CSE 4062 - Semester 2021

Intro to Data Science and Analytics

Delivery #3: Exploring your data

Group 2 Members:

Aleyna BOZACI 150319630 - aleynabozacii@gmail.com

Bekir ÖZKAN 150319557 - bekirozkan9698@gmail.com

Merve YAYIN 150116051 - yayinm8@gmail.com

Metehan ERTAN 150117051 - metehan.ertan@hotmail.com

Mustafa USCA 150416054 - uscamustafa64@gmail.com

1- Dataset

1 - Deceptive Opinion Spam Corpus v1.4s (shortly Turk). We will be using negative reviews for testing.

- 400 truthful positive reviews from TripAdvisor
- 400 deceptive positive reviews from Mechanical Turk
- 400 truthful negative reviews from Expedia, Hotels.com, Orbitz, Priceline, TripAdvisor and Yelp
- 400 deceptive negative reviews from Mechanical Turk

2 - We extracted a smaller dataset from YELP Dataset as it was too large. We are going to use 800 truthful and 800 deceptive reviews for the learning part and 200 reviews for the testing part.

Deceptive Opinion Spam Corpus v1.4s has 800 rows for learning and 800 rows for testing. For the Yelp dataset there are 1600 rows for learning and 200 rows for learning. Target attribute is truthful or deceptive as boolean.

2 - Statistics

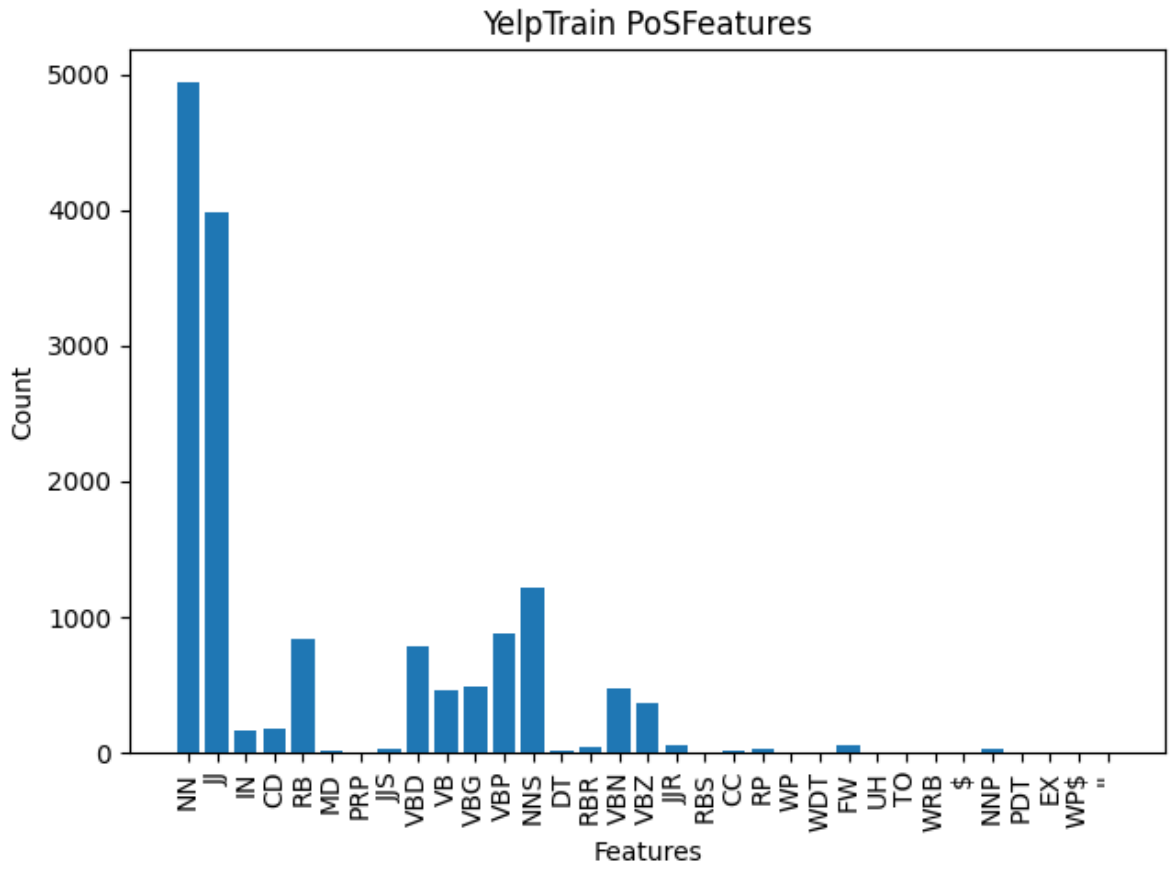
To create charts from attributes we used nltk and spacy libraries. We started with nltk and found PoS tags and IoB tags, later discovered the spacy library and found NER tags using spacy.

