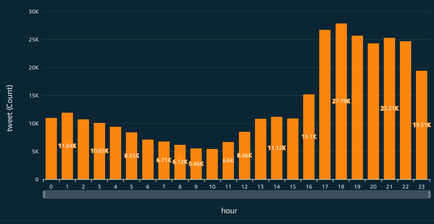
Sentiment Analysis - Elon Musk dataset

1. Data analysis

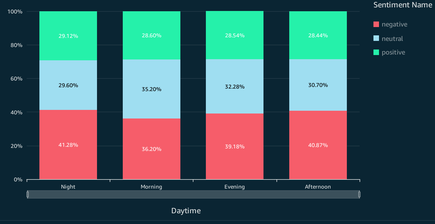
Analysis related to time of publishing tweets

- The highest number of tweets was posted between 5 PM and 11 PM, likely because we are analyzing Monday and Tuesday – times when people are typically back from work.
- There are no major differences in sentiment distribution across dayparts, although negative tweets are more frequent at night.
- In this dataset, the number of tweets on Tuesday was twice as high as on Monday. It's unclear whether this reflects a specific pattern.
- There are no significant differences in sentiment distribution between the two weekdays, although slightly more positive tweets were posted on Monday.

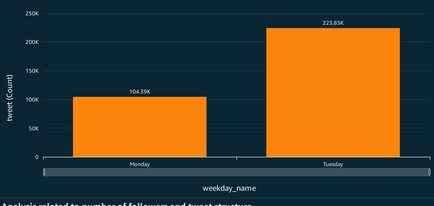
Number of tweets generated in particular hours



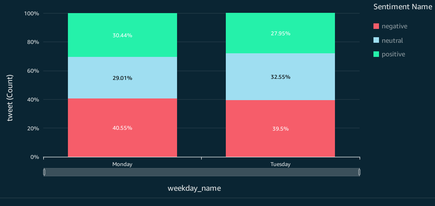
Distribution of the sentiment in selected dayparts



Number of tweets by day of week



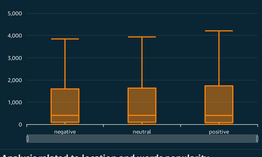
Distribution of the sentiment by day of week



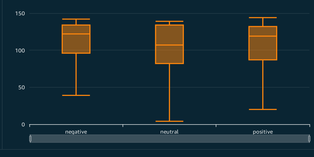
Analysis related to number of followers and tweet structure

- There are no major differences in the number of followers based on sentiment, although positive tweets tend to come from users with slightly more followers.
- Tweets labeled as negative tend to be longer. Neutral tweets are slightly shorter, and positive ones are in between.
- When analyzing the number of words, the pattern is similar to tweet length, although the differences are less pronounced.

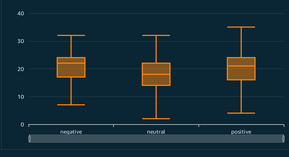
Number of followers by sentiment



Tweet length by sentiment



Number of words used in tweets by sentiment

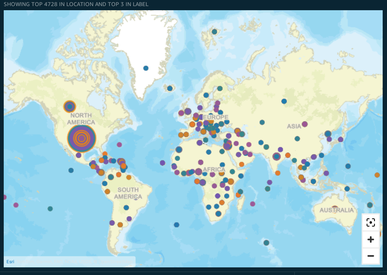


Analysis related to location and words popularity

- The highest number of tweets is generated in the US, which is natural given Elon Musk's strong connection to this region.
- Besides obvious words like *Twitter* and *Elon Musk*, frequent terms also include *Trump* — due to his connection to Elon Musk — as well as *employees* and *people*, especially in tweets with negative sentiment.

