FAKE NEWS DETECTION

CONTENT

- Introduction
- Dataset and Data Cleaning
- Tools
- EDA
- Topic Modeling
- Recommendation

INTRODUCTION

- We consume news through several mediums throughout the day in our daily routine, but sometimes it becomes difficult to decide which one is fake and which one is authentic.

- Do you trust all the news you consume from online media?

DATASET AND DATA CLEANING

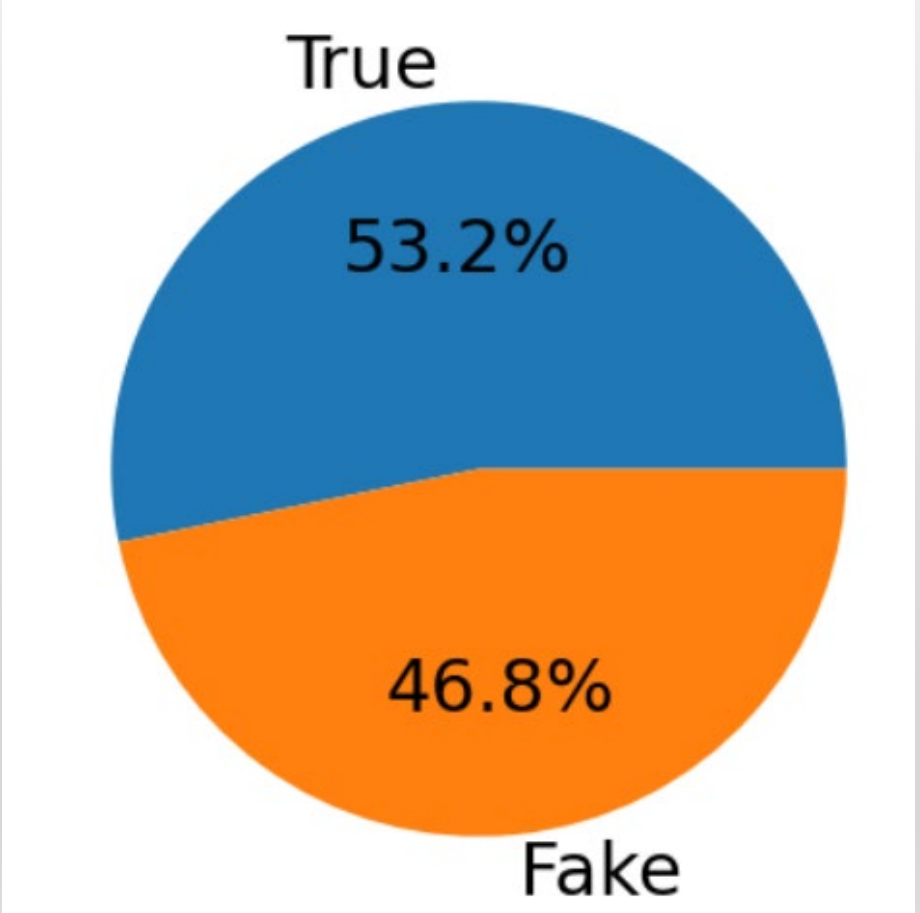
- The dataset has download from Kaggle website
- It has 1000 rows
- It is balanced
- Tokenization
- Stop words
- Lemmatization

