REPORT

UNIVERSITY: ADANA ALPARSLAN TÜRKEŞ SCIENCE and TECHNOLOGY UNIVERSITY

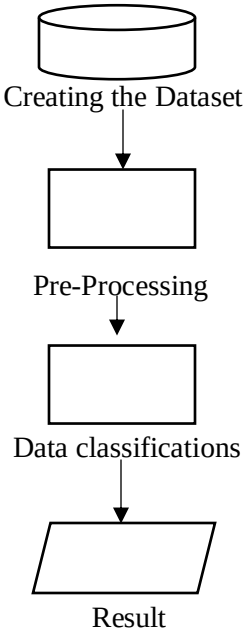
COURSE NAME: INTRODUCTION TO DATA SCIENCE

SUBJECT OF THE PROJECT: ANALYSIS OF COMMENTS WITH TWITTER FANATICISM HASHTAGS

STUDENT'S NAME and NUMBER: Gamze GENÇ – 180101011

In this project, the social media platform Twitter by applying text mining methods to the received data. It is aimed to analyze the comments.

Processes Steps:



1.Tools and Libraries Used

As the development environment during the project Spyder, Python is used as the software language.

In the study, numpy, pandas, nltk, tweepy, configparser,string,re libraries were used.

- Pandas simplifies data analysis and preprocessing. It was used in the project to receive, process, and analyze the data kept in the csv file.
- Nltk is an open-source library developed in python to work with human language data. It was used to remove the Turkish stop-word in the project.
- Tweepy Module is a library for twitter in python. It is written using the Twitter API. We can do many things through this library (Twitting, reading the timeline, followers, followed, etc.).
- There is a config.ini file in my work. This file is the configuration file containing some data of the program (API key, Access token etc.).In my API code, I will use this data. I parsed this config file with the configparser module. After parsing, I have the necessary data read from this file.

2.Create a dataset:

Twitter API (Twitter Apps) was used to collect data from Twitter. Necessary permissions (api key and access token) were created and tweets containing the keyword 'Fanatic' were collected. These processes are using the Tweepy library in Python. A total of 1000 data sets were obtained.

3.Pre-processing:

Twitter messages can contain usernames starting with @, hashtags starting with #, and punctuation marks. These characters are removed from the tweets and actions are taken. Before proceeding to the classification process, these features in the messages must be pre-processed. For these operations, the punctation function (string.punctuation) in the String library in Python is used.

The stop-word function in the NLTK library in Python and the Counter function from the Collection library were used to remove the words (often repetitive, stop-words, meaningless Turkish stop words) and characters from the tweet comments and thus to make the analysis easier in the datasets.

The TurkishStemmer function from the snowballstemmer library in Python was used to take the root of the words in the text.

The *re* (regular expression) library was also used to clean the emojis used in the comments.

In this context, the following procedures were applied to the comments.
•Removal of stop words.
• Cleaning hashtags, usernames, repetitive words, punctuation marks and emojis from the text
• Any non-letter characters are removed from messages.

Rare words in the project: {'kaliteli', 'olurum', 'toplarım', 'yaşar', 'seviyorum', 'kötü.', 'tanıştığım', 'Rizespor', 'teklif', '@ceydamaalesef', 'zirve', 'çayda', "'koç'u", 'bana;', 'alamam'}

Stopwords in the project: {'mü', 'nerede', 'niçin', 'biz', 'acaba', 'mı', 'nasıl', 'mu', 'için', 'da', 'eğer', 'aslında', 'nerde', 'de', 'kim', 'çünkü', 'veya', 'ise', 'daha', 'belki', 'biri', 'ki', 'bu', 'ile', 'birşey', 'şu', 'hiç', 'yani', 'hem', 'bazı', 'siz', 'çok', 'nereye', 'o', 'hepsi', 'şey', 'diye', 'niye', 'hep', 'her', 'az', 'defa', 'neden', 'sanki', 'ama', 'gibi', 'ne', 'tüm', 'ya', 've', 'en', 'kez', 'birkaç'}

Frekans in the Project:{'https://t.co/2cvdPlppdM', 'istatistik:', 'bir', '📊', 'Nisa', 'İlgili', 'haberler:', "Bölükbaşı'dan", 'RT', 'Acun', 'Ilıcalı', 'Fanatik:', '📰', 'fanatik', 'Detaylı'}

Here is the detailed version of the data frame before saving it as a csv file:
It's have 7 attribute and 1000sample.