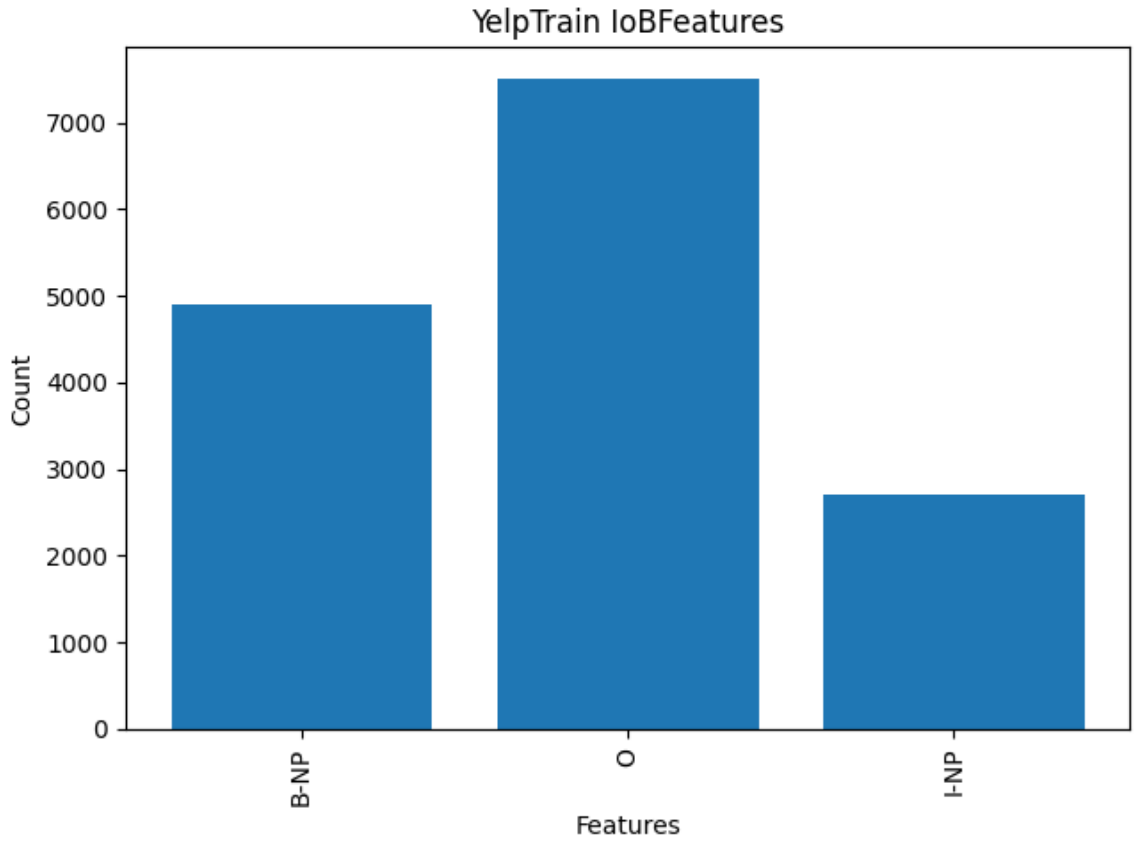
Firstly we tokenized the whole dataset and removed stopwords. Removing stop words will increase the accuracy and remove noise. Later PoS tagged the dataset. Lastly used this information to find IoB tags. After all this processes used the spacy library to find NER tags.

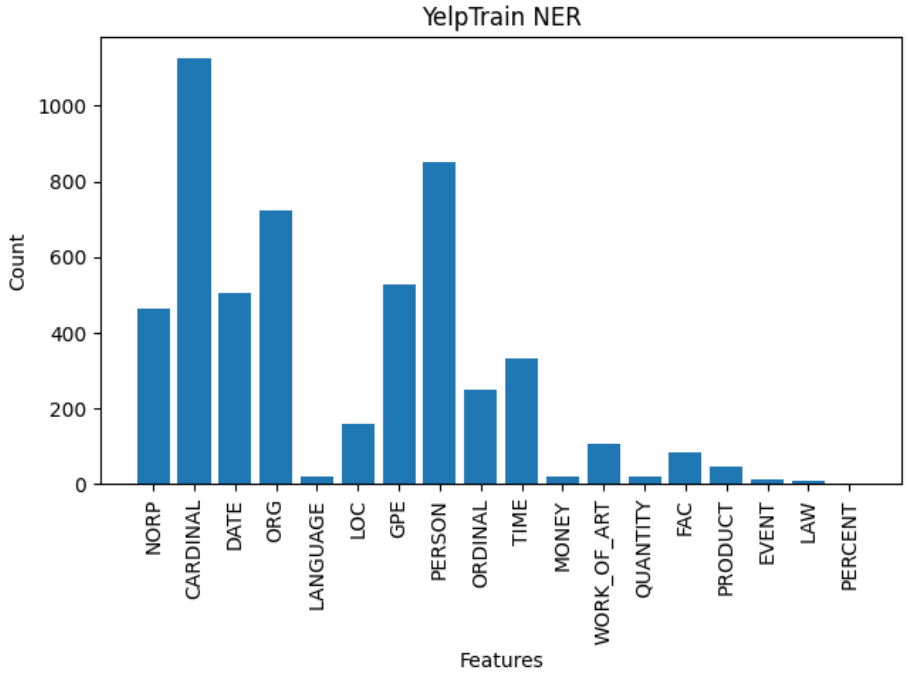
Currently we have 18 attributes : 'NORP', 'CARDINAL', 'DATE', 'ORG', 'LANGUAGE', 'LOC', 'GPE', 'PERSON', 'ORDINAL', 'TIME', 'MONEY','WORK_OF_ART', 'QUANTITY', 'FAC', 'PRODUCT', 'EVENT', 'LAW', 'PERCENT'. Result is 0 if review is fake, 1 if it’s true.

