

T.C MARMARA UNIVERSITY FACULTY OF ENGINEERING

Course Code: CSE4062 Introduction To Data Science

Semester: 2021 Spring

Group Number: 4

Delivery: Delivery #3- Exploring Data Part 2

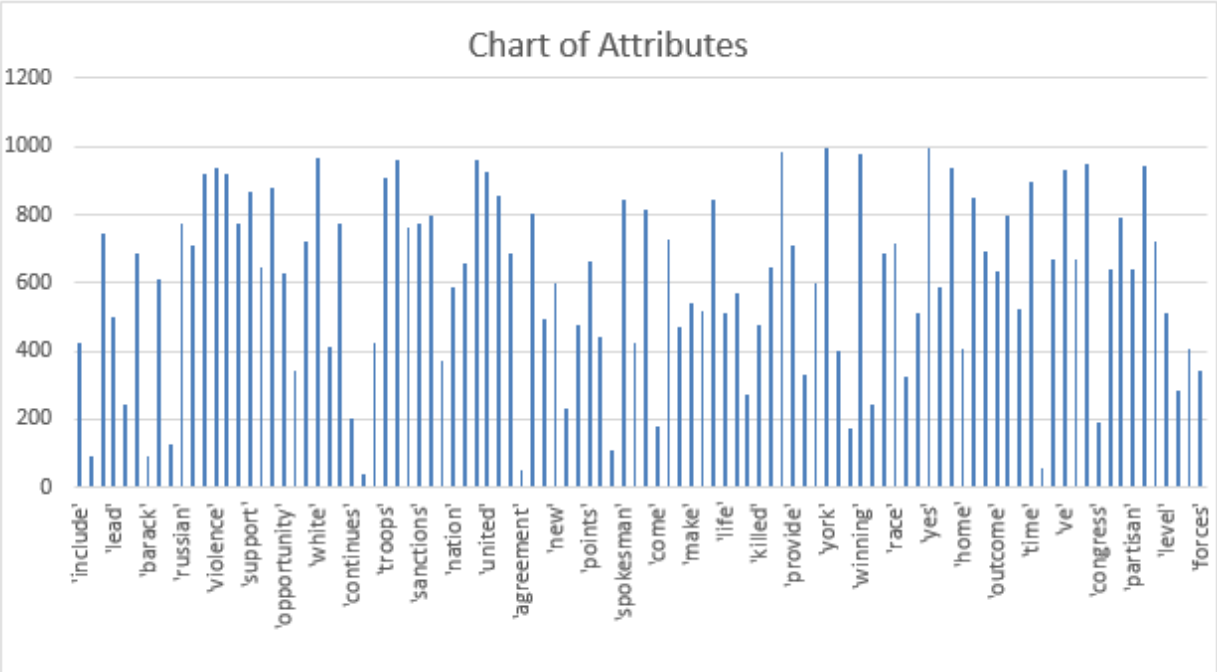
Title Of Project: Detecting Fake News

Group Members:

150319629	Seda Nur Yıldız	Industrial Engineering	sedaanur.yildiz@gmail.com
150319616	Şaziye Eren	Industrial Engineering	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

Chart Of Our Attributes:

'good'	367	'actually'	40	'force'	339
'doing'	263	'fact'	309	'nomination'	604
'course'	208	'majority'	537	'secretary'	790
'focused'	336	'2014'	17	'self'	794
'using'	929	'finally'	331	'field'	325
'case'	136	'obamacare'	613	'attention'	81
'media'	555	'attack'	78	'following'	338
'explain'	306	'worse'	988	'expected'	303
'able'	31	'executive'	302	'view'	937
'washington'	953	'administration'	43	'article'	73
'fall'	313	'history'	405	'look'	525
'2008'	12	'june'	471	'll'	521
'crisis'	216	'announced'	61	'second'	788
'things'	893	'running'	771	'half'	382
'really'	728	'did'	250	'supporting'	871
'politicians'	667	'need'	597	'wilson'	973
'living'	520	'potential'	681	'great'	374
'bush'	123	'gop'	368	'movement'	584
'energy'	286	'trust'	913	'progressive'	702
'hope'	410	'didn't'	251	'response'	754
'change'	147	'speak'	835	'americans'	60
'turned'	919	'working'	986	'mcconnell'	551
'11'	3	'saying'	782	'wrote'	993
'muslim'	587	'china'	153	'society'	826
'america'	58	'told'	902	'common'	185
'poor'	672	'say'	781	'hand'	384
'health'	393	'months'	581	'today'	901
'care'	133	'later'	493	'george'	357
'base'	93	'speaker'	836	'joe'	464
'worst'	989	'john'	465	'20'	11
'confirmed'	192	'efforts'	277	'leadership'	500
'2009'	13	'freedom'	347	'nominee'	605
'responsibility'	755	'caucus'	138	'possible'	678
				'ted'	885