df - DataFrame

Index	User	Tweet	ozel_karaktersiz	stop_word	sık_kullanılan	kelime_kok	emojisiz
133	86efsaneusta	⚫⚪ danimarkal…	⚫⚪ danimarkalı iş…	cointerbiyecis…	cointerbiyec…	cointerbiyec…	cointerbiyecis son dal
134	akfcfatih	cok fanatik ani…	cok fanatik anilci …	cok fanatik an…	cok anilci o…	cok anilci o…	cok anilci olduk anil
135	revizyonist…	@comradeyutyo k…	comradeyutyo kendis…	comradeyutyo k…	comradeyutyo…	comradeyutyo…	comradeyutyo kendi aş
136	ozanozarar	@crypto_amante …	cryptoamante ayhand…	cryptoamante a…	cryptoamante…	cryptoamante…	cryptoamante ayhandog
137	geniuslabAJ	@cum3169 @saova…	cum3169 saovaradx t…	cum3169 saovar…	cum3169 saov…	cum3169 saov…	cum3169 saovaradx tae
138	fanatik_olma	@cumdullahgul @…	cumdullahgul ugrrtr…	cumdullahgul u…	cumdullahgul…	cumdullahgul…	cumdullahgul ugrrtrk sen mi ekonomist
139	DeportesRep…	danimarkalı iş …	danimarkalı iş ve s…	danimarkalı iş…	danimarkalı …	danimarkal i…	danimarkal iş spor in
140	adnbyrm	@dastan2004 64 …	dastan2004 64 doğum…	dastan2004 64 …	dastan2004 6…	dastan2004 6…	dastan2004 64 doğumlu
141	kayiplarday…	o bir deniz yan…	o bir deniz yanılır…	bir deniz yanı…	deniz yanılı…	de yanılır b…	de yanılır be güneş s
142	M_EOFT	bu adam fb için…	bu adam fb için tv …	degajsportscom…	degajsportsc…	degajsportsc…	degajsportsco ezber s
143	Maliceheuse	fanatik delici…	fanatik delici abin…	fanatik delici…	delici abin…	delic ap var…	delis ap varsa

Then, the comments were saved in a csv file with two attributes (user and tweets) and 1000 examples with the help of the pandas library.

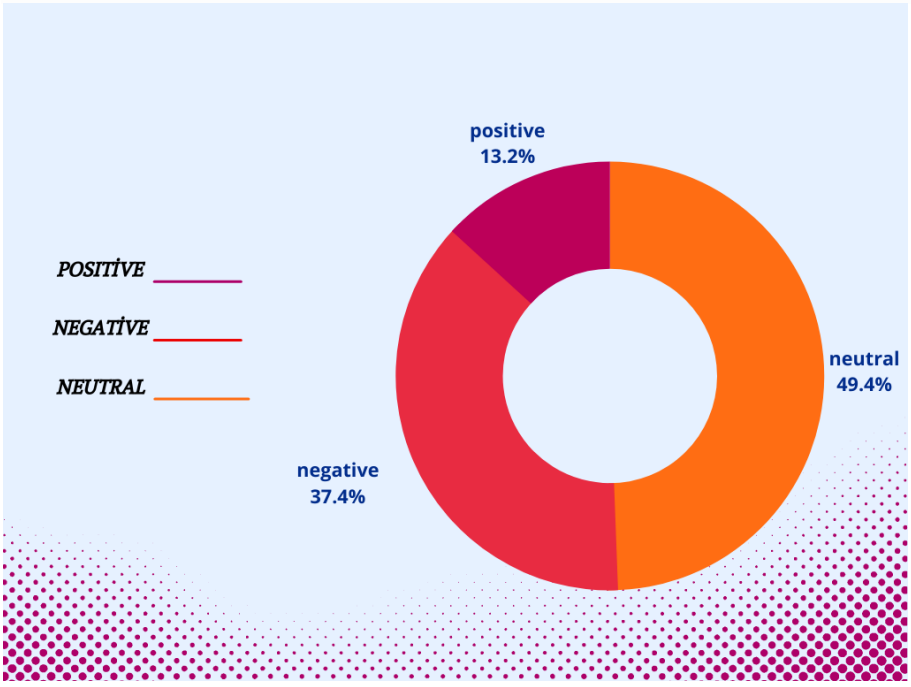
4.Data classifications:

The comments obtained from Twitter were evaluated according to their polarity score, positivity, negativity, and neutrality.

5. Result:

A data set containing 1000 tweets about Fanatik collected on Twitter was used. The results obtained as a result of the study in Figure-1 polarity results are shown.

According to the results, it is seen that the tweets are 13,2% positive, 37,4% negative and 49.4% neutral. With these results, most of those who wrote about fanaticism shared neutrality. The number of those who share positively also shows those who have such an expectation. Negatively loaded shares are the ones that clearly show that this expectation is not true.



-Figure1-