

T.R.
BOLU ABANT İZZET BAYSAL UNIVERSITY
INSTITUTE OF GRADUATE STUDIES
Department of Mathematics



ON TWITTER DATA ANALYSIS

MASTER OF SCIENCE

MUHAMMET KARADAĞ

ACADEMIC SUPERVISOR
Assistant Professor Sibel CANSU

BOLU, OCAK - 2023

APPROVAL OF THE THESIS

ON TWITTER DATA ANALYSIS submitted by **Muhammet KARADAĞ** and defended before the Examining Committee Members listed below in partial fulfillment of the requirements for the degree of **Master of Science** in **Department of Mathematics, Institute of Graduate Studies of Bolu Abant Izzet Baysal University** in **18.01.2023** by

Examining Committee Members

Signature

Supervisor

ASST.PROF. SİBEL CANSU

Bolu Abant Izzet Baysal University

.....

Member

PROF. DR. HALİM ÖZDEMİR

Sakarya University

.....

Member

ASST. PROF. UĞUR USTAOĞLU

Bolu Abant Izzet Baysal University

.....

Prof. Dr. İbrahim KÜRTÜL

Director of Institute of Graduate Studies

ETHICAL DECLARATION

In this thesis dissertation that was properly prepared according to the Thesis Writing Rules of Bolu Abant İzzet Baysal University of the Institute of Graduates Studies, I hereby declare that;

- All data, information, and documents presented in the thesis were obtained in accordance with the academic and ethical rules,
- All data, documents, assessments, and results were presented in accordance with the scientific ethical and moral rules,
- All works that were benefitted in the thesis were appropriately cited,
- No alteration was made in the data used,
- Study presented in this thesis is original,

Otherwise, I declare that I accept the loss of all my rights in case any contradiction that may arise against me.

Based on the plagiarism report that was generated on the date of ___31___/___01___/___2023___ by using predetermined filtrations set by Directorate of Institute of Graduate Studies of the Turnitin programme, a plagiarism detection software, the similarity index detected was ___12___ %.

Muhammet Karadağ

ABSTRACT

**ON TWITTER DATA ANALYSIS
MSC THESIS
MUHAMMET KARADAĞ
BOLU ABANT İZZET BAYSAL UNIVERSITY
INSTITUTE OF GRADUATE STUDIES
DEPARTMENT OF MATHEMATICS
(SUPERVISOR: ASST. PROF. SİBEL CANSU)**

BOLU, JANUARY 2023

ix + 35

With the development of technology, the internet has reached our pockets and innovations have come with it. Social media has become a very important media tool with the development of the internet. Since people from all over the world use this medium, it has become a huge data source. Twitter accounts for a large part of this data formation. With text mining, we can take the raw text from Twitter, preprocess it and get it structured. After taking the structured form of the text, it is passed to the sentiment analysis part. Sentiment analysis grades a text as positive, negative, or neutral based on the emotion given by the words. By looking at the text in general, we decide whether the positive aspect of this text is more weighted or the negative aspect of it. In this thesis, first text mining and then sentiment analysis of the data we pulled from Twitter were made. We can easily pull tweets under the name of Twitter's Covid19 and data mining hashtags by opening a developer account. The text mining of the data collected for about a month was performed, and then the sentiment analysis step was started and the sentiment scores of the data were examined. While performing text mining and sentiment analysis, Python software language and Jupyter Notebook were used. To conduct the research, lessons were taken from an online site and learned to a certain point. In the application part of the thesis, how we pull the data with Python coding, which date range is used, the codes we write for the preprocessing of the data are mentioned.

KEYWORDS:Data Mining, Sentiment Analysis, Twitter, Social Media

ÖZET

TWİTTER DATA ANALİZİ
YÜKSEK LİSANS TEZİ
MUHAMMET KARADAĞ
BOLU ABANT İZZET BAYSAL ÜNİVERSİTESİ
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ
MATEMATİK ANABİLİM DALI
(TEZ DANIŞMANI:DR. ÖĞR. ÜYESİ SİBEL CANSU)
BOLU, OCAK - 2023
ix + 35

Teknolojinin gelişmesi ile internet ceplerimize kadar inmiştir ve yeniliklerde beraberinde gelmiştir. Sosyal medya internetin gelişmesi ile çok önemli bir medya aracı olmuştur. Bu mecraı dünyanın her yerinden insanlar kullandıkları için büyük bir veri kaynağı olmuştur. Twitter bu veri oluşumunun büyük bir kısmını oluşturmaktadır. Metin madenciliği ile Twitter'daki işlenmemiş metinleri alıp ön işlemlerden geçirip yapılandırılmış halini alabiliriz. Metnin yapılandırılmış halini aldıktan sonra duygu analizi kısmına geçilir. Duygu analizi bir metnin kelimelerinin verdiği duyguya göre pozitif, negatif ya da nötr diye puanlama yapar. Metnin geneline bakarak bu metnin genel olarak pozitif yönü mü daha ağırlıklı yoksa negatif yönü mü buna karar veririz. Bu tezde Twitter'dan çektiğimiz verilerin önce metin madenciliği daha sonra duygu analizi yapılmıştır. Twitter'ın Covid19 ve datamining hashtagleri adı altındaki tweetleri developer hesabı açılarak rahatlıkla çekilebilir. Yaklaşık bir ay boyunca çekilen verilerin metin madenciliği yapıldı ve daha sonra duygu analizi basamağına geçildi ve verilerin duygu puanlarına bakıldı. Metin madenciliği ve duygu analizi yapılırken Python yazılım dilinden ve Jupyter Notebook'tan yararlanıldı. Tezin uygulama bölümünde Python kodlamaları ile veriyi nasıl çektiğimizden, hangi tarih aralığının kullanıldığından, verilerin ön işleme için yazdığımız kodlardan bahsedilmiştir.

ANAHTAR KELİMELELER: Veri Madenciliği, Duygu analizi, Twitter, Sosyal Medya

TABLE OF CONTENTS

	Page
APPROVAL OF THE THESIS	iii
ETHICAL DECLARATION.....	iv
ABSTRACT	v
ÖZET	vi
TABLE OF CONTENTS	vii
LIST OF FIGURES.....	viii
LIST OF TABLES.....	ix
LIST OF PICTURES	ix
1. INTRODUCTION	1
2. AIM AND SCOPE OF THE STUDY.....	3
3. TEXT MINING.....	4
3.1 Areas Of Text Mining.....	4
3.2 Text Mining And Preprocessing.....	5
4. SENTIMENT ANALYSIS	6
4.1 Document Level Sentiment Analysis	6
4.2 Sentence Level Sentiment Analysis.....	6
4.3 Aspect Level Sentiment Analysis	6
5. TWITTER.....	8
5.1 Hashtag	8
5.2 Trend Topic	8
5.3 Retweet	9
5.4 Mention.....	9
5.5 Twitter Developer Account	9
5.6 Data Extraction From Twitter.....	11
6. MATERIALS AND METHODS.....	13
6.1 Covid19 Data Set.....	13
6.1.1 Sentiment Analysis Of Covid19	18
6.2 Datamining Data Set.....	21
6.2.1 Sentiment Analysis OfDatamining.....	26
7. CONCLUSION	30
8. REFERENCES	31
9. APPENDIX.....	32

LIST OF FIGURES

Figures6.1. The Figures of the word with the highest frequencies.	17
Figures6.2. Figures of sentiment analysis.	19
Figures6.3. Tweet times.	20
Figures6.4. Figures of days	21
Figures6.5. The Figures of the word with the highest frequencies.	25
Figures6.6. Sentiment classes of data set by numbers.	27
Figures6.7. Tweet times.	28
Figures6.8. Tweet days.....	29

LIST OF TABLES

Table 8.1. First five lines of the data set.....	14
Table 8.2. Text column of the data set.	14
Table 8.3. First five lines of the data set after preprocessing.	14
Table 8.4. Text column of the data set after preprocessing.....	15
Table 8.6. First five lines of the most favorited tweets.	16
Table 8.7. Word frequencies of tweets.....	16
Table8.8. Frequency one standard deviation away from the mean.	17
Table8.10. Sentiment classes of the tweets.	18
Table8.11. Text, sentiment score and sentiment class.....	19
Table8.12. Tweet time and days.....	20
Table8.13. First five lines of the data set.....	22
Table8.14. Text column of the data set.	22
Table8.15. First five lines of the data set after preprocessing.	22
Table8.16. Text column of the data set after preprocessing.....	23
Table8.17. First five lines of the most favorited tweets.	24
Table8.18. Word frequencies of tweets.....	24
Table8.19. Frequency one standard deviation away from the mean.	25
Table8.21. Sentiment classes of the tweets.	26
Table8.22. Text, sentiment score and sentiment class.....	27
Table8.23. Tweet time and days.....	28

LIST OF PICTURES

Picture 7.1. First step of opening account.....	99
Picture 7.2. Second step of opening account	10
Picture 7.3. Third step of opening account	10
Picture 7.4. Confirmation e-mail.....	11
Picture 8.1. Users who tweet more than once.	15
Picture 8.2. Wordcloud of the most used words.	18
Picture 8.3. Wordcloud of the most used words.	26

1. INTRODUCTION

The development and change of technology brings innovations in all areas of life. Especially in the internet environment, there is a constant change. With the use of social media too much, a huge amount of data is created on the internet. Social media are platforms and blogs where billions of users share their ideas, complaints and comments today. This change in the internet has also caused changes in data mining.

Data mining uses structured data. The enormous increase in data on social media and the internet has shown that unstructured data needs to be analyzed. Text mining emerged in this process and has been one of the most researched areas of recent years.¹

One of the areas where text mining is used is sentiment analysis. Sentiment analysis can be used to reveal the sentiment of a large text book, short text, or tweet. Everyone from celebrities to politicians who want to follow the public is aware of the importance of sentiment analysis. There is little information on sentiment analysis on Twitter. The purpose of writing this thesis is to contribute to the elimination of this deficiency in the literature.

The development of the internet and technology today has brought us social media. Social media is a form of internet that has emerged as an alternative media tool. Especially during the pandemic period, people had to use social media too much because they could not go out. People who use social media and smartphones can easily become content producers. The use of social media by people from past to present has attracted the attention of researchers and there are many studies on social media. Therefore, there are many studies in the literature. However, there is not enough information in the field of sentiment analysis on social media. Most of the studies are on English texts as they are easier to apply. Sentiment analysis of Turkish texts is not very common.

In a way, this thesis consists of four main parts mainly. In the first part, information about text mining is given and information about the usage areas, steps and preprocessing of text mining are also given.

In the second part, information about sentiment analysis is given and an explanation is also given about the importance of emotion in texts. Then, the stages and usage areas of sentiment analysis are explained. In these explanations, the relevance of the stages of sentiment analysis to each other is mentioned.

