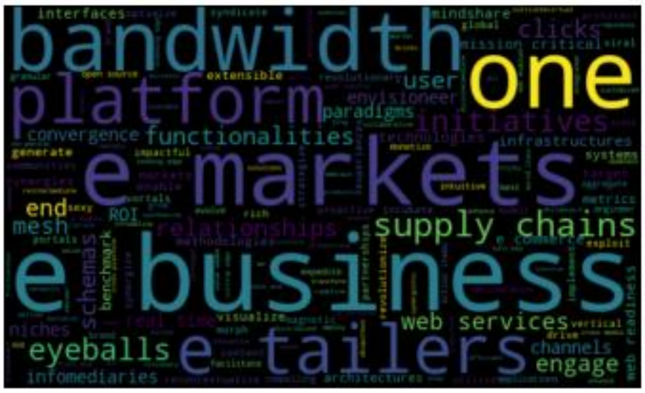


Figure 13: WordCloud Representation using Mock Data from Covid-19 Tweets.

From the figure A it can be observed that, 'e businesses has the biggest size in the wordcloud which shows that it is the most used word in the tweets during the pandemic between the relevant years considering the dataset. Also, the words, 'e markets', 'platform', 'e tailers', 'one' draw the attention which are the other frequent words in the tweets.

9.3 Centrality Measures

A mock data was used to evaluate centrality measures and detecting important nodes, which can be tought as a representation of influential people, in a graph. The mock data is a part of a Twitter network data which was referred from a data science github repository [31]. Python was selected as the programming language and the Networkx library was used in order to analyze and evaluate the network. The data includes some users' occupation levels as celebrity, politician and scientist. From the network data, the nodes were extracted which have the metadata occupation as celebrity beside their neighbors and the networks were visualized.

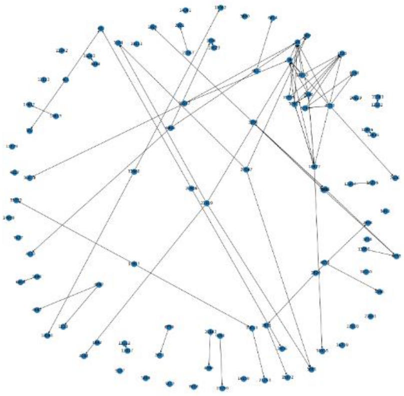
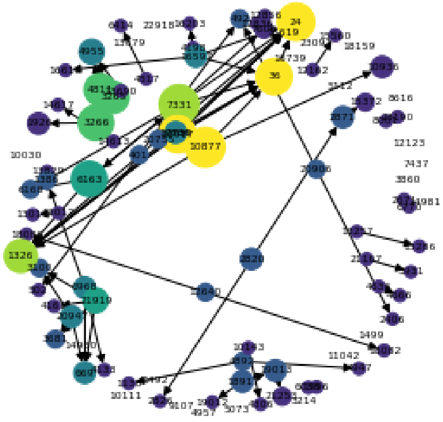
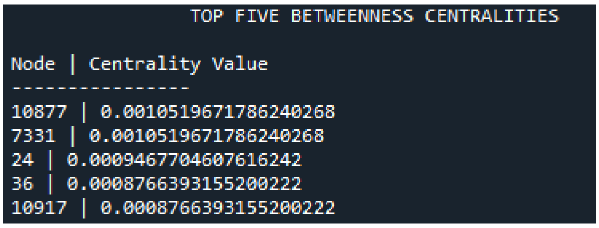


Figure 14: The Default Network Graph without Labels Using the Mock Data, Occupation as Celebrity.

The centrality measures were obtained from the network and the nodes with the top five scores were recorded among 192 nodes. In order to observe the network from a better perspective, Figure 14 was visualized as; node colors varies with degree(number of connection/neighbor a node has) and the node size with betweenness, degree, closeness and pagerank measures.

9.3.1 Betweenness Centrality Measure



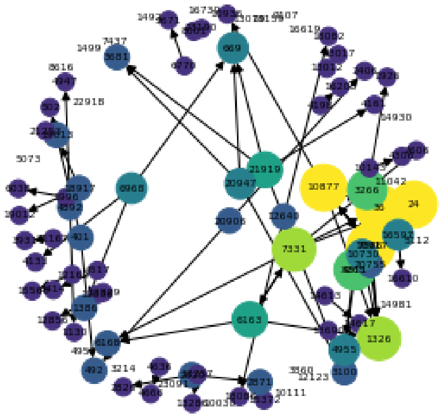


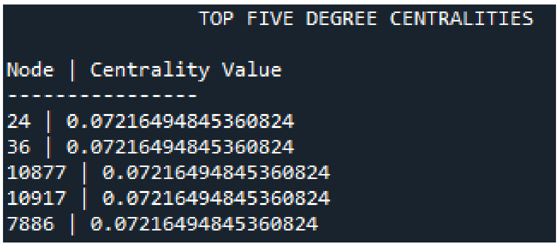
```
                TOP FIVE BETWEENNESS CENTRALITIES

Node | Centrality Value
----------------
10877 | 0.0010519671786240268
7331 | 0.0010519671786240268
24 | 0.0009467704607616242
36 | 0.0008766393155200222
10917 | 0.0008766393155200222
```

Figure 15: Betweenness Centrality Measure Result with Mock Data.

Nodes with the top five values were recorded as 10877,7331,24,36 and 10917. The score of the node 10877 is same with 7331 and also the score of the node 36 is same with 10917, which shows they have the same importance in the network considering the betweenness measure. However, as can be seen the colours of 10877 and 7331 are different which shows they have different degree which means the number of neighbors of them are different. For these five nodes considering the betweenness centrality it can be concluded that, the frequent of the nodes in the shortest path among the edges are higher than the others. They can be assumed as an important bridges that allows to flow essential information in this dataset. In the graph it can be observed that, these nodes are labelled and have bigger size compared to the others considering the betweenness measure. The node 10917 may not be seen clearly due to the density of the nodes in that part of the network.

9.3.2 Degree Centrality Measure



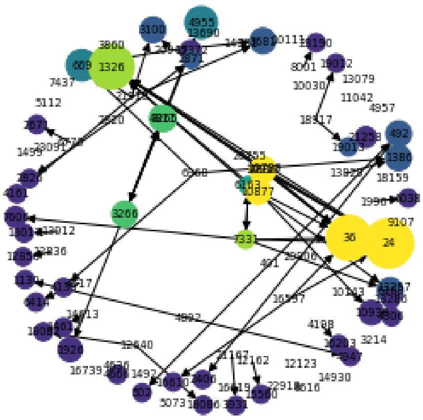


```
                    TOP FIVE DEGREE CENTRALITIES

Node | Centrality Value
----------------
24 | 0.07216494845360824
36 | 0.07216494845360824
10877 | 0.07216494845360824
10917 | 0.07216494845360824
7886 | 0.07216494845360824
```

Figure 16: Degree Centrality Measure Result with Mock Data.

Nodes with the top five values were recorded as 24,36,10877,10917 and 7886. The score of the top five nodes are equal which shows they have the same importance in the network considering the degree centrality measure. Thus, it can be said that these nodes are the most connected nodes among 192 nodes because the degree centrality measure assumes that the important nodes have many connections and the number of the edges attached to these nodes are higher than the others, In the graph it can be observed that, these nodes are labelled and have bigger size compared to the others considering the degree centrality measure. Some nodes may not be seen clearly due to the density of the nodes in that part of the network.

9.3.3 Closeness Centrality Measure



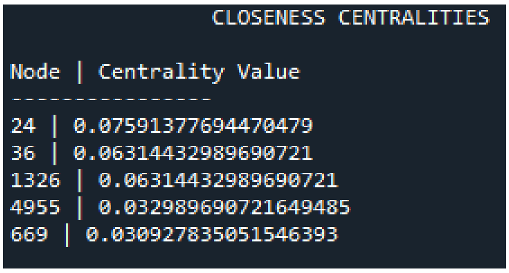
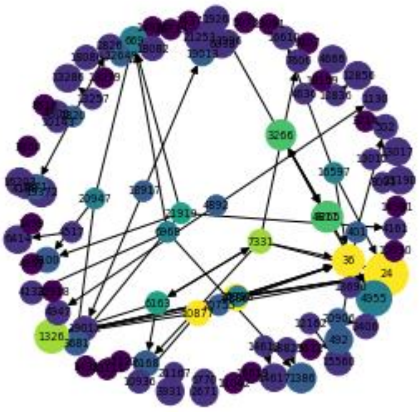

```
                    CLOSENESS CENTRALITIES

Node | Centrality Value
----------------
24 | 0.07591377694470479
36 | 0.06314432989690721
1326 | 0.06314432989690721
4955 | 0.032989690721649485
669 | 0.030927835051546393
```

Figure 17: Closeness Centrality Measure Result with Mock Data.

Nodes with the top five values were recorded as 24,36, 1326, 4955 and 669. In the graph it can be observed that, these five nodes are labelled and have bigger size compared to the others considering the closeness centrality measure. Thus, they are the top five central nodes among the 192 nodes. The measure can be used identify how quickly an event or a word may be spreaded by the important nodes(influential people) in the network because the important nodes are evaluated as closer ones to the other nodes. In this dataset it is the node 24.

9.3.4 PageRank



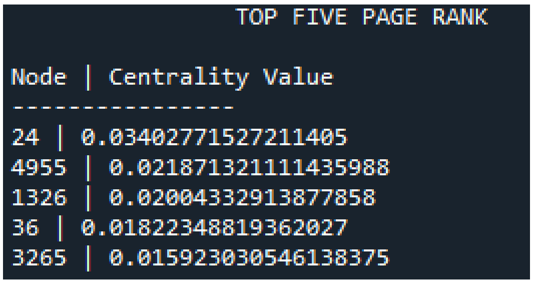

```
                  TOP FIVE PAGE RANK

Node | Centrality Value
----------------
24 | 0.03402771527211405
4955 | 0.021871321111435988
1326 | 0.02004332913877858
36 | 0.01822348819362027
3265 | 0.015923030546138375
```

Figure 18: PageRank Centrality Measure Result with Mock Data.

Nodes with the top five values were recorded as 24,4955,1326,36 and 3265. In the PageRank measure, the direction of the links and the importance of those links are taken into account, then an importance score is assigned to the nodes. Thus, these nodes are the top five important nodes and includes important inlinks. The most important node relative to the others is 24 considering the PageRank measurement. As can be seen, node 36 has the same degree number with the node 24(the colour of the node represents the degree value), but the importance of the node 24 is higher than the node 36 which shows that the number of connections or the number of neighbors may not reflect the importance of a node in a graph or the influential people in a social network.

9.4 Preliminary Result with Collected Data

The data is collected as a tweet which is collected according to different subject categories. The data is included people's ids, their tweets, reply to information etc. In this result part, the n-gram algorithm, which is written and applied to sample data, is used for some part of real data.

The selected subject is sma_ilac_saglık for this preliminary result to observe if this algorithm is work on real data. Since the collected data is not include sentiment analysis, there is no positive, negative, and neutral separation. As a next step, all categorized data will be combined, and sentimental analysis applied to combined data.

