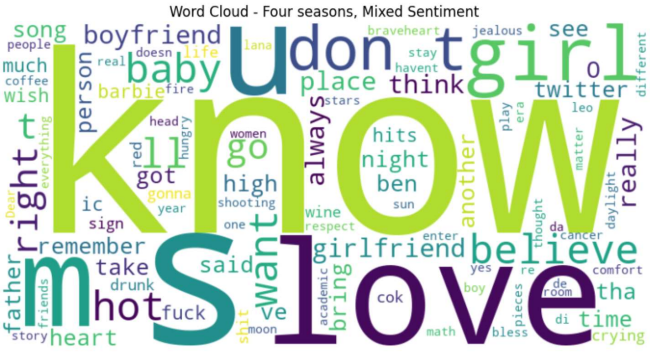
Didem Iris Topbas

30654

CS210

Tweet-Art



(the words I highly use in my tweets)

Hypothesis:

The sentiment of my tweets is influenced by various factors such as personal achievements, various life events, and personal interactions. My hypothesis is that the beginning of holidays during the summer season is likely to lead to more positive and upbeat expressions in my tweets, so I expect that the highest sentiment month will be in the summer seasons compared to other seasons.

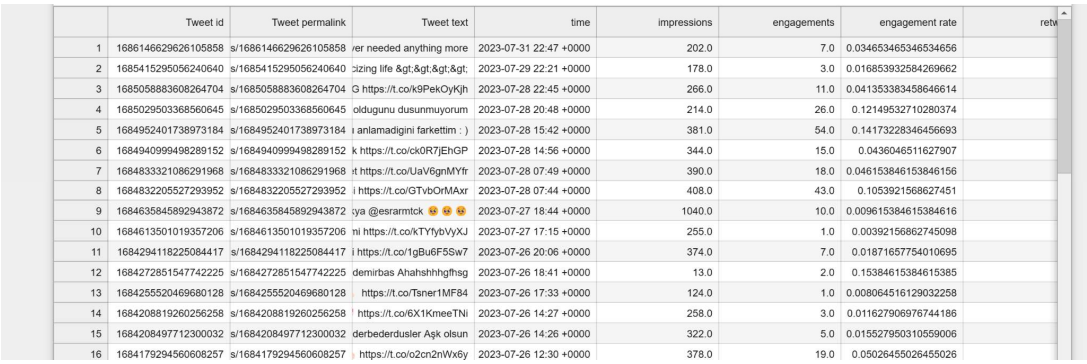
Motivation:

My motivations for this project was to explore the sentiment of my tweets (which I post very frequently and with random emotions) throughout the seasons and analyze the seasons by exploring when I posted negative tweets the most or positive tweets the most, which would create a sense of overall average sentiment for my tweets. I was very curious for the result of this analysis and whether my hypothesis would be proved correct or not. I also wanted to delve into diverse set of techniques throughout the project such as using Natural Language Processing (NLP) and creative visualization of my twitter data.

Data source

I have retrieved my Twitter data by first accessing my Twitter Analytics and exported all my data based on tweets for every month of 2023 in csv format. Then I read the .csv files using

pandas library and stored every month into a data frame; also merged the month data frames into season data frames. After that, I organized the data by preprocessing and extracting the necessary columns from the data such as ‘Twitter data’, ‘Tweet text’, ‘time’, ‘impressions’, ‘likes’ and dropped the rows from the data with missing values. I have eliminated other columns such as ‘Tweet permalink’, ‘retweets’, ‘replies’, ‘user profile clicks’ from the data since they weren’t necessary for my exploratory data analysis topic and also it was crucial to eliminate irrelevant columns from the data to make a simple graph and efficient analysis. In addition to preprocessing, I have also eliminated the links and mentions in tweets for further analysis.