In the third part, detailed information about Twitter, which is one of the social media tools, is given. Since the application area of this thesis is Twitter, it is written under the main title. It has been mentioned that Twitter is an important social media tool to follow daily life. In addition, since a Twitter developer account must be opened to extract data, a detailed explanation is given in the subtitle about how it is opened. After the developer account is opened, it is explained how to extract data from Twitter.

The application, which is the last part of the thesis, is the fourth main title. In the application chapter the application part, the data taken from Twitter has been processed. The tweets posted under the hashtags Covid19 and datamining were taken for about 1 month. Then all the extracted data were processed.

2. AIM AND SCOPE OF THE STUDY

Large sizes and variety of data are produced on the Internet and social media, and some of the content produced is visual and sensory, but only a small percentage. It is the texts that have the main percentile. These texts are very important, they can be analyzed.

The texts processed in the text analysis area are first pre-processed and simplified. Text analysis is related to sentiment analysis because the processed data is analyzed by sentiment analysis and a result can be found. However, sentiment analysis cannot be done without data processing. This is because there are unnecessary words or characters in the text, such as unnecessary emojis, misspelled words, so an accurate result cannot be obtained if sentiment analysis is performed without data processing in the texts.

Most of the texts on social media and the internet are in English. Therefore, researches from the past to the present have generally been made on texts written in English. In other words, there are not many studies on the texts written in Turkish in the literature. Another purpose of writing the thesis is that text mining and sentiment analysis researches by pulling data from Twitter do not have enough research in the literature.

In this thesis, the texts under the Covid19 and datamining hashtags were taken. The steps of preprocessing, tokenization, stemming and stopwords, which are the stages of text mining, have been applied to the simplest form. After the text mining was finished, the sentiment analysis part was started and it was examined whether the Covid19 and datamining texts had positive emotions, negative emotions or neutral emotions.

3. TEXT MINING

In recent years, with the development of technology, the internet has become one of the indispensable things that people use constantly. The fact that people, institutions and organizations share their ideas has made social media a great source of information. Unfortunately, data on social media is very difficult to process as it is. These data often appear to consist of misspelled words, abbreviations, and slang or nonsense words that are not in everyday speech.²

For this reason, the data in social media needs to be filtered in order to be processed and it needs to be converted into a proper form using text mining methods.

Text mining targets studies such as classification, clustering, text summary and sentiment analysis of texts.³

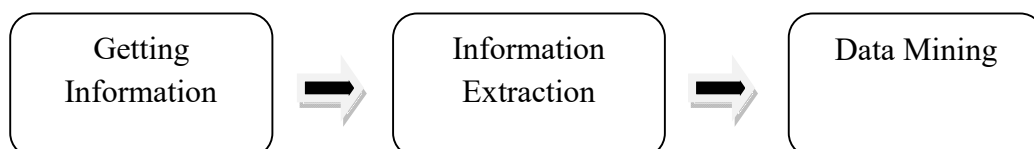
3.1 Areas Of Text Mining

The cost of using data mining is quite high. It happens with many applications. At the same time, there are many companies that make and sell ready-made software for the data mining application area. For this reason, desired data mining applications can be made at attractive prices through shopping. This eliminates the difference between small business and large business and allows small businesses to benefit from such services, allowing them to gain a competitive advantage alongside the big ones. There are many fields that use data mining. If we give an example of some areas where they are mostly used;

- Sales and Marketing
- Education
- Production and planning
- Accommodation
- Banking: Credit approval

Stages of text mining

There are three stages of text mining. These are divided into information retrieval, information extraction and data mining.⁴



3.2 Text Mining And Preprocessing

The data taken from the internet by any method is not structured. In order for these data to be used, they are converted into a smoother form using text mining methods. The following steps are followed while preprocessing the texts;

Normalization

Tokenization

Cleaning small words of three letters or less

Removing punctuation marks in text

Removing numbers

Lemmatization

In the normalization step, we aim to correct misspelled words and convert all text to lowercase. Unnecessary spaces at the beginning, end or even inside the sentence are removed.

In tokenization, we split our text into segments based on its targeted properties. By typing a different command, unnecessary characters, numbers and words such as and, which are frequently used in the language, are removed.

In the Lemmatization step, the words are separated to their roots and remain in their simplest form.³

4. SENTIMENT ANALYSIS

According to the Turkish Language Association, emotion is defined as "the ability to evaluate objects or events in terms of moral and aesthetic" and "the impression that certain objects, events or individuals create in the inner universe of man". As it is said in the definition, the concept of emotion includes negative, positive and neutral elements. For example, words such as love, affection, interest, fun, loyalty, hope can be positively charged, while words such as fight, hate, war can be negatively charged.⁵

The Internet has become a tool that we use in almost every field in our daily lives, and people share their feelings, thoughts and ideas in virtual environments such as social media, blogs, forums. Therefore, there is a large amount of data on people's emotions on the internet.⁶

4.1 Document Level Sentiment Analysis

Sentiment analysis is examined under three main headings. These are called document level, sentence level and aspect level.

It categorizes positive or negative emotions by obtaining a general idea without going into too much detail at the document level. The text is evaluated as a whole and classified as positive or negative, whichever opinion is more weighty. At this level sentiment analysis yields a result and is not used for multiple comparisons.⁷

4.2 Sentence Level Sentiment Analysis

The sentiment analysis at the document level is insufficient for some data because all the sentences at the document level sentiment analysis are classified according to the dominant sentiment score in the whole data without going into the details of the sentiment score. Some of the comments made by people have a high positive score, some have a high negative score, and some sentences may be neutral. In sentence-level sentiment analysis, we first check whether there is any emotion in each sentence, not as a whole. If emotion is expressed, our data is classified as positive or negative. Research on whether to express emotion is called subjectivity classification.⁶

4.3 Aspect Level Sentiment Analysis

Doing sentiment analysis at the document level causes emotional expressions to be ambiguous. When a person comments on an entity, even if

document-level sentiment analysis classifies it as positive, it does not mean that the person has made a completely positive comment about that entity. For example, if a commenter has made a positive comment about 5 characteristics of a person or entity and a negative comment about 7 properties, that comment is called negative at the document level. As can be seen from this example, there is a loss of information here. The person made a positive comment about the 5 features of the asset, but this information did not emerge at the document level.

In order to get the best information about the interpretation, it is important to know which positive and negative features of the asset the interpretation is made for. This is done in aspect level sentiment analysis. It can find the properties of the target entity. This process is called feature extraction.⁷

5. TWITTER

Twitter is an American microblogging and social service where users interact with each other and post messages called "tweets". Founded by Jack Dorsey in March 2006, Twitter has limited its members to a maximum of 280 characters when tweeting. (Until 2017, the character limit was 140, then with new updates it became 280 characters.). It offers the opportunity to share and send direct messages from private.⁸

Users of Twitter can become a free member and create their own profiles and communicate with other profiles. The site, which has a very simple interface, is attractive because it can be used easily. Twitter has about 50 different language options. Those who want to become a member must first determine their username when creating their profile. If the selected username was previously taken by another user, the new member should choose another username. Your username is important for other members to find you. Moreover, the @ symbol is automatically preceded by the usernames of all members. Members can change their chosen username at any time.¹

5.1 Hashtag

A user of the Twitter platform uses hashtags when tweeting to express their feelings and thoughts about a particular topic. The main purpose of the hashtag is to gather people around a certain topic. The hashtag is denoted by the # symbol. Users who want to use hashtags simply put the # symbol at the beginning of a word or sentence. Then, Twitter converts the word with this hashtag into a link, and if other users want to see what is written under this hashtag, they just need to enter this link.⁹

5.2 Trend Topic

If there are too many tweets about a topic during the day, Twitter pins these tweets as a trend topic (TT) to the top of the search section and makes it easy for other users to see.

Usually trending topics have words with hashtags. The more tweets are tweeted for that hashtag, the higher the ranking will be, so users who do not want to stay behind the agenda can follow the current issues of daily life by looking at the trending topics. If a user's tweet is favored or retweeted more than other users on trending topics, Twitter will show those tweets first.¹⁰

5.3 Retweet

A retweet (RT) is the sharing of any tweet posted by another user on their own page without any changes. Retweeted tweets are shared with the name of the user who first shared the tweet.¹¹

5.4 Mention

When mentioning a user on Twitter, the @ symbol is placed in front of the username and a space is placed, then the tweet is written and sent to the user. In this way, as it is shown who is mentioned in the tweet, the person mentioned will receive a notification and can reply to the sent tweet in the same way.¹²

When two people reply to each other, only the relevant people, such as the respondent's followers and the person in the chat, will see the reply in their own timeline.¹²

5.5 Twitter Developer Account

Normally a profile on apps.twitter.com was required to use the Twitter Search API, but now you need to create an individual developer account on developer.twitter.com.

In the first part, you are asked for what purpose you will use the Twitter developer tools. Since we will do research, we can go to the next step by choosing the Academic research section.

What is your primary reason for using Twitter developer tools?

We'll help you on your path to getting the most out of Twitter APIs and data.

Professional ...for commercial uses	Hobbyist ...for a personal project	Academic ...for education or research	Other I don't fit any of those
☐ Building B2B products	☐ Making a bot	☒ Doing academic research	☐ Embedding Tweets on a website
☐ Building consumer products	☐ Building tools for Twitter users	☐ Teaching	☐ Doing something else
☐ Build customized solutions in-house	☐ Exploring the API	☐ Student	
☐ Publishing ads programmatically			

Next

Picture5.1. First step of opening account

The second part asks if you are really the right person to associate the developer account you created with your personal twitter account. Here, we choose our name and the country we are in and go to the next page.

This is you, right?

 This @username will be the login for your developer account.

[Switch @username](#)
[Create new @username](#)

Individual developer account
[Switch to a team developer account](#)

You are signing up for an individual developer account. ⓘ

 We'll send important communications about your account to this email. ⓘ

[Change email address](#)

What country do you live in? Turkey

What would you like us to call you?

[Back](#) [Next](#)

Picture5.2. Second step of opening account

In the third part, we are asked for what purpose we will use the Twitter API. They want us to write in detail what we will research about and where we will use it with our own sentences.

How will you use the Twitter API or Twitter data?

In your words

In English, please describe how you plan to use Twitter data and/or APIs. For students and teachers, please include the name of the school, the name of the instructor and the course number (if available). The more detailed the response, the easier it is to review and approve.

I would like to use Twitter's Search API for collecting developer salary tweets that are hashtagged with #KnowYourWorth. And extracting this info to google spreadsheets for grouping each salary info based on education, experience, job title, location and given salary amount.

Response must be at least 200 characters

The specifics

Please answer each of the following with as much detail and accuracy as possible. Failure to do so could result in delays to your access to the Twitter developer platform or rejected applications.

