
# MEASURING POLITICAL POLARIZATION USING BIG DATA : THE CASE OF TURKISH ELECTIONS

SELİM SÜRÜCÜ

AUGUST 2020

# MEASURING POLITICAL POLARIZATION USING BIG DATA : THE CASE OF TURKISH ELECTIONS

A THESIS SUBMITTED TO THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
ÇANKAYA UNIVERSITY

BY

SELIM SÜRÜCÜ

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
COMPUTER ENGINEERING

AUGUST 2020

## STATEMENT OF NON-PLAGIARISM

I hereby declare that all information in this study has been obtained and bestowed in conformance with educational rules and ethics provided in educational medium. I can also state that, all materials and outcomes used in this thesis are referenced and cited in accordance with the educational rules.

**Name, Last Name:** Selim SÜRÜCÜ

**Signature:**

**Date:** 17.09.2020

## ABSTRACT


### MEASURING POLITICAL POLARIZATION USING BIG DATA : THE CASE OF TURKISH ELECTIONS

SÜRÜCÜ, Selim

M.Sc., Computer Engineering Department

Supervisor: Asst. Prof. Dr. Roya CHOUPANI

Co-Supervisor: Prof Dr. Erdoğan DOĞDU

AUGUST 2020, 56 pages.

Big data has been the driving force behind the latest machine learning and deep learning accomplishments in many learning tasks. Social media data, as a big data resource, has recently been used in many social studies to understand the social movements and political and social changes. In this study, we will analyze social media (Twitter) data to measure political polarization, which is one of the recent concerns in politics. This study made use of Twitter data collected in the 2019 elections in Turkey; new metrics are developed to measure the political polarization. We analyzed the political groups in the social network and then measure political polarization overtime during the election period. By applying community detection algorithms, we first identify communities based on the interactions among users. Then, we measure the interaction among user groups (communities) to successfully show the existence and growth of political polarization using big data during a general election process. To the best of our knowledge, this is the first wide-scale big data study on political polar-

ization in a political election process.

**Keywords:** Political Polarization, Community Detection, Big Data, Social Media Analysis.

# ÖZ

## BÜYÜK VERİ KULLANARAK SİYASİ KUTUPLAŞMAYI ÖLÇME: TÜRK SEÇİMLERİ ÖRNEĞİ

SÜRÜCÜ, Selim

Yüksek Lisans, Bilgisayar Mühendisliği Bölümü

Tez Danışman: Dr. Öğr. Üyesi Roya CHOUPANI

Ortak Danışman: Prof. Dr. Erdoğan DOĞDU

Ağustos 2020, 56 sayfa

Büyük veri, birçok öğrenme görevinde en son makine öğrenimi ve derin öğrenme başarılarının arkasındaki itici güç olmuştur. Sosyal medya verileri, büyük bir veri kaynağı olarak, sosyal hareketleri, politik ve sosyal değişiklikleri anlamak için birçok sosyal çalışmada kullanılmıştır. Bu çalışmada, siyasetin son dönem endişelerinden biri olan siyasi kutuplaşmayı ölçmek için sosyal medya (Twitter) verilerini analiz edeceğiz. Bu çalışmada, Türkiye'de 2019 seçimlerinde toplanan Twitter verilerinden yararlanılmıştır; siyasi kutuplaşmayı ölçmek için yeni ölçütler geliştirilmiştir. Sosyal ağdaki siyasi grupları analiz ettik ve ardından seçim döneminde zaman içindeki siyasi kutuplaşmayı ölçtük. Topluluk algılama algoritmalarını uygulayarak, önce toplulukları kullanıcılar arasındaki etkileşimlere göre belirleriz.Ardından, genel bir seçim sürecinde büyük verileri kullanarak siyasi kutuplaşmanın varlığını ve büyümesini başarılı bir şekilde göstermek için kullanıcı grupları (topluluklar) arasındaki etkileşimi ölçüyoruz. Bildiğimiz kadarıyla bu, siyasi bir seçim sürecinde ilk geniş ölçekli siyasi kutuplaşmaya ilişkin veri

çalışmasıdır.

**Anahtar Kelimeler:** Politik Kutuplaşma, Topluluk Tespiti, Büyük Veri, Sosyal Medya Analizi

## ACKNOWLEDGEMENT

Firstly, I would like to express my sincere appreciation and gratitude to Asst. Prof Dr. Roya CHOUPANI and Prof. Dr. Erdoğan DOĞDU as my supervisor and co-supervisors respectively, for their continuous guidance and counseling, supervision, suggestions and immense knowledge throughout the process of this masters study. Without your precious supports, it would not be possible to conduct this research, and make this thesis reach its conclusion.

My sincere acknowledgement also goes my thesis committee members Asst. Prof. Dr. Abdül Kadir GÖRÜR and Asst. Prof. Dr. Seda ŞAHİN for their motivation as well as guidance towards the end of writing this thesis. Your insightful motivational comments encouraged me throughout this process.

I would like to thank my friends Mr. Hamza Haruna MOHAMMED and Miss. Selin HELVACIOĞLU, who supported this study with their thoughts, opinions and suggestions.

Finally, I most express my profound sincere and gratitude to my family, especially my parents for their unparalleled love, encouragement, help and continuous support both financially and emotionally throughout my years of study and the process of this thesis. This attainment would not have been possible without them. I dedicate this accomplishment to them.

# TABLE OF CONTENTS

## LIST OF FIGURES

## LIST OF TABLES

## LIST OF ABBREVIATIONS

**UGC:** User Generated Content

**TB:** Terabayt

**G:** Graph

**V:** Vertices

**E:** Edges

**LPA:** Label Propagation Algorithm

**RDD:** Resilient Distributed Dataset

**API:** Application Programming Interface

**etc:** Et cetera, other similar things

# CHAPTER 1
# INTRODUCTION

## 1.1 Motivations

With the development of communication technologies, radical changes have been observed in the methods of interaction between individuals. Internet usage comes first among these changes. Over the years, worldwide Internet usage has increased every year since 1995 [1]. Likewise, Internet usage has also been increasing in Turkey [2]. According to TUIK's data in 2019, 88.3% of the households have Internet access and 75.3% of the total population uses the Internet in Turkey. Furthermore, the development of Internet technologies and infrastructures resulted in the expansion of Internet access, which is accessible to almost every location in this country. The usage of the Internet spans from online games to shopping, music services to social media.

With this increase in Internet usage, the use of social media also increases with direct proportionality. According to January 2020 statistics, 49% of the world's population and approximately 84% of the Internet users make use of social media[3]. Social media are the Internet applications that allow sharing and dissemination of any content created by users. Roughly, there exist 52 million active users of social media in Turkey, which constitutes 63% of its population. Additionally, statistics show that social media usage will keep increasing exponentially in this country. This illustrates the importance of analysis, analytic applications, and systems to collect and extract meaningful information from this data. Recently, big data emerged due to high volume of the social media data.

[1] *https://www.internetworldstats.com/emarketing.htm* (*accessed* : 01.07.2020)
[2] $http://www.tuik.gov.tr/PreIstatistikTablo.do?istab\_id = 382$ (*accessed* : 01.07.2020)
[3] *https://www.slideshare.net/DataReportal/digital-2019-turkey-january-2019-v01*

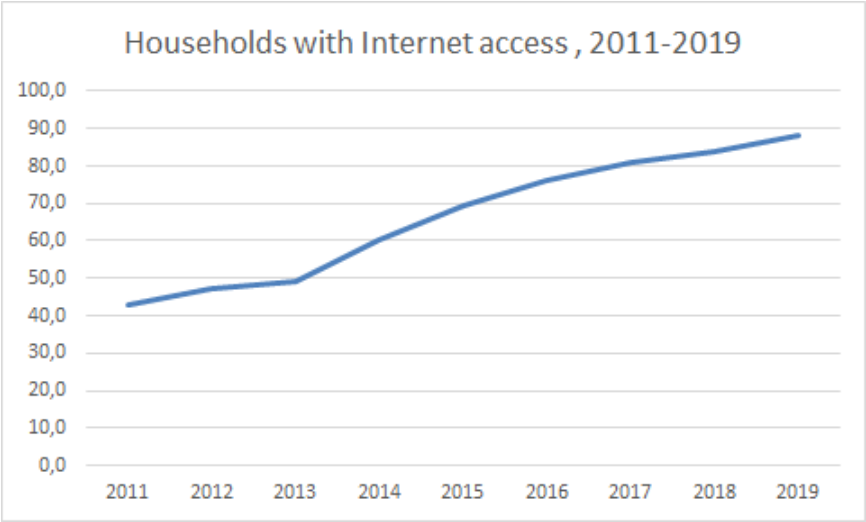


Figure 1.1: Households with Internet access, 2011-2019

In social media, users share emotions, thoughts, photos, etc. about any event or situation, themselves, or other people, through various means (photo, text, video, check-in, etc.). These sharing and dialogs between users are called user-based content (UGC-User Generated Content) and the number of these contents is increasing each year.

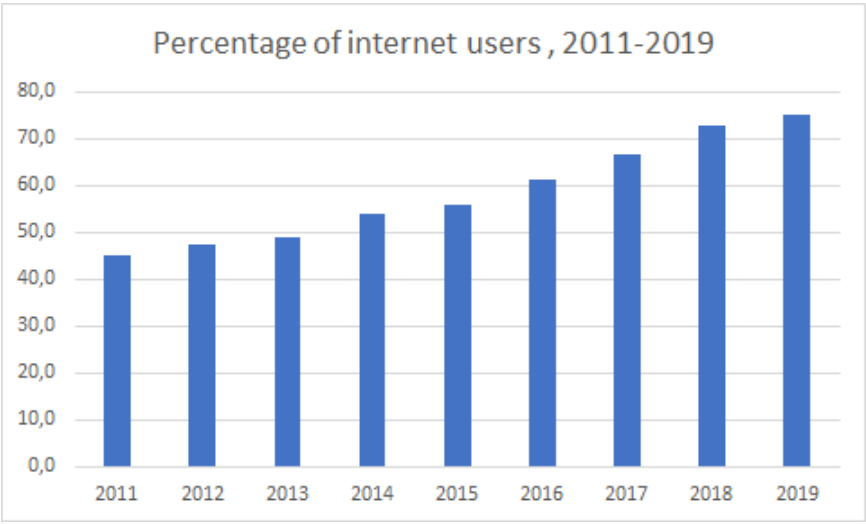


Figure 1.2: Percantage of internet Users in Turkey, 2011-2019

Users can interact with other users on social media using different methods that vary according to social media channels. Interaction between users can be short- term (Retweet, Like, Sharing, etc.), especially messaging, and long-term ones (such as adding friends and following). Along with these interactions, data is accumulated by social

media providers. These data are presented to developers through APIs for research or even the development of independent tools by students, firms, institutions, and researchers.