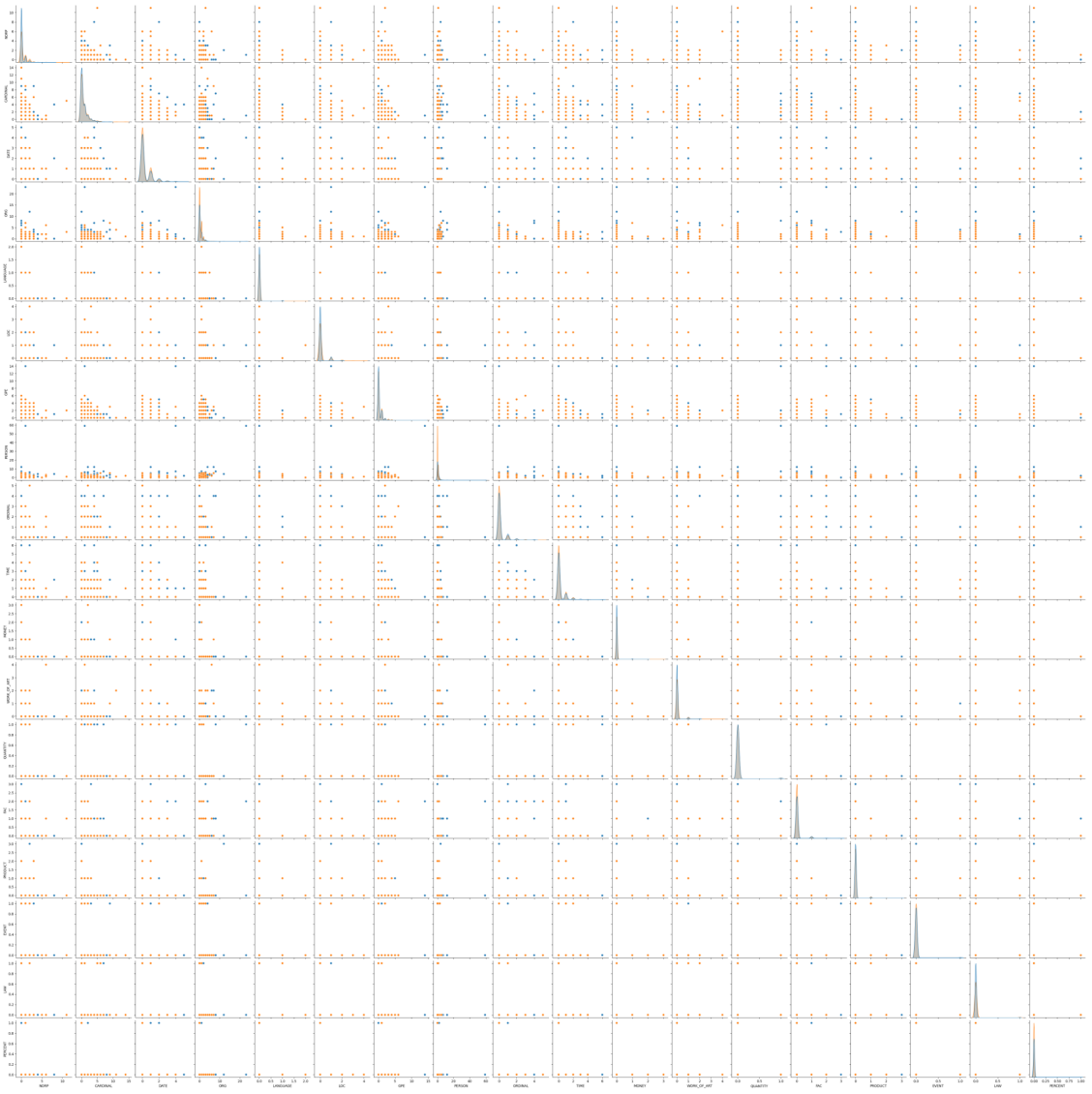
Created scatter plot with seaborn library. All charts are uploaded on github with higher resolution.