Are you planning to analyze Twitter data? ☒ Yes

Please describe how you will analyze Twitter data including any analysis of Tweets or Twitter users.

I will use Twitter Search API with twurl CLI tool and download json info and extract some elements with jq CLI tool.

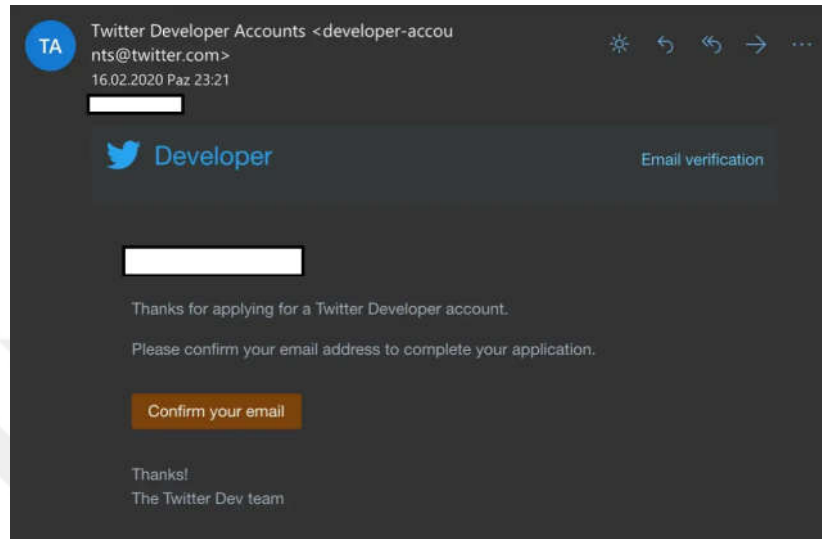
Response must be at least 200 characters

Picture5.3. Third step of opening account

In the fourth part, we are asked to review the information we have filled in.

In the fifth part, you have to read and accept the developer agreement.

In the sixth part, a confirmation e-mail is sent to the e-mail address associated with your twitter account.



Picture5.4. Confirmation e-mail

After clicking the link in the message sent to your e-mail address, there is a confirmation period for your account to be opened. Responses are made via e-mail during the day. After approval, we can create our application via <https://developer.twitter.com/en/apps>.¹⁴

5.6 Data Extraction From Twitter

There are tools necessary to extract data from Twitter. These;

- Twitter API
- Python
- Jupyter Notebook

The interface used to extract data from Twitter is Twitter Api.

The programming language used for data analysis, processing and data visualization is Python.

The name of the program used to analyze the texts taken from Twitter and run Python codes is Jupyter notebook.¹⁵

Libraries used;

- Tweepy (Twitter API)
- Pandas (Data analysis tool)
- Matplotlib (Data visualization)
- Seaborn (Data visualization)
- WordCloud (Data visualization)
- TextBlob (Text data processing)
- NLTK (Natural language processing package)
- Scikit-learn (Artificial learning)



6. MATERIALS AND METHODS

When starting the text mining phase, we need to have a dataset. The data set in this thesis is taken from Twitter with hashtagsCovid19 and datamining. The time interval of tweets is between 02.11.2022 and 27.11.2022, and a total of 2977 tweets were taken.

The data pulled from Twitter contains too many abbreviations, photos and emojis. Using the data without preprocessing will change the effectiveness of the result.

6.1 Covid19 Data Set

The data set containing Covid19 hashtag has nine columns which is shown below.

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1684 entries, 0 to 1683
Data columns (total 9 columns):
#   Column                Non-Null Count  Dtype
---  -
0   id                    1684 non-null  int64
1   text                  1684 non-null  object
2   created_at            1684 non-null  object
3   retweeted             1684 non-null  bool
4   retweet_count         1684 non-null  int64
5   user_screen_name      1684 non-null  object
6   user_followers_count  1684 non-null  int64
7   user_location         1152 non-null  object
8   Hashtags              1684 non-null  object
dtypes: bool(1), int64(3), object(5)
memory usage: 120.1+ KB
```

An example of unstructured data containing Covid19 hashtag before preprocessing is as follows. In the Python programming language, the head command displays the first five lines of the dataset. So the first five lines are shown.

	id	text	created_at	retweeted	retweet_count	user_screen_name	user_followers_count	user_location	Hashtags
0	1589442840031268864	Hey Elon Musk—remember when you said "Kids are...	2022-11-07 02:21:18+00:00	False	4350	DrEricDing	721979	Washington DC Virginia	['text': 'COVID19', 'indices': [100, 108]]
1	1589558667485261824	Nigeria is not failing because of "global" con...	2022-11-07 10:01:33+00:00	False	2694	MoghaluKingsley	271929	Nigeria	['text': 'COVID19', 'indices': [62, 70]]
2	1589524162904698880	12) Let's cross over the pond to UK—more kids...	2022-11-07 07:44:27+00:00	False	227	DrEricDing	721979	Washington DC Virginia	['text': 'COVID19', 'indices': [60, 68]]
3	1589765896327008257	RT @maduro_en: RT @CancilleriaVE: President @N...	2022-11-07 23:45:00+00:00	False	164	OliverF51985344	63	NaN	['text': 'COVID19', 'indices': [87, 95]]
4	1589765894112092160	New—The impact of the #COVID-19 pandemic on #d...	2022-11-07 23:45:00+00:00	False	0	TheLancetEndo	34456	London	['text': 'COVID', 'indices': [22, 28]], ['tex...

Table 6.1. First five lines of the data set.

Also the text column of the data set before preprocessing is below.

```

0      Hey Elon Musk—remember when you said "Kids are...
1      Nigeria is not failing because of "global" con...
2      12) Let's cross over the pond to UK— more kids...
3      RT @maduro_en: RT @CancilleriaVE: President @N...
4      New—The impact of the #COVID-19 pandemic on #d...

1479   RT @bankimooncentre: "Building on what we have...
1480   RT @Alexblx: @hugh_mankind WHAT'S in the #mRNA...
1481   In this review,the authors detail the #neurolo...
1482   DCPS requires students, staff to take #Covid19...
1483   RT @RizzaIslam: Well well well did you know th...
Name: text, Length: 1484, dtype: object

```

Table 6.2.Text column of the data set.

Punctuation marks and emojis in tweets have been removed in the preprocessing step. All tweets have been converted to lowercase. The numbers have been removed. Words that are included in the sentence but have no meaning for analysis, such as prepositions and conjunctions, which are called stopwords in English, have been removed from the tweet. The first five lines of the data set and the text column after preprocessing are given below.

	id	text	created_at	retweeted	retweet_count	user_screen_name	user_followers_count	user_location	Hashtags
0	1589442840031268864	hey elon muskremember said kid essentially imm...	2022-11-07 02:21:18+00:00	False	4350	DrEricDing	721979	Washington DC Virginia	['text': 'COVID19', 'indices': [100, 108]]
1	1589558667485261824	nigeria failing global condition covid global ...	2022-11-07 10:01:33+00:00	False	2694	MoghaluKingsley	271929	Nigeria	['text': 'COVID19', 'indices': [62, 70]]
2	1589524162904698880	let cross pond uk kid died covid england first...	2022-11-07 07:44:27+00:00	False	227	DrEricDing	721979	Washington DC Virginia	['text': 'COVID19', 'indices': [60, 68]]
3	1589765896327008257	maduro_en cancelleriave president nicolasmad...	2022-11-07 23:45:00+00:00	False	164	OliverF51985344	63	NaN	['text': 'COVID19', 'indices': [87, 95]]
4	1589765894112092160	newthe impact covid pandemic diabetes service ...	2022-11-07 23:45:00+00:00	False	0	TheLancetEndo	34456	London	['text': 'COVID', 'indices': [22, 28]], ['tex...

Table 6.3. First five lines of the data set after preprocessing.

```

0      hey elon muskremember said kid essentially imm...
1      nigeria failing global condition covid global ...
2      let cross pond uk kid died covid england first...
3      maduro_en cancelleriave president nicolasmad...
4      newthe impact covid pandemic diabetes service ...
...
1479   bankimooncentre building already learned must...
1480   alexblx hugh_mankind whats mrna vaccine since...
1481   reviewthe author detail neurological complicat...
1482   dcps requires student staff take covid test re...
1483   rizzaislam well well well know theyre getting...
Name: text, Length: 1484, dtype: object

```

Table 6.4.Text column of the data set after preprocessing.

In data set user_screen_name column represents the user name of the Twitter user. When we check the data set there are 1445 unique user screen name.

```

user_screen_name
11VanessaH      1
11thHourLib     1
16thStMALL      1
1_stlight      1
202hanyu        1
..
zhengyizhizhan1 3
zin_zah         1
zoecttrll       1
zoo_bear        1
zulaani         1
Name: id, Length: 1445, dtype: int64

```

The list of the users who have more than one tweet with Covid19 hashtag has 120 members which are:

```

['504_dr', 'ACHCCOVID', 'AVINASH27551234', 'AilsaForshaw', 'AugustadSays', 'B0tSci', 'BCDFoundation', 'BadjiYoussou1', 'Biscuit
sCat', 'BrianKiely2', 'BringBackIso', 'BuildnBlaze', 'BuildnBlazeTV', 'CDCgov', 'CGombatto', 'COSseaton', 'COVID_Scale', 'Covid
19Travel', 'CrispinTeddy', 'Crof', 'DaAnsahonSports', 'DanaGarrison', 'Deradevill', 'DrEricDing', 'DrJacobsRad', 'DureKhan3',
'ERN_Malleyscrub', 'EpochTimes', 'GothicLady7', 'GrandmasRus', 'Greg40800368', 'Grignwan', 'In_Phy_T', 'IsolationBot', 'JaniceB
72627796', 'KateEmerson88', 'Kazemakase13', 'KoparaFallsKid', 'LLynneIrwin', 'LaferriereTrina', 'LovelyMsNet', 'LukhmanM', 'Mar
iaMcElligott1', 'MariaSammumMasc', 'MatthewJLadd', 'MicrobesInfo', 'MissyCooperBro', 'MoghaluKingsley', 'MsAvaArmstrong', 'NIHR_
ARC_NENC', 'NYKKargil', 'PhilMarchand88', 'R5akanyemini', 'Rosex77388548', 'Rpen85815419', 'RuralHomeCareBG', 'SafetyFirst_
', 'ServerlessFan', 'Stephan90411631', 'Sue_Sullivan', 'TCGerry', 'TaNaishaJ', 'TalktoBeverley', 'Tammocanada', 'TheMemeticis
t', 'Tian_Ai', 'UNOPS', 'UofU31', 'Vasia_efst', 'WubneshGebreye7', 'Yomama2397', '_Jakashi', '_OfficialECI', '_marching_Ents',
'adriangherghe12', 'amaarali772969', 'ankitsa64753824', 'ashamedaustrali', 'beverleyturner', 'bharat_rising', 'bradboyqld', 'br
o06818129', 'catoubzh', 'chirontervs', 'deepgreendesign', 'doritmi', 'epstein_dan', 'erwinloh', 'gaudiumeng', 'grumman600', 'gu
stavokralj', 'ipattorneyliza', 'jazmasigan_2', 'kelly_cc21', 'krishna444_test', 'LCP4gGvNS9FuZI', 'lizditz', 'mamba_aka', 'moh
ammed2021mm7', 'myclothesbrush', 'n_khajazada', 'nilikm', 'nobody68260234', 'oldriccid', 'outbreak_india', 'picardonhealth', 'p
riyank1454202', 'rakan2yemini', 'sallbee', 'seanzeng1989', 'stevecrripe57', 'tyler_nedrud', 'valurank', 'veg_biryani', 'vera_t
enacious', 'vicmidnightsun', 'white21candy', 'wizn_21', 'xxmv2', 'zhengyizhizhan1']

```

Picture6.1. Users who tweet more than once.

By using count() function we can identify hashtag top contributors.