However, various problems arise with the excess of users on social media such as unreal accounts, fake news, offensive posts, etc. These problems can be given as examples. Efforts are being made to prevent such problems automatically [11, 4]. In addition to that, users can take individual measures, by blocking the person or page they do not want to interact with or receiving unwanted messages. They can block them in order not to see any activities in the timeline or alike in their social media.

In social media, users interact with people with whom they share the same emotion, thoughts, and usually do not interact and communicate with people having different emotions and thoughts. This is also applied to social media tool usage. For example, a person sharing or liking content/tweet/post on Facebook, re-tweeting on Twitter, or tweeting with the same hashtag, shows the same feeling and thought. Twitter is the best example of the new media in which social, political, and cultural issues, news, and announcements are shared among users. Twitter founded in the USA (Francisco) in 2006 has 340 million active users by 2020. It's believed that an average of approximately 6,000 tweets is sent per second as of the year 2020 [4]. The use of Twitter ranks 4th with 9.5% between social media use in Turkey [5].

Furthermore, data on Twitter has been collected for many areas and topics, and analyses have been made on this collected data. The studies in recent years, include in social events and movements [47, 3, 28, 56], analyzing users' behavior [8, 10, 64], analyzing brands [42, 12], classifying them by doing emotional and sentimental analysis on twitter data (how happy, crime, etc.) [55], analyzing how users are affected by a social event (terror, pandemic, etc.) [44, 34, 27, 16, 25, 32]. These studies dealt with different topics.

[4] *https: //www.oberlo.com/blog/twitter-statistics* (*accessed* : 11.07.2020)
[5] *https: / / gs . statcounter.com/social-media-stats/all/turkey* (*accessed* : 11.07.2020)

With the shift of social media tools, especially twitter, to the field of political activity; political parties, politicians, and candidates moved away from classical propaganda means of sharing information (main media interviews, discussion programs, press conferences) and started to use online media (new media) more actively, which leads these platforms to become the trend means of political propaganda discussion platforms. Political posts made here can reach large masses in a short time. Users can also interact with these tweet messages and the ecosystem created and shared by other users.

In this study, we analyzed the political polarization by using the twitter data we collected during the 2019 local election period, without knowing the personal information and political orientations of the users, by looking at the interactions between the users. Possible communities will be identified from the collected data using community detection algorithms, and it will be affirmed that each member is associated with only one community. When determining communities, an analysis will be made based on the idea that users who interact with each other will be in the same community. Using the detected communities, changes in their interactions with each other will be analyzed. According to our hypothesis, as the election period approaches, we expect the interaction between the communities to decrease and the interaction to be higher before or after the election period. For our hypothesis to be correct, it is very important to identify the communities correctly. Besides, this study will show that users with close feelings and thoughts can be detected by following the movements of the users on social media without having prior knowledge.

In previous studies, communities were identified by using candidate or party accounts and prior knowledge about users was obtained. In this study, we tried to make community determination without needing any information. Also in the elections in Turkey, the amount of social media data in other studies was substantially less than the amount of data we use.

## 1.2 Big Data

Before delving into the definition of big data, it is necessary to define the data. Generally speaking, data is the smallest piece of unprocessed information. Any meaningful or meaningless thing is considered data. With the development of Internet technology and the growing number of people using the Internet, the data produced is increasing every year. Approximately 1 000 000 TB (quintillion bytes) data stack is generated daily [24]. The classified, meaningful, and machine-processable form of this stack of data is called big data. Big data is expected to contain concepts commonly known as 3V.

- **Volume** = The smallest unit for data is considered TB. TB and above data stacks are considered to contain this concept.
- **Variety** = Data sources and formats are diverse. For example social media, e-commerce transactions, social media data, financial transactions, etc. are used as data sources. Nowadays, the data is not just text, but in addition to text data, it appears in various formats such as video, photo.
- **Velocity** = The time taken for the data to be generated, collected, and processed decreases gradually.

In addition to this 3V, there is also 2V.

- **Verification** = Another component that occurs when it is necessary to check whether the growing data is safe during the formation and collection. It is also determined by processing this component if the correct data is delivered, or it has refrained from the undesired people.
- **Value** = The component expresses if the data obtained as a result of the production and processing of big data represents a value.

## 1.3 Social Networks as Graphs

Social Networks are structures that make it easy to connect users and keep them up to date with news and a wide variety of other content effectively. Graphs are

used to display complex structures such as social networks, communication networks, technological networks, biological networks, and chemical interaction networks. Graphs are structures formed by a set of vertices (also called nodes) and a set of edges that are connections between pairs of vertices.

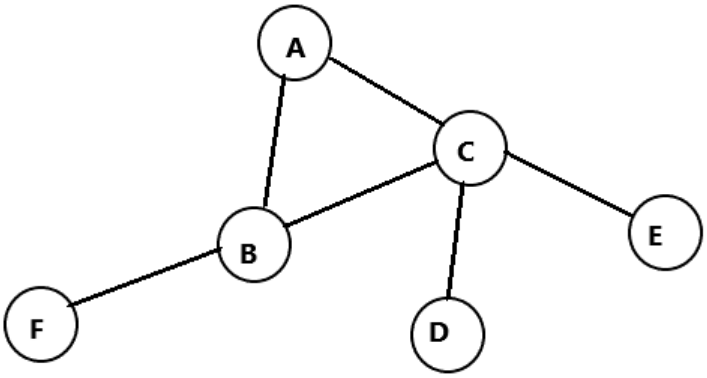


Figure 1.3: Graph example

While creating a graphical representation of social networks, the asset nodes in the network create the relationships edges between them [46]. The graph may differ depending on the structure of the network. If the relationship between the two nodes is one-sided, directional graphs can be used to represent them. For instance, the follow-up system on Twitter can be given as an example. Because when User A follows User B, User B doesn't have to follow User A.

If the interaction in the structure of the network is bilateral, there is no need to use direction for it you can use a non-directional graph. The friendship system on Facebook can be given as an example. When User A adds User B as a friend, User B doesn't need to add User A separately, and he automatically becomes a friend of User A. Graphs can also be classified according to whether the edges are weighted or not. If there is a weight on the sides, it is called a weighted graph. For example, the number of re-tweeting each other between the two users is given on the sides as weights. Let User A re-tweet User B 4 times and User B re-tweet User A 1 time.

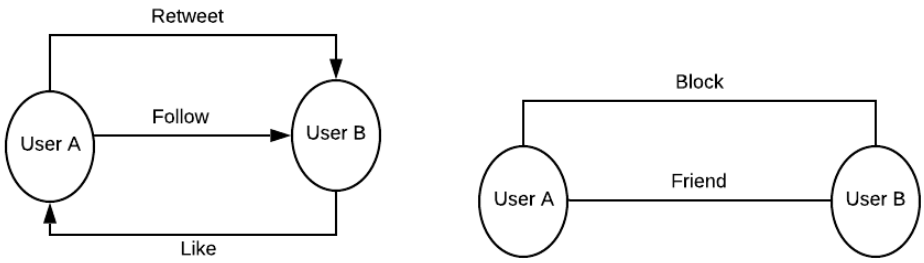


Figure 1.4: representation of social media interactions with graph

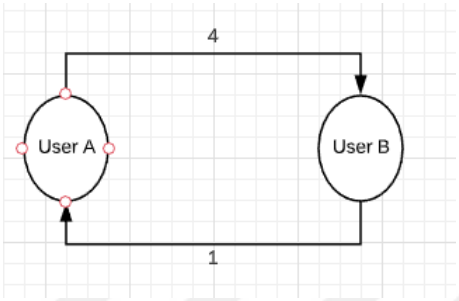


Figure 1.5: Weighted graph example: Retweet counts

## 1.4 Community Detection in Social Networks

Community detection is a concept that identifies subsystems interaction between nodes with regards to the network structure of a complex system. These subsystems are called communities or clusters [23]. Community detection is used in different areas such as Social media [58, 5, 39, 40], Communications Networks [9, 50, 37], Biological Systems and Healthcare [60, 65], Economics [19, 48], Academia and Scientometrics [38, 54, 24], E-commerce [2, 43, 6]. Many algorithms, techniques, and tools have been developed and used for community detection. Communities identified as a result of community detection are classified into two different types according to the status of the nodes. If nodes are members of only one community, they are called disjoint communities. If a node can subscribe to two or more communities, these communities are called overlapping communities [15].

## 1.5 Political Polarization

In politics, polarization (or polarisation) can refer to the divergence of political attitudes to ideological extremes. Almost all discussions of polarization in political science consider polarization in the context of political parties and democratic systems

of government [6].


Figure 1.6: Political Polarization - Yes or No

The political polarization in Turkey is similar to those in other countries. In addition, polarization has a high effect on society in certain periods. These periods are; election periods, social events, foreign country relations, etc.

[6] *https://en.wikipedia.org/w/index.php?title=Political_polarization* (*accessed* : 28.07.2020)

# CHAPTER 2
# BACKGROUND

In this section, information is given about the technologies used in this study, which we think should be known about big data discussions.

## 2.1 Map Reduce

MapReduce is a system developed by Google that allows easy analysis of very large data in a distributed manner. In this system, two functions namely map and reduce, are used. The other two functions included in the figure were used to help map-reduce functions (Splitting, Reducing). These functions are not essentially required.

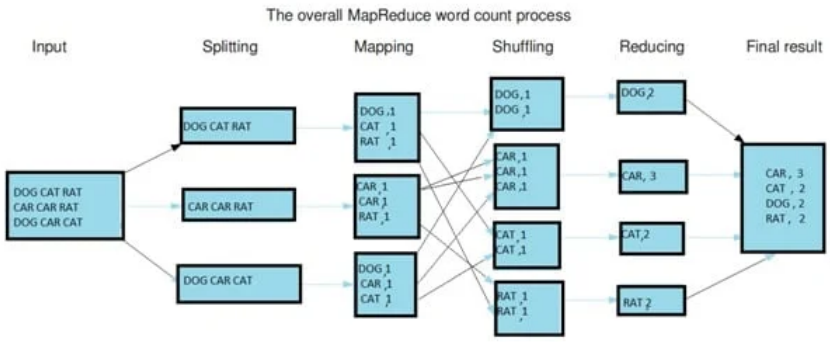


Figure 2.1: Representation of the map-reduce function

During the map phase, the data is taken and sent to the workers in smaller pieces. This process works in parallel. In the Map function, we expect the data to be brought into compliance status, which we can use in studies. In addition, in the map function, we can make restrictions on the data and choose from the data. The object consists of the map function value and key returns.

