T.C MARMARA UNIVERSITY FACULTY OF ENGINEERING

Course Code: CSE4062 Introduction To Data Science

Semester: 2021 Spring

Group Number: 4

Delivery: Delivery #4- Predictive Analytics

Title Of Project: Detecting Fake News

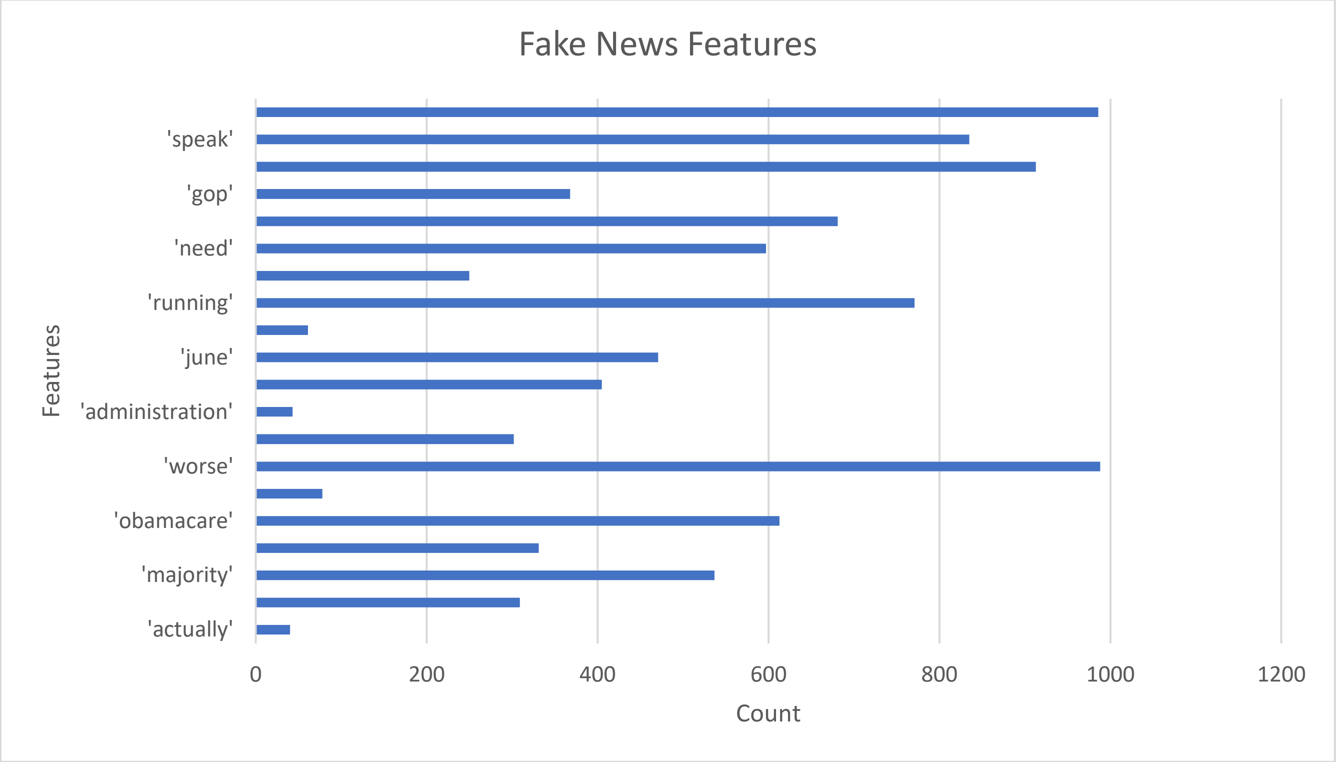
Group Members:

150319629	Seda Nur Yıldız	Industrial Engıneering	sedaanur.yildiz@gmail.com
150319616	Şaziye Eren	Industrial Engıneering	szyeren@gmail.com
150415051	Muhammet Salman	Mechanical Engineering	muhammetsalman69@gmail.com
150319559	Fatih Özudok	Industrial Engineering	fatihozudok@gmail.com
150116042	Celil Mete	Computer Engineering	celilmete64@gmail.com

Statistics

We have a text-based dataset. Our all data is features. We choose 20 attributes.

‘Working, speak, trust, gop, potential, need, did, running, announced, June, history, administration, executive, worse, attack, Obamacare, finally, majority, fact, actually’



Classifacition

method	pac	pac2	Mnb	mnb3	sgdc	sgdc4
f1_score macro	0.5462827097089753	0.8886666130711891	0.5781526545441801	0.8428091643791462	0.5236880709579157	0.8973948470048839
f1_score micro	0.5469613259668509	0.8887134964483031	0.5951065509076559	0.8429360694554064	0.5240726124704025	0.8973954222573007
accuracy score	54.7%	88.87%	59.51%	84.29%	52.41%	89.74%
featuere selection	count vectizer	tfidf vectorizer	count vectizer	tfidf vectorizer	count vectizer	tfidf vectorizer

1.1. PAC Method

F1 Score Macro : 0.546282
F1 Score Micro : 0.546961
Accuracy Score : 54.7
Feature Selection : count vectizer

1.2. PAC 2 Method

F1 Score Macro : 0.888666
F1 Score Micro : 0.888713
Accuracy Score : 88.87
Feature Selection : tfidf vectizer

1.3. MNB Method

F1 Score Macro : 0.578152
F1 Score Micro : 0.59510
Accuracy Score : 59.51
Feature Selection : count vectizer

1.4. MNB 2 Method

F1 Score Macro : 0.842809
F1 Score Micro : 0.842936

Accuracy Score : 84.29
Feature Selection : tdidf vectizer

1.5. SGDC Method

F1 Score Macro : 0.523688
F1 Score Micro : 0.524072
Accuracy Score : 52.41
Feature Selection : count vectizer

1.6. SGDC 2 Method

F1 Score Macro : 0.897394
F1 Score Micro : 0.897395
Accuracy Score : 89.74
Feature Selection : tdifd vectizer

Confusion Matrix

	Accuracy Score	F1 Score Macro	F1 Score Micro
PAC	54.7	0.54628	0.54696
PAC2	88.87	0.88866	0.88871
MNB	59.51	0.5781	0.5951
MNB3	84.29	0.84280	0.84293
SGDC	52.41	0.52368	0.52407
SGDC4	89.74	0.897394	0.897395

Result of the best model is above. Algorithm of the best resulting model is SGD Classifier. TfIdf vectorizer is used for the model. Accuracy of the model is 89.74%.

Description of Results

Our dataset is a text dataset containing different news which are either real or fake. The dataset is labeled as “FAKE” or “REAL”.

We used 3 different algorithms and 2 different feature selection methods. We tested 6 different models. As we use text-based dataset we cannot change the features. This limits our flexibility.

Among the models as it can be seen from the tables above SGD Classifier with tfidf vectorizer gave the best results.
As SGD classifier is a linear model it works fast. The complexity of the algorithm is O(knp) where n,p are the matrix size and the k is number of epochs.

According to our experience from this project tfidf vectorizer gives better results than count vectorizer for all three algorithms.