The n-gram algorithm is used after sentiment part is removed from the code. The algorithm is edit based on new data. The column names and filtering the words section is changed. The filtering part is changed since the sample data was included English comments. However, real data includes Turkish tweets. The filtering language is changed English to Turkish. The results can be shown on below figures. The selected n-grams are unigram and bigram. Because these n-grams more convenient for this project. Trigram can be applied as an additional result. This will be no problem since the algorithm includes for loop for n-gram. If for loop change as a three, the trigram result will be obtained.

In this process, currently there is a problem about Turkish character. In data there are symbols rather than some special letters for Turkish such as ğ, ç, I, ş, ü, ö. First, replacing method was tried on this problem. The characters (i, ç) are changed with c. However, the method is not worked. This problem will be tried to solve with some encryption or another method.

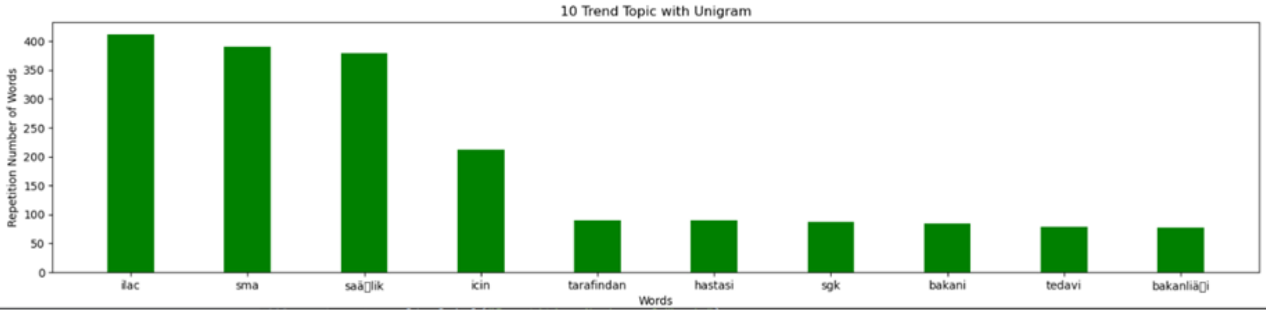


Figure 19: Unigram Result with SMA_Ilac_Saglık Tweet Dataset

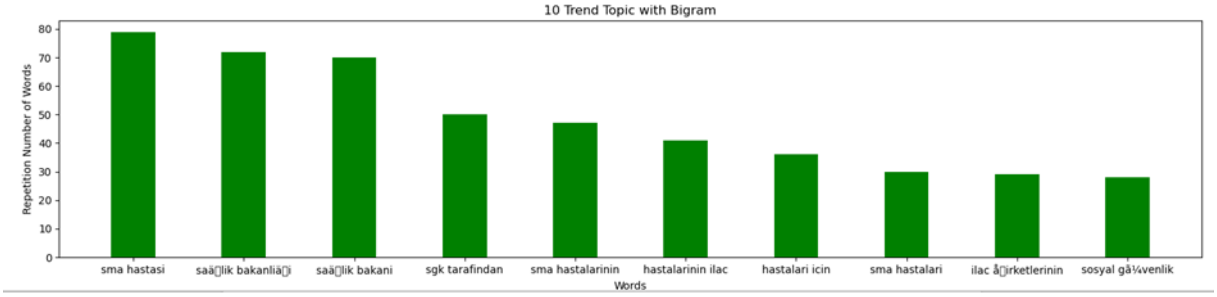


Figure 20: Bigram Result with SMA_Ilac_Saglık Tweet Dataset

Sentimental analysis was performed for the sample data used for the preliminary result. Sentimental analysis was not performed for the new data to be used for the application. Therefore, there are some differences. Data were collected from Turkey, as the project focused on identifying trending topics and influential people in health in Turkey. For this reason, tweets contain Turkish characters. Turkish characters in the data are seen as some symbols. This problem will be fixed by encryption or other methods. Tweets divided into separate topics from the collected data will be evaluated both separately and collectively. According to this evaluation, an analysis will be performed on how to obtain the most efficient result.

The results were obtained for the application carried out for the pre-delivery. However, there are some deficiencies mentioned above. These deficiencies will be completed by the final reports and improvements will be realized. Thus, the results will be more efficient. The application work for the n gram algorithm we use to find trending topics can be summarized in this way.

FINAL RESULTS

9.5 FINAL RESULTS WITH COLLECTED DATA

9.5.1 Trend Topics Results

In the project n-gram algorithm is used to find trend topics. Four type of the n-gram which are unigram, bigram, trigram and fourgram, is used to present more detailed and logical data analysis. The n-gram algorithm method is explained with detailed in the methods part of the report. The n-gram algorithm gets the n-gram type as a variable. The four different types of the n-gram results are obtain with using for loop to automate the process.

The collected tweets consist of seven different health-area topics. The collected tweets are written with Turkish characters, because of it collected from Turkey subject. The data includes various features (created date, tweet text, id, user_id etc.). In the n-gram, the essential features are tweet text and created date. First since the data collected as separately based on topic, seven file is merge as a one combine data file. At the end of this process, there are 6767 number of row tweets to analyze. Since, the data files collected as a JSON file, these files read on python and converted, saved as a CSV file.

Encoding

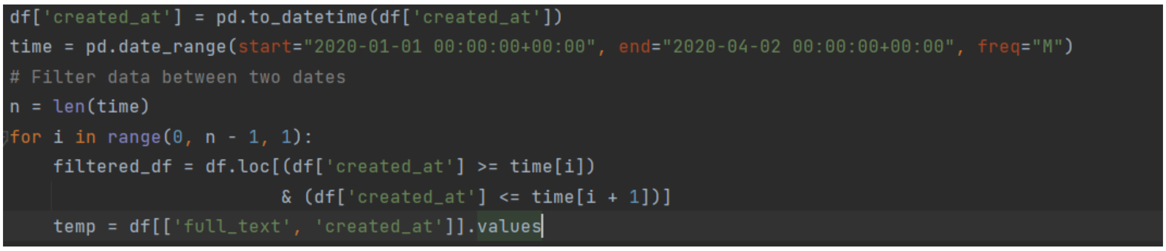
The data is included Turkish characters, it caused some symbol problems. Some of the special Turkish characters are shown in the result as symbols. This problem solved with special encoding which is utf-8. UTF-8 is used for electronic communication, and it known as variable-width character encoding [34]. While reading csv file, read with the encoding features and this solution is worked properly. All the characters shown in the result without any error.

Stemming

After merging, reading, and encoding the data, stemming must be done. Stemming is the process of producing morphological variants of a root/base word. Mainly, it returns the words root form which is without the word suffix. The reason to apply this process is in the result of trending topics, similar words are occurred such as 'bakan'-'bakanlar', 'aşı'-'aşının'. Since these similar words give the same topic, the results have become repetitive. The known stemming libraries written for English or other most-used languages. The TurkishStemmer[35] library is a special library for Turkish language stemming. For this reason, TurkishStemmer library is implemented on the algorithm. The stemmer library implemented to n-gram algorithm part from the program to stemming each word. After the stemming, some words showed wrong. The reason of this error is the stemming get the root of the word and some of the words in Turkish can be different meaning with word suffix and without suffix. The example of these words is 'aş-aşı', 'bak-bakım' etc. Even after removing the suffix 'bak' word is meaningful, the original of the word which is 'bakım' has different meaning. For this reason, some customization is done. This customization is done by special python function which is replace(). In the algorithm firstly get the words list and map process, after that replacement with the wrong and original form process is done. This customization worked on this project without any error.

Date Selection

Date specify is an additional feature for this project. The data includes created date of the tweet, this additional extension is suitable for this project. The results are present based on specified date with this feature. Special python function is used to get and specify the date. Firstly, tweet's dates (created_at column name) are collected with datetime function. After that, a variable defines to store date range. This date range defined with date_range function. 'date_range' function take the start, end date and the frequency of the date range. The frequency can be month, year, week, day even it can be specified as a custom such as 3 months or 6 days. Each of the date interval are represent different date range result. For this reason, for loop is used to obtain four types of the n-gram result based on each date interval. In this loop loc() function is used to compare the created date and define date. The date range assign to a variable to store as a array and in the for loop this array index compared with the created date.



```
df['created_at'] = pd.to_datetime(df['created_at'])
time = pd.date_range(start="2020-01-01 00:00:00+00:00", end="2020-04-02 00:00:00+00:00", freq="M")
# Filter data between two dates
n = len(time)
for i in range(0, n - 1, 1):
    filtered_df = df.loc[(df['created_at'] >= time[i])
                         & (df['created_at'] <= time[i + 1])]
    temp = df[['full_text', 'created_at']].values
```

Figure 21: Snapshot of the Date Specify Code.

In each of the array index, there is one date, and this dates frequency is a month. For example; time [2020-01-01, 2020-02-01,2020-03-01,...]. For this reason, in the for loop the created date is compared current index and next index. In each iteration tweets that their created date is same as specified month, used as a data frame to n-gram analysis. As a result of this, after all the iterations finished, multiple results are obtained for each month of each year between 2020 and 2022-03.

Data Visualization

Plotly is a graphing library for python. It is open-source, interactive and browser-based library. In this project, plotly is chosen to visualize result. The reason of this, it is browser-base and interactive. The charts can be managed by user and user can zoom in or zoom out. The properties of the plotly charts are various. This gives an advantage to present the results more user friendly. At the same time, since the results are present on the webpage, it can be implemented on the webpage more easily with some extensions.

The a few of the results for trending topics are shown on the below. Two different date is selected to present. However, in the real result include figures for each month between 2020.01 to 2022.03. 26 figures are present on the webpage.