On the other hand, Reduce is the part where the data returned by the map function

is combined with the logic we will use. Reduce operation contains the operations performed on the value by looking at the keys returned from the Map function. The basic reduce operation is the sum operation.

## 2.2 PySpark

Spark is developed and distributed by Apache. Spark is a data processing engine that is faster than Hadoop. Apache Spark performs much better than Hadoop in the distributed implementation of artificial learning algorithms. Apache Spark has a high-level operator called RDD (Resilient Distributed Dataset). This operator facilitates operations on unstructured data. Apache Spark can be developed with Scala, Java, R, and Python languages.

The spark library used by Python is called pyspark. It was preferred in this study because Spark is fast and developed with python via pyspark.

# CHAPTER 3
# LITERATURE REVIEW

Social media trends lead to more exploratory research in social media analysis, event management, and so on. In the recent Turkish 2019 local election, many studies have been done to analyze candidates in terms of different criteria using the election tweets on different twitter accounts. The most adopted approaches in these studies were in terms of; profile management, tweets management, and event management perspectives. In the tweet management analysis, the twitter accounts of the candidates were registered earlier but happened to be actively used during the election period to reach out to voters (citizens) [63].

Furthermore, in the past the US general election, the process was analyzed in real-time, an emotional analysis was carried out with classification algorithms, as a result of the examination of over 100.000 Tweets, political orientations were estimated with 1.65% error [53].

In another study conducted using twitter data during the Presidency of June 24, 2018, political parties and candidates emphasized that a political agenda could be created in online networks, the propaganda made was rapidly reached to large masses, and that this propaganda was repeatedly shared on the social network by re-tweeting people [20].

In another study, a comprehensive study was carried out on the general elections of June 7, 2015, in Turkey. In this study, 18 columnists, and 810 tweets published by previously determined authors were analyzed. This work considered more media agendas [22].

In another study related to the 2019 local elections, by using thematic analysis method, the researchers tried to understand what strategies the political propaganda activities in social media contained. This study was carried out by examining the considerations of political leaders, and it was seen that the leaders' communication styles were mostly addressed to the voters on the subject tags and a one-sided interaction method was adopted [26].

Recently, Twitter has been used by local governments to inform citizens about services and works and to get feedback from citizens about these services and works which made it easier to reach citizens [30]. Local governments can communicate with social media and citizens and can be informed with new activities, and their work is directed by taking their opinions and suggestions [7]. Another study on social media structures provided unique opportunities for e-government structures of governments to emphasize that institutions can interact with citizens using these channels [21].

A study was carried out to determine which crime report was taken from 3 countries (India, England, America). Incident reports that did not have a crime tag were labeled by using graph and graph clustering methods. The graph creation process was tried to be determined by creating the words Person - Person (PER-PER), Person - Location (PER - LOC), Organization – Person (ORG - PER), which are the words in the report. At the worst case on average values, the results were 75.65% in the UK for PER - PER, 77.99% in India for PER - LOC, and 73.08% in India for ORG - PER. The results were better in the Precision metric than other metrics (Recall, F-measure) [13].

The results on ready datasets (Football, Karate, Dolphins, Books) with Louvain and LPA community detection algorithms and their own BCD (Brainstorming for Community Detection) algorithm using 3 Metrics (Normalized Mutual Information - NMI, Max Modularity - Qmax, Average Modularity - Qavg) were compared. For the Karate

dataset, BCD produced the best result for all metrics. For the other two datasets (Dolphins and Books), BCD produced good results in NMI however it failed to do so in the remaining datasets [66].

By analyzing emotions of the tweets sent by candidates in the American elections, it was tried to determine whether the discourses of the candidates affected more than one million users. 3 hypotheses were analyzed and 2 of these hypotheses were confirmed to be correct and 1 hypothesis was found to be inaccurate. The first hypothesis "There is a relationship between daily news and events and the words popular on Twitter" was found to be correct. The second hypothesis suggested that the tweets sent by the candidates had an effect on the users, but it was revealed that this claim was not supported due to the fact that the data of the candidates in the dataset was quite small and its effect was negligible. The last hypothesis was that users did not generate new ideas and did not interact with each other. This hypothesis was confirmed because the re-tweets were too many and the tweets and direct messages sent at first were small in number [62].

In another study, a method called vision intensity function was proposed to measure polarization. It has been shown that using the proposed method of tweets sent by Hugo Chavez in the Venezuelan elections and the interaction of users with these tweets, different degrees of polarization depending on the structure of the network were present [35].

In another study conducted on the 2016 election in America and the 2017 election in England with location-based emotion analysis studies, unlike the data collection method, similar trends were observed towards political candidates and parties [61].

It was conducted in a study on the roles and behaviors of government-citizen and local government-citizen interaction. It was emphasized that better conditions should be created and more research should be done to improve the communication of each of the European Union member giants [51].

The findings obtained from another research by using the location variable to analyze Twitter data in more detail have emphasized that user sensitivity and behavior analysis will reach more accurate results. A web-based application that facilitates data analysis and local-based analysis has been developed and its application has been tested in the 2016 election in America [62].

In Saudi Arabia, it has been analyzed by analyzing the tweet patterns, types, content, and interactions of the municipalities. Municipalities have been observed to have the highest number of reactions with users about projects and actions [1].

In the study on Facebook messages, researches and studies were conducted on the communication strategies of the institutions on social media. As a result of this study, it has been observed that there are differences in the interaction of institutions with citizens depending on their mission and goals [14].

# CHAPTER 4
# COMMUNITY DETECTION

In this section, the metrics used in the evaluation of graphs and the algorithms used in community detection are explained. Metric formulas and pseudo-codes of algorithms are included.

## 4.1 Community Detection

It is the name given to the units formed by people that can be put together in various ways such as certain criteria and features. Although it is a concept used in different disciplines such as sociology and biology, it also entered into Computer Science using graphs in network analysis (social networks, network networks, etc.).

Figure 4.1: Community

In order to find the communities in the data that can be displayed with a graph, the process of clustering the nodes called community detection is used. Communities are also an indicator of the presence of interrelated, and/or affected nodes [59].

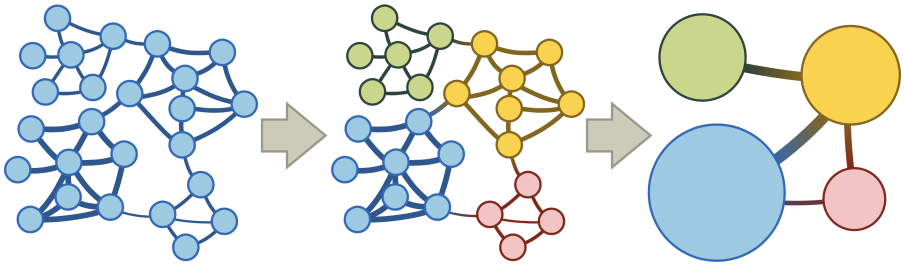

Figure 4.2: Community Detection

Numerous community detection algorithms have been developed to detect communities. The basic starting point for algorithms is finding interactions between nodes, group nodes, and merging those nodes (Figure 4.2).

## 4.2 Graph Measures

Community detection algorithms work based on the metrics used for graphs. These metrics vary widely. Here are the most common metrics in community detection algorithms.

### 4.2.1 Modularity

Modularity is the value that expresses the power of a graph to be divided into lower parts [29]. It can take values between (-1, +1). If the value of 1 is taken, the connection between the graph is strong.

$$Q = \frac{1}{2m} \sum_{i,j \in N} \left[ A_{ij} - \frac{k_i * k_j}{2m} \right] \tag{4.1}$$

$$m = \sum_{i,j \in N} A_{ij} \tag{4.2}$$

where:

$N$ : set of nodes

$A_{ij}$ : weighted of the edge between $i$ and $j$

$k_i$ : degree of node $i$

$k_j$ : degree of node $j$

Table 4.1: Worst - Best Values for Modularity Metrics

| ***Worst Value*** | ***Best Value*** |
|---|---|
| $-1$ | $+1$ |

### 4.2.2 Closeness Centrality

Closeness is calculated by calculating the sum of the inverse distances of a node's shortest paths to other nodes [18].

$$C(x) = \frac{1}{\sum_{y} d(y,x)} \tag{4.3}$$

where:

$d(y,x)$ : Distance value from $x$ to all nodes

Table 4.2: Worst - Best Values for Closeness Centrality Metrics

| ***Worst Value*** | ***Best Value*** |
|---|---|
| 0 | $+1$ |

### 4.2.3 Edge Betweenness Centrality

It is the calculation of the centrality value for the edges. It is calculated by dividing the number of paths between nodes by the shortest path [31].

$$C_b(e) = \sum_{s,t \in V} \frac{\sigma(s,t|e)}{\sigma(s,t)} \tag{4.4}$$

where:

$V$ : set of nodes

$\sigma(s,t|e)$ : number of those paths ($s$-$t$ nodes)

$\sigma(s,t)$ : number of shortest paths ($s$-$t$ nodes)

Table 4.3: Worst - Best Values for Edge Betweenness Centrality Metrics

| ***Worst Value*** | ***Best Value*** |
|---|---|
| 0 | +1 |

### 4.2.4 Erdos Renyi Modularity

It is known as Newman-Girvan Modularity. It is a metric developed for random graphs. It assumes that the vertex points in the graph are connected by a random probability p. It is tried to calculate the density with the communities in the graph [33]. This metric produces a value between 0 and 1.

$$Q(S) = \frac{1}{m} * \sum_{s \in S} \left[ m_s - \frac{m * n_s * (n_s - 1)}{n * (n - 1)} \right] \tag{4.5}$$

where:

$m$ : the number of graph's edges

$m_s$ : the number of community's edge

$n_s$ : the number of edges from nodes in S to nodes outside S.