2.1- Yelp Train

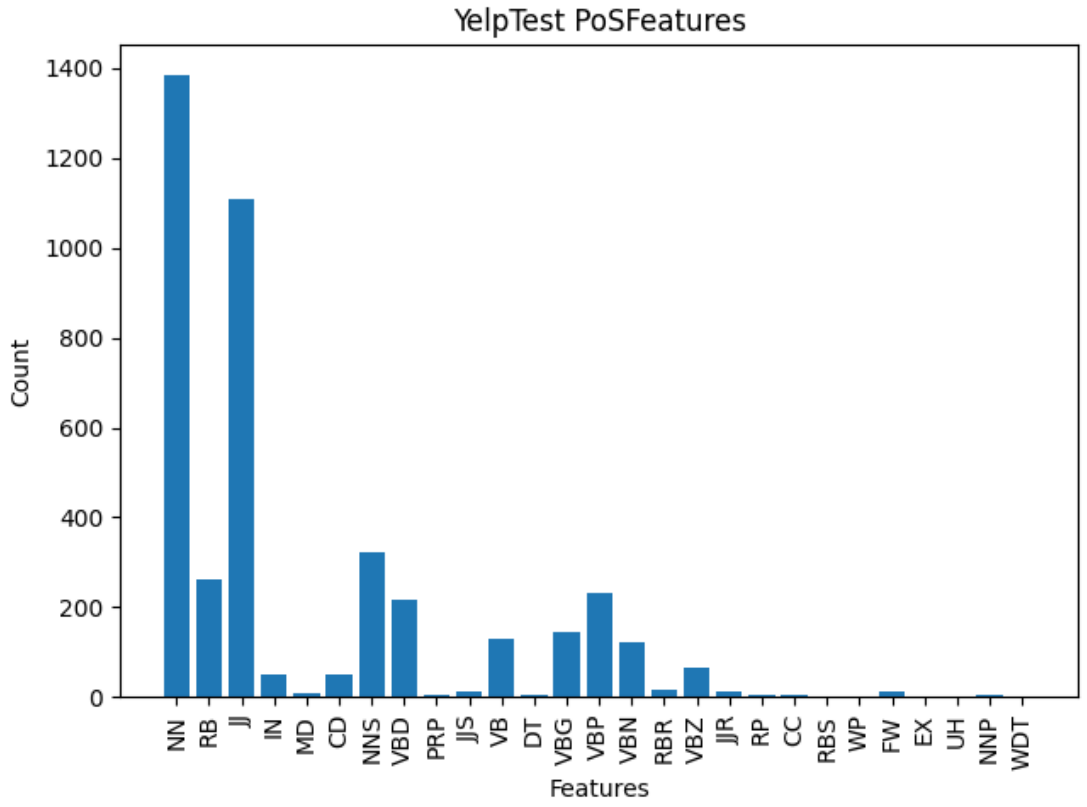


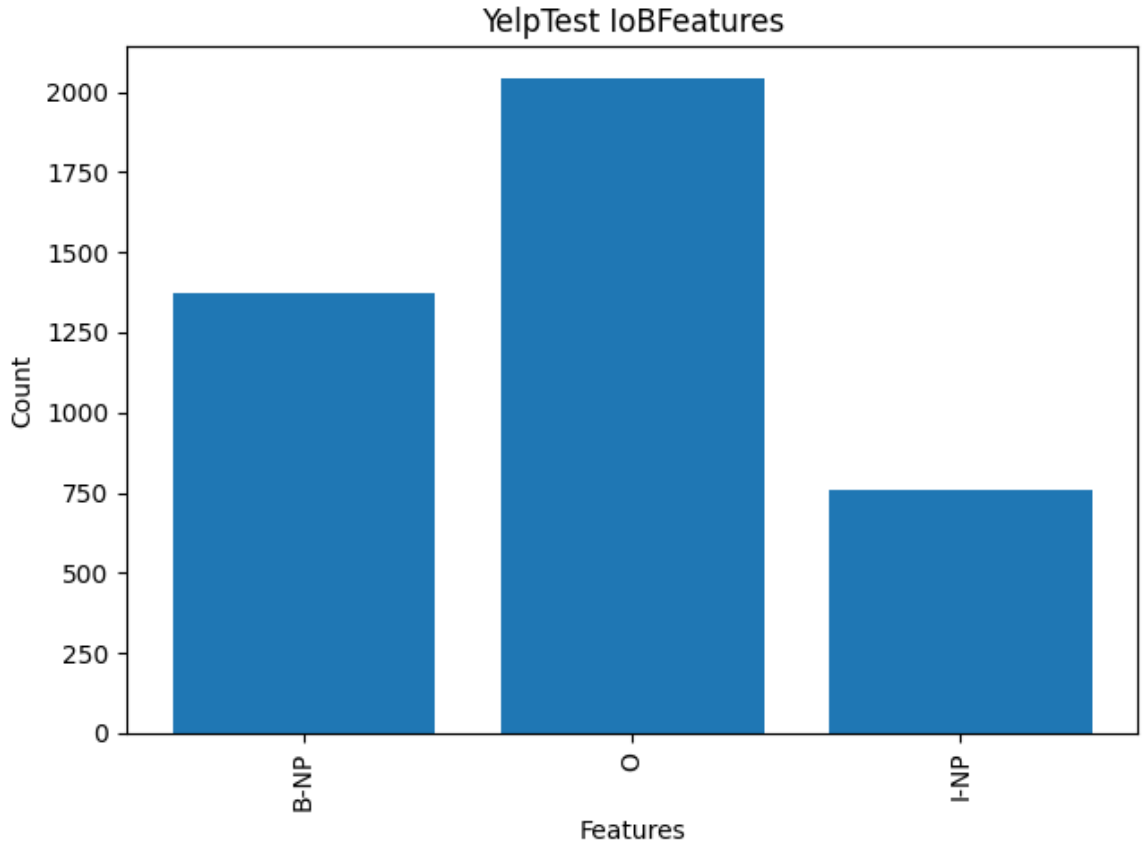


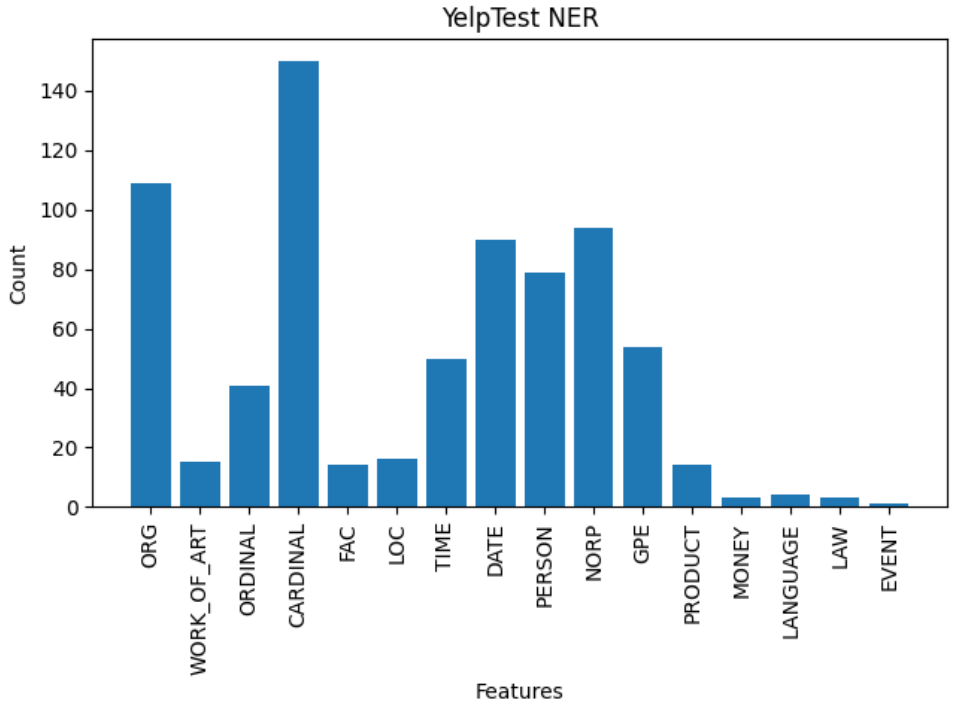

YelpTrain NER
Count
1000
800
600
400
200
0
NORP
CARDINAL
DATE
ORG
LANGUAGE
LOC
GPE
PERSON
ORDINAL
TIME
MONEY
WORK_OF_ART
QUANTITY
FAC
PRODUCT
EVENT
LAW
PERCENT
Features