```

user_screen_name
LCP4gGvNS9FuZI      17
DrEricDing           14
adriangherghe12      12
_OfficialECI         11
deepgreendesign       8
COVID_Scale          7
CrispinTeddy          7
krishna444_test       6
BringBackIso         6
n_khajazada           5
Name: id, dtype: int64

```

When the tweets were examined, it was seen that there were 1546 tweets that did not start with RT. That's about 92 percent of the tweets in the dataset. The top five most favorited tweets are shown below.

	id	text	created_at	retweeted	retweet_count	user_screen_name	user_followers_count	user_location	Hashtags
1179	1594120597608763395	gavinnewsom average caseload positivity rate ...	2022-11-20 00:09:02+00:00	False	7979	JessicaMeeker17	618	Georgia, USA	[]
1428	1596065567739289600	kimdotcom watch documentary died suddenly nov...	2022-11-25 08:57:39+00:00	False	6427	zulaani	20797	NaN	[[{"text": 'DiedSuddenly', 'indices': [91, 104]]]
1574	1596623833624367105	chelsea_wats thread omicron surge system stra...	2022-11-26 21:56:00+00:00	False	5001	Meredi13821396	2	NaN	[[{"text": 'Omicron', 'indices': [29, 37]], {'t...
301	1589442840031268864	hey elon muskremember said kid essentially imm...	2022-11-07 02:21:18+00:00	False	4865	DrEricDing	721413	Washington DC Virginia	[[{"text": 'COVID19', 'indices': [100, 108]]]
396	1590108886350303233	dreericding hey elon muskremember said kid ess...	2022-11-08 22:27:55+00:00	False	4865	SwearyHaridan	10443	On Mastodon & Counter.Social	[[{"text": 'COVID19', 'indices': [116, 124]]]

Table 6.5. First five lines of the most favorited tweets.

When the frequencies of the words in the tweets are examined, the following data set is obtained.

	words	frequency
0	in	59.0
1	reform	7.0
2	preston	7.0
3	underway	7.0
4	party	8.0
...	...	...
5431	httpstcoemcnsry	1.0
5432	dcps	1.0
5433	httpstcojtcunbwx	1.0
5434	httpstcowxjqngg	1.0
5435	requires	1.0

5436 rows x 2 columns

Table 6.6. Word frequencies of tweets.

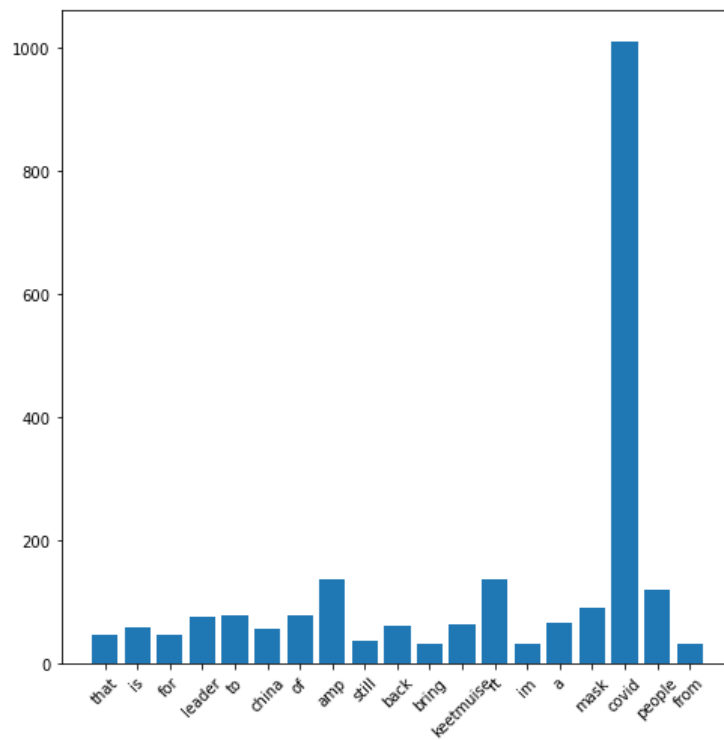
There are 106 words whose word frequency is bigger than one standard deviation away from the frequency mean.

	words	frequency
0	in	59.0
7	that	46.0
10	is	59.0
14	for	46.0
17	leader	75.0
...	...	...
2235	alyamani	28.0
2363	voice	54.0
2693	bin	50.0
2694	image	51.0
2695	salman	50.0

106 rows × 2 columns

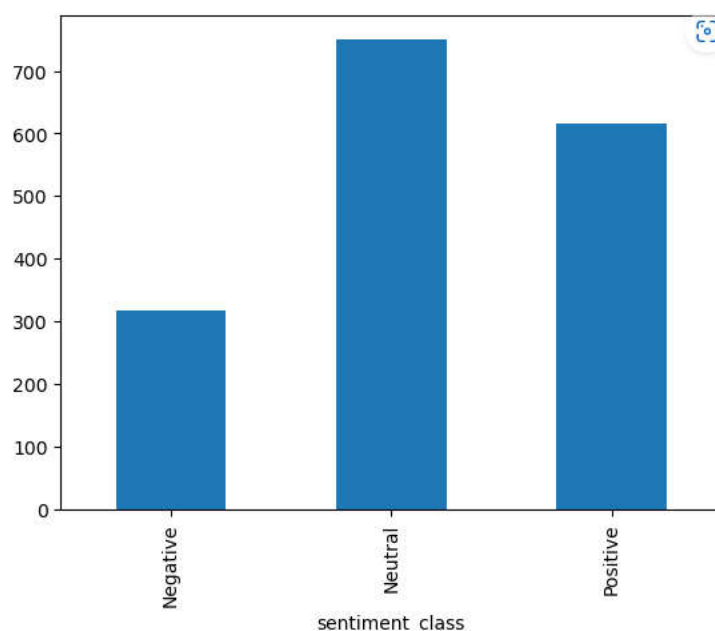
Table6.7. Frequency one standard deviation away from the mean.

The Figures of the words with high frequency in the data set containing the hashtag Covid19 is given below.



Figures6.1. The Figures of the word with the highest frequencies.

In data set there are 288 negative, 711 neutral and 685 positive tweets. Barplot of sentiment classes is shown below.



Figures6.2. Figures of sentiment analysis.

If we want to see tweet text, sentiment score and sentiment class together, a data set like the one below can be created. The first 25 rows of the obtained data set are given below.

	text	sentiment_score	sentiment_class
0	former reform party leader preston manning ann...	0.000000	Neutral
1	according to bennsteil only two major nations ...	0.031250	Positive
2	our neighbourhood paediatricians office answer...	0.000000	Neutral
3	rt keetmuise bring back masks bring back masks...	0.000000	Neutral
4	rt cpac_tv former reform party leader preston ...	0.000000	Neutral
5	rt mfwatches im quite happy for people to back...	0.400000	Positive
6	locked down chinese signing jie mi give me ric...	-0.077778	Negative
7	rt pfizer today we announced a phase study wi...	0.000000	Neutral
8	rt thelastperson under matt hancock there we...	0.500000	Positive
9	rt hervorrager covid is not harmless maybe you...	0.000000	Neutral
10	rt sachintaparia timesofindia amp others need ...	0.000000	Neutral
11	rt pfizer today we announced a phase study wi...	0.000000	Neutral
12	im very late to the covid party but it really ...	-0.147500	Negative
13	daisymay she can fro after almost years of li...	-0.500000	Negative
14	rt talkerteacher oh my heart as one of my stud...	0.000000	Neutral
15	rt keetmuise people being treated in their car...	0.000000	Neutral
16	rt drtedros shukran naser abduallah alhayen kuw...	0.000000	Neutral
17	today we announced a phase study with biontec...	0.000000	Neutral
18	rt uhccommunities chws are essential in rapidl...	0.000000	Neutral
19	rt cpac_tv former reform party leader preston ...	0.000000	Neutral
20	rt cpac_tv former reform party leader preston ...	0.000000	Neutral
21	rt talkerteacher oh my heart as one of my stud...	0.000000	Neutral
22	rt mohfw_india amritmahotsav unitefightcorona ...	0.000000	Neutral
23	rt drtedros shukran naser abduallah alhayen kuw...	0.000000	Neutral
24	rt talkerteacher oh my heart as one of my stud...	0.000000	Neutral

Table6.9. Text, sentiment score and sentiment class.

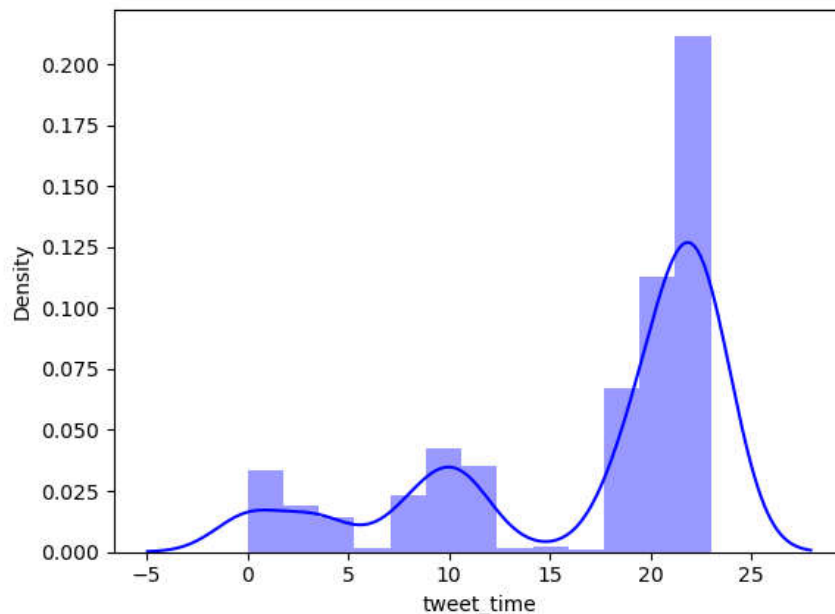
'according to bennsteil only two major nations russia amp china have avoided monetary belt tightening to stave of ' is an example of positive text, 'our neighbourhood paediatricians office answering machine says masks still required here some of us take preve ' is an example of neutral and 'locked down chinese signing jie mi give me rice jiemi covidisnotover givemerice jimmyjimmy china lockdown covid discodancer' is an example of negative text.

In order to examine which hours and days the most tweets were tweeted, the tweet time and the day the tweet was posted were added to the data set. In this context, the datetime library is used.

	id	text	created_at	sentiment_class	tweet_time	days
0	1587831675978805248	former reform party leader preston manning ann...	2022-11-02 15:39:06+00:00	Neutral	15	Wednesday
1	1587901948040093696	according to bennsteil only two major nations ...	2022-11-02 20:18:20+00:00	Positive	20	Wednesday
2	1587604429921984520	our neighbourhood paediatricians office answer...	2022-11-02 00:36:07+00:00	Neutral	0	Wednesday
3	158812655757895169	rt keetmuis bring back masks bring back masks...	2022-11-03 11:10:51+00:00	Neutral	11	Thursday
4	1588126552730271744	rt cpac_tv former reform party leader preston ...	2022-11-03 11:10:50+00:00	Neutral	11	Thursday
...	...	...	...	...	...	...
1679	1597076721408233474	bankimooncentre building already learned must...	2022-11-28 03:55:37+00:00	Neutral	3	Monday
1680	1597076692623060993	alexblx hugh_mankind whats mrna vaccine since...	2022-11-28 03:55:30+00:00	Positive	3	Monday
1681	1597076684045684736	reviewthe author detail neurological complicat...	2022-11-28 03:55:28+00:00	Neutral	3	Monday
1682	1597076682376155136	dcps requires student staff take covid test re...	2022-11-28 03:55:27+00:00	Neutral	3	Monday
1683	1597076669466292224	rizzaislam well well well know theyre getting...	2022-11-28 03:55:24+00:00	Positive	3	Monday

Table6.10. Tweet time and days.

The tweet time was visualized using the Seaborn library.

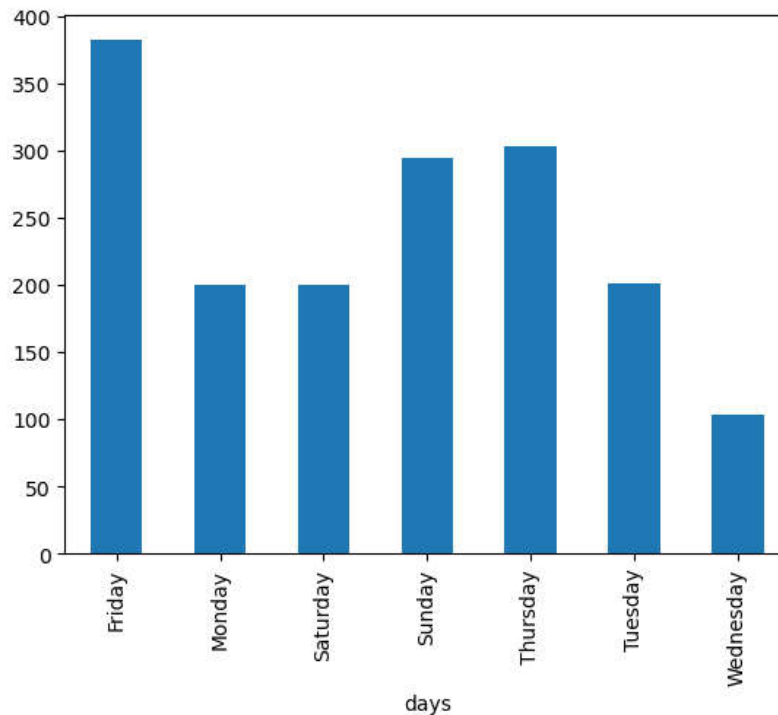


Figures6.3. Tweet times.

The distribution of days in the tweets is as follows.