$n$ : the number of nodes

Table 4.4: Worst - Best Values for Erdos Renyi Modularity Metrics

| ***Worst Value*** | ***Best Value*** |
|---|---|
| 0 | +1 |

### 4.2.5 Conductance

It is a measure of how well-knit a graph is. It is the volume of elements that show out from a group. Small values are desired [49].

$$f(s) = \frac{c_s}{2 * m_s + c_s} \tag{4.6}$$

where:

$c_s$ : the number of community

$m_s$ : the number of community's edges

Table 4.5: Worst - Best Values for Conductance Metrics

| ***Worst Value*** | ***Best Value*** |
|---|---|
| +1 | 0 |

### 4.2.6 Modularity Density

Modularity density is a new process that uses the number of nodes in each cluster to normalize the objective value rather than the total number of edges. It is aimed to prevent small and dense communities from being neglected [45].

$$Q(S) = \sum_{c \in S} \frac{1}{n_c} \left[ \sum_{i \in C} k_{iC}^{in} - \sum_{i \in C} k_{iC}^{out} \right] \tag{4.7}$$

where:

$S$ : set of communities

$n_C$ : the number of nodes in C

$k_{iC}^{in}$ : the degree of node i within C

$k_{iC}^{out}$ : the degree of node i outside C

Table 4.6: Worst - Best Values for Erdos Renyi Modularity Metrics

| ***Worst Value*** | ***Best Value*** |
|---|---|
| *Lowest value* | *Highest value* |

## 4.3 Community Detection Algorithms

### 4.3.1 Louvin Algorithm

It was built on the faster community in large networks in 2008 and emerged with the work of Blondel and his friends at Louvain University [9]. It is an algorithm developed using the modularity metric. It is aimed to reach the most effective (high) value

of the modularity value. We try to maximize modularity value for each community individually. The algorithm continues to run until the modularity values of the communities remain unchanged (See: Algorihms 1). The algorithm is used in weighted but undirected graphs.

```
function LouvainPruneMoveNodes(Graph G)
  C the index of communities for each nodes of G
  P = V(G)
  while P ≠ ∅ do
    v = random node x ∈ P
    P = P − {v}
    best_q = −∞
    best_c = community of v
    for all neighboring nodes n of v do
        gain_q = ΔQ between v and n
      if best_q < gain_q then
        best_q = gain_q
        best_c = community of n
      end if
    end for
    C = Place v the best_c
    for all neighboring nodes n of v do
      if n is not in community of v then
        P = P + {n}
      end if
    end for
  end while
  return C
end function
```

Figure 4.3: Louvin Algorithm

### 4.3.2 Label Propagation Algorithm (LPA)

It is an algorithm that can quickly detect communities on the network without having any knowledge of communities in a network structure and was recommended by Raghavan et al in 2007 [41]. Initially, a unique tag is assigned to each node. The closeness value is calculated for each node and those with the same closeness value are assigned the same label. Operations continue until the labels on the nodes remain stable [57].

**Input**: $G = (V, E)$.
**Output**: the result of community detection.
(1) Initialization: assign a unique label to each node in the network, $c_i(0) = i$.
(2) Calculate the node influence value for each node and arrange nodes in descending order of NI storing the results in the vector $X$.
(3) Iteration of label propagation:
(a) Set $t = 1$;
(b) For each node $v_i \in X$, let $c_i(t) = f(c_{i1}(t), \ldots, c_{im}(t), c_{i(m+1)}(t-1), \ldots, c_{ik}(t-1))$, where $v_{i1}, \ldots, v_{im}$ are neighbors of $v_i$ that have already been updated in the current iteration and $v_{i(m+1)}, \ldots, v_{ik}$ are neighbors that are not yet updated in the current iteration. The function $f$ here returns the label that the maximum number of their neighbors has. If multiple labels simultaneously contained by the greatest number of nodes, then we recalculate each of the values of labels contained by greatest number nodes according to (5) and choose the label with maximum value to assign to the node $v_i$.
(c) If the label of every node does not change anymore, then stop the algorithm. Else, set $t = t + 1$ and go to Step (b).
(4) Community division: assign all nodes share the same label into a community; the type of labels indicates the number of communities.

Figure 4.4: Label Propagation Algorithm

### 4.3.3 Girvan–Newman algorithm

It is a hierarchical community detection algorithm, using the edge betweenness metric. It works with an approach that seeks to identify the potential likely to be among communities. In other words, it focuses on possible factors among communities rather than the ones that should be in the community. It ensures that the communities are distinguished by erasing the edges at each iteration [17]. It works with weightless and non-directional graphs.

**Pseudo code:**

Step 1. Calculate edge-betweenness for all edges.

Step 2. Remove the edge with highest betweenness.

Step 3. Recalculate betweenness.

Step 4. Repeat until all edges are removed, or modularity function is optimized (depending on variation)

Figure 4.5: Girman-Newman Algorithm

### 4.3.4 Eigenvector algorithm

It is an algorithm that tries to make community detection according to the modularity metric [36]. The large network is aimed at identifying small communities with renewal until maximum modularity is reached.

```
Let v = [1, 1, 1, 1, ... 1].
Repeat 100 times {
  Let w = [0, 0, 0, 0, ... 0].
  For each person i in the social network
    For each friend j of i
      Set w[j] = w[j] + v[i].
  Set v = w.
}
Let S be the sum of the entries of v.
Divide each entry of v by S.
```

Figure 4.6: Eigenvector Algorithm

### 4.3.5 Leiden algorithm

It is an algorithm that emerged with the development of the Louvain Algorithm. It is a very new algorithm proposed in 2018 [52]. It is a more complex algorithm than the Louvain algorithm, but it guarantees a better community detection. The ability to work with directional and weighted graphs is a big plus. Smaller communities are united to create a maximum of modularity metrics for communities divided by movements within the community, and larger communities are created.

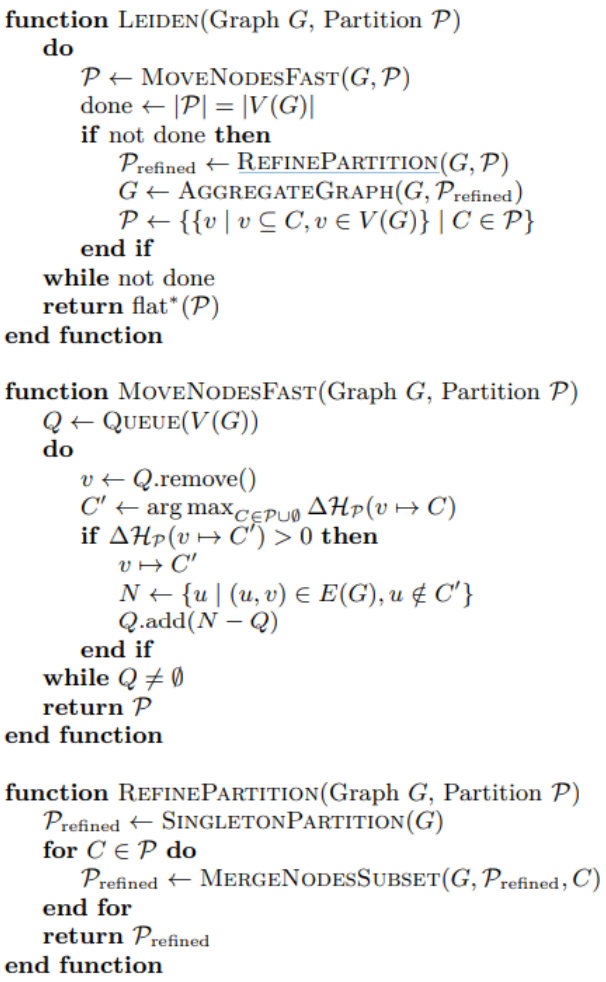


Figure 4.7: Leiden Algorithm

# CHAPTER 5
# MEASURING POLITICAL POLARIZATION USING BIG DATA

## 5.1 Dataset

We use Twitter data in our study and we make use of Twitter API (Application Programming Interface) to collect Twitter data (Tweet, retweet, etc.). We collected data following the steps below.

- We have become a member of the API site [1], which Twitter has prepared specifically for developers.
- After the account was created, we created ***apps*** for our app. When created, we need to save the values of ***Consumer Key***, ***Consumer Secret***, ***Access Token*** and ***Access Token Secret*** that are exclusive to us for use in our application.
- We used the ***Twitter Streaming Api*** because we wanted to reach all the tweets that were posted regardless of the user.
- We have collected the data by defining the ***StreamListener*** in the ***Tweepy*** library for Python.
- • We recorded the collected data by writing to the .txt file. We have given the date of the day taken as the file name. For example: *20190221.txt*

Since there were provinces repeated in the elections, data had to be collected over a long period. It was the most important Istanbul election among the repeated elections.

**157 days** of data were collected between **14/02/2019 - 20/07/2019**. We collected about **90,005,000 datas**.

[1] *https://developer.twitter.com/*

Table 5.1: Data Information

| | ***Counts*** |
|---|---|
| **Retweeet count** | $68,074,679$ |
| **Tweet count** | $21,930,465$ |
| **TOTAL** | $90,005,144$ |

## 5.2 Data Preparation and Pre-processing

After collecting the data, we had files with .txt extension with tweets per day. By reading these files with Python code, using the map-reduce structure in pyspark, we transformed the tweets to be used by our algorithm.

In the map function, we reached the user who posted the tweet and the user/users who retweeted this tweet, from within the json object.

**Algorithm 1** Map and reduce function for map-reduce(pyspark)

```
1: procedure MapTweet(tweet)
2:     if tweet is retweeted then
3:         user1 ← user who tweeted
4:         user2 ← user who retweeted the tweet
5:         EMIT([user1, user2], 1)                    ▷ user1 retweets user2
6: procedure Reduce([user1, user2], counts[])
7:     sum ← 0
8:     while all count c ∈ counts [c1, c2, c3] do
9:         sum ← sum + c
10:        EMIT([user1, user2], count sum)
```

Our map function ran each line in parallel. Our map function returned a 1 for each user who retweeted and tweeted. We combined the values returned from the map function with the reduce function. Here we used the reduce function in its basic form by doing a sum.

We placed the data we obtained as a result of Reduce under another folder in the form of .txt, in the same way as the name of the file we read.

## 5.3 Preference of Community Detection Algorithm

We had to select one of the 5 community detection algorithms (LPA, Louvain, Leiden, Girvan–Newman, EigenVector). When making this choice; we considered that an algorithm working on directional and weighted graphs is required. Community detection algorithms are generally directional but take weight into account.