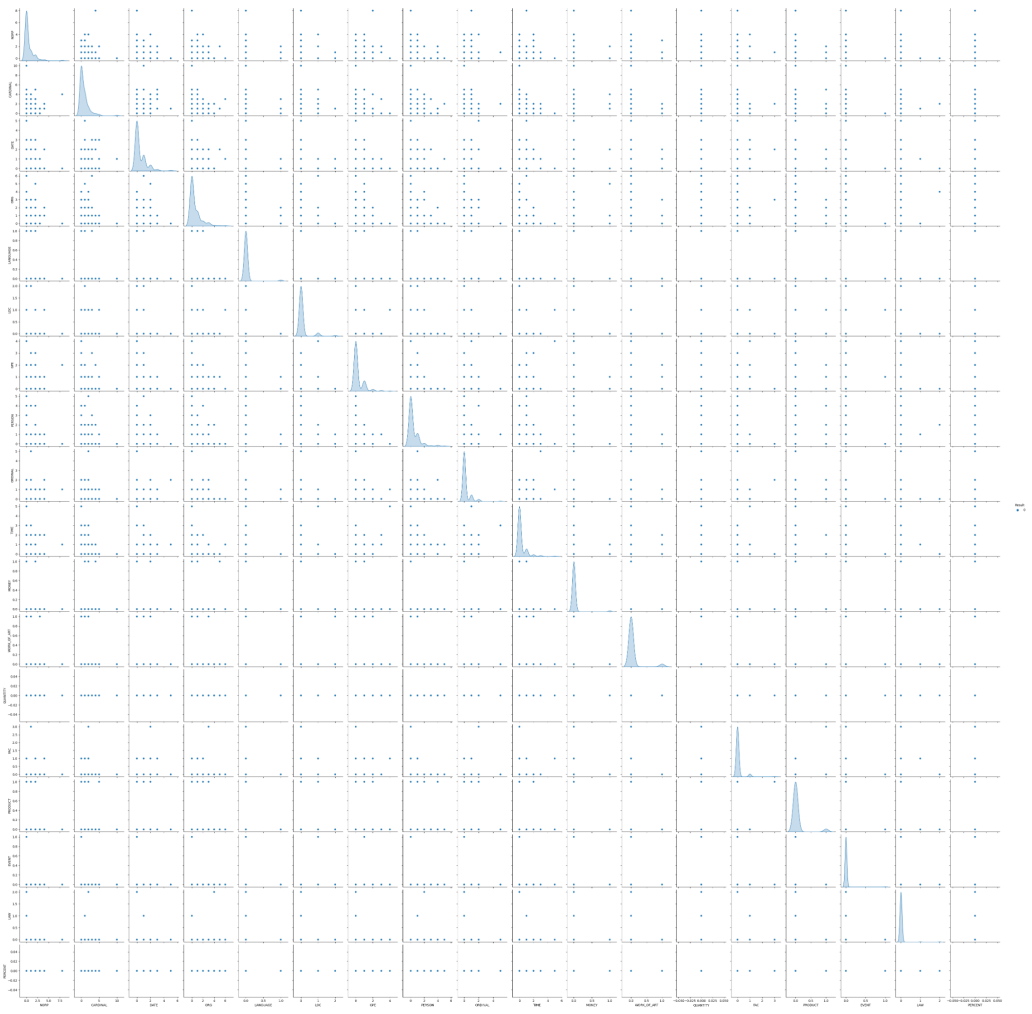


2.2- Yelp Test

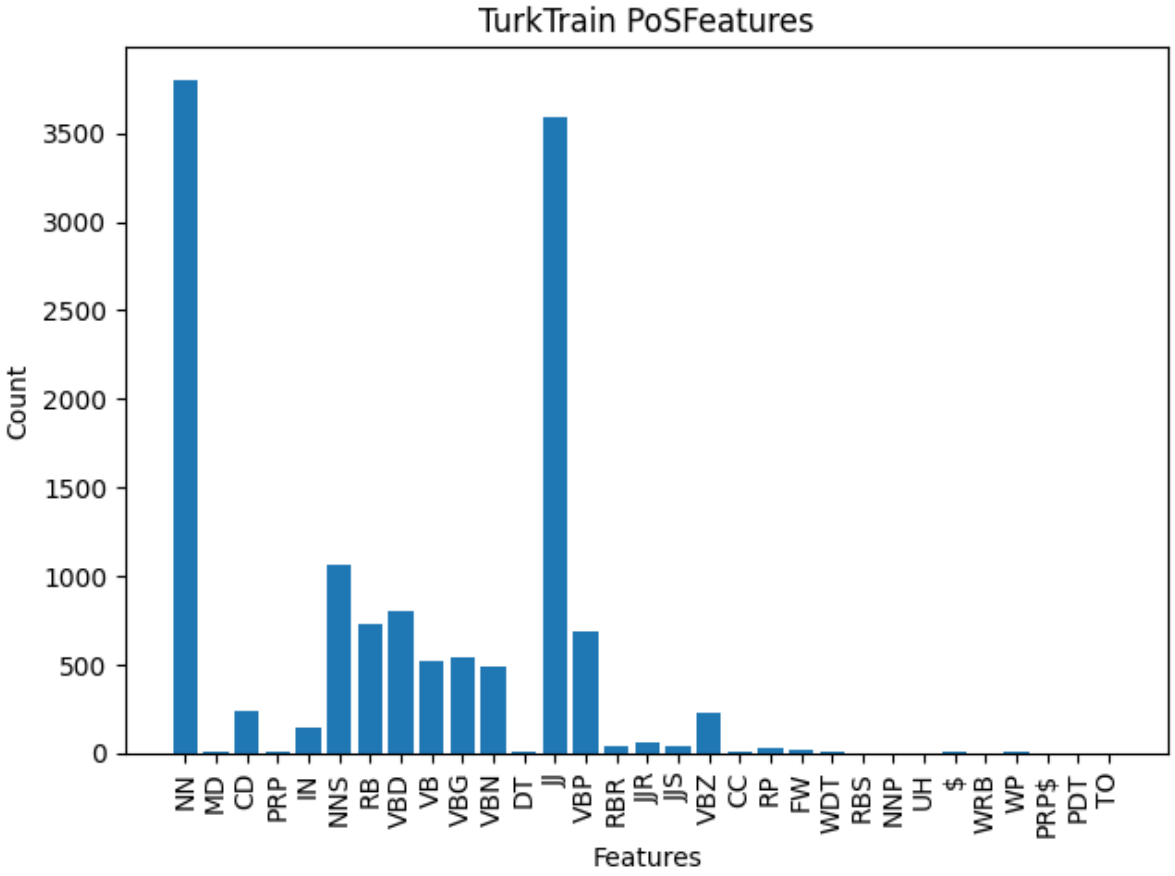

YelpTest PoSFeatures
Count
Features
0
200
400
600
800
1000
1200
1400
NN
RB
JJ
IN
MD
CD
NNS
VBD
PRP
JJS
VB
DT
VBG
VBP
VBN
RBR
VBZ
JJR
RP
CC
RBS
WP
FW
EX
UH
NNP
WDT