```
days
Friday      382
Monday       200
Saturday     200
Sunday       294
Thursday     303
Tuesday      201
Wednesday    104
Name: id, dtype: int64
```

The barplot of the distribution of days will be



Figures6.4. Figures of days

6.2 Datamining Data Set

The data set containing datamining hashtag has nine columns which is shown below.

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1693 entries, 0 to 1692
Data columns (total 9 columns):
#   Column                Non-Null Count  Dtype
---  -
0   id                     1693 non-null   int64
1   text                   1693 non-null   object
2   created_at             1693 non-null   object
3   retweeted              1693 non-null   bool
4   retweet_count          1693 non-null   int64
5   user_screen_name       1693 non-null   object
6   user_followers_count   1693 non-null   int64
7   user_location          1200 non-null   object
8   Hashtags               1693 non-null   object
dtypes: bool(1), int64(3), object(5)
memory usage: 120.7+ KB
```

The first five lines of the data set before preprocessing is as follows.

	id	text	created_at	retweeted	retweet_count	user_screen_name	user_followers_count	user_location	Hashtags
0	1589465562346041345	*Online* eBook [open access; NO DOWNLOAD] — #M...	2022-11-07 03:51:35+00:00	False	88	KirkDBorne	369505	Maryland, USA	[{"text": "MachineLearning", "indices": [44, 6...
1	1589614601909731333	100+ Free #DataScience eBooks (downloadable PD...	2022-11-07 13:43:49+00:00	False	11	KirkDBorne	369505	Maryland, USA	[{"text": "DataScience", "indices": [10, 22]]]
2	1587679041553039361	2010 #DataMining and Knowledge Discovery Handb...	2022-11-02 05:32:35+00:00	False	49	KirkDBorne	369505	Maryland, USA	[{"text": "DataMining", "indices": [5, 16]]]
3	1589755982561021952	RT @HakomTimeSeries: RT:How to visualize the p...	2022-11-07 23:05:37+00:00	False	7	PythonBot_By_PM	1352	REMOTE	[{"text": "stats", "indices": [97, 103]}, {"te...
4	1589755961702490115	RT @HakomTimeSeries: RT:How to visualize the p...	2022-11-07 23:05:32+00:00	False	7	BotSci	4617	NaN	[{"text": "stats", "indices": [97, 103]}, {"te...

Table6.11. First five lines of the data set.

Also the text column of the data set before preprocessing is below.

```

0      *Online* eBook [open access; NO DOWNLOAD] — #M...
1      100+ Free #DataScience eBooks (downloadable PD...
2      2010 #DataMining and Knowledge Discovery Handb...
3      RT @HakomTimeSeries: RT:How to visualize the p...
4      RT @HakomTimeSeries: RT:How to visualize the p...
...
1488   RT @RosanaFerrero: 📄 R Reference Card for Data...
1489   CursorTek is #selling #bulk #data of #MessageT...
1490   Are you looking to fetch/scrape big data from ...
1491   RT @RosanaFerrero: 📄 R Reference Card for Data...
1492   CursorTek is #selling #bulk #data of #RealEsta...
Name: text, Length: 1493, dtype: object

```

Table6.12. Text column of the data set.

The first five rows of the dataset and the text column after preprocessing are given below.

	id	text	created_at	retweeted	retweet_count	user_screen_name	user_followers_count	user_location	Hashtags
0	1589465562346041345	online ebook open access download machinelearn...	2022-11-07 03:51:35+00:00	False	88	KirkDBorne	369505	Maryland, USA	[{"text": "MachineLearning", "indices": [44, 6...
1	1589614601909731333	free datascience ebooks downloadable pdfs begi...	2022-11-07 13:43:49+00:00	False	11	KirkDBorne	369505	Maryland, USA	[{"text": "DataScience", "indices": [10, 22]]]
2	1587679041553039361	datamining knowledge discovery handbook recent...	2022-11-02 05:32:35+00:00	False	49	KirkDBorne	369505	Maryland, USA	[{"text": "DataMining", "indices": [5, 16]]]
3	1589755982561021952	hakomtimeseries how visualize percentage vkan...	2022-11-07 23:05:37+00:00	False	7	PythonBot_By_PM	1352	REMOTE	[{"text": "stats", "indices": [97, 103]}, {"te...
4	1589755961702490115	hakomtimeseries how visualize percentage vkan...	2022-11-07 23:05:32+00:00	False	7	BotSci	4617	NaN	[{"text": "stats", "indices": [97, 103]}, {"te...

Table6.13. First five lines of the data set after preprocessing.

```

0      online ebook open access download machinelearn...
1      free datascience ebooks downloadable pdfs begi...
2      datamining knowledge discovery handbook recent...
3      hakomtimeseries how visualize percentage vkan...
4      hakomtimeseries how visualize percentage vkan...
...
1488   rosanaferreiro r reference card data mining ya...
1489   cursoek selling bulk data massagetherapy k plu...
1490   looking fetchscrape big data website excel dat...
1491   rosanaferreiro r reference card data mining ya...
1492   cursoek selling bulk data realestateagents k p...
Name: text, Length: 1493, dtype: object

```

Table6.14.Text column of the data set after preprocessing.

In data set `user_screen_name` column represents the user name of the Twitter user. When we check the data set there are 694 unique user screen name.

```

user_screen_name
100dayscode      54
16thStMALL        1
2007Alzate        1
202hanyu          1
20BlueCrowns     1
..
xuebinwei         1
yayayarndiva      1
yuanjianbo        1
zarishmakane      1
zs746196          1
Name: id, Length: 694, dtype: int64

```

By using `count()` function we can identify hashtag top contributors.

```

user_screen_name
ricardo_ik_ahau   126
cursortek         114
corpus_news       61
100dayscode      54
KirkDBorne        52
PythonBot_By_PM   44
godfrey_G         38
Analytics_699     38
o_fabriciosouza   30
JonTheDataMan     23
Name: id, dtype: int64

```

When the tweets were examined, it was seen that there were 1594 tweets that did not start with RT. That's about 94 percent of the tweets in the dataset. The top five most favorited tweets are shown below.

	id	text	created_at	retweeted	retweet_count	user_screen_name	user_followers_count	user_location	Hashtags
393	1589936769323057152	colleague across canada never seen many child ...	2022-11-08 11:04:00+00:00	False	840	nilikm	52701	Ottawa, Ontario	[[{"text": 'COVID19', 'indices': [93, 101]]]
434	1590492331886940160	billius new worldwide expe panel study nature...	2022-11-09 23:51:36+00:00	False	772	pjavidan	1372	Muwekma Ohlone land	[]
489	1590490285833756672	picardonhealth kid suffering covid way never ...	2022-11-09 23:43:28+00:00	False	650	Blackbeau5	930	United States	[[{"text": 'COVID19', 'indices': [44, 52]], {"t...
397	1590493864250351616	whiterose_ private person dislike social medi...	2022-11-09 23:57:41+00:00	False	613	horseplayer7	1625	New York	[]
467	1590490963562369024	thebma ahead government official inquiry hand...	2022-11-09 23:46:10+00:00	False	543	ListenVery	4243	NaN	[[{"text": 'COVID19', 'indices': [72, 80]]]

Table6.15. First five lines of the most favorited tweets.