Table 5.2: Algorithms that work in graph types

| ***Algorithms*** | ***Directed Graph*** | ***Weighted Graph*** |
|---|---|---|
| *LPA* | *X* | – |
| *Leiden* | *X* | *X* |
| *Louvain* | *X* | *X* |
| *Girman – Newman* | *X* | – |
| *Eigenvector* | *X* | – |

[X] supported by the algorithm.
[-] not supported by the algorithm.

Since the directional and weighted graph of the graph will be used in our study, two algorithms (Leiden, Louvain) can be preferred when Table 5.2 is analyzed.

In addition to Table 5.2, we obtained results by using some metrics for ***3500000 datas*** and ***98000 nodes***.

Table 5.3: Algorithm results by metrics for 3500000 datas and 98000 nodes

| ***Algorithms*** | ***Erdos Renyi Modularity*** | ***Conductance*** | ***Modularity Density*** |
|---|---|---|---|
| *LPA* | 0.7829 | 0.1619 | 4190.1575 |
| *Leiden* | 0.8127 | 0.0061 | 3530.6777 |
| *Louvain* | 0.7887 | 0.0123 | 3498.3589 |
| *Girman – Newman* | 0.8111 | 0.0949 | 4395.4571 |
| *Eigenvector* | 0.3031 | 0.1120 | 3448.5386 |

It is necessary to examine the results in Table 5.3 separately according to metrics.

- The best result is 1 for Erdos Renyi Modularity. When the algorithms are examined, the algorithm that produces the closest result to the value 1 is **Leiden**

**Algorithm**.

- For the **Conductance** metric, , we expect the best result to be close to 0. When the results are examined, the algorithm that gives the closest result to 0 is **Leiden Algorithm**.

- There is no ideal upper or lower value for the **Modularity Density** metric. The density value is expected to be high for this metric. The best result for this metric is the **Girman - Newman Algorithm**.

We also found the number of communities that these 5 algorithms can detect for the same data (3500000 data and 98000 nodes) and the number of members of the largest community.

The effect of Table 5.4 on our choice is that we do not want the number of members in 1 community to be overly high. This situation is not a good result. We do not want the number of communities to be small and the number of members too high. 2 algorithms (Leiden, Louvain) produce the result we want.

It has been observed in the studies that Leiden and LPA algorithm produced very similar results for the Networkx library. The slight differences seen do not matter for our study. We want the number of communities in social media to be neither too small nor too much. For this, the LPA algorithm is used in this study.

Table 5.4: Results of Algorithms

| ***Algorithms*** | ***Size of Communities*** | ***the number of users in the largest community*** |
|---|---|---|
| *LPA* | 11242 | 28919 |
| *Leiden* | 6601 | 21423 |
| *Louvain* | 6637 | 22711 |
| *Girman − Newman* | 11433 | 21506 |
| *Eigenvector* | 6263 | 82434 |

### 5.4 Identifying communities

Using the LPA algorithm we chose, we enabled the detection of communities. We wrote the communities we identified to .txt files.

When determining the communities, all data in March were used. The reason for choosing March was that the election process would make itself felt strongly in this month.

**Algorithm 2** Identification of communities

```
1: procedure FindCommunities(retweet_folder, communnity_size)        ▷ .txt files(tweet) in folder
2:     Graph ← Null
3:     while all the files in the retweet_folder are readed do ▷ read all files in folder
4:         add to Graph - user1,user2 and retweet Count
5:     lpa_com ← Find Community for Graph according to LPA algorithm
6:     write lpa_com to file
```

### 5.5 Political Polarization Detection and Measurement

For interaction between communities, we calculated the inter re-tweet ratio metric found by dividing the number of interactions within the community by the total number of interactions between other communities.

**Algorithm 3** Found Communities Relations

```
1: procedure found_com_relations(file)
2:     while For the whole communities do
3:         users ← Separately for each community
4:         while All users in the community do
5:             while again all users in Communities do   ▷ The user here is named as other_user
6:                 if the user Is herself then                ▷ user equal to other_user
7:                     continue
8:                 else
9:                     if There is a relationship between user and other_user then
10:                        add to relationship matrix   ▷ add to matrix(user,other_user) + retweet_count
```

The low inter re-tweet ratio value is an indication of political polarization. Inter re-tweet ratio is the metric we use for political polarization. This is a measure showing the ratio of re-tweets between groups to all re-tweets between all users including re-tweets with groups, also called inter group re-tweets.

$$inter\_retweet\_ratio = \frac{intra}{inter + intra} \tag{5.1}$$

- **Intra =** Intra is the total number of interactions within community.
- **Inter =** Inter is the total number of interactions between groups.

**Algorithm 4** Calculation of Inter Retweet Ratio

```
1: procedure CAL_RATIO(file)
2:     while read all lines in file do          ▷ line in file and line is a community
3:         interrt ← 0                          ▷ relationship in the community
4:         intrart ← 0                          ▷ relationship between other communities
5:         while line in fh : do
6:             if user1 and user2 are in the same community then
7:                 interrt += int(weighted_value)      ▷ weighted_value = Total interactions within 2 users
8:             else
9:                 intrart += int(weighted_value)      ▷ weighted_value = Total interactions within 2 users
10:    inter_retweet_ratio ← intrart/interrt + intrart
```

The algorithm we used to measure inter retweet ratio is LPA algorithm as implemented in networkx package.

# CHAPTER 6

# EXPERIMENTS

This section describes the necessary configuration setups and the experimental results of this thesis study.

## 6.1 Experimental Setup

We used our application on a server to take advantage of the power of spark and to process the data quickly. Our Ubuntu server is a system consisting of ***190 GB RAM, 40 cores*** and ***20 workers***. The properties of each core are given in Table 5.1.

Table 6.1: Properties of each core

| | |
|---|---|
| $cpu\ family$ | 6 |
| $model\ name$ | $Intel(R)Xeon(R)CPUE5-2630v4@2.20GHz$ |
| $CPU\ MHz$ | 2296.022 |
| $Cache\ size$ | $25600KB$ |
| $siblings$ | 20 |
| $cpu\ cores$ | 10 |
| $fpu$ | $yes$ |
| $cpuid\ level$ | 20 |
| $bogomips$ | 4399.64 |
| $clflush\ size$ | 64 |
| $cache\ alignment$ | 64 |
| $address\ sizes$ | $46bitsphysical, 48bitsvirtual$ |

The code and experiment of this studies can be found in GitHub repository [1].

[1] $https://github.com/selimsurucu/community\_detection/$

## 6.2 Experimental Results

Due to the large size of our data and the fact that we collect data in a wide range, we have determined a large number of users and a large community.

Table 6.2: Number of users and communities

| | |
|---|---|
| **Number of unique users** | 384272 |
| **Number of users in all communities** | 384272 |
| **Number of communities** | 4767 |

We calculated the daily and monthly inter retweet ratio values over a long period, taking into account the repeated selection in cities like Istanbul. In order for us to see the impact of political polarization, the country had to leave the election period entirely. For this, we kept our analysis in a wide range of dates.

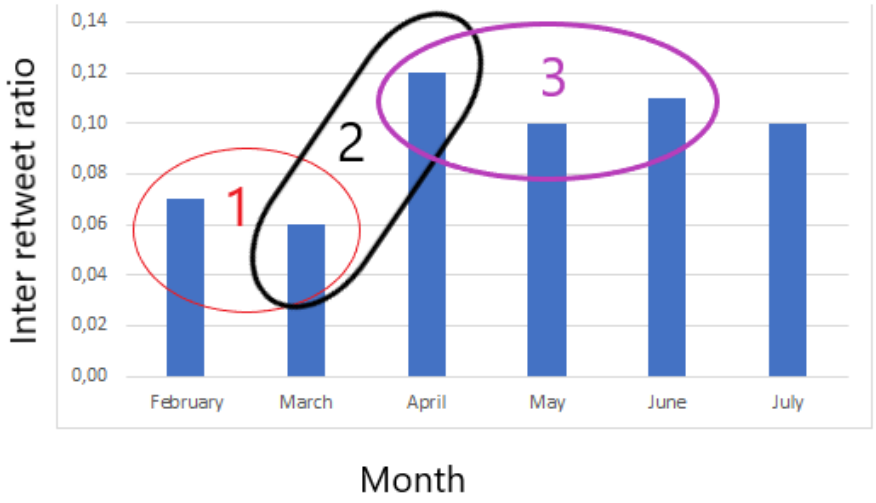


Figure 6.1: Montly Results

When we examine Figure 6.1 according to the areas on it:

- **For area 1 (red circle)** = It is the period in which the least interaction (inter retweet ratio) is observed in 6 months. The reason for this is the selection effect.
- **For area 2 (black circle)** = In April, we see the highest interaction (inter retweet ratio) in 6 months. Since the election took place on the last day of March, it could mean that the country is out of the election mood.

- **When area 1 (red circle) and area 2 (black circle are considered together)** = We can clearly see the political polarization during these periods.

- **For area 3 (lilac circle)** = Interaction (inter retweet ratio) between communities is very close to each other in these 3 months (April, May, and June). We think that the reason for the small decrease in May and June is due to the elections repeated in some places, especially in Istanbul.

Looking at Figure 6.1 showing the monthly results, it was seen that the minimum inter retweet ratio value was low in February and March. Considering that the election is at the end of March, we can assume that March is the month when the election was the most contested among users. Monthly results confirm this opinion as the least interaction period is March.

Immediately after the election, there was a remarkable increase, but after this increase, there is a decrease again. We think that this decline is due to the re-election in some provinces. Re-entering an election period may have reduced interaction between users.