TOOLS

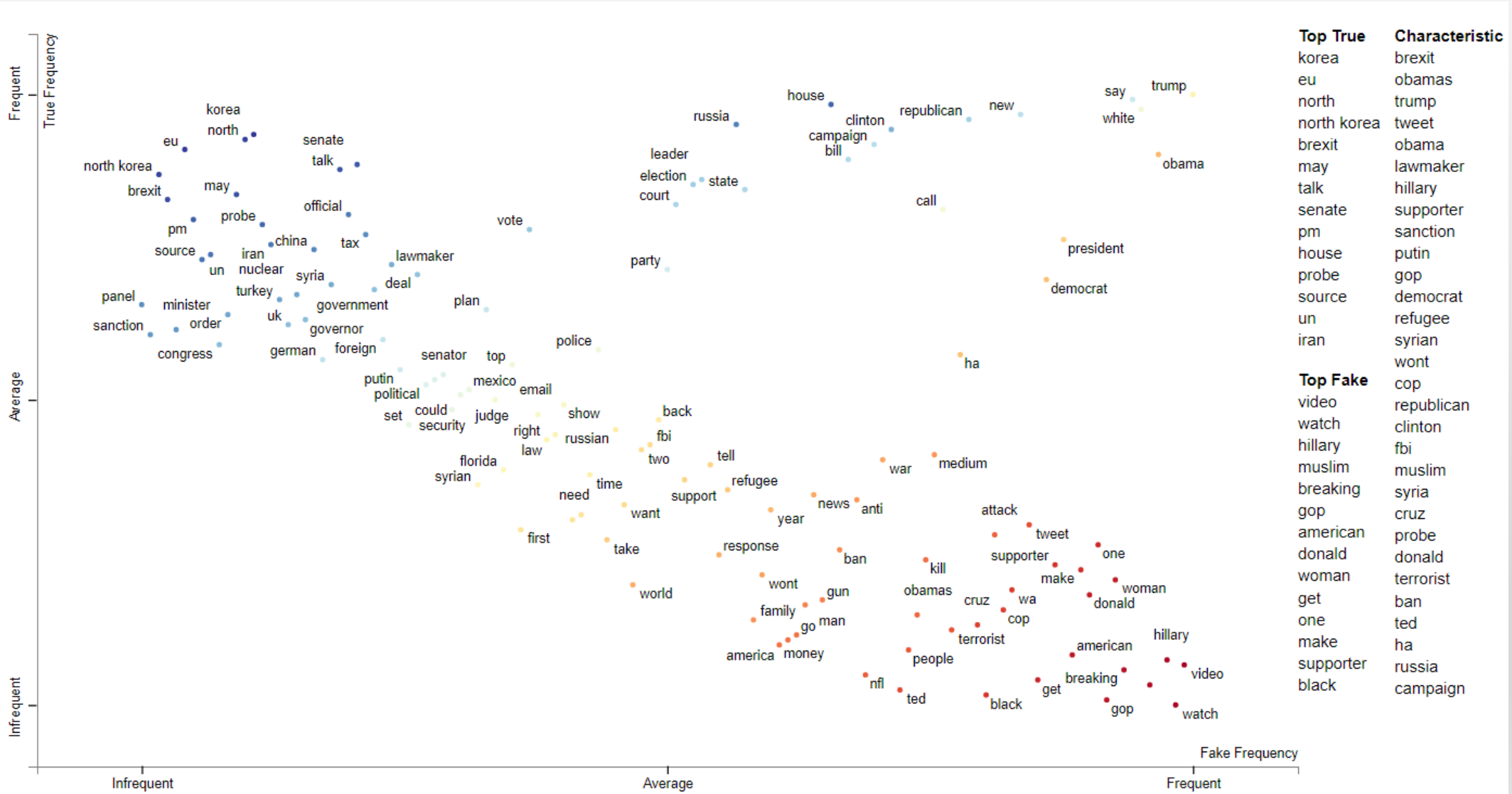
- Pandas
- Matplotlib
- Wordcloud
- Sklearn
- NMF
- CountVectorizer, TfidfVectorizer
- CorEx
- LSA

EDA

- The dataset has 532 true news and 468 fake news, which indicates that the dataset is balanced



EDA



Top True	Characteristic
korea	brexit
eu	obamas
north	trump
north korea	tweet
brexit	obama
may	lawmaker
talk	hillary
senate	supporter
pm	sanction
house	putin
probe	gop
source	democrat
un	refugee
iran	syrian
	wont
Top Fake	cop
video	republican
watch	clinton
hillary	fbi
muslim	muslim
breaking	syria
gop	cruz
american	probe
donald	donald
woman	terrorist
get	ban
one	ted
make	ha
supporter	russia
black	campaign