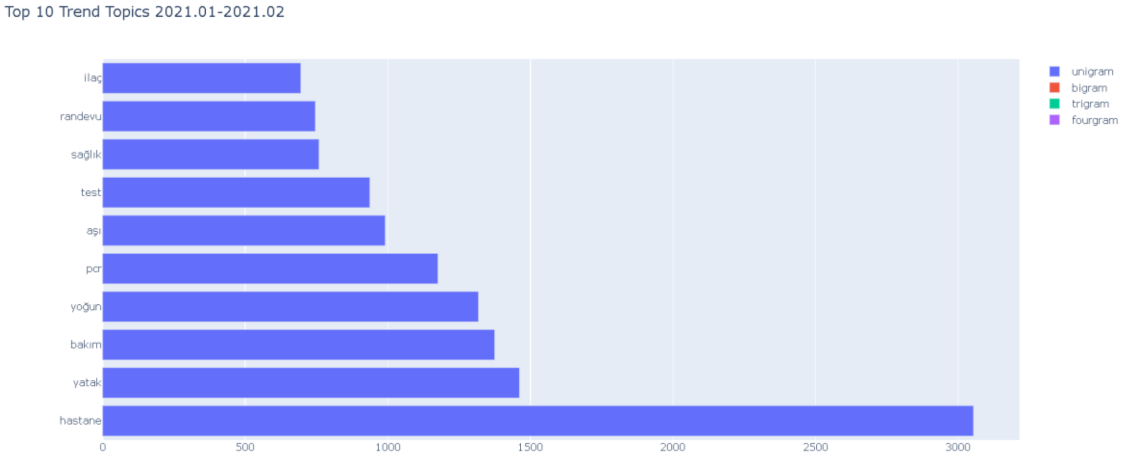


Figure 22: Trend Topics Result Unigram (2021.01-2021.02)

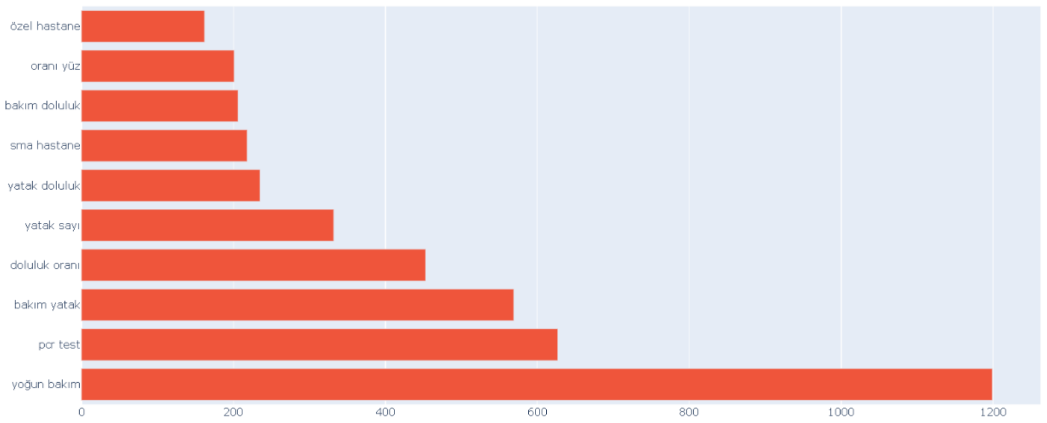


Figure 23: Trend Topics Result Bigram (2021.01-2021.02)

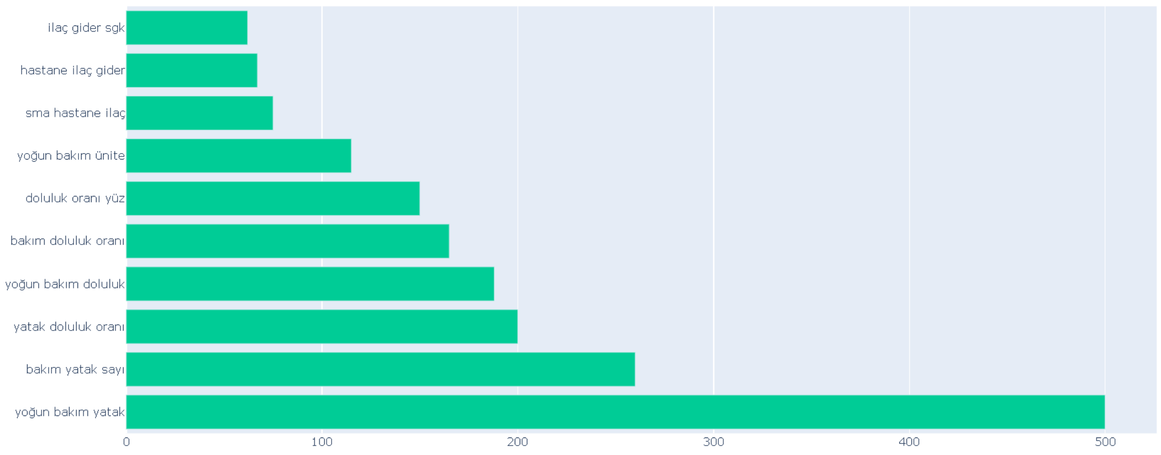


Figure 24: Trend Topics Result Trigram (2021.01-2021.02)

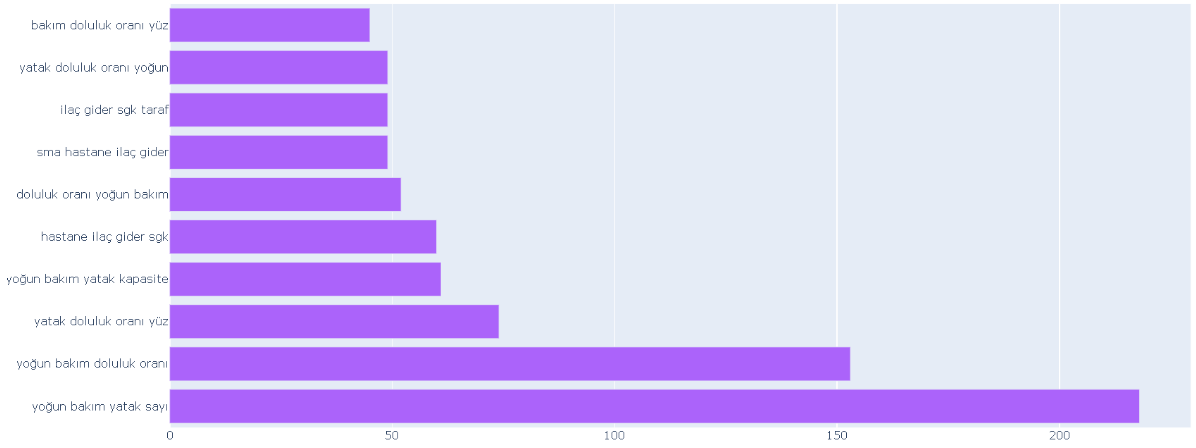


Figure 25: Trend Topics Result Fourgram (2021.01-2021.02)

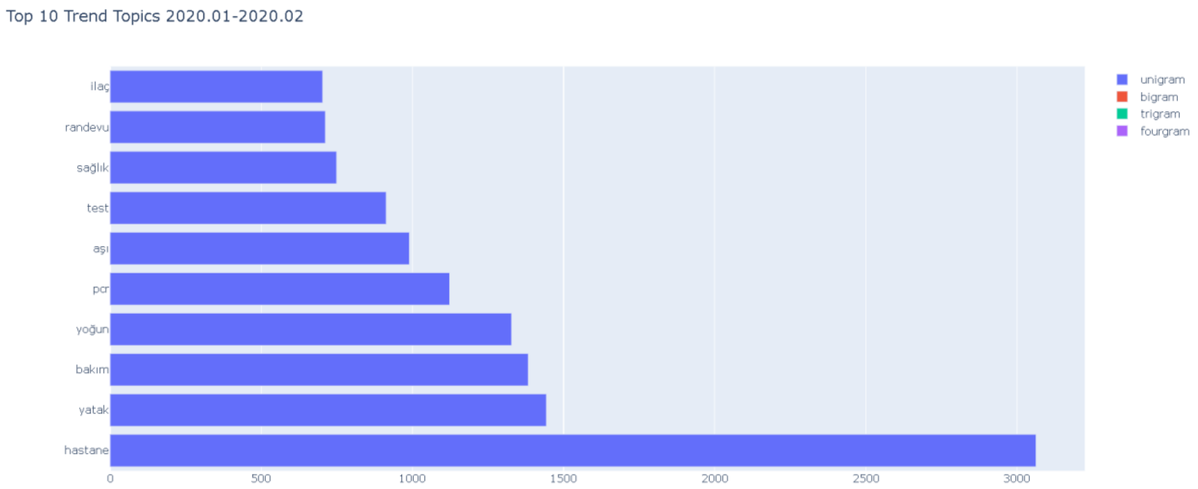


Figure 26: Trend Topics Result Unigram (2020.01-2020.02)

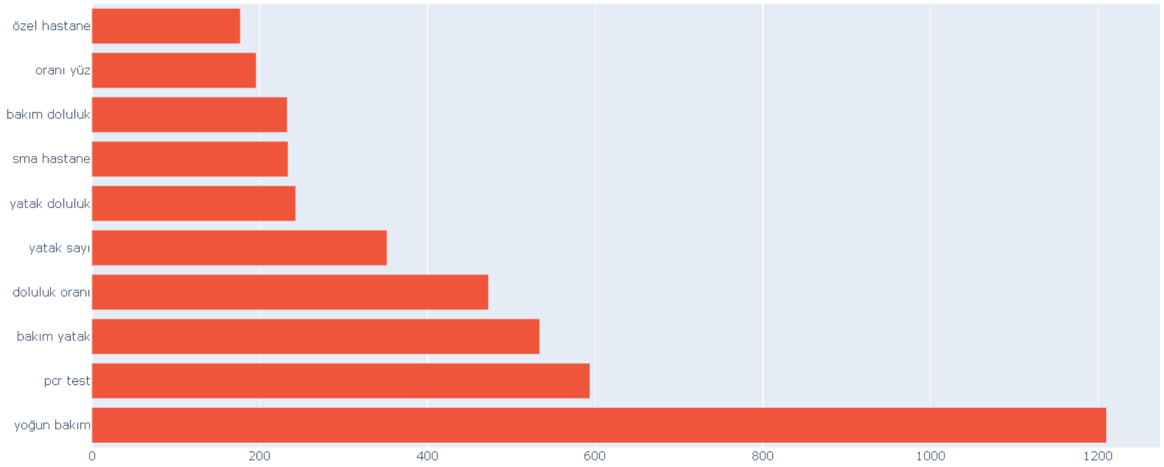


Figure 27: Trend Topics Result Bigram (2020.01-2020.02)

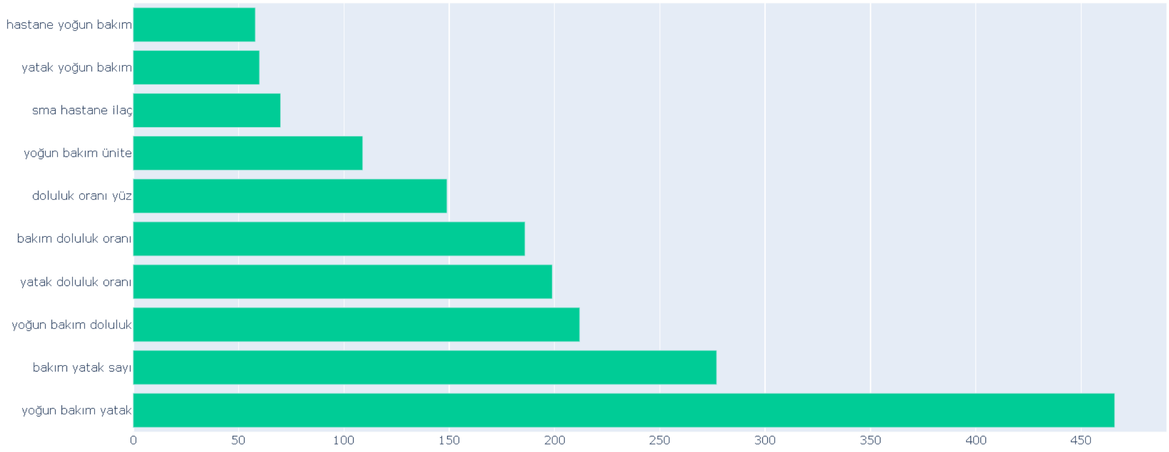


Figure 28: Trend Topics Result Trigram (2020.01-2020.02)