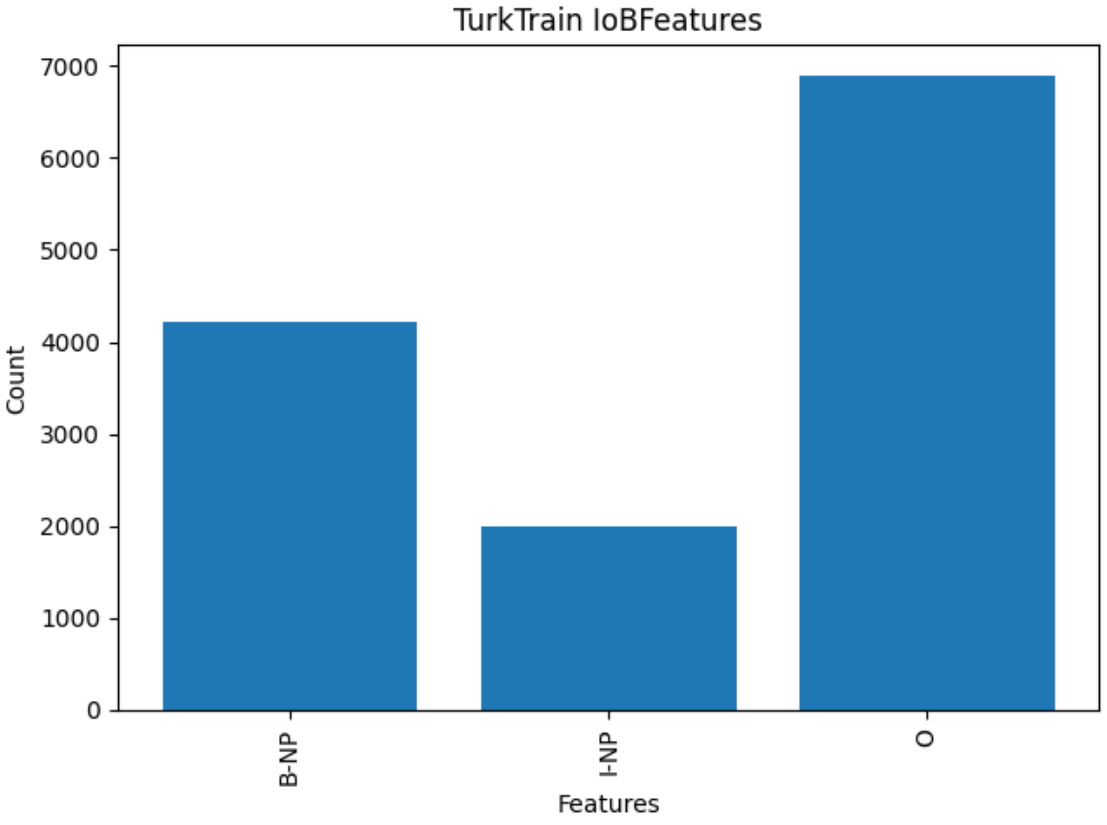

YelpTest IoBFeatures
Count
Features
0
250
500
750
1000
1250
1500
1750
2000
B-NP
O
I-NP

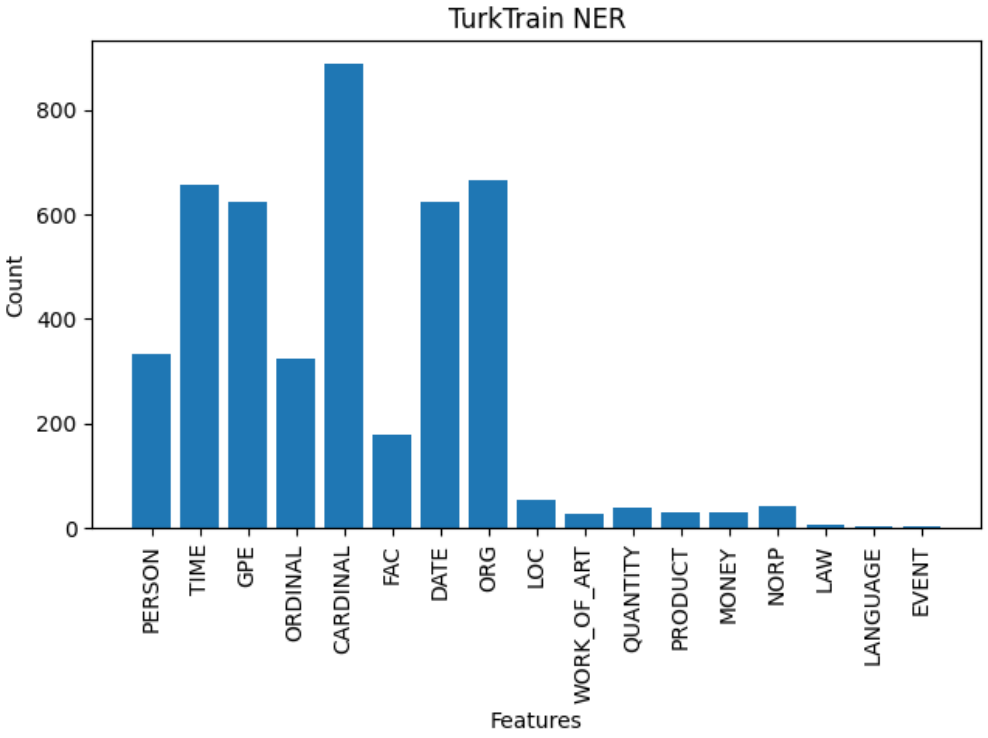

YelpTest NER
Count
0
20
40
60
80
100
120
140
ORG
WORK_OF_ART
ORDINAL
CARDINAL
FAC
LOC
TIME
DATE
PERSON
NORP
GPE
PRODUCT
MONEY
LANGUAGE
LAW
EVENT
Features