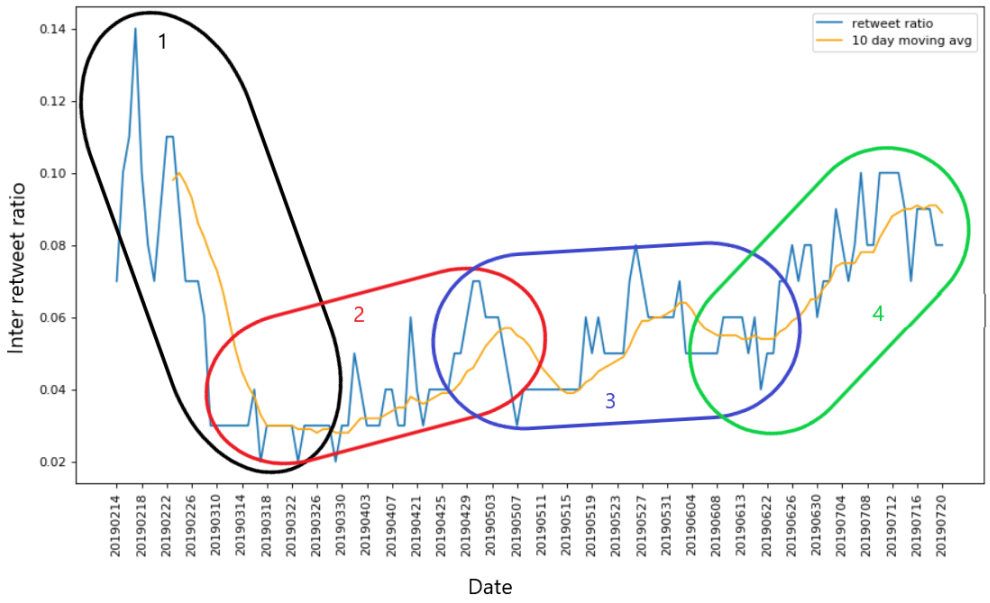


Figure 6.2: Daily Results

When we examine Figure 6.2 according to the areas on it we can make the following observations:

- **For area 1 (black circle) =** There is clearly a sharp decrease in interactions This situation shows us that the election environment has been entered. This is one of the areas showing us that the political polarization took place in Turkey.

- **For area 2 (red circle) =** There is little interaction until the election date (31 March). After the election, at the beginning of April, there are jumps in interactions. This shows us that the effect of the election has decreased.

- **For area 3 (blue circle) =** It was observed that interactions increased during this period. But there are sharp declines from time to time.

- **For area 4 (green circle) =** Until the repeated selection of istanbul (date = 23 June), it is seen that there is little interaction.After repeated selection of Istanbul, which exited during the election period in Turkey, it is evident from the increasing interaction between the communities.

When the daily results in Figure 6.2 are analyzed, it is seen that the inter retweet ratio value decreased significantly in the last month before the election.

It is observed that there was an increase in April. But when the days of this month are examined, it is seen that the fluctuations are deep. We think that this is due to the continuation of speculation about the election, and the decision to repeat the election in some cities with the third week of April. We see that the inter retweet ratio value does not decrease at a rate similar to that in March. We think that this is due to the low number of provinces where the election will be repeated.

It is understood from the increase in the inter retweet ratio of the rate that the country left the political period after the second election in Istanbul. After 23 June, when the renewed election took place in Istanbul, the inter retweet ratio value increased above the 10-day average.

# CHAPTER 7
# CONCLUSIONS AND FUTURE WORK

With this study, we have shown that there is a political polarization among social media users and this can be proven by quantitative measurements. In addition, our hypothesis has been confirmed that users will have less interaction with users who have different views and opinions during the political propaganda period.

This study shows that without having prior knowledge of users on social media, political communities in which users are located can be identified. In other words, the daily movements we make on social media / Internet have shown that meaningful results can be produced when examined for a long time.

In future studies, the meaning of the messages in the collected data can be analyzed by emotion analysis methods. As a result of this review, the message may have an impact on users.

It can also be checked if there are political parties and candidates among the members of the identified communities. By looking at these, additional observations can be made in the analysis of the communities.

If other social media accounts of users can be detected in future studies, better community results can be determined in more complex data. But in order to do this, users of social media tools such as Twitter and Facebook must be matched. This pairing can be done even with mail accounts, and consequently more data will be obtained and more accurate analysis can be carried out by analyzing it.

The detected communities can be compared one by one by looking at their interactions with each other. As a result, communities with high and low interactions can be identified. Communities that have different thoughts with one another and communities with the same idea can be identified.

## REFERENCES


[1] Arwa Al-Aama. "The use of Twitter to Promote E-participation: Connecting Government and People". In: *International Journal of Web Based Communities* 11 (Jan. 2015), p. 73. DOI: 10.1504/IJWBC.2015.067082.

[2] G. Adomavicius and A. Tuzhilin. "Toward the Next Generation of Recommender Systems: a Survey of the State-of-the-art and Possible Extensions". In: *IEEE Transactions on Knowledge and Data Engineering* 17.6 (2005), pp. 734–749.

[3] B. Ayan, B. Kuyumcu, and B. Ciylan. "Detection of Islamophobic Tweets on Twitter Using Sentiment Analysis". In: *Fen Bilimleri Dergisi , Gazi Universitesi.* 2019, pp. 495–502.

[4] A. Azab et al. "Fake Account Detection in Twitter Based on Minimum Weighted Feature set". In: *World Academy of Science, Engineering and Technology, International Journal of Computer and Information Engineering* 10 (2015), pp. 13–18.

[5] Punam Bedi and Chhavi Sharma. "Community Detection in Social Networks". In: *WIREs Data Mining and Knowledge Discovery* 6.3 (2016), pp. 115–135. DOI: 10.1002/widm.1178.

[6] Ghazaleh Beigi et al. "Leveraging Community Detection for Accurate Trust Prediction". In: (2014).

[7] J. C. Bertot et al. "Social Media Technology and Government Transparency". In: *Computer* 43.11 (2010), pp. 53–59. DOI: 10.1109/MC.2010.325.

[8] David R. Bild et al. "Aggregate Characterization of User Behavior in Twitter and Analysis of the Retweet Graph". In: *ACM Trans. Internet Technol.* 15.1 (Mar. 2015). ISSN: 1533-5399. DOI: 10.1145/2700060.

[9] Vincent D Blondel et al. "Fast Unfolding of Communities in Large Networks". In: *Journal of Statistical Mechanics: Theory and Experiment* 2008.10 (Oct. 2008), P10008. ISSN: 1742-5468. DOI: 10.1088/1742-5468/2008/10/p10008.

[10] Francesco Buccafurri et al. "Comparing Twitter and Facebook User Behavior: Privacy and Other Aspects". In: *Computers in Human Behavior* 52 (2015), pp. 87–95. ISSN: 0747-5632. DOI: https://doi.org/10.1016/j.chb.2015.05.045.

[11] Qiang Cao et al. "Aiding the Detection of Fake Accounts in Large Scale Social Online Services". In: Apr. 2012, pp. 15–15.

[12] E. Connors. "Trademarks and Twitter: The Costs and Benefits of Social Media on Trademark Strength, and What This Means for Internet-Savvy Celebs Notes". In: *Knowledge-Based Systems* 52 (2018), p. 189.

[13] Priyanka Das and Asit Kumar Das. "Graph-Based Clustering of Extracted Paraphrases for Labelling Crime Reports". In: *Knowledge-Based Systems* 179 (2019), pp. 55–76. ISSN: 0950-7051. DOI: https://doi.org/10.1016/j.knosys.2019.05.004.

[14] Nic DePaula and Ersin Dincelli. “An Empirical Analysis of Local Government Social Media Communication: Models of E-government Interactivity and Public Relations”. In: June 2016, pp. 348–356. DOI: 10.1145/2912160.2912174.

[15] Santo Fortunato and Darko Hric. “Community Detection in Networks: A User Guide”. In: *Physics Reports* 659 (Nov. 2016), pp. 1–44. ISSN: 0370-1573. DOI: 10.1016/j.physrep.2016.09.002.

[16] P. Garg, H. Garg, and V. Ranga. “Sentiment Analysis of the Uri Terror Attack Using Twitter”. In: *2017 International Conference on Computing, Communication and Automation (ICCCA)*. 2017, pp. 17–20.

[17] M. Girvan and M. E. J. Newman. “Community Structure in Social and Biological Networks”. In: *Proceedings of the National Academy of Sciences* 99.12 (2002), pp. 7821–7826. ISSN: 0027-8424. DOI: 10.1073/pnas.122653799.

[18] Jennifer Golbeck. “Chapter 3 - Network Structure and Measures”. In: *Analyzing the Social Web*. Ed. by Jennifer Golbeck. Boston: Morgan Kaufmann, 2013, pp. 25–44. ISBN: 978-0-12-405531-5. DOI: https://doi.org/10.1016/B978-0-12-405531-5.00003-1.

[19] Xiangquan Gui et al. “Dynamic Communities in Stock Market”. In: *Abstract and Applied Analysis* 2014 (May 2014), pp. 1–9. DOI: 10.1155/2014/723482.

[20] N. Güz, C. Yeğen, and B.O. Aydın. “Dijital Propaganda ve Politik Başarı: 24 Haziran 2018 Cumhurbaşkanlığı Seçiminin Twitter Analizi”. In: *Erciyes İletişim Dergisi* 6.2 (2019), pp. 1461–1482.

[21] Sara Hofmann et al. “Old Blunders in New Media? How Local Governments Communicate with Citizens in Online Social Networks”. In: Jan. 2013, pp. 2023–2032. ISBN: 978-1-4673-5933-7. DOI: 10.1109/HICSS.2013.421.

[22] N. Bilge İspir and Deniz Kılıç. “Defining Twitter Agenda During June 2015 Turkey General Election, A Social Network Analysis Application”. In: *Karadeniz Teknik Üniversitesi İletişim Araştırmaları Dergisi* 7 (2017), pp. 2–10. ISSN: 2146-3212.

[23] Muhammad Aqib Javed et al. “Community Detection in Networks: A Multidisciplinary Review”. In: *Journal of Network and Computer Applications* 108 (2018), pp. 87–111. ISSN: 1084-8045. DOI: https://doi.org/10.1016/j.jnca.2018.02.011.

[24] Pengsheng Ji and Jiashun Jin. *Coauthorship and Citation Networks for Statisticians*. 2014. arXiv: 1410.2840 [stat.AP].

[25] Kyujin Jung and Han Woo Park. “Citizens’ Social Media Use and Homeland Security Information Policy: Some Evidences from Twitter Users During the 2013 North Korea Nuclear Test”. In: *Government Information Quarterly* 31.4 (2014), pp. 563–573. ISSN: 0740-624X. DOI: https://doi.org/10.1016/j.giq.2014.06.003.

[26] Deniz KILIÇ and Gökhan GÖKULU. “Sosyal Medyada Siyasal İletişim: 31 Mart 2019 Yerel Seçimlerinde Siyasi Liderlerin Twitter Kullanımı Örneği”. In: *Journal of Awareness* 5 (May 2020), pp. 213–136. DOI: 10.26809/joa.5.017.

[27] Bumsoo Kim. “Effects of Social Grooming on Incivility in COVID-19”. In: *Cyberpsychology, Behavior, and Social Networking* 23 (Apr. 2020). DOI: 10.1089/cyber.2020.0201.