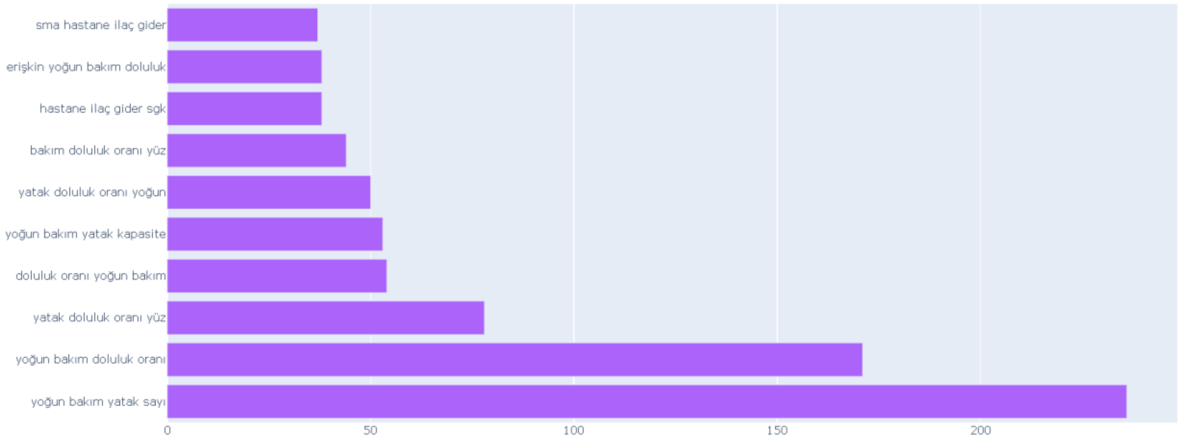
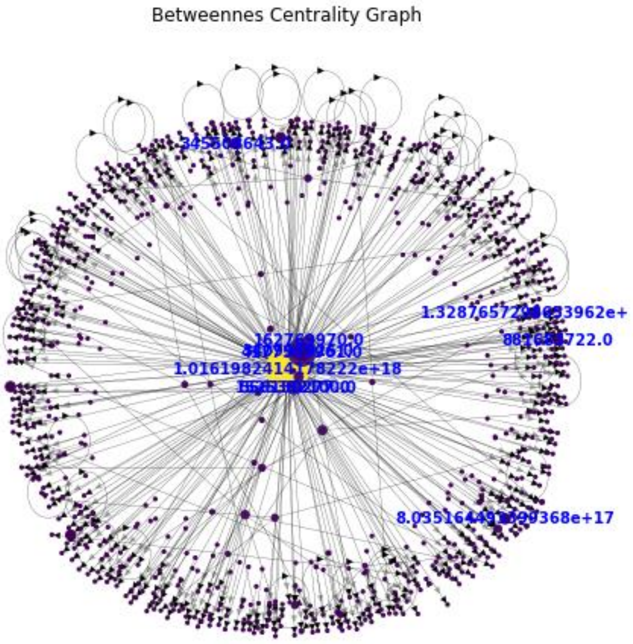


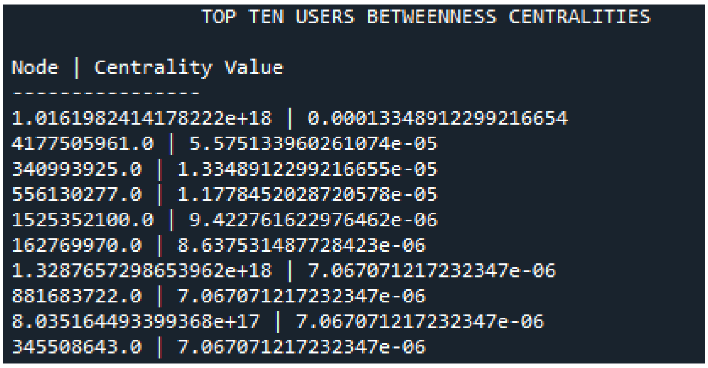
Figure 29: Trend Topics Result Fourgram (2020.01-2020.02)

9.5.2 Social Network Analysis Result

Using the collected tweet data from 2020.01.01 to 22.03.2022 the most influential users(as a node) were found in different topics about health. In order to observe the network from a better perspective the node colors varies with degree(number of connection/neighbor a node has) and the node size with betweenness, degree, closeness and pagerank measures. In the graph it can be observed that, the important top ten nodes are labelled and have bigger size compared to the others in each measures. In some of the graphs the labels of the nodes may not seen clearly because some of the networks of a topic include more than 2000 nodes but the importance of the nodes are also recorded as a table in a descending order. While evaluating the users(nodes) in the network , the reply relation between users , reply counts of a user, retweet , favourite and quote count of the tweets were taken into account.

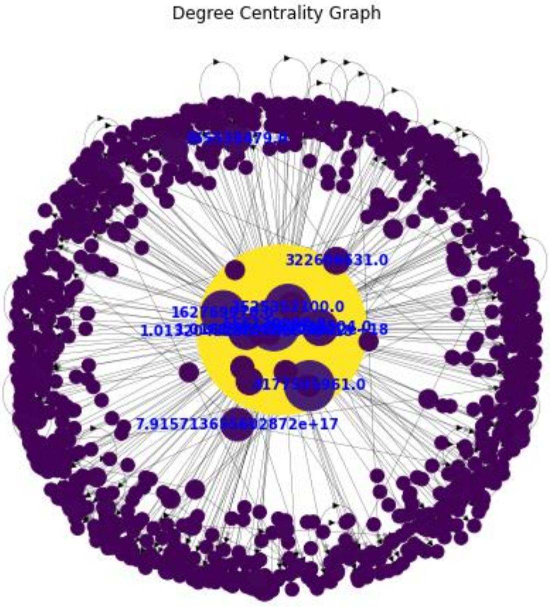
Topic 1 (hastane,aşı,randevu)

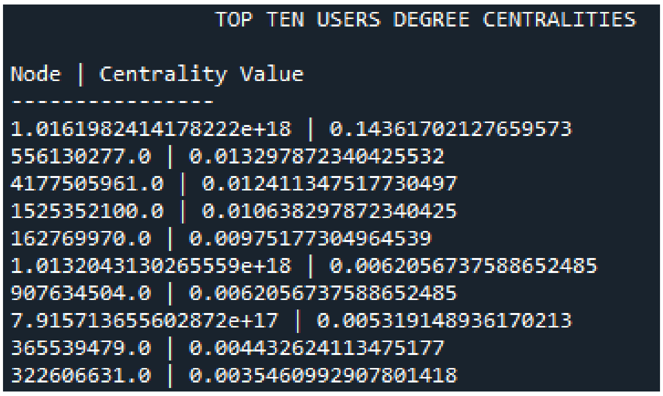




```
                    TOP TEN USERS BETWEENNESS CENTRALITIES

Node | Centrality Value
----------------
1.0161982414178222e+18 | 0.00013348912299216654
4177505961.0 | 5.575133960261074e-05
340993925.0 | 1.3348912299216655e-05
556130277.0 | 1.1778452028720578e-05
1525352100.0 | 9.422761622976462e-06
162769970.0 | 8.637531487728423e-06
1.3287657298653962e+18 | 7.067071217232347e-06
881683722.0 | 7.067071217232347e-06
8.035164493399368e+17 | 7.067071217232347e-06
345508643.0 | 7.067071217232347e-06
```

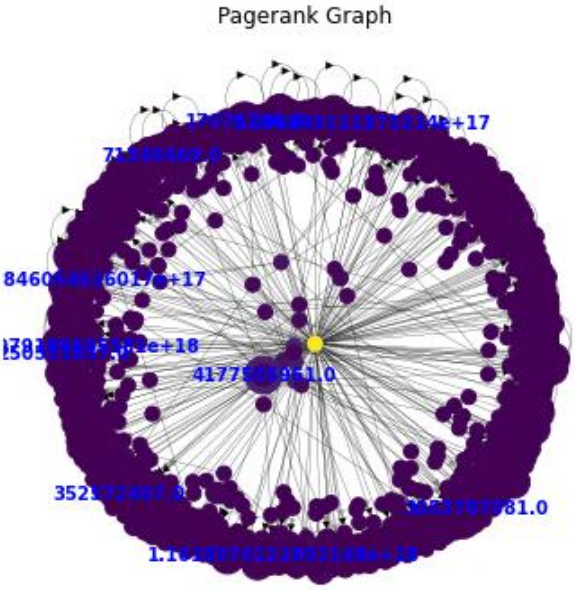
Figure 30: Betweennes Centrality Analysis of Network for Topic 1

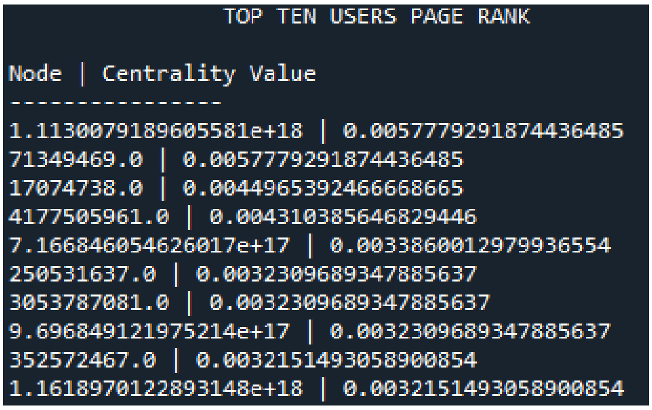




```
                TOP TEN USERS DEGREE CENTRALITIES

Node | Centrality Value
----------------
1.0161982414178222e+18 | 0.14361702127659573
556130277.0 | 0.013297872340425532
4177505961.0 | 0.012411347517730497
1525352100.0 | 0.010638297872340425
162769970.0 | 0.00975177304964539
1.0132043130265559e+18 | 0.0062056737588652485
907634504.0 | 0.0062056737588652485
7.915713655602872e+17 | 0.005319148936170213
365539479.0 | 0.004432624113475177
322606631.0 | 0.0035460992907801418
```

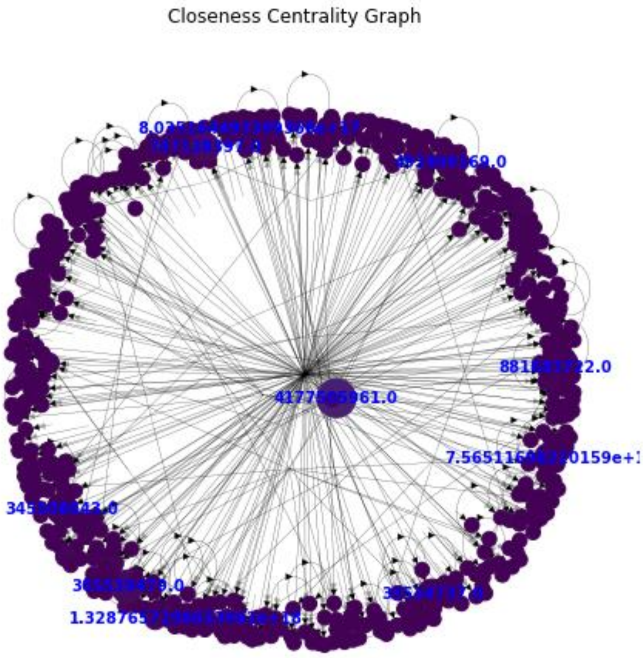
Figure 31: Degree Centrality Analysis of Network for Topic 1