Number of tweets by location and sentiment

SHOWING TOP 4728 IN LOCATION AND TOP 3 IN LABEL

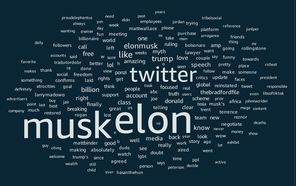


Top 200 most popular words in whole dataset

SHOWING TOP 200 IN WORD

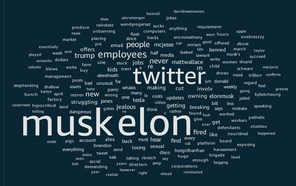
Top 200 most popular words with positive sentiment

SHOWING TOP 200 IN WORD



Top 200 most popular words with negative sentiment

SHOWING TOP 200 IN WORD



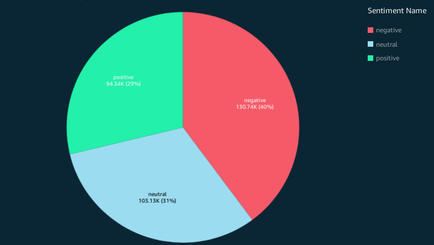
Top 200 most popular words with neutral sentiment

SHOWING TOP 200 IN WORD

Distribution of the sentiment and retweets in dataset

- Particular classes representing sentiment are slightly inbalanced - in dataset there is the highest number of tweets with negative sentiment (40%)
- There is only 87k unique tweets, so around 27% of whole dataset, the rest 73% these are retwees.
- Some retweets were published over 7.5k times, which means that random data splitting could lead to data leakage during model building. That's why I decided to split the data chronologically (based on time).- in the test set there is 20% of the most actual tweets; there is still a risk of appearing retweets, but smaller.

Distribution of tweets by sentiment



Comparison of unique tweets to retweets

Top 10 most popular retweets with the sentiment

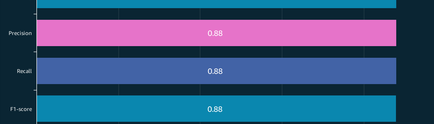
tweet	sentiment_name	count
RT @Mc.Jesse: I'm jealous of Elon Musk's kids. They're the only people who never have to hear from him.	negative	7,490
RT @MattWallace888: The Twitter employees Elon Musk fired are struggling to find new jobs because all of the offers they are getting involv…	negative	6,339
"RT @the_meghaning: Elon Musk called Trent Reznor a \"crybaby\" for deleting his Twitter account. Wil Wheaton then wrote this incisive summary…"	neutral	5,553
RT @gmadoesthings: Elon Musk & Jordan Peterson, neither of whom have been "cool" a single day in their lives, raging on Trent Reznor, who…	neutral	5,397
RT @MattGertz: Elon Musk interacting with sycophantic right-wing influencers this weekend, a thread.	neutral	4,910
RT @OccupyDemocrats: BREAKING: In another blow to Elon Musk, White Stripes frontman Jack White ditches Twitter over Musk's "assh*ole move"…	neutral	3,541
RT @hotgirlhanfran: elon musk did to twitter what brendon urie did to panic! at the disco	negative	3,500
RT @disclosetv: JUST IN - Elon Musk has reinstated Rep. Marjorie Taylor Greene's (R-GA) personal Twitter account.	neutral	3,374
RT @thebradfordfile: It's amazing the media is more focused on Elon Musk than a $15 billion ponzi scheme for the global ruling class. LOL.	positive	3,338
RT @ACTBrigitte: If Apple takes Twitter out of the app store, Elon Musk should buy them next.	neutral	3,061

2. Model metrics

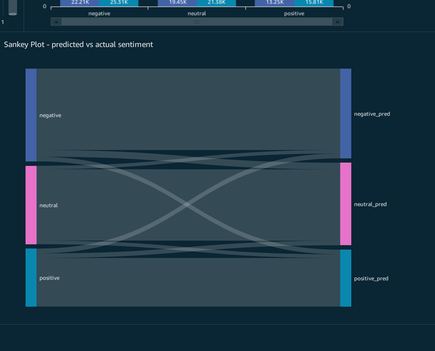
Summary of the model performance

- The best result gave Logistic Regression with features based on tweet text.
- The model correctly predicts sentiment in most cases, reaching 88% accuracy.
- It performs best on neutral tweets, with an accuracy of 90%.
- The model makes the most mistakes between positive and neutral tweets, which are sometimes hard to distinguish.

Model Evaluation - Logistic Regression



Accuracy level by the sentiment

Matrix Plot - predicted vs actual sentiment

Sankey Plot - predicted vs actual sentiment