[28] Rohan Kshirsagar et al. *Predictive Embeddings for Hate Speech Detection on Twitter*. 2018. arXiv: 1809.10644.

[29] Zhenping Li et al. "Quantitative Function and Algorithm for Community Detection in Bipartite Networks". In: *Information Sciences* 367-368 (2016), pp. 874–889. ISSN: 0020-0255. DOI: https://doi.org/10.1016/j.ins.2016.07.024.

[30] David Lorenzi et al. "Utilizing Social Media to Improve Local Government Responsiveness". In: *Proceedings of the 15th Annual International Conference on Digital Government Research*. dg.o '14. New York, NY, USA: Association for Computing Machinery, 2014, pp. 236–244. ISBN: 9781450329019. DOI: 10.1145/2612733.2612773.

[31] LongJason Lu and Minlu Zhang. "Edge Betweenness Centrality". In: *Encyclopedia of Systems Biology*. Ed. by Werner Dubitzky et al. New York, NY: Springer New York, 2013, pp. 647–648. ISBN: 978-1-4419-9863-7. DOI: 10.1007/978-1-4419-9863-7_874.

[32] David Mair. "Westgate: A Case Study: How al-Shabaab used Twitter during an Ongoing Attack". In: *Studies in Conflict  Terrorism* 40 (Apr. 2016), pp. 1–20. DOI: 10.1080/1057610X.2016.1157404.

[33] Colin McDiarmid and Fiona Skerman. *Modularity of Erdős-Rényi Random Graphs*. 2018. arXiv: 1808.02243 [math.CO].

[34] Richard J. Medford et al. "An "Infodemic": Leveraging High-Volume Twitter Data to Understand Early Public Sentiment for the Coronavirus Disease 2019 Outbreak". In: *Open Forum Infectious Diseases* 7.7 (June 2020), ofaa258. ISSN: 2328-8957. DOI: 10.1093/ofid/ofaa258.

[35] A. J. Morales et al. "Measuring Political Polarization: Twitter Shows The Two Sides of Venezuela". In: *Chaos: An Interdisciplinary Journal of Nonlinear Science* 25.3 (2015). ISSN: 1089-7682. DOI: 10.1063/1.4913758.

[36] M. E. J. Newman. "Finding Community Structure in Networks Using the Eigenvectors of Matrices". In: *Physical Review E* 74.3 (Sept. 2006). ISSN: 1550-2376. DOI: 10.1103/physreve.74.036104.

[37] Nam P. Nguyen et al. "Dynamic Social Community Detection and Its Applications". In: *PLOS ONE* 9.4 (Apr. 2014), pp. 1–18. DOI: 10.1371/journal.pone.0091431.

[38] Jelili Oyelade, Olufunke Oladipupo, and Ibidun Obagbuwa. "Application of k Means Clustering algorithm for prediction of Students Academic Performance". In: *International Journal of Computer Science and Information Security* 7 (Feb. 2010).

[39] Mert Ozer, Nyunsu Kim, and Hasan Davulcu. "Community Detection in Political Twitter Networks Using Nonnegative Matrix Factorization Methods". In: *Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2016*. 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2016 ; Conference date: 18-08-2016 Through 21-08-2016. United States: Institute of Electrical and Electronics Engineers Inc., Nov. 2016, pp. 81–88. DOI: 10.1109/ASONAM.2016.7752217.

[40] Symeon Papadopoulos et al. "Image Clustering Through Community Detection on Hybrid Image Similarity Graphs". In: *2010 IEEE International Conference on Image Processing*. Nov. 2010, pp. 2353–2356. DOI: 10.1109/ICIP.2010.5653478.

[41] Usha Nandini Raghavan, Réka Albert, and Soundar Kumara. "Near Linear Time Algorithm to Detect Community Structures in Large-Scale Networks". In: *Physical Review E* 76.3 (Sept. 2007). ISSN: 1550-2376. DOI: 10.1103/physreve.76.036106.

[42] L.P. Ramsey. "Brandjacking on Social Networks: Trademark Infringement by Impersonation of Markholders". In: *Buffalo Law Review*. 2010, p. 851.

[43] Sebastián A. Ríos and Ivan F. Videla–Cavieres. "Generating Groups of Products Using Graph Mining Techniques". In: *Procedia Computer Science* 35 (2014). Knowledge-Based and Intelligent Information Engineering Systems 18th Annual Conference, KES-2014 Gdynia, Poland, September 2014 Proceedings, pp. 730–738. ISSN: 1877-0509. DOI: https://doi.org/10.1016/j.procs.2014.08.155.

[44] M. Roy et al. "Ebola and Localized Blame on Social Media: Analysis of Twitter and Facebook Conversations During the 2014–2015 Ebola Epidemic". In: *Culture, Medicine, and Psychiatry* 44.1 (Mar. 2020), pp. 56–79. ISSN: 1573-076X. DOI: 10.1007/s11013-019-09635-8. URL: https://doi.org/10.1007/s11013-019-09635-8.

[45] Rafael Santiago and Luís C. Lamb. "Efficient Modularity Density Heuristics for Large Graphs". In: *European Journal of Operational Research* 258.3 (2017), pp. 844–865. ISSN: 0377-2217. DOI: https://doi.org/10.1016/j.ejor.2016.10.033.

[46] Satu Elisa Schaeffer. "Graph Clustering". In: *Computer Science Review* 1.1 (2007), pp. 27–64. ISSN: 1574-0137. DOI: https://doi.org/10.1016/j.cosrev.2007.05.001.

[47] R Schroeder, S. Everton, and R. Shepherd. "Mining Twitter Data from the Arab Spring". In: *Calhoun Faculty and Researchers' Publication*. 2012, pp. 54–64.

[48] Elena Claudia Serban, Alexandru Bogeanu, and Eugeniu Tudor. "Clustering Techniques In Financial Data Analysis Applications On The U.S. Financial Market". In: *Annals - Economy Series* 4 (Aug. 2013), pp. 176–194.

[49] Jianbo Shi and Jitendra Malik. "Normalized Cuts and Image Segmentation". In: *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE* 22 (2020).

[50] Sara Soliman, Maged El-Sayed, and Yasser Hassan. "Semantic Clustering of Search Engine Results". In: *The Scientific World Journal* 2015 (Dec. 2015), pp. 1–9. DOI: 10.1155/2015/931258.

[51] Tina Tomazic and Katja Mišič. "Parliament-Citizen Communication in Terms of Local Self-government and Their use of Social Media in the European Union". In: *Lex Localis* 17 (Oct. 2019), pp. 1057–1079. DOI: 10.4335/17.4.1057-1079(2019).

[52] V. A. Traag, L. Waltman, and N. J. van Eck. "From Louvain to Leiden: Guaranteeing Well-Connected Communities". In: *Scientific Reports* 9.1 (Mar. 2019), p. 5233. ISSN: 2045-2322. DOI: 10.1038/s41598-019-41695-z.

[53] Andranik Tumasjan et al. "Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment". In: vol. 10. Jan. 2010.

[54] Qisen Wang et al. "Academic Paper Recommendation Based on Community Detection in Citation-Collaboration Networks". In: vol. 9932. Sept. 2016, pp. 124–136. ISBN: 978-3-319-45816-8. DOI: 10.1007/978-3-319-45817-5_10.

[55] Xiaofeng Wang, Matthew Gerber, and Donald Brown. "Automatic Crime Prediction Using Events Extracted from Twitter Posts". In: Apr. 2012, pp. 231–238. DOI: 10.1007/978-3-642-29047-3_28.

[56] H. Watanabe, M. Bouazizi, and T. Ohtsuki. "Hate Speech on Twitter: A Pragmatic Approach to Collect Hateful and Offensive Expressions and Perform Hate Speech Detection". In: *IEEE Access* 6 (2018), pp. 13825–13835.

[57] Yan Xing et al. "A Node Influence Based Label Propagation Algorithm for Community Detection in Networks". In: *The Scientific World Journal* 2014 (June 2014), p. 627581. ISSN: 2356-6140. DOI: 10.1155/2014/627581.

[58] Bo Yang, Dayou Liu, and Jiming Liu. "Discovering Communities from Social Networks: Methodologies and Applications". In: *Handbook of Social Network Technologies and Applications*. Ed. by Borko Furht. Boston, MA: Springer US, 2010, pp. 331–346. ISBN: 978-1-4419-7142-5. DOI: 10.1007/978-1-4419-7142-5_16.

[59] Jaewon Yang, Julian McAuley, and Jure Leskovec. "Community Detection in Networks with Node Attributes". In: *2013 IEEE 13th International Conference on Data Mining* (Dec. 2013). DOI: 10.1109/icdm.2013.167.

[60] Ding Yanrui et al. "Identifying the Communities in the Metabolic Network Using 'Component' Definition and Girvan-Newman Algorithm". In: *Proceedings of the 2015 14th International Symposium on Distributed Computing and Applications for Business Engineering and Science (DCABES)*. USA: IEEE Computer Society, 2015, pp. 42–45. ISBN: 9781467365932. DOI: 10.1109/DCABES.2015.18.

[61] Ussama Yaqub et al. "Analysis and Visualization of Subjectivity and Polarity of Twitter Location Data". In: *Proceedings of the 19th Annual International Conference on Digital Government Research: Governance in the Data Age*. Delft, The Netherlands: Association for Computing Machinery, 2018. ISBN: 9781450365260. DOI: 10.1145/3209281.3209313. URL: https://doi.org/10.1145/3209281.3209313.

[62] Ussama Yaqub et al. "Analysis of Political Discourse on Twitter in the Context of the 2016 US Presidential Elections". In: *Government Information Quarterly* 34 (Nov. 2017). DOI: 10.1016/j.giq.2017.11.001.

[63] B. Yetkin. "2019 Yerel Seçimlerin Adayların Twitter Kullanımı". In: *Hacettepe Üniversitesi İletişim Fakültesi Kültürel Çalışmalar Dergisi* 6.2 (), pp. 382–405.

[64] Zeynep Zengin Alp and Şule Gündüz Öğüdücü. "Identifying Topical Influencers on Twitter Based on User Behavior and Network Topology". In: *Knowledge-Based Systems* 141 (2018), pp. 211–221. ISSN: 0950-7051. DOI: https://doi.org/10.1016/j.knosys.2017.11.021.

[65] Yun Zhang. "Cluster Analysis and Network Community Detection with Application to Neuroscience". Feb. 2017.