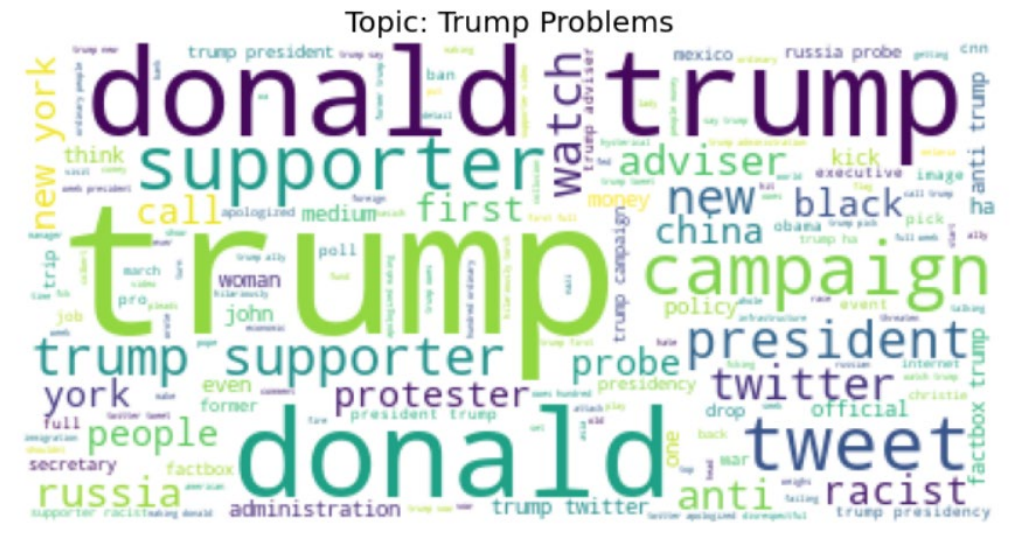
True document count: 532; word count: 4,209
Fake document count: 468; word count: 4,825

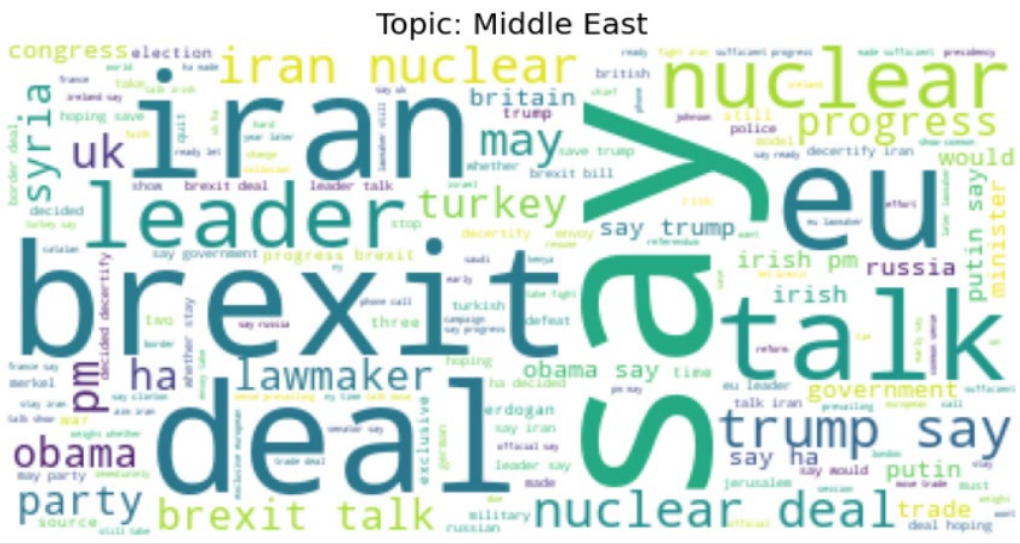
TOPIC MODELING

Several topic modeling algorithms was tried, including:

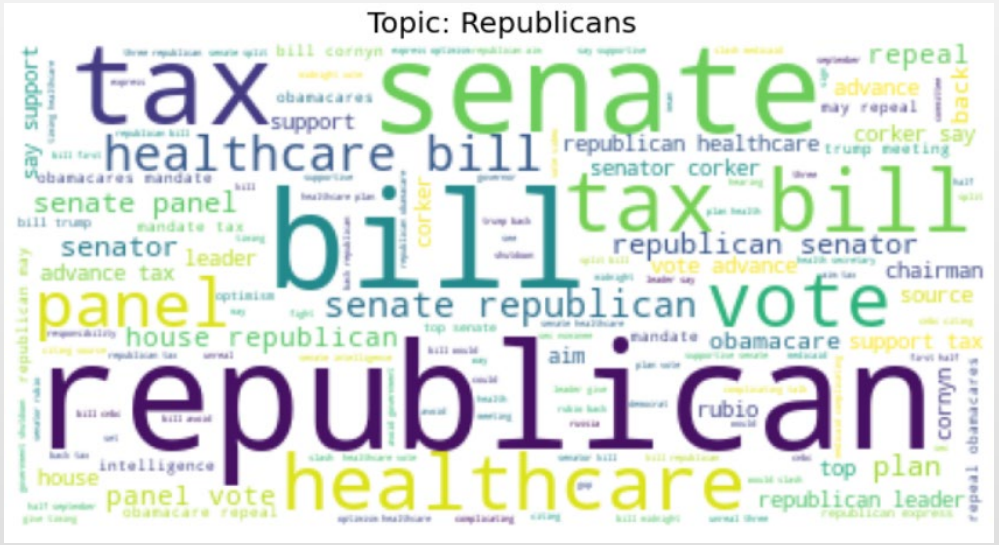
- LSA
- CorEx
- NMF (which is the best choice with 7 topics)

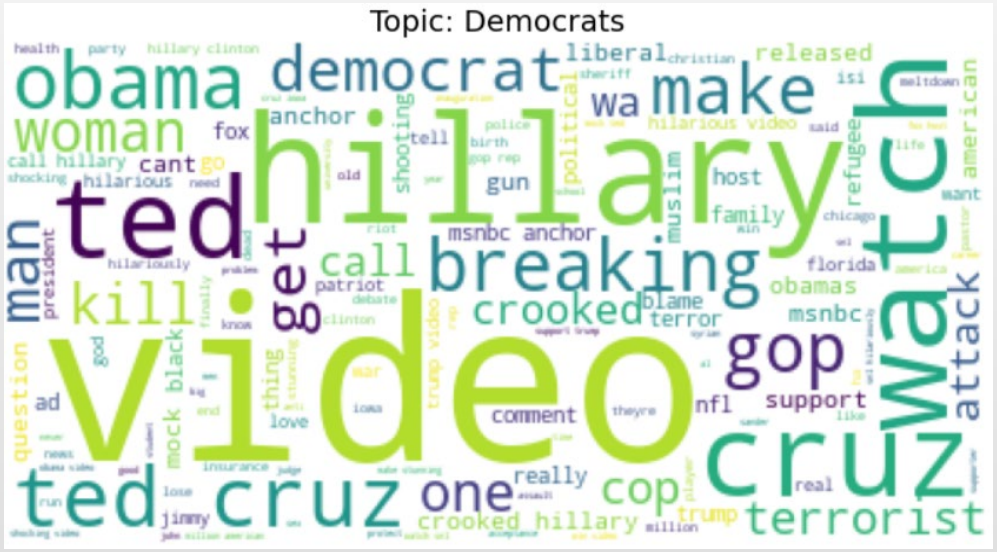
TOPIC MODELING