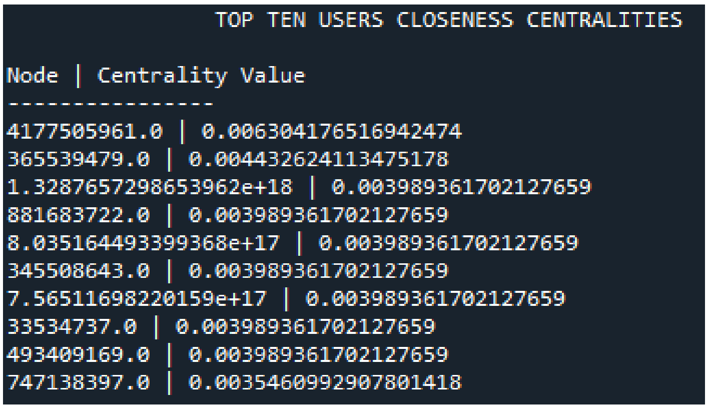




```
                TOP TEN USERS PAGE RANK

Node | Centrality Value
----------------
1.1130079189605581e+18 | 0.0057779291874436485
71349469.0 | 0.0057779291874436485
17074738.0 | 0.0044965392466668665
4177505961.0 | 0.004310385646829446
7.166846054626017e+17 | 0.0033860012979936554
250531637.0 | 0.0032309689347885637
3053787081.0 | 0.0032309689347885637
9.696849121975214e+17 | 0.0032309689347885637
352572467.0 | 0.0032151493058900854
1.1618970122893148e+18 | 0.0032151493058900854
```

Figure 32: PageRank Analysis of Network for Topic 1

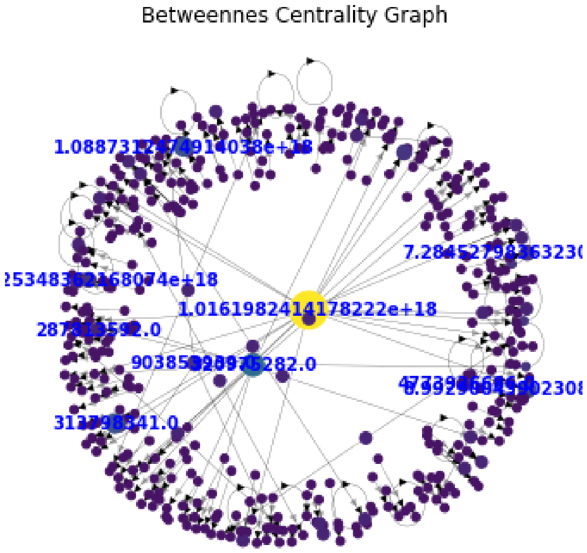


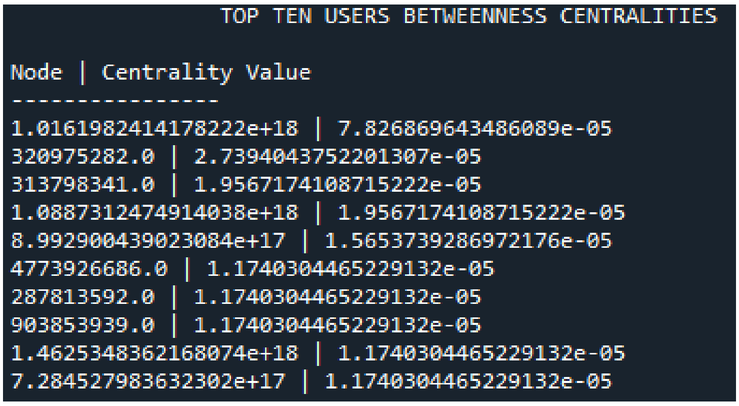


```
                    TOP TEN USERS CLOSENESS CENTRALITIES

Node | Centrality Value
----------------
4177505961.0 | 0.006304176516942474
365539479.0 | 0.00443262411347 5178
1.3287657298653962e+18 | 0.003989361702127659
881683722.0 | 0.003989361702127659
8.035164493399368e+17 | 0.003989361702127659
345508643.0 | 0.003989361702127659
7.56511698220159e+17 | 0.003989361702127659
33534737.0 | 0.003989361702127659
493409169.0 | 0.003989361702127659
747138397.0 | 0.0035460992907801418
```

Figure 33: Closeness Centrality Analysis of Network for Topic 1

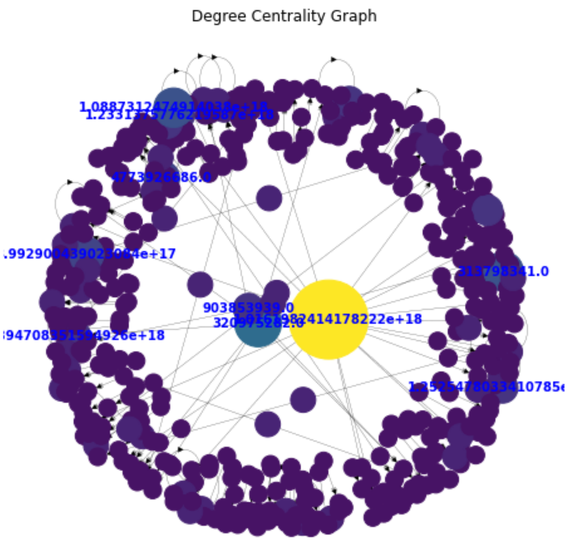
Topic 2 (hastane,entübe)





```
                    TOP TEN USERS BETWEENNESS CENTRALITIES

Node | Centrality Value
----------------
1.0161982414178222e+18 | 7.826869643486089e-05
320975282.0 | 2.7394043752201307e-05
313798341.0 | 1.9567174108715222e-05
1.0887312474914038e+18 | 1.9567174108715222e-05
8.992900439023084e+17 | 1.5653739286972176e-05
4773926686.0 | 1.1740304465229132e-05
287813592.0 | 1.1740304465229132e-05
903853939.0 | 1.1740304465229132e-05
1.4625348362168074e+18 | 1.1740304465229132e-05
7.284527983632302e+17 | 1.1740304465229132e-05
```

Figure 34: Betweennes Centrality Analysis of Network for Topic 2



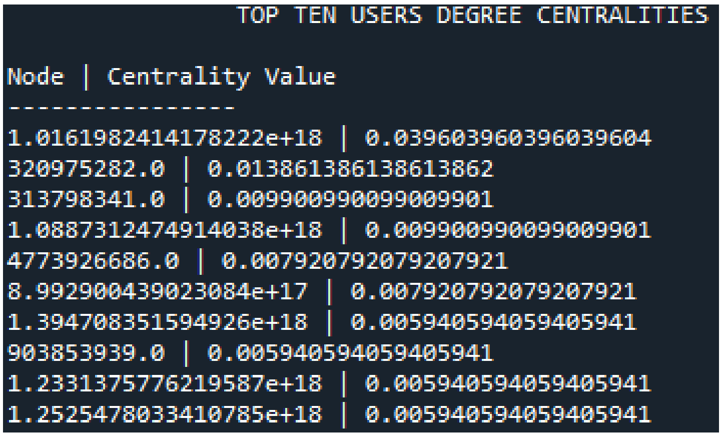
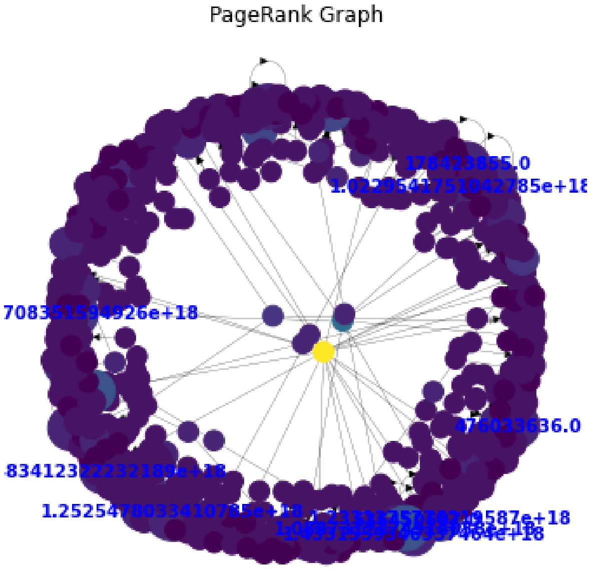

```
                 TOP TEN USERS DEGREE CENTRALITIES

Node | Centrality Value
----------------
1.0161982414178222e+18 | 0.039603960396039604
320975282.0 | 0.013861386138613862
313798341.0 | 0.009900990099009901
1.0887312474914038e+18 | 0.009900990099009901
4773926686.0 | 0.007920792079207921
8.992900439023084e+17 | 0.007920792079207921
1.394708351594926e+18 | 0.005940594059405941
903853939.0 | 0.005940594059405941
1.2331375776219587e+18 | 0.005940594059405941
1.2525478033410785e+18 | 0.005940594059405941
```

Figure 35: Degree Centrality Analysis of Network for Topic 2



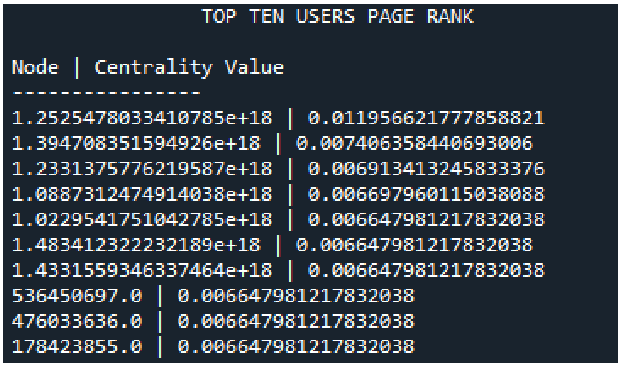
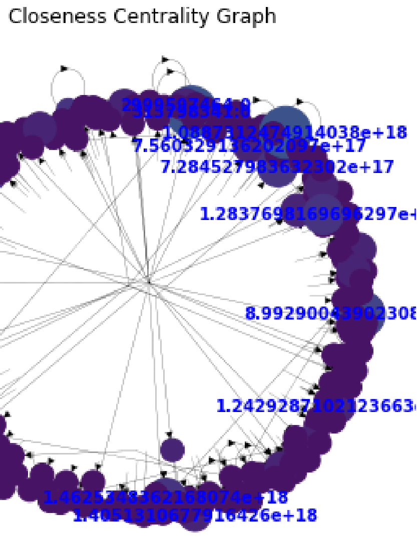
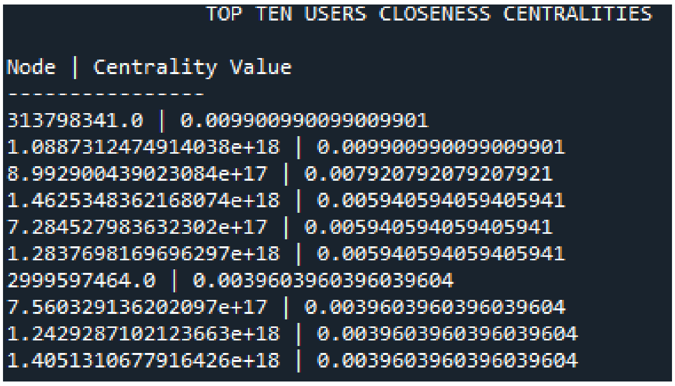

```
             TOP TEN USERS PAGE RANK

Node | Centrality Value
----------------
1.2525478033410785e+18 | 0.011956621777858821
1.394708351594926e+18 | 0.007406358440693006
1.2331375776219587e+18 | 0.006913413245833376
1.0887312474914038e+18 | 0.006697960115038088
1.0229541751042785e+18 | 0.006647981217832038
1.483412322232189e+18 | 0.006647981217832038
1.4331559346337464e+18 | 0.006647981217832038
536450697.0 | 0.006647981217832038
476033636.0 | 0.006647981217832038
178423855.0 | 0.006647981217832038
```