Scatter Plot

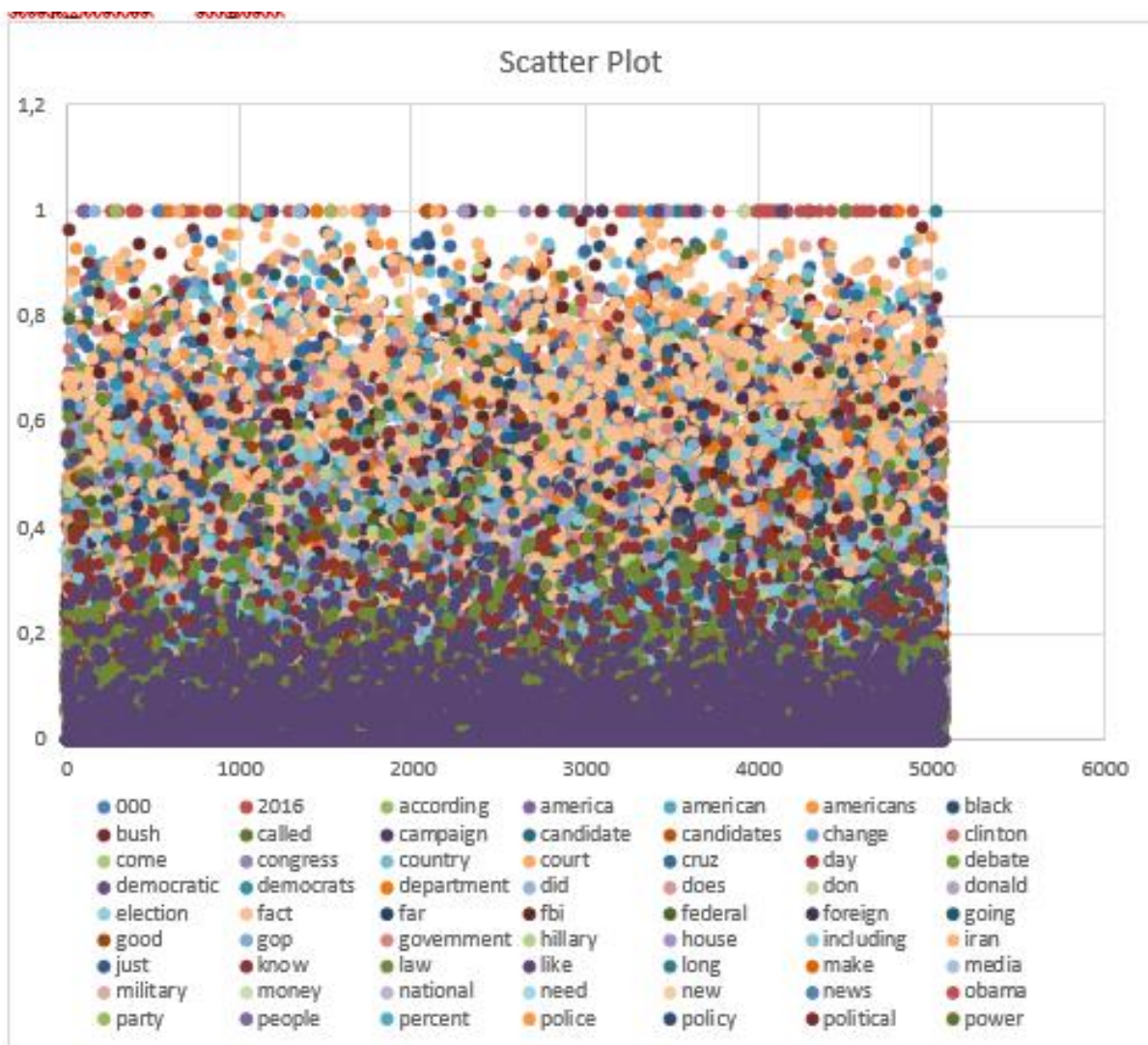
Below is the scatter plot of the inverse document frequencies of most used 100 words in our dataset.

The data for the graph is gathered with - sklearn tfidf_vectorizer -class.

Max_df = 0.7

Max_features = 100

Stop_words = "english"



Result

We investigated our dataset and extracted above features from dataset.

Since we are using a text dataset our attributes are the words. And extracted features are inverse document frequencies of the words and counts of the word.