	Tweet id	Tweet permalink	Tweet text	time	impressions	engagements	engagement rate	retw
1	1686146629626105858	s/1686146629626105858	/er needed anything more	2023-07-31 22:47 +0000	202.0	7.0	0.034653465346534656	
2	1685415295056240640	s/1685415295056240640	izing life >>>	2023-07-29 22:21 +0000	178.0	3.0	0.016853932584269662	
3	1685058883608264704	s/1685058883608264704	G https://t.co/k9PekOyKjh	2023-07-28 22:45 +0000	266.0	11.0	0.041353383458646614	
4	1685029503368560645	s/1685029503368560645	oldugunu dusunmuyorum	2023-07-28 20:48 +0000	214.0	26.0	0.12149532710280374	
5	1684952401738973184	s/1684952401738973184	ı anlamadigini farkettim :)	2023-07-28 15:42 +0000	381.0	54.0	0.14173228346456693	
6	1684940999498289152	s/1684940999498289152	k https://t.co/ck0R7jEhGP	2023-07-28 14:56 +0000	344.0	15.0	0.0436046511627907	
7	1684833321086291968	s/1684833321086291968	t https://t.co/UaV6gnMYfr	2023-07-28 07:49 +0000	390.0	18.0	0.046153846153846156	
8	1684832205527293952	s/1684832205527293952	i https://t.co/GTvbOrMAxr	2023-07-28 07:44 +0000	408.0	43.0	0.1053921568627451	
9	1684635845892943872	s/1684635845892943872	;ya @esrarmtck 😍😍😍	2023-07-27 18:44 +0000	1040.0	10.0	0.009615384615384616	
10	1684613501019357206	s/1684613501019357206	ni https://t.co/kTYfybVyXJ	2023-07-27 17:15 +0000	255.0	1.0	0.003921568627450980	
11	1684294118225084417	s/1684294118225084417	i https://t.co/1gBu6F5Sw7	2023-07-26 20:06 +0000	374.0	7.0	0.01871657754010695	
12	1684272851547742225	s/1684272851547742225	demirbas Ahahshhhgfhsg	2023-07-26 18:41 +0000	13.0	2.0	0.15384615384615385	
13	1684255520469680128	s/1684255520469680128	https://t.co/Tsner1MF84	2023-07-26 17:33 +0000	124.0	1.0	0.008064516129032258	
14	1684208819260256258	s/1684208819260256258	https://t.co/6X1KmeeTNi	2023-07-26 14:27 +0000	258.0	3.0	0.011627906976744186	
15	1684208497712300032	s/1684208497712300032	derbederdusler Aşk olsun	2023-07-26 14:26 +0000	322.0	5.0	0.015527950310559006	
16	1684179294560608257	s/1684179294560608257	https://t.co/o2cn2nWx6y	2023-07-26 12:30 +0000	378.0	19.0	0.05026455026455026	

	Tweet id	Tweet text	time	impressions	likes	Language
0	1619931943021457409	but we were something, don't u think so? 🎶	2023-01-30 05:34 +0000	125.0	1.0	en
1	1616814002671345665	@unknownobiwan ✨ real love ✨	2023-01-21 15:04 +0000	45.0	0.0	en
2	1616773936372666368	@unknownzem reaction atacak flörtü var demek ki	2023-01-21 12:25 +0000	101.0	0.0	en
3	1616732605117874177	@ysinanoz 🌞 ✨ you are the main character ✨ 🌞	2023-01-21 09:41 +0000	90.0	1.0	en
4	1614350782974812162	the moral of the story https://t.co/oJzXTrOpFM	2023-01-14 19:56 +0000	520.0	7.0	en
...	...	...	...	...	...	...
576	1720558225261498463	birileriyle tanismadan once soyle bir testten ...	2023-11-03 21:47 +0000	275.0	8.0	NaN
577	1720535537046503933	bu aralar ben 🌹 https://t.co/vO404AlxXi	2023-11-03 20:16 +0000	212.0	3.0	NaN
578	1720159462453825751	but do, mi, ti, why not me?	2023-11-02 19:22 +0000	162.0	1.0	NaN
579	1720146093059571988	bir tane daha chandler friends editini kalbim ...	2023-11-02 18:29 +0000	230.0	5.0	NaN
580	1720018358232289619	boyle bir gun https://t.co/xuUJEuHCED	2023-11-02 10:01 +0000	340.0	9.0	NaN

581 rows × 6 columns

Data analysis

The first technique I used for my data analysis was the Language Detection method, ‘langdetect’, that created a function to detect the language of a tweet. The reason I used this model was for considering only my tweets that were written in English since the sentiment analysis I performed later on my tweets, worked accurately only on the tweets that were written in English and did not accurately analyze the sentiment of the tweets that were written in Turkish. After detecting the language of my tweets, I created new data frames for every month and season that contained only my tweets that were written in English.

Furthermore, for the second stage of my analysis, I have proceeded with the Sentiment Analysis of my tweets using ‘textblob’, which is a library for processing textual data, and analyzing the sentiment of a given text by providing a polarity score. The polarity ranged from -1 (most

negative) to 1 (most positive). In addition to the sentiment analysis, I have also added a user interaction, where the user is asked to select a season (winter, spring, fall, summer, four seasons) and a sentiment range (positive, negative, mixed), and if the user selects a sentiment range other than mixed (positive, negative), the user is also asked whether to display top 5 positive/negative tweets or not, in which the top tweets are based on their sentiment polarity (the highest/ the lowest).