Figure 36: PageRank Analysis of Network for Topic 2

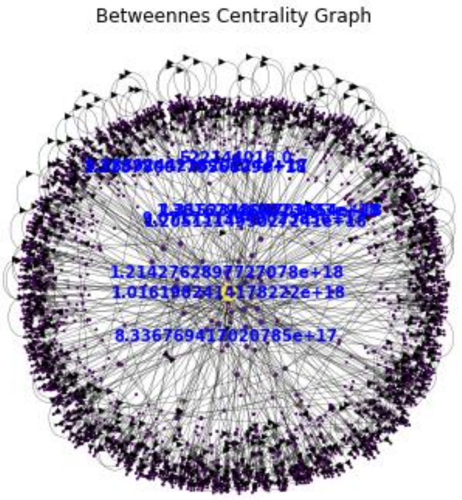


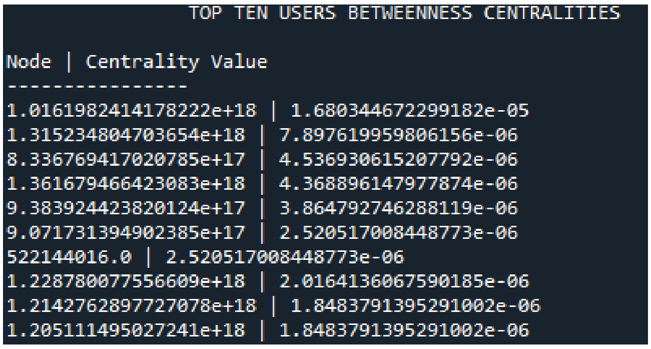


```
                    TOP TEN USERS CLOSENESS CENTRALITIES

Node | Centrality Value
-----------------
313798341.0 | 0.009900990099009901
1.0887312474914038e+18 | 0.009900990099009901
8.992900439023084e+17 | 0.007920792079207921
1.4625348362168074e+18 | 0.005940594059405941
7.284527983632302e+17 | 0.005940594059405941
1.2837698169696297e+18 | 0.005940594059405941
2999597464.0 | 0.0039603960396039604
7.560329136202097e+17 | 0.0039603960396039604
1.2429287102123663e+18 | 0.0039603960396039604
1.4051310677916426e+18 | 0.0039603960396039604
```

Figure 37: Closeness Centrality Analysis of Network for Topic 1

Topic 3 (hastane,PCR)

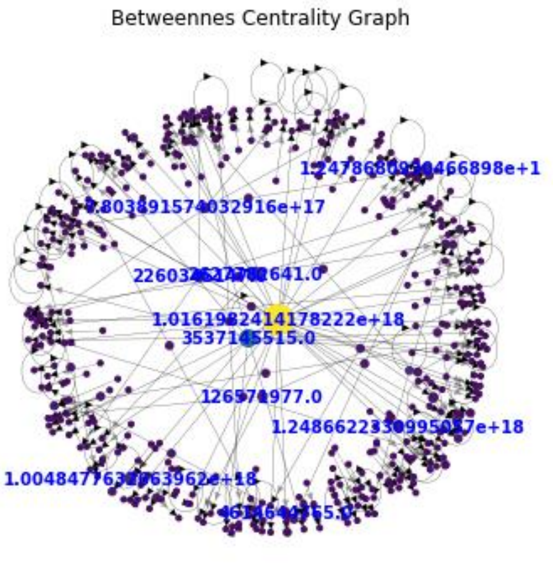


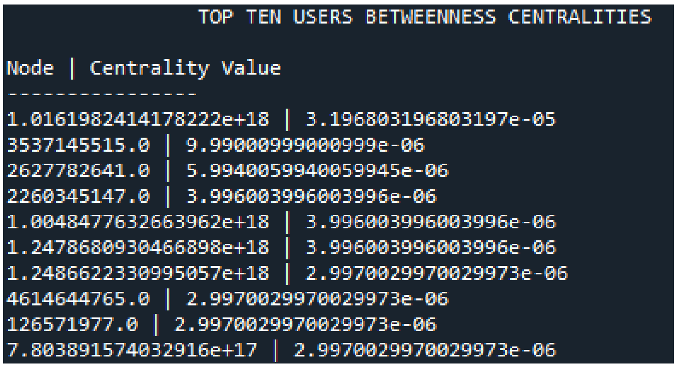


```
                    TOP TEN USERS BETWEENNESS CENTRALITIES

Node | Centrality Value
----------------
1.0161982414178222e+18 | 1.680344672299182e-05
1.315234804703654e+18 | 7.897619959806156e-06
8.336769417020785e+17 | 4.536930615207792e-06
1.361679466423083e+18 | 4.368896147977874e-06
9.383924423820124e+17 | 3.864792746288119e-06
9.071731394902385e+17 | 2.520517008448773e-06
522144016.0 | 2.520517008448773e-06
1.228780077556609e+18 | 2.0164136067590185e-06
1.2142762897727078e+18 | 1.8483791395291002e-06
1.205111495027241e+18 | 1.8483791395291002e-06
```

Figure 38: Betweennes Centrality Analysis of Network for Topic 3

Topic 4 (kanser,sma,ilaç,sağlık,bakanlık)

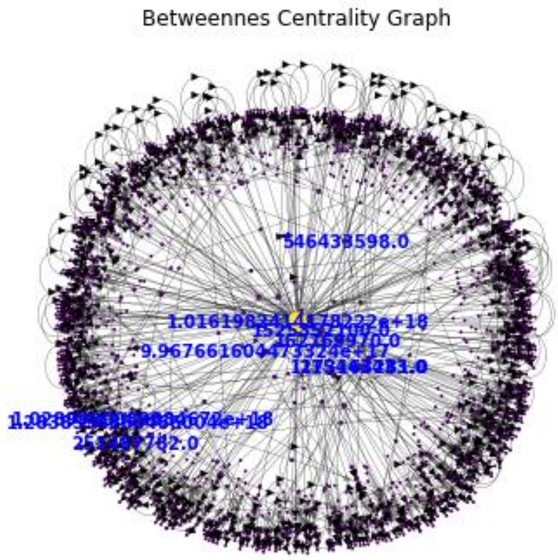

Betweennes Centrality Graph

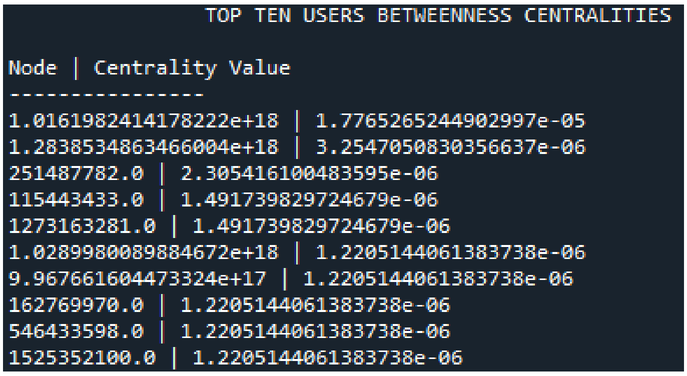


```
                 TOP TEN USERS BETWEENNESS CENTRALITIES

Node | Centrality Value
----------------
1.0161982414178222e+18 | 3.196803196803197e-05
3537145515.0 | 9.99000999000999e-06
2627782641.0 | 5.994005994005994 5e-06
2260345147.0 | 3.996003996003996e-06
1.0048477632663962e+18 | 3.996003996003996e-06
1.2478680930466898e+18 | 3.996003996003996e-06
1.2486622330995057e+18 | 2.9970029970029973e-06
4614644765.0 | 2.9970029970029973e-06
126571977.0 | 2.9970029970029973e-06
7.803891574032916e+17 | 2.9970029970029973e-06
```

Figure 39: Betweennes Centrality Analysis of Network for Topic 4

Topic 5(yoğun bakımda yer)

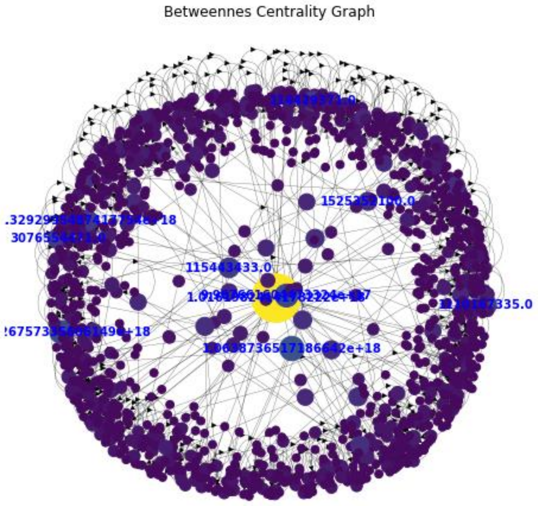

Betweennes Centrality Graph

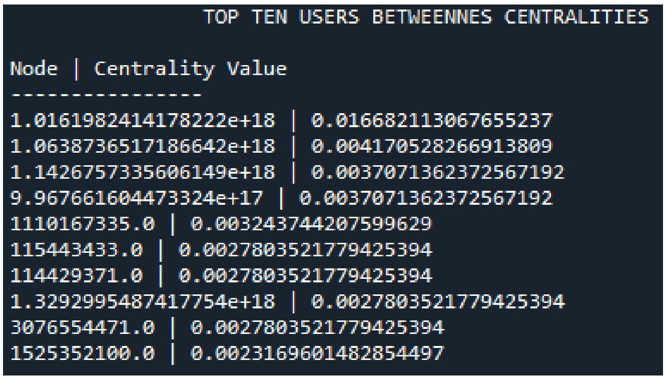


```
                 TOP TEN USERS BETWEENNESS CENTRALITIES

Node | Centrality Value
----------------
1.0161982414178222e+18 | 1.7765265244902997e-05
1.2838534863466004e+18 | 3.2547050830356637e-06
251487782.0 | 2.305416100483595e-06
115443433.0 | 1.491739829724679e-06
1273163281.0 | 1.491739829724679e-06
1.0289980089884672e+18 | 1.2205144061383738e-06
9.967661604473324e+17 | 1.2205144061383738e-06
162769970.0 | 1.2205144061383738e-06
546433598.0 | 1.2205144061383738e-06
1525352100.0 | 1.2205144061383738e-06
```