[66] Beldi Zohra and Malika Bessedik. "A New Brainstorming Based Algorithm for the Community Detection Problem". In: *2019 IEEE Congress on Evolutionary Computation (CEC)*. June 2019, pp. 2958–2965. DOI: 10.1109/CEC.2019.8789897.

# APPENDIX A
# DATA SET (DAILY)


Figure A.1: Information for displaying daily data

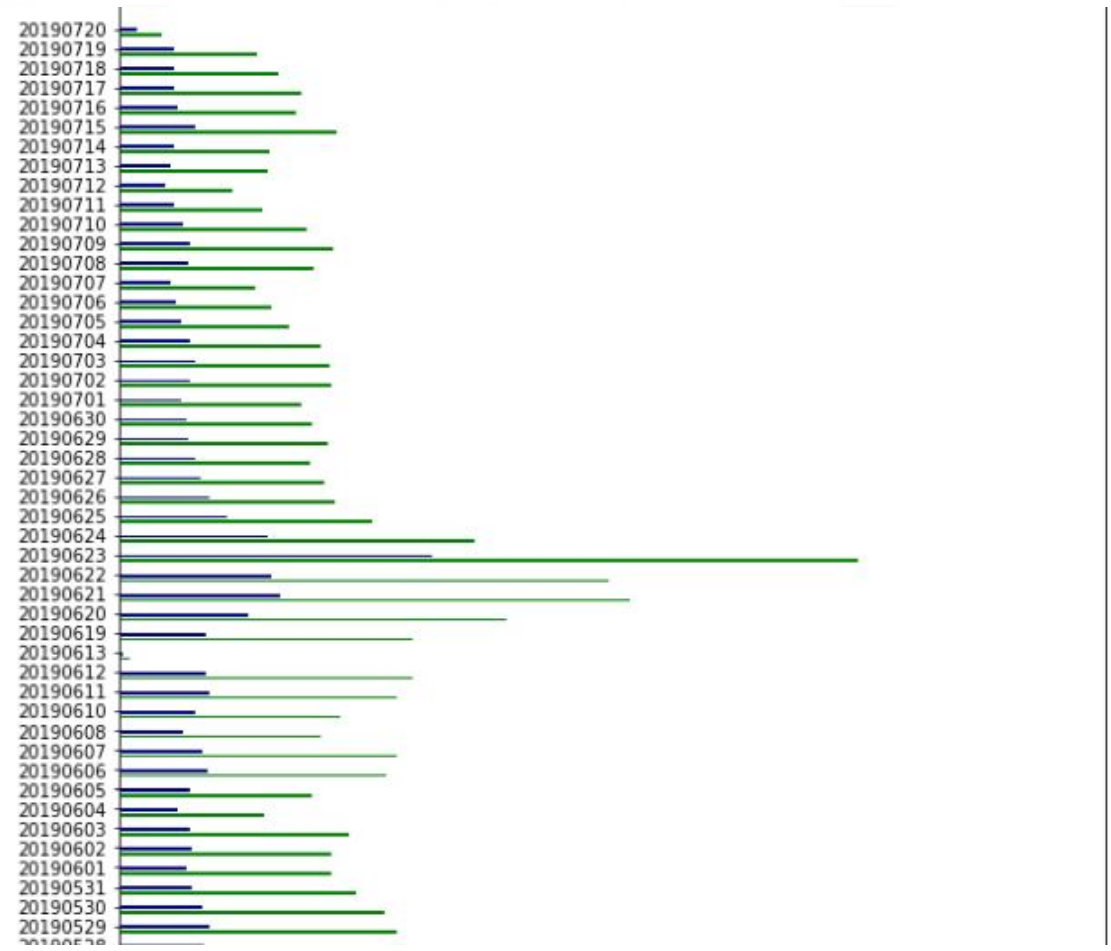


Figure A.2: 1st part of daily data

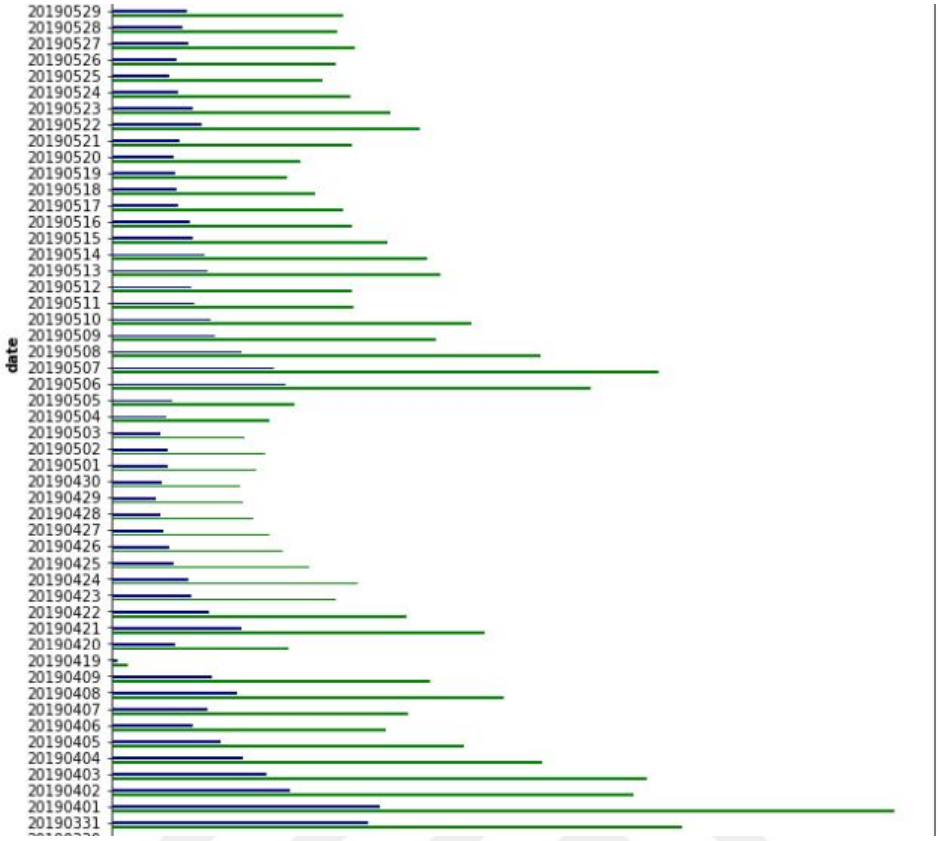


Figure A.3: 2nd part of daily data

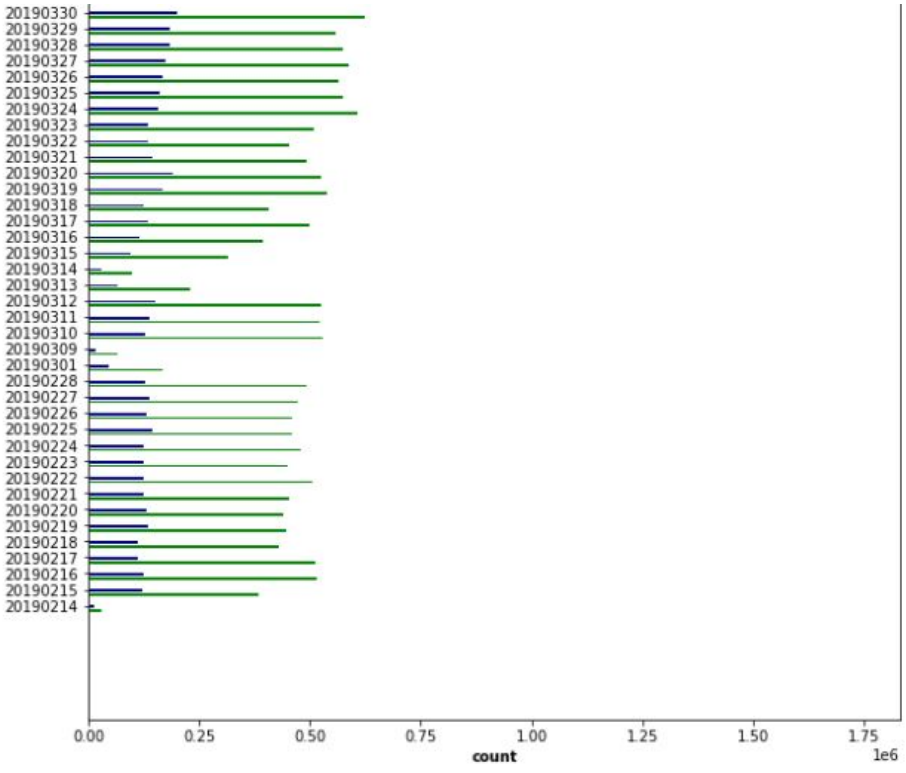


Figure A.4: 3rd part of daily data

# Selim Sürücü

## Personal Information


| | |
|---|---|
| Date of birth: | Çankırı,Çerkeş — April 26, 1993 |
| Adrress: | Yayla Mah. Yozgat Bul. No : 32 / 3, CP 06020, Keçiören, ANKARA |
| Mobile: | 5549443866 |
| email: | selimsurucuu@gmail.com |
| Driving License: | B ( 11.11.2015) |
| Orcid: | 0000-0002-8754-3846 |


## Work Experience

| | |
|---|---|
| Actual | Çankiri Karatekin University<br>*Researcher* |
| 01.02.2017 – 20.02.2020 | Keydata Bilgi İşlem Teknolojileri A.Ş.<br>*Software Developer* |
| 15.06.2016– 01.02.2017 | Nanotürk Yazilim<br>*Software Developer* |

## Education

2018-2020 Master of Computer Engineering, **Çankaya University**, ANKARA

2011 – 2016 Bachelor of **Computer Engineering**, Karabük University, KARABÜK

## Curses

2019 TensorFlow in Practice (COURSERA)
2019 Introduction to TensorFlow for Artificial Intelligence (COURSERA)
2019 NLP in Tensorflow (COURSERA)
2019 Introduction to TensorFlow for Artificial Intelligence (COURSERA)
2019 Sequences, Time Series and Prediction (COURSERA)

## Languages

English: B2 Level

## Made Worrk

Traffic Education Project with MS Kinect
Server Based Calorie Calculator Android Application
Smart Parking System
Child Education AR application (Unity- Vuforia)

## Interests and Extracurricular Activities

Watching movie, Reading books, Finance, Statistical, Games Theory, Programming, Sports, Soccer, Literature, Technology.