When the frequencies of the words in the tweets are examined, the following data set is obtained.

	words	frequency
0		1228.0
1	httpstcoxiiocpidbx	28.0
2	bigdata	496.0
3	download	127.0
4	ebook	96.0
...	...	...
3080	exceldataentry	2.0
3081	roofing	1.0
3082	httpstcolinanpfyj	1.0
3083	httpstcohtbhsrh	1.0
3084	httpstcoaomnwvti	1.0

3085 rows × 2 columns

Table6.16. Word frequencies of tweets.

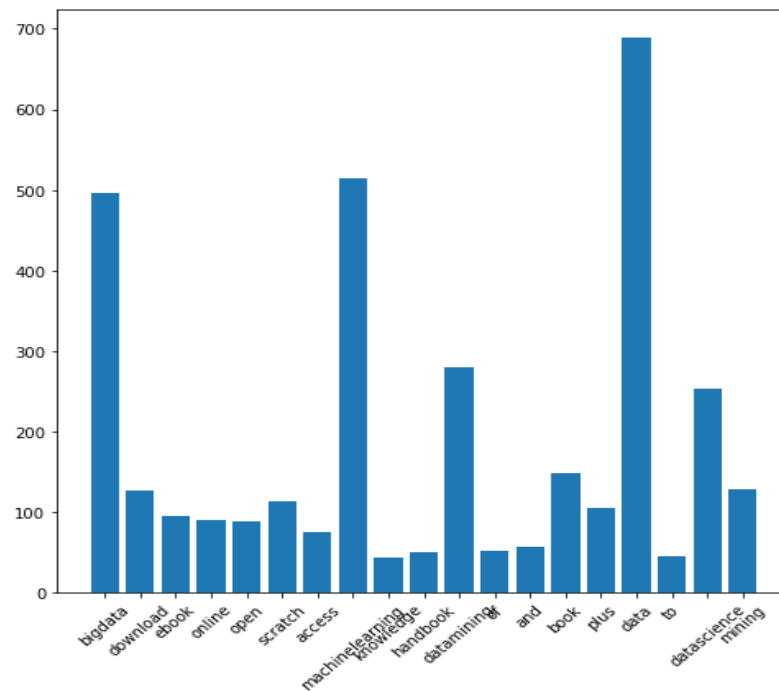
There are 64 words whose word frequency is bigger than one standard deviation away from the frequency mean.

	words	frequency
0		1228.0
2	bigdata	496.0
3	download	127.0
4	ebook	96.0
6	online	91.0
...	...	...
2396	bot	72.0
2414	httpstcodbekuxdgzm	60.0
2416	covering	62.0
2417	favorite	62.0
2418	machinelearnin	54.0

64 rows × 2 columns

Table6.17. Frequency one standard deviation away from the mean.

The Figures of the words with high frequency in the data set containing the hashtag datamining is given below.



Figures6.5. The Figures of the word with the highest frequencies.

A wordcloud of the most used words in tweets under the hashtag datamining has been prepared.



Picture6.3. Wordcloud of the most used words.

6.2.1 Sentiment Analysis OfDatamining

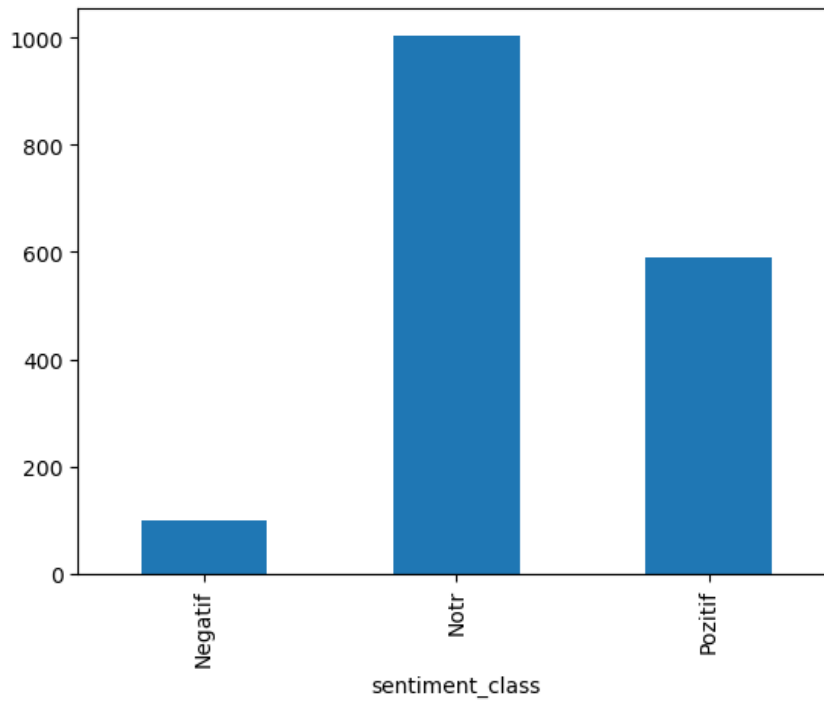
After adding sentiment class and sentiment score columns to the data set we have:

	id	text	created_at	sentiment_class
0	1587533760664580096	online ebook open access no download machine...	2022-11-01 19:55:18+00:00	Notr
1	1587679041553039361	datamining and knowledge discovery handbook ...	2022-11-02 05:32:35+00:00	Positif
2	1587534842799218688	data science from scratch plus more datascienc...	2022-11-01 19:59:36+00:00	Positif
3	1588121207790075906	dm to recover your gmail machinelearning from...	2022-11-03 10:49:36+00:00	Notr
4	1588118260360400897	design pandaspython dataprocessing databasesto...	2022-11-03 10:37:53+00:00	Notr
...	...	...	...	...
1688	1597064240682373120	rosanaferrero r reference card data mining ya...	2022-11-28 03:06:01+00:00	Notr
1689	1597062733291089920	cursoek selling bulk data massagetherapy k plu...	2022-11-28 03:00:02+00:00	Notr
1690	1597060409454125057	looking fetchscrape big data website excel dat...	2022-11-28 02:50:48+00:00	Notr
1691	1597059895332773888	rosanaferrero r reference card data mining ya...	2022-11-28 02:48:45+00:00	Notr
1692	1597054935102693383	cursoek selling bulk data realestateagents k p...	2022-11-28 02:29:03+00:00	Notr

1693 rows x 4 columns

Table6.18. Sentiment classes of the tweets.

In data set there are 100 negative, 1004 neutral and 589 positive tweets. Barplot of sentiment classes is shown below.



Figures6.6. Sentiment classes of data set by numbers.

The first 25 rows of the obtained data set are given below.

	text	sentiment_score	sentiment_class
0	online ebook open access no download machinel...	0.000000	Notr
1	datamining and knowledge discovery handbook ...	0.222222	Pozitif
2	data science from scratch plus more datascienc...	0.500000	Pozitif
3	dm to recover your gmail machinelearning from...	0.000000	Notr
4	design pandaspython dataprocessing databasesto...	0.000000	Notr
5	i will do any kind of data entry scraping mini...	0.600000	Pozitif
6	httpscobynunjtng ace the data science interv...	0.200000	Pozitif
7	checkout cursortek is selling bulk data of ins...	0.000000	Notr
8	rt corpus_news httpstcoovomewriqk datamining d...	0.000000	Notr
9	httpstcoovomewriqk datamining design decisionm...	0.000000	Notr
10	we offer professional datacleansing services t...	0.100000	Pozitif
11	rt corpus_news pythonprogramming databasestora...	0.000000	Notr
12	rt corpus_news pythonprogramming databasestora...	0.000000	Notr
13	rt corpus_news pythonprogramming databasestora...	0.000000	Notr
14	rt corpus_news pythonprogramming databasestora...	0.000000	Notr
15	rt corpus_news pythonprogramming databasestora...	0.000000	Notr
16	pythonprogramming databasestorage python dayso...	0.000000	Notr
17	rt bhavana what is computer science visit htt...	0.000000	Notr
18	rt bhavana what is computer science visit htt...	0.000000	Notr
19	what is computer science datamining via httpst...	0.000000	Notr
20	rt technative data governance amp bi steps to...	0.750000	Pozitif
21	rt technative data governance amp bi steps to...	0.750000	Pozitif
22	rt technative data governance amp bi steps to...	0.750000	Pozitif
23	rt pisquare_io pitalents success amp roi more ...	0.400000	Pozitif
24	rt kanezadiane data governanceampbi steps to s...	0.750000	Pozitif

Table6.19.Text, sentiment score and sentiment class.

'data science from scratch plus more datascience books to help you go deeper [httpstcomgakkjetq](#) [httpstcovffwxawivx](#)' is an example of positive text, 'online ebook open access no download machinelearning from scratch [httpstcoxiiocpldbx](#) bigdata [httpstcoodgcqlgaej](#)' is an example of neutral and 'cursortek is selling bulk data of massagetherapy and k plus other businesses catagoires like medicalspas [httpstcoqxofafirmwt](#)' is an example of negative text.