Figure 40: Betweennes Centrality Analysis of Network for Topic 5

Topic 6 (yoğun,bakım,yatak)

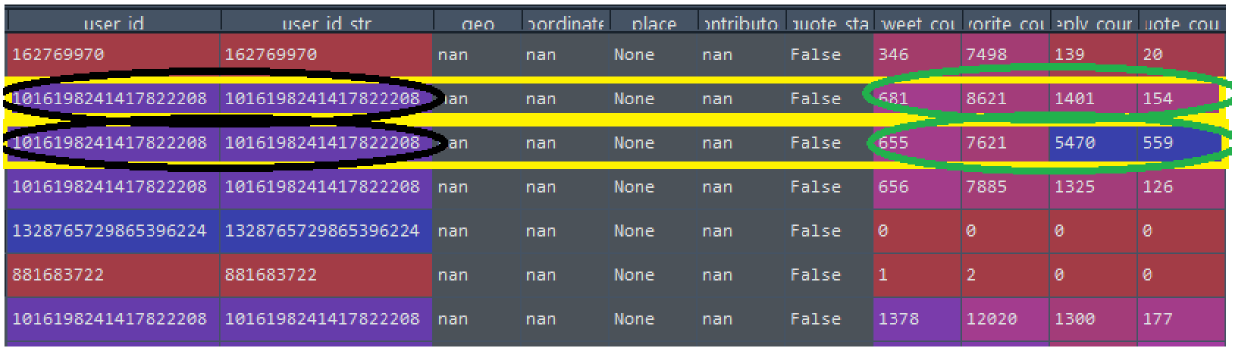




```
                    TOP TEN USERS BETWEENNES CENTRALITIES

Node | Centrality Value
----------------
1.0161982414178222e+18 | 0.01668211306765523 7
1.0638736517186642e+18 | 0.004170528266913809
1.1426757335606149e+18 | 0.003707136237256719 2
9.967661604473324e+17 | 0.003707136237256719 2
1110167335.0 | 0.003243744207599629
115443433.0 | 0.0027803521779425394
114429371.0 | 0.0027803521779425394
1.3292995487417754e+18 | 0.0027803521779425394
3076554471.0 | 0.0027803521779425394
1525352100.0 | 0.0023169601482854497
```

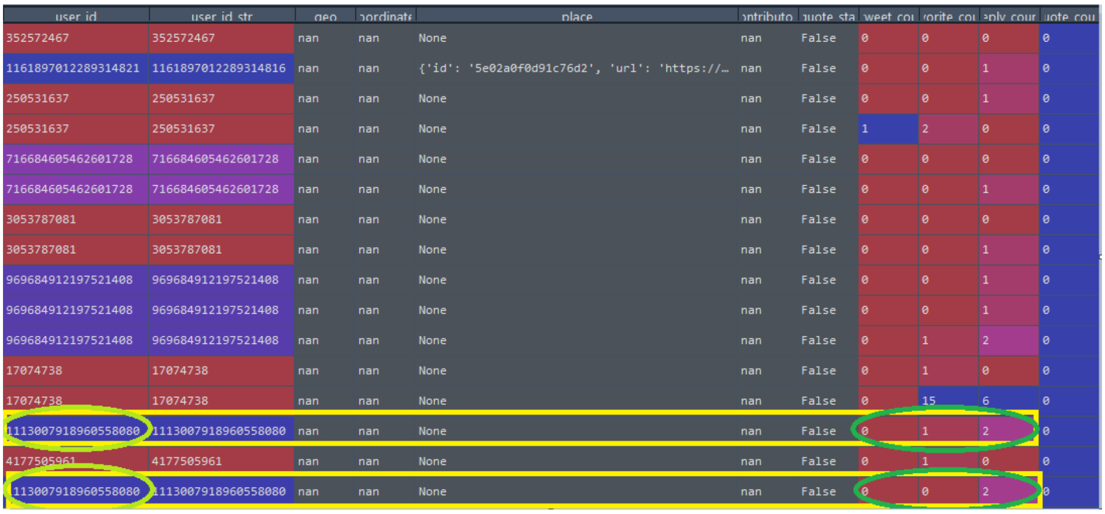
Figure 41: Betweennes Centrality Analysis of Network for Topic 6

For the topic 1 and topic 2 all the centrality measurements results were added to the report in order to show that all the measurements were studied and the most efficient one tried to be found. It was observed that the most reliable measurement is betweennes centrality compared to the others. For example, for the analysis in topic 1 the same tweet data was used in Figure 30(betweennes) and Figure 32(PageRank) however when the nodes were investigated in the real dataframe it was found that the connection between nodes and the user's tweet features are neglible for pagerank measures compared to the betwennes measure. For all the topics, the betweennes measure performs the best efficiency compared to the other measures. As mentioned before this was also investigated in the dataframes of all the measures and an example can be described as;



user id	user id str	geo	oordinate	place	ontributo	uote sta	weet cou	orite cou	eply cour	uote cou
162769970	162769970	nan	nan	None	nan	False	346	7498	139	20
1016198241417822208	1016198241417822208	nan	nan	None	nan	False	681	8621	1401	154
1016198241417822208	1016198241417822208	nan	nan	None	nan	False	655	7621	5470	559
1016198241417822208	1016198241417822208	nan	nan	None	nan	False	656	7885	1325	126
1328765729865396224	1328765729865396224	nan	nan	None	nan	False	0	0	0	0
881683722	881683722	nan	nan	None	nan	False	1	2	0	0
1016198241417822208	1016198241417822208	nan	nan	None	nan	False	1378	12020	1300	177

Figure 42: A piece of dataframe from Betweennes Centrality Analysis for topic1



user id	user id str	geo	oordinate	place	ontributo	uote sta	weet cou	orite cou	eply cour	uote cou
352572467	352572467	nan	nan	None	nan	False	0	0	0	0
1161897012289314821	1161897012289314816	nan	nan	{'id': '5e02a0f0d91c76d2', 'url': 'https://…	nan	False	0	0	1	0
250531637	250531637	nan	nan	None	nan	False	0	0	1	0
250531637	250531637	nan	nan	None	nan	False	1	2	0	0
716684605462601728	716684605462601728	nan	nan	None	nan	False	0	0	0	0
716684605462601728	716684605462601728	nan	nan	None	nan	False	0	0	1	0
3053787081	3053787081	nan	nan	None	nan	False	0	0	0	0
3053787081	3053787081	nan	nan	None	nan	False	0	0	1	0
969684912197521408	969684912197521408	nan	nan	None	nan	False	0	0	1	0
969684912197521408	969684912197521408	nan	nan	None	nan	False	0	0	1	0
969684912197521408	969684912197521408	nan	nan	None	nan	False	0	1	2	0
17074738	17074738	nan	nan	None	nan	False	0	1	0	0
17074738	17074738	nan	nan	None	nan	False	0	15	6	0
1113007918960558080	1113007918960558080	nan	nan	None	nan	False	0	1	2	0
4177505961	4177505961	nan	nan	None	nan	False	0	1	0	0
113007918960558080	1113007918960558080	nan	nan	None	nan	False	0	0	2	0

Figure 43: A piece of dataframe from Pagerank Analysis for topic1

From Figure 30 it can be observed that the most important node was recorded and marked as an example in Figure 42. Its every feature such as retweet count, favorite count etc. , are more than 600 which effects a user's influence in a network. On the other hand, the most important node in Figure 32 does not have considerable feature counts as marked in Figure 43. The same comparison were done for all the measures and the results showed that the betweennes centrality measure can be used to find influential users in a network efficiently.

9.5.3 Web Page

The results are present on the webpage. Webpage is built with flask, python, basic html and css. The flask is used to implement python code to webpage. Since both results which are influencers and trending topics, are obtained from python, the best option to present this result on the webpage is implementing the python codes to webpage.

The webpage mainly includes three-page, home, trending topics and influencers. In the home page, there are some details about project, some information about project members. In the trending page, there are figures to present trending topics based on dates. In the influencers page, most influencer peoples are present based on centrality algorithms.

10. DISCUSSION

Different methods that will be used in the project were explained in Part 4 of the report considering the literature and the results were obtained using collected datasets by applying these methods. Detecting trending topics and the influencers has the highest percentage for the success criteria of the project.

Thus, the preliminary works were performed for these criteria to understand the subject and bring the methods' application to a certain level. After the preliminary results, it can be observed that, the decided methods gave the optimal results for this project. As a result of this, same methods are used on the collected real data as a final result of the project.

The proposed methods include the n-grams algorithm to identify trending topics and social network analysis to identify influencers. To identify trending topics, it was used n-gram method because the data that is be used is a text called as tweet which were be gathered from the Twitter. The n-grams algorithm was used to analyze sentences as shown in the result part. Even sentiment analysis is applied on the preliminary result with mock data, it is not used on the real data. The reason of this, when the collected tweets were examined, it is observed that tweet's topics did not give the optimal analysis. At the same time, this analysis was not crucial for the project, it was an additional property. So, sentiment analysis not implemented on the result of the project.

Dataset, which is consist of collected tweets, is examined on unigram, bigram, trigram and four gram. Since the words used in the tweets can have different structure such as, single word, compound word or two words, it makes sense when put them together. Thus, four type of the n-gram results were analyzed jointly. The results for the n-grams algorithm shows that, the methods can work proper on sentences. Consequently, n-grams can be helpful while detecting the trending topics.

A WordCloud was constructed to visualize most common words from a collected tweet data. WordCloud can be used as a colored and attractive representation of the frequent words when a user wants to look at trending topics. For this reason, WordCloud is present on the home page of the webpage.

In order to identify influential people on a social network, various centrality measures were applied to the real twitter data. The network was analyzed and visualized based on the graph theory because a user in a social network can be considered as a node and the relationship between the users can be thought as edge in a graph. Important nodes, which also means influential people, and their centrality values were recorded and visualized in graphs. As can be seen in the results, important nodes are not same in each measure for a topic. The reason is the processes of the evaluation are different for the measurements as explained in the methodology. When all the measurements were compared for all the topics it was concluded that the betweennes measure is the best choice in order to analyze and find influential people in a network. In other measures when the dataframes were investigated , which was also shown as an example , it was found that the features of a tweet that are required to be an influencer , should have significant numbers and counts in order to access and influence other users in a network. In some of the measures, the inefficient results may attributed to numbers of tweet . The result may obtained more accurately with more number of tweets. Also , more users can be added to the system , so the measures can evaluate more connection. Additionally , the tweets that is related to the topics may be increased and more words may included while gathering them, so more specific users can be obtained as influencers.