Moreover, the third stage of my analysis included Exploratory Data Analysis, where I created a scatter plot figure to display the sentiment of my tweets for the chosen season and sentiment range. The figure is based on the date (x-axis) and tweet impressions (y-axis), and markers represent the tweets along with their color representing their sentiment polarity (blue for positive, red for negative). I have used 'matplotlib.pyplot' to accomplish displaying the result of this analysis. Additional to the scatter plot, by using the 'wordcloud' library, I have generated a word cloud that spotlights 100 most frequently used words in positive, negative or mixed tweets, which provided an insight into my interests, and their color being based on their sentiment polarity, is why we call it an artwork.

For the last stage of my analyses, I calculated the highest and lowest sentiment month according to their sentiment polarity and displayed the most positive and most negative tweets again based on their sentiment polarity. At last, I have conducted a Hypothesis Testing based on the results of t-test and p-values on the sentiment scores of summer season and other seasons which determined whether to reject the null hypothesis based on the result.

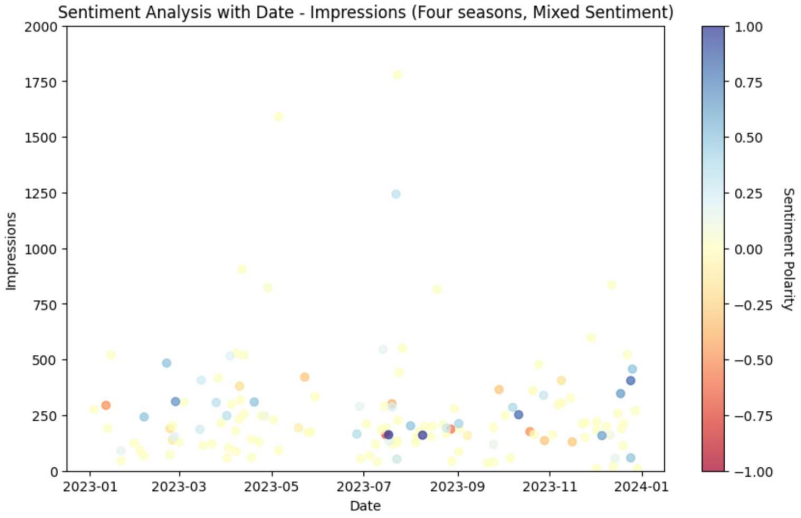
Findings

```
Highest Sentiment Month: June
Lowest Sentiment Month: January

Most Positive Tweet:
constructions have started to build the best me ✨

Most Negative Tweet:
and I'm bad like the barbie

Average Total Sentiment: 0.060484153979051936

Hypothesis Testing:
T-statistic: 0.29799807361852393
P-value: 0.7668719966143014
The hypothesis that the highest sentiment month is in the summer season is not rejected.
```

Finding out that the average sentiment of my tweets being positive with an average of 0.06 was actually a big relief since I was writing my tweets unbeknownslty that I would analyze their sentiment one day. Also, I was surprized with the result that my most positive tweet was 'constructions have started to build the best me' and that my most negative tweet was 'and I'm bad like the barbie', which was funny.

```
Select a season (winter, spring, fall, summer, four seasons):  summer
Please select a sentiment range (positive/negative/mixed):  positive
Do you want to show top 5 positive tweets? (yes/no):  yes
Top 5 Tweets:
1. constructions have started to build the best me ✨
2. when he brings out the best version of you &gt;&gt;&gt;&gt;
3. salt air, and the rust on your door
I never needed anything more
4. a friendly reminder to not play cakmak oyunu with strangers
5. Dear {inner} child,
Its me.
I know you havent always been able to trust me.
I get it.
I havent always seen you.

I am here.
I love you.
I hear you.
```

Proving that my hypothesis was correct, that the highest sentiment month would be in the summer season (result was June) based on my predictions of the external factors that would influence this, was proved both by our calculation result from Hypothesis Testing and by the scattered plot of Sentiment Analysis (Four seasons, Mixed Sentiment), where we could visually see the markers having the darkest tone of blue and the highest sentiment. It was also interesting, how the visualization of different sentiment of my tweets scattered across the year is also representing how my emotions changed throughout the year as well.



Limitations and Future Work:

Some of the limitations I faced across this project was not being able to find the correct Natural Language Processing model for analyzing the sentiment of my tweets that were in Turkish. That was the reason I focused on my tweets that were in English since the models available showed accurate results. In addition to this problem, the sentiment analysis approach used for my tweets did not handle sarcasm, irony, or complex language structures effectively.

The future plan for my project is to explore the use of deep learning models and train my own data for sentiment analysis using models such as recurrent neural networks or transformer-based models, which would be more powerful approach to improve accuracy and handle complex language structures.