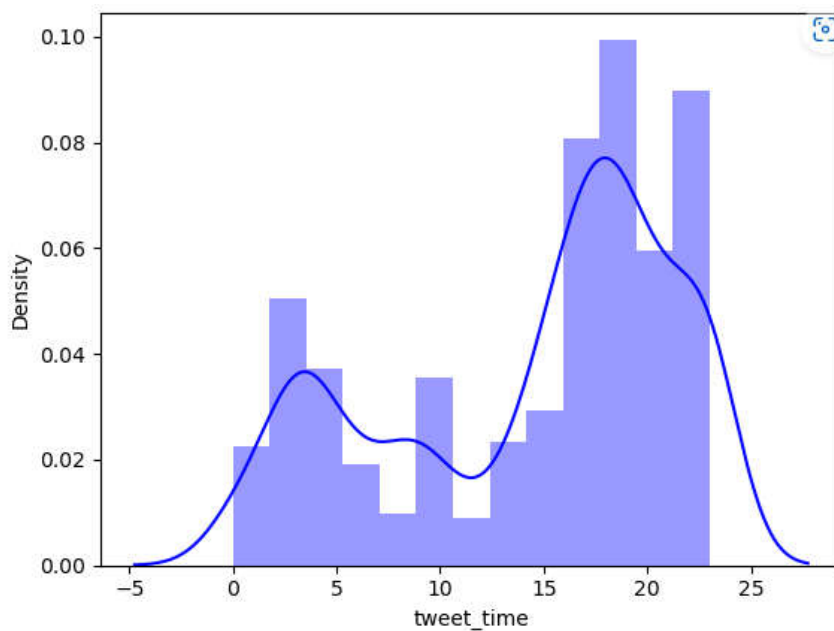
In order to examine which hours and days the most tweets were tweeted, the tweet time and the day the tweet was posted were added to the data set. In this context, the datetime library is used.

	id	text	created_at	sentiment_class	tweet_time	days
0	1587533760664580096	online ebook open access no download machinel...	2022-11-01 19:55:18+00:00	Notr	19	Tuesday
1	1587679041553039361	datamining and knowledge discovery handbook ...	2022-11-02 05:32:35+00:00	Pozitif	5	Wednesday
2	1587534842799218688	data science from scratch plus more datascienc...	2022-11-01 19:59:36+00:00	Pozitif	19	Tuesday
3	1588121207790075906	dm to recover your gmail machinelearning from...	2022-11-03 10:49:36+00:00	Notr	10	Thursday
4	1588118260360400897	design pandaspython dataprocessing databasesto...	2022-11-03 10:37:53+00:00	Notr	10	Thursday
...	...	...	...	...	...	...
1688	1597064240682373120	rosanaferrero r reference card data mining ya...	2022-11-28 03:06:01+00:00	Notr	3	Monday
1689	1597062733291089920	cursoek selling bulk data massagetherapy k plu...	2022-11-28 03:00:02+00:00	Notr	3	Monday
1690	1597060409454125057	looking fetchscrape big data website excel dat...	2022-11-28 02:50:48+00:00	Notr	2	Monday
1691	1597059895332773888	rosanaferrero r reference card data mining ya...	2022-11-28 02:48:45+00:00	Notr	2	Monday
1692	1597054935102693383	cursoek selling bulk data realestateagents k p...	2022-11-28 02:29:03+00:00	Notr	2	Monday

1693 rows x 6 columns

Table6.20. Tweet time and days.

The tweet time was visualized using the Seaborn library.

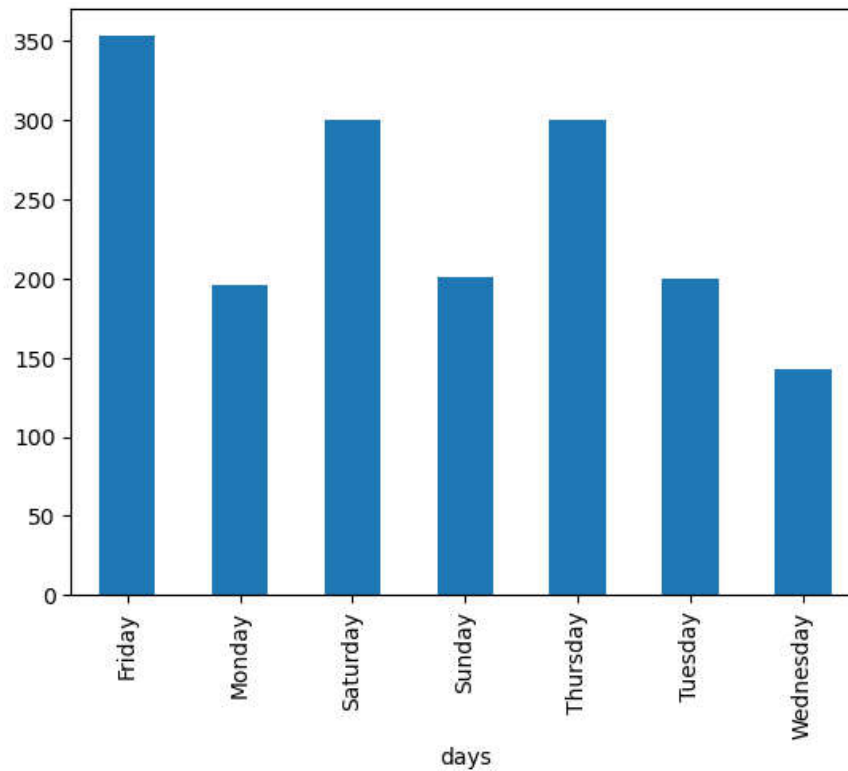


Figures6.7. Tweet times.

The distribution of days in the tweets is as follows.

```
days
Friday      353
Monday      196
Saturday     300
Sunday      201
Thursday     300
Tuesday     200
Wednesday   143
Name: id, dtype: int64
```

The barplot of the distribution of days will be



Figures6.8. Tweet days.

7. CONCLUSION

The purposes of text mining and sentiment analysis are to make an unstructured text structured and calculate the sentiment score. For this, a developer account was opened on Twitter. The tweets of Covid19 and datamining hashtags were pulled, and then the text mining phase was started. In text mining, tweets were tried to be brought into the simplest possible form, and then the sentiment analysis step was started.

In sentiment analysis, the positive, negative and neutral states of the texts were examined. The days and times of tweets were examined. The frequencies of the words were checked. In our study of the datamining hashtag, neutral sentences are more dominant, followed by positive sentences and finally negative sentences. Our ranking does not change in our study of the Covid 19 hashtag. The results related to the studies carried out are supported by Figuresics. Wordcloud of the most used words in the texts was created. Although certain hashtags have been researched in our study, it is possible to work with texts belonging to another hashtag. It can even be examined directly in the tweets of a specific twitter user without using hashtags.

8. REFERENCES

Vancouver citation system was used in this thesis.

1. Kızılkaya, Y. M. (2018). Duygu analizi ve sosyal medya alanında uygulama.
2. Akgül, E. S., Ertano, C., & Banu, D. İ. R. İ. (2016). Twitter verileri ile duygu analizi. Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi, 22(2), 106-110.
3. Kumaş, E. (2021). Türkçe Twitter Verilerinden Duygu Analizi Yapılırken Sınıflandırıcıların Karşılaştırılması. Eskişehir Türk Dünyası Uygulama ve Araştırma Merkezi Bilişim Dergisi, 2(2), 1-5.
4. Dolgun, M. Ö., Özdemir, T. G., & Oğuz, D. (2009). Veri madenciliğinde yapısal olmayan verinin analizi: Metin ve web madenciliği. İstatistikçiler Dergisi: İstatistik ve Aktüerya, 2(2), 48-58.
5. <https://sozluk.gov.tr/>
6. Özyurt, B., & Akçayol, M. A. (2018). Fikir Madenciliği ve Duygu Analizi, Yaklaşımlar, Yöntemler Üzerine Bir Araştırma. Selçuk Üniversitesi Mühendislik, Bilim ve Teknoloji Dergisi, 6(4), 668-693.
7. Ekim, H. E., & İner, A. B. (2021). Duygu Analizi ve Fikir Madenciliği Uygulamaları Üzerine Literatür Taraması. Kahramanmaraş Sütçü İmam Üniversitesi Mühendislik Bilimleri Dergisi, 24(2), 93-114.
8. Twitter'ın tarihi ve kurucusu.[Internet].[cited 2022 Dec 20]. Available from: <https://www.britannica.com/topic/Twitter>
9. Etiketlerin Tweetleri anahtar kelimelere göre kategorize etmekte kullanılması.[Internet].[cited 2022 Dec 20].Available from:<https://help.twitter.com/tr/using-twitter/how-to-use-hashtags#:~:text=Bir%20etiket%2C%20%23%20sembol%C3%BC%20ile%20yaz%C4%B1ld%C4%B1%20kolayca%20takip%20edebilmelerini%20sa%C4%9Flar>
10. [Internet].[cited 2022 Dec 22]. Available from: <https://wmaraci.com/nedir/trend-topic>
11. [Internet].[cited 2022 Dec 21]. Available from: <https://help.twitter.com/tr/using-twitter/retweet-faq#:~:text=Retweet%20nedir%3F,veya%20ba%C5%9Fka%20birinin%20Tweetlerini%20Retweetleyebilirsiniz>
12. [Internet].[cited 2022 Dec 17]. Available from: <https://help.twitter.com/tr/using-twitter/mentions-and-replies>
13. [Internet].[cited 2022 Dec 20]. Available from: <https://zaferayan.medium.com/twitter-search-api-ile-tweet-analizi-nas%C4%B1l-yap%C4%B1l%C4%B1r-88b28901d022>
14. Küçük kartal, H. K. (2020). Twitter'daki Verilere Metin Madenciliği Yöntemlerinin Uygulanması. Eskişehir Türk Dünyası Uygulama ve Araştırma Merkezi Bilişim Dergisi, 1(2), 10-13.

9. APPENDIX

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
```

Data Mining data set

```
data=pd.read_csv("7Kasım.csv")
```

```
sekiz=pd.read_csv("8Kasım.csv")
dokuz=pd.read_csv("9Kasım.csv")
on=pd.read_csv("10Kasım.csv")
onbir=pd.read_csv("11Kasım.csv")
oniki=pd.read_csv("12Kasım.csv")
onuc=pd.read_csv("13Kasım.csv")
onyedi=pd.read_csv("17Kasım.csv")
onsekiz=pd.read_csv("18Kasım.csv")
ondokuz=pd.read_csv("19Kasım.csv")
yirmi=pd.read_csv("20Kasım.csv")
yirmiki=pd.read_csv("22Kasım.csv")
yirmibes=pd.read_csv("25Kasım.csv")
yirmialti=pd.read_csv("26Kasım.csv")
yirmiyedi=pd.read_csv("27Kasım.csv")
```

```
df=pd.concat([data,sekiz,dokuz,on,onbir,oniki,onuc,onyedi,onsekiz,ondokuz,yirmi,yirmiki,yirmibes,yirmialti,yirmiyedi], axis=0)
```

```
df.info()
```

```
liste=[]
for i in range(1493):
    liste.append(i)
```

```
len(liste)
```

```
1493
```

```
df.index=liste
```

```
df.info()
```

Pre-Processing

```
!pip install -U textblob
```

```
Requirement already satisfied: textblob in c:\users\user\anaconda3\lib\site-packages (0.17.1)
Requirement already satisfied: nltk>=3.1 in c:\users\user\anaconda3\lib\site-packages (from textblob) (3.7)
Requirement already satisfied: regex>=2021.8.3 in c:\users\user\anaconda3\lib\site-packages (from nltk>=3.1->textblob) (2022.7.9)
Requirement already satisfied: click in c:\users\user\anaconda3\lib\site-packages (from nltk>=3.1->textblob) (8.0.4)
Requirement already satisfied: joblib in c:\users\user\anaconda3\lib\site-packages (from nltk>=3.1->textblob) (1.1.0)
Requirement already satisfied: tqdm in c:\users\user\anaconda3\lib\site-packages (from nltk>=3.1->textblob) (4.64.1)
Requirement already satisfied: colorama in c:\users\user\anaconda3\lib\site-packages (from click->nltk>=3.1->textblob) (0.4.5)
```