11. CONCLUSION

The two main objectives of the project, which are determining the trending topics and influential people in the field of health, were performed. The project was carried out in line with two main objectives. The first part is the identification of trend topics. The n gram algorithm was used to identify trending topics in the field of health. According to this algorithm, results were obtained from the data collected as unigram, bigram, trigram, and fourgram. The second part is identifying influential people. Social network analysis was used to identify people who are influential in the field of health. Centrality Measures, one of the most used concepts in social network analysis, was used. Results were obtained from data collected using Betweenness Centrality Measure, Degree Centrality Measure, Closeness Centrality Measure, and PageRank. Thus, a method model was proposed that includes these techniques in identifying trend topics and influential people. A web interface was designed to present the results obtained and a website was built using flask framework. As a useful tool, the results are presented and visualized. As can be understood from the explanations, all the items specified in the success criteria were realized and the project was completed according to the work-time plan and work packages.

12. PLAN FOR FUTURE STUDIES

The project was completed in accordance with the time planning and work packages in line with the determined objectives and targets. If it is thought to continue in the future and build on the work done, it could be as follows: The first suggestion is related to the subject and area of the project. The specified techniques, that is the proposed method model, were applied in the field of health in Turkey. Studies can also be carried out for other subjects or areas that need to be determined and important to be specified for our country. Trend topics and influential people can be determined within topics such as political, economic, etc. In this way, analyzes can be made. Another suggestion may be related to interface design. The user interface that assists in the presentation and visualization of the results was designed and set up as a web page. If it is designed and built as a mobile application, it can be useful for different user profiles. It can also be operated in a different area.

13. REFERENCES

[1] L. O. Adegboyega, “Influence of Social Media on the Social Behavior of Students as Viewed by Primary School Teachers in Kwara State , Nigeria,” vol. 7, no. 1, pp. 43–53, 2020, doi: 10.17509/mimbar-sd.

[2] A. W. K. Yeung *et al.*, “Implications of Twitter in Health-Related Research: A Landscape Analysis of the Scientific Literature,” *Front. Public Heal.*, vol. 9, no. July, pp. 1–9, 2021, doi: 10.3389/fpubh.2021.654481.

[3] Y. Albalawi and J. Sixsmith, “Identifying Twitter influencer profiles for health promotion in Saudi Arabia,” *Health Promot. Int.*, vol. 32, no. 3, pp. 456–463, 2017, doi: 10.1093/heapro/dav103.

[4] J. Fiaidhi, S. Mohammed, and A. Islam, “Towards identifying personalized twitter trending topics using the twitter client RSS feeds,” *J. Emerg. Technol. Web Intell.*, vol. 4, no. 3, pp. 221–226, 2012, doi: 10.4304/jetwi.4.3.221-226.

[5] Y. A. Winatmoko and M. L. Khodra, “Automatic Summarization of Tweets in Providing Indonesian Trending Topic Explanation,” *Procedia Technol.*, vol. 11, pp. 1027–1033, 2013, doi: 10.1016/j.protcy.2013.12.290.

[6] T. Yang and Y. Peng, “The Importance of Trending Topics in the Gatekeeping of Social Media News Engagement: A Natural Experiment on Weibo,” *Communic. Res.*, 2020,

doi: 10.1177/0093650220933729.

[7] C. P. M. R. Ermelinda Ora, "A Methodology for Identifying Influencers and their Products Perception on Twitter," in *Proceedings of the 20th International Conference on Enterprise Information Systems*, Italy, 2018.

[8] T. M. Swit Phuvipadawat, "Breaking News Detection and Tracking in Twitter," in 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, Japan, 2010.

[9] A. S. Somya Jain, "Identification of influential users on Twitter: A novel weighted," Chaos, Solitons and Fractals, vol. 139, 2020.

[10] G. P. C. M. D. C. S. P. R. S. A. Luca Maria Aiello, "Sensing trending topics in Twitter," 2013.

[11] M. O. V. L. Saˇsa Petrovi´, "Streaming First Story Detection with application to Twitter," in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL*, Los Angeles, California, 2010.

[12] D. M. M. H.-P. Ahmad Alsolami, "A Structured Mechanism for Identifying Political," in *World Academy of Science, Engineering and Technology*, 2021.

[13] Oo, Myat Mon, and May Thu Lwin. "Detecting Influential Users in a Trending Topic Community Using Link Analysis Approach."

[14] Bethu, Srikanth, et al. "Data science: Identifying influencers in social networks." *Periodicals of Engineering and Natural Sciences (PEN)* 6.1 (2018): 215-228.

[15] https://help.twitter.com/tr/using-twitter

[16] https://help.twitter.com/tr/rules-and-policies/twitter-api

[17] C. Salman, “Göreceli Kenar Önemi Metodu(A New Network Centrality Measure : Relative Edge Importance Method),” 2018.

[18] A. Hacı, B. Veli, Ü. İletişim, F. Süreli, and E. Dergi, “Sosyal Medyanın Gündem Belirleyicileri: Twitter’da Gündem Belirleme Süreci Üzerine Bir Sosyal Ağ Analizi Agenda Setters of Social Media: A Social Network Analysis on the Agenda Setting Process of Twitter,” 2020, [Online]. Available: https://orcid.org/0000-0003-0476-8394.

[19] U. Bakan, Z. E. T. Facebook, and A. Kelimeler, “Sanat Okullarının Twitter Kullanım Karakteristiklerine İlişkin Bir Sosyal Ağ Analizi Perspektifi,” vol. 1, pp. 138–155, 2020.

[20] https://web.stanford.edu/~jurafsky/slp3/3.pdf

[21] https://www.analyticsvidhya.com/blog/2021/09/what-are-n-grams-and-how-to-implement-them-in-python/

[22] https://realpython.com/python-nltk-sentiment-analysis/

[23] https://www.datacamp.com/community/tutorials/simplifying-sentiment-analysis-python

[24] https://www.nltk.org

[25] https://www.geeksforgeeks.org/generating-word-cloud-python/

[26] Z. Eren, and E.Kıral, “Socıal Network Analysıs And Usage In Educatıonal Research”, Frames From Educatıon Internatıonal Chapter Book, pp. 308-353, 2019.

[27 https://pypi.org/project/stop-words/#available-languages [28] M. Gençer, “Sosyal Ağ Analizi Yöntemlerine Bir Bakış,” *Yildiz Soc. Sci. Rev.*, vol. 3, no. 2, pp. 19–34, 2017.

[29] C. Salman, “Göreceli Kenar Önemi Metodu(A New Network Centrality Measure : Relative Edge Importance Method),” 2018.

[30] U. Bakan, Z. E. T. Facebook, and A. Kelimeler, “Sanat Okullarının Twitter Kullanım

Karakteristiklerine İlişkin Bir Sosyal Ağ Analizi Perspektifi,” vol. 1, pp. 138–155, 2020.
[31] https://github.com/trenton3983/DataCamp/tree/master/data
[32] Gupta, R., Vishwanath, A., and Yang, Y. (2020), COVID-19 Twitter Dataset with Latent Topics, Sentiments and Emotions Attributes
[33] M. R. Mufid, A. Basofi, M. U. H. Al Rasyid, I. F. Rochimansyah, and A. Rokhim, “Design an MVC Model using Python for Flask Framework Development,” *IES 2019 - Int. Electron. Symp. Role Techno-Intelligence Creat. an Open Energy Syst. Towar. Energy Democr. Proc.*, no. Mvc, pp. 214–219, 2019, doi: 10.1109/ELECSYM.2019.8901656.
[34] https://www.utf8-chartable.de/
[35] https://github.com/otuncelli/turkish-stemmer-python

Istanbul Medipol University

School of Engineering and Natural Sciences

Graduation Project

PROJECT ACTIVITIES AND WORK PLAN

For first semester

Work and Activity	**Responsible Group Member**	**Timeline**													
		1. week	2. week	3. week	4. week	5. week	6. week	7. week	8. week	9. week	10. week	11. week	12. week	13. week	14. week
1. Literature Review	Miray Ceren Feridun	■	■	■	■	■	■	■	■	■	■	■	■	■	■
2. Basic Terminology Research	Miray Ceren Feridun	■	■	■	■	■	■	■							
3. Comparison of Different Method Types	Miray Ceren Feridun					■	■	■	■	■	■	■			
4. Deciding of Method	Miray Ceren Feridun						■	■	■	■	■	■	■		
5. Investigation of Decided Method Examples	Miray Ceren Feridun											■	■	■	■

Istanbul Medipol University

School of Engineering and Natural Sciences

Graduation Project

For second semester

Work and Activity	**Responsible Group Member**	**Timeline**													
		1. week	2. week	3. week	4. week	5. week	6. week	7. week	8. week	9. week	10. week	11. week	12. week	13. week	14. week
6. Literature Review	Miray Ceren Feridun	■	■	■	■	■	■	■	■	■	■	■	■	■	■
7. Data Collection	Miray Ceren Feridun	■	■	■	■	■	■	■							
8. Investigate another additional methods	Miray Ceren Feridun	■	■	■	■	■	■	■	■	■					
9. Applied methods on collected data	Miray Ceren Feridun					■	■	■	■	■	■	■	■		
10. User interface design	Miray Ceren Feridun								■	■	■	■	■	■	■
11. Implementation results to user interface	Miray Ceren Feridun								■	■	■	■	■	■	■

LIST OF WORK PACKAGES

WP No	Detailed Definition of Work and Activity
1 and 6	Investigating and analyzing related articles and studies
2	Investigating related basic terminology
3	Compare different methods in various studies to decide our proposed method
4	Proposing efficient model
5	Examined proposed method's algorithm examples and implementations
7	Data collection from real tweets
8	Literature review for other methods and similar studies was realize method investigation was performed.
9	Implementation of determined methods to the data
10	Decision of interface design
11	Implementation of interface design on web page in order to display the results

Work package	Target	Measurable outcome	Contribution to overall success(%)
WP 1 and 6	Investigate various methods, process, and their implementation subjects	Analyze at least 8 articles related to project	5%
WP 2	Learn and examine basic terminology which are used in the project to understand	Present background information about related concepts	2%
WP 3	Get a comparison based on application of methods that are investigated from different studies	Determination of suitable and usable methods after comparison	3%
WP 4	Proposing an efficient model based on project target	Proposed one efficient model for project	10%
WP 5	Increase information about the proposed method and use the algorithm methodology	Present detailed explanation about method and examples related to this method	10%
WP 7	Collection of enough data from real tweets to get efficient result	To get features of tweets	25%
WP 8	Examine additional methods, process, and their implementation subjects	Analyze at least 2 different methods	2%
WP 9	Getting results from real data with application of determined methods	To determine at least five trending topics and influencers with %80 accuracy	35%
WP 10	Determinaton of useful interface design	Creating design template	2%
WP 11	Showing the results of the completion of the project in the interface	Display the results in a proper way	8%
			Total: 100%

WORK PACKAGE DISTRUBUTION										
Project Member	WP1 and 6	WP2	WP3	WP4	WP5	WP7	WP8	WP9	WP10	WP11
Miray	100/3	100/3	100/3	100/3	100/3	100/3	100/3	100/3	100/3	100/3
Ceren	100/3	100/3	100/3	100/3	100/3	100/3	100/3	100/3	100/3	100/3
Feridun	100/3	100/3	100/3	100/3	100/3	100/3	100/3	100/3	100/3	100/3
Total	100	100	100	100	100	100	100	100	100	100

CURRICULUM VITAE

Istanbul Medipol University

School of Engineering and Natural Sciences

Graduation Project

SUPPORT LETTERS (if any)