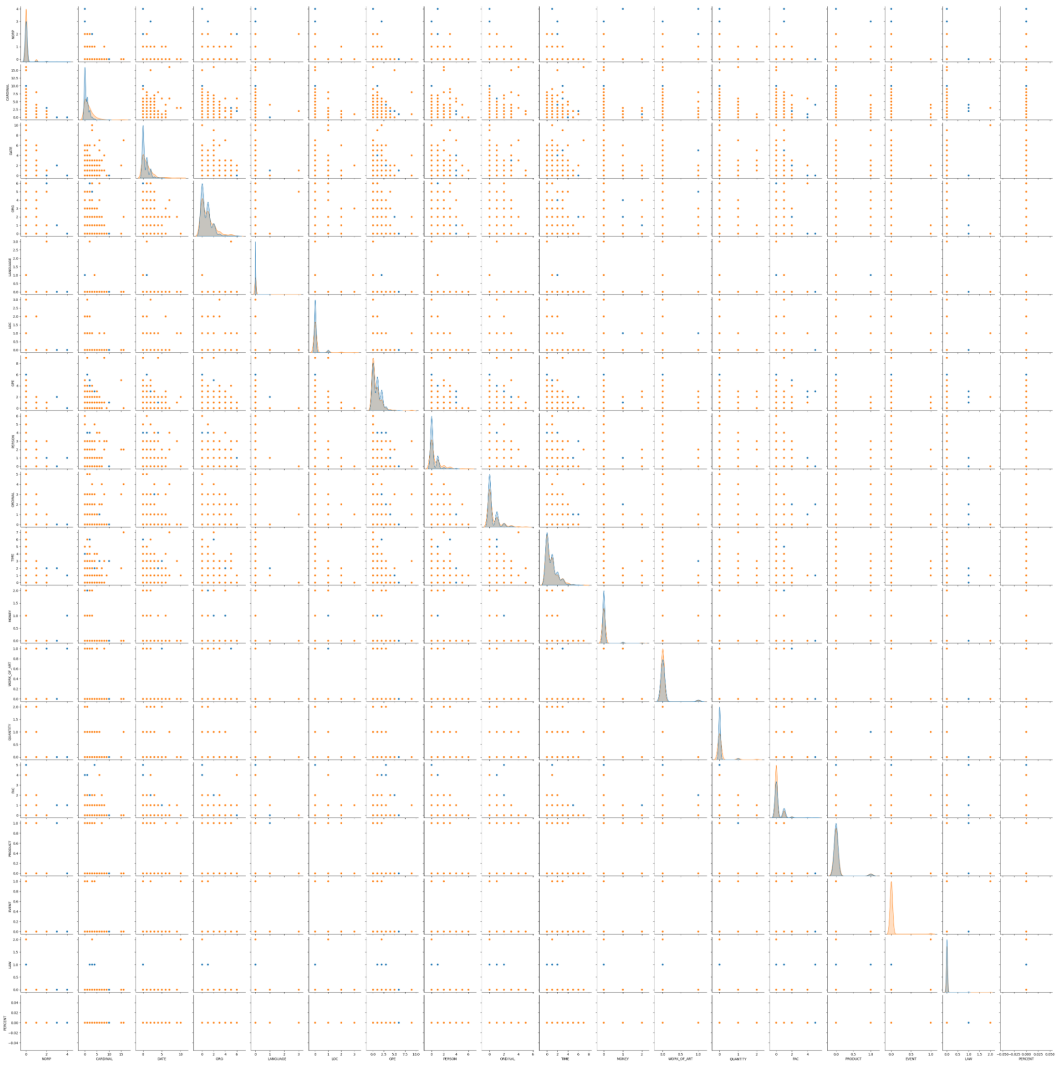


2.3- Turk Train

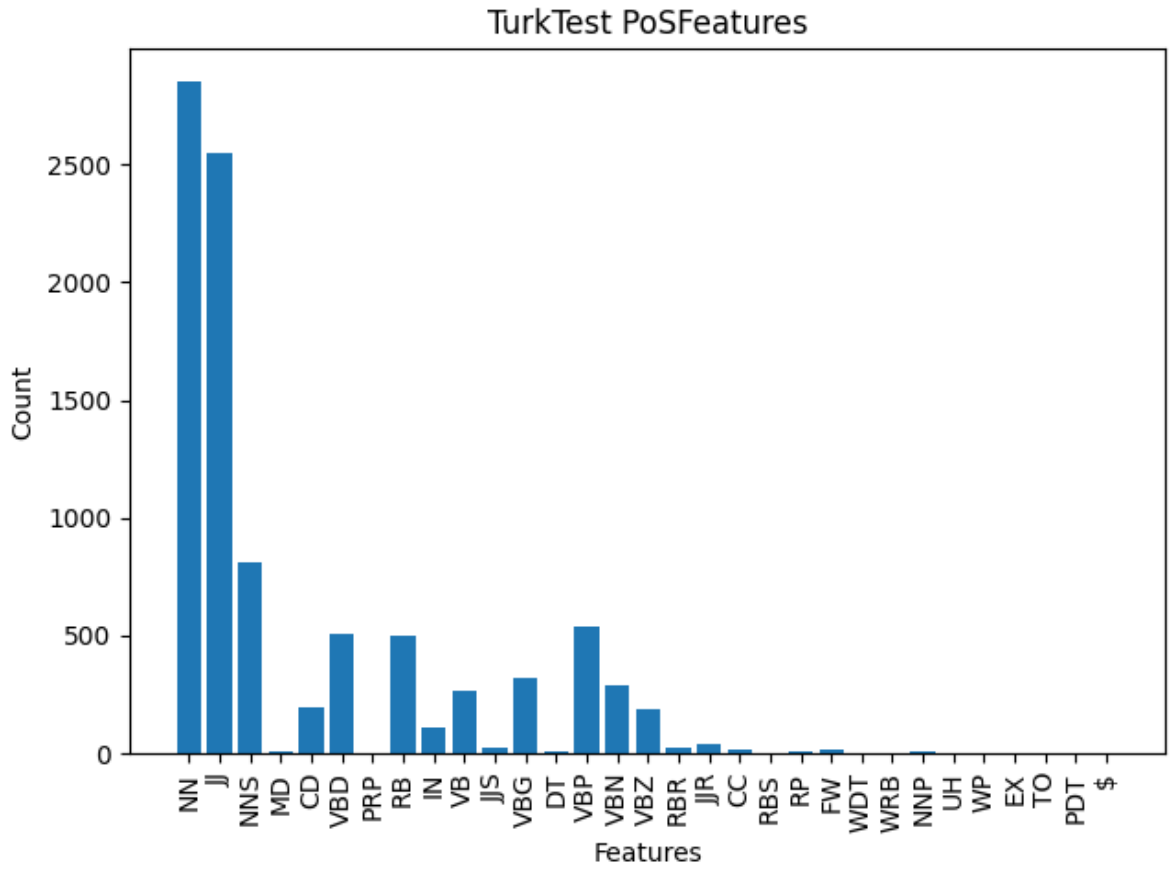


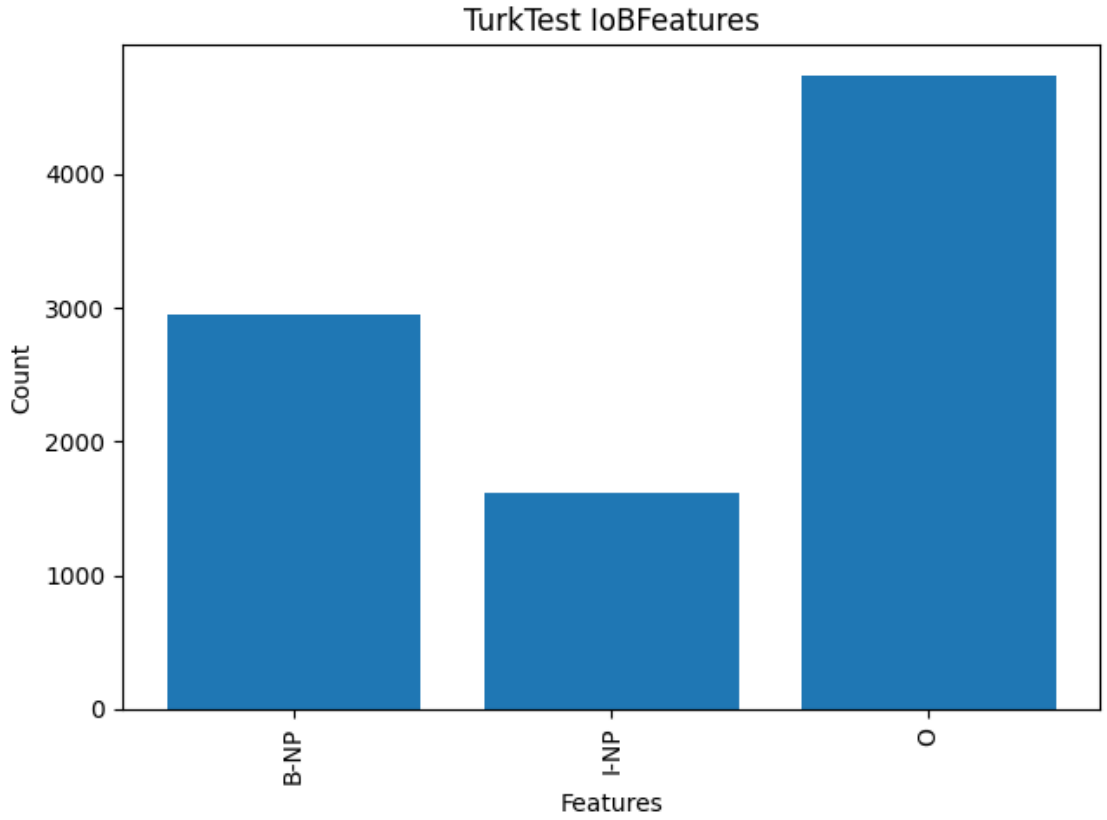


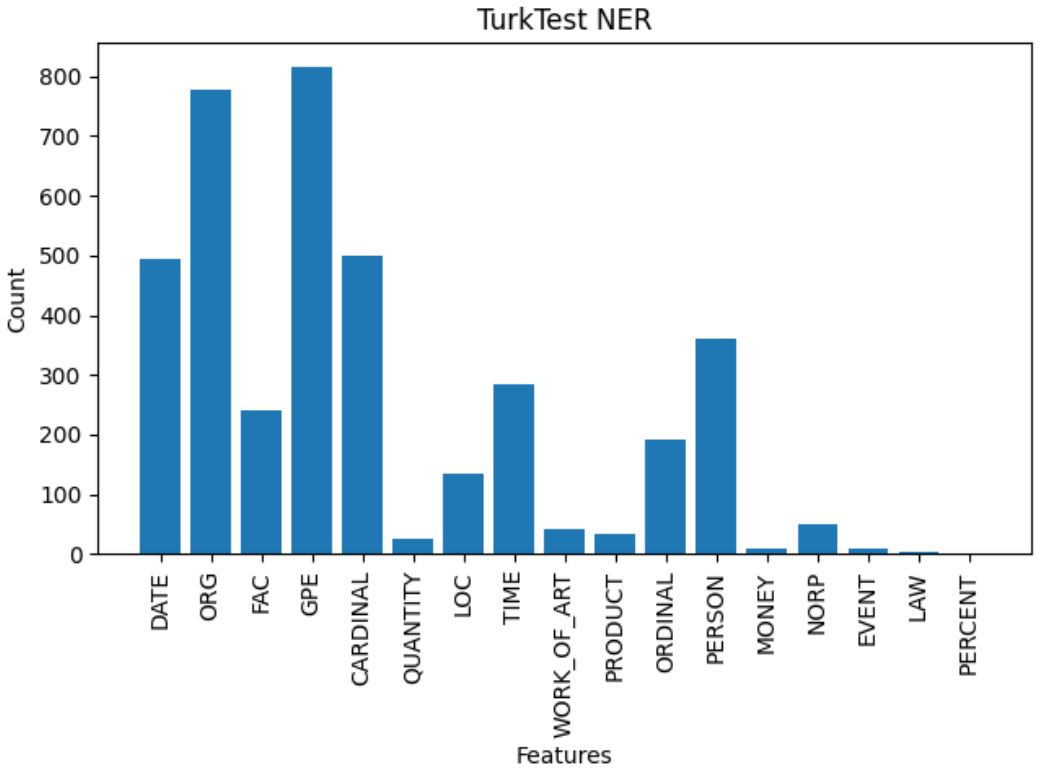

TurkTrain NER
Count
800
600
400
200
0
PERSON
TIME
GPE
ORDINAL
CARDINAL
FAC
DATE
ORG
LOC
WORK_OF_ART
QUANTITY
PRODUCT
MONEY
NORP
LAW
LANGUAGE
EVENT
Features



2.4- Turk Test


TurkTest PoSFeatures
Count
0
500
1000
1500
2000
2500
NN
JJ
NNS
MD
CD
VBD
PRP
RB
IN
VB
JJS
VBG
DT
VBP
VBN
VBZ
RBR
JJR
CC
RBS
RP
FW
WDT
WRB
NNP
UH
WP
EX
TO
PDT
$
Features


TurkTest IoBFeatures
Count
0
1000
2000
3000
4000
B-NP
I-NP
O
Features


TurkTest NER
Count
0
100
200
300
400
500
600
700
800
DATE
ORG
FAC
GPE
CARDINAL
QUANTITY
LOC
TIME
WORK_OF_ART
PRODUCT
ORDINAL
PERSON
MONEY
NORP
EVENT
LAW
PERCENT
Features