```
#big-small transformation
df['text'] = df['text'].apply(lambda x: " ".join(x.lower() for x in x.split()))

#punctuation
df['text'] = df['text'].str.replace('[^\w\s]','')

#numbers
df['text'] = df['text'].str.replace('\d','')

#stopwords
import nltk
#nltk.download('stopwords')
from nltk.corpus import stopwords
sw = stopwords.words('english')
df['text'] = df['text'].apply(lambda x: " ".join(x for x in x.split() if x not in sw))

#lemmi
from textblob import Word
#nltk.download('wordnet')
df['text'] = df['text'].apply(lambda x: " ".join([Word(word).lemmatize() for word in x.split()])))
```

```
df['text'] = df['text'].str.replace('rt','')
```

```
C:\Users\User\AppData\Local\Temp\ipykernel_11492\21392015.py:5: FutureWarning: The default value of regex will change from True to False in a future version.
```

```
df['text'] = df['text'].str.replace('[^\w\s]','')
```

```
C:\Users\User\AppData\Local\Temp\ipykernel_11492\21392015.py:8: FutureWarning: The default value of regex will change from True to False in a future version.
```

```
df['text'] = df['text'].str.replace('\d','')
```

text column after preprocessing

```
df["text"]
```

```
#### First five lines after preprocessing
```

```
df.head()
```

```
kasim3=pd.read_csv("3Kasim.csv")
kasim3.drop("Unnamed: 0", axis=1, inplace=True)
```

```
kasim3=kasim3.loc[:, "id": "Hashtags"]
```

```
kasim4=pd.read_csv("4Kasim.csv")
kasim4.drop("Unnamed: 0", axis=1, inplace=True)
kasim4=kasim4.loc[:, "id": "Hashtags"]
```

```
df_new=pd.concat([kasim3, kasim4, df], axis=0)
```

```
df_new.info()
```

```
df_new["user_screen_name"].unique().size
```

```
694
```

```
user_names=df_new.groupby("user_screen_name")["id"].count()
```

```
user_names
```

```
# list of multiple tweeters
names=[]
for i in range(len(user_names)):
    if user_names[i]!=1:
        names.append(user_names.index[i])

print(names)
```

```
#We save the user names series as a dataframe|.
names=user_names.to_frame()
names
```

```
#### hashtag top contributors
```

```
df_new.groupby("user_screen_name").count()["id"].sort_values(ascending = False)[0:10]
```

```
#### Original Contributors
```

```
df_new["text"].str.startswith("rt").head(10)
```

```
df_new[~df_new["text"].str.startswith("rt")].count()["id"]
```

```
1594
```

```
df_new[~df_new["text"].str.startswith("rt")].count()["id"] / len(df_new)
```

```
0.941523922031896
```

```
#Ratio of originals to retweets (comparison)
```

```
df_new[~df_new["text"].str.startswith("rt")].count()["id"] / df_new[df_new["text"].str.startswith("rt")].count()["id"]
```

```
16.1010101010101
```

```
#### Most favorite retweets
```

```
df_new.sort_values("retweet_count", ascending = False).head()
```

```
### frequencies of words in tweets
```

```
freq_df = df_new["text"].apply(lambda x: pd.value_counts(x.split(" ")).sum(axis = 0).reset_index())
```

```
freq_df.columns= ["words", "frequency"]
```

```
freq_df
```

```
a=freq_df[freq_df.frequency> freq_df.frequency.mean()+freq_df.frequency.std()]
```

```
# Figure Size
```

```
fig = plt.figure(figsize =(8, 8))
plt.bar(a["words"][1:20], a["frequency"][1:20])
plt.xticks(rotation=45)
plt.show()
```

WordCloud

```
!pip install wordcloud
```

```
Requirement already satisfied: wordcloud in c:\users\user\anaconda3\lib\site-packages (1.8.2.2)
Requirement already satisfied: pillow in c:\users\user\anaconda3\lib\site-packages (from wordcloud) (9.2.0)
Requirement already satisfied: numpy>=1.6.1 in c:\users\user\anaconda3\lib\site-packages (from wordcloud) (1.21.5)
Requirement already satisfied: matplotlib in c:\users\user\anaconda3\lib\site-packages (from wordcloud) (3.5.2)
Requirement already satisfied: packaging>=20.0 in c:\users\user\anaconda3\lib\site-packages (from matplotlib->wordcloud) (21.3)
Requirement already satisfied: kiwisolver>=1.0.1 in c:\users\user\anaconda3\lib\site-packages (from matplotlib->wordcloud) (1.4.2)
Requirement already satisfied: cycler>=0.10 in c:\users\user\anaconda3\lib\site-packages (from matplotlib->wordcloud) (0.11.0)
Requirement already satisfied: pyparsing>=2.2.1 in c:\users\user\anaconda3\lib\site-packages (from matplotlib->wordcloud) (3.0.9)
Requirement already satisfied: python-dateutil>=2.7 in c:\users\user\anaconda3\lib\site-packages (from matplotlib->wordcloud) (2.8.2)
Requirement already satisfied: fonttools>=4.22.0 in c:\users\user\anaconda3\lib\site-packages (from matplotlib->wordcloud) (4.25.0)
Requirement already satisfied: six>=1.5 in c:\users\user\anaconda3\lib\site-packages (from python-dateutil->matplotlib->wordcloud) (1.16.0)
```

```
from os import path
from PIL import Image
from wordcloud import WordCloud, STOPWORDS, ImageColorGenerator
import matplotlib.pyplot as plt
```

```
text = " ".join(i for i in df.text)
```

```
wordcloud = WordCloud(background_color = "white").generate(text)
plt.imshow(wordcloud, interpolation = "bilinear")
plt.axis("off")
plt.tight_layout(pad= 0)
plt.show()
```

Sentiment Analysis

```
df_new.head()
```

	id	text	created_at	retweeted	retweet_count	user_screen_name	user_followers_count	user_location	Hashtags
0	1587533760664580096	online ebook open access no download machinel...	2022-11-01 19:55:18+00:00	False	122	KirkDBorne	368975	Maryland, USA	[{"text": 'MachineLearning', 'indices': [44, 6...]
1	1587679041553039361	datamining and knowledge discovery handbook ...	2022-11-02 05:32:35+00:00	False	47	KirkDBorne	368975	Maryland, USA	[{"text": 'DataMining', 'indices': [5, 16]]]
2	1587534842799218688	data science from scratch plus more datascienc...	2022-11-01 19:59:36+00:00	False	28	KirkDBorne	368975	Maryland, USA	[{"text": 'DataScience', 'indices': [39, 51]]]
3	1588121207790075906	dm to recover your gmail machinelearning from...	2022-11-03 10:49:36+00:00	False	0	GigsonHac	1262	London, Ontario	[{"text": 'MachineLearning', 'indices': [28, 4...]
4	1588118260360400897	design pandaspython dataprocessing databasesto...	2022-11-03 10:37:53+00:00	False	0	corpus_news	1131	North Carolina, USA	[{"text": 'design', 'indices': [0, 7]}, {"text": ...]

Adding sentiment score and sentiment class features to the dataset.

```
## Adding emotion score and emotion class features to the dataset.from textblob import TextBlob
```

```
def sentiment_skorla(df):
    text = df["text"]

    for i in range(0,len(text)):
        textB = TextBlob(text[i])
        sentiment_skoru = textB.sentiment.polarity
        df.at[i, 'sentiment_score'] = sentiment_skoru

        if sentiment_skoru < 0.00:
            duygu_sinifi = 'Negatif'
            df.at[i, 'sentiment_class'] = duygu_sinifi

        elif sentiment_skoru > 0.00:
            duygu_sinifi = 'Pozitif'
            df.at[i, 'sentiment_class'] = duygu_sinifi

        else:
            duygu_sinifi = 'Notr'
            df.at[i, 'sentiment_class'] = duygu_sinifi

    return df
```

```
sentiment_skorla(df_new)
```

```
df_new[["id", "text", "created_at", "sentiment_class"]]
```

```
df_new.groupby("sentiment_class").count()["id"]
```

```
duygu_freq=df_new.groupby("sentiment_class").count()["id"]
duygu_freq.plot.bar(x="sentiment_class", y="id");
```

```
df1=df_new[["text", "sentiment_score", "sentiment_class"]]
df1.head(25)
```

Adding Tweet Times and Days to the Dataset

```
df_new.head()

from datetime import datetime
import datetime

"""
d=df_new["created_at"][0][0:-6]
d=datetime.datetime.strptime(d, '%Y-%m-%d %H:%M:%S').strftime("%H")
d
"""

'\nd=df_new["created_at"][0][0:-6]\nd=datetime.datetime.strptime(d, \'%Y-%m-%d %H:%M:%S\').strftime("%H")\nd\n'
```

```
df_new["tweet_time"]=df_new["created_at"]
df_new["tweet_time"]=df_new["tweet_time"].apply(lambda x: x[0:-6])
df_new.head()

for i in range(1693):
    df_new["tweet_time"][i]=datetime.datetime.strptime(df_new["tweet_time"][i], '%Y-%m-%d %H:%M:%S').strftime("%H")
```

C:\Users\user\AppData\Local\Temp\ipykernel_11492\4117623610.py:2: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy

```
df_new["tweet_time"][i]=datetime.datetime.strptime(df_new["tweet_time"][i], '%Y-%m-%d %H:%M:%S').strftime("%H")
```

```
df_new.head(10)

df_new["tweet_time"].unique()

array(['19', '05', '10', '09', '08', '07', '06', '04', '03', '02', '01',
       '00', '12', '11', '23', '22', '21', '20', '13', '18', '17', '14',
       '16', '15'], dtype=object)
```

```
df_new["tweet_time"] = pd.to_numeric(df_new["tweet_time"])
df_new.info()
```

Visualization of Tweet Times

```
import seaborn as sns
sns.distplot(df_new["tweet_time"], kde = True, color = "blue");
```

Code to add days

```
df_new["days"]=df_new["created_at"]
df_new["days"]=df_new["days"].apply(lambda x: x[0:-6])

for i in range(1693):
    df_new["days"][i]=datetime.datetime.strptime(df_new["days"][i], '%Y-%m-%d %H:%M:%S').strftime("%A")

df_new.head()
```

Day distributions in tweets

```
df_new.groupby("days").count()["id"]
```

Visualization

```
gun_freq = df_new.groupby("days").count()["id"]
gun_freq.plot.bar(x = "days", y = "id");

df_new.groupby(["tweet_time", "days"])[["tweet_time"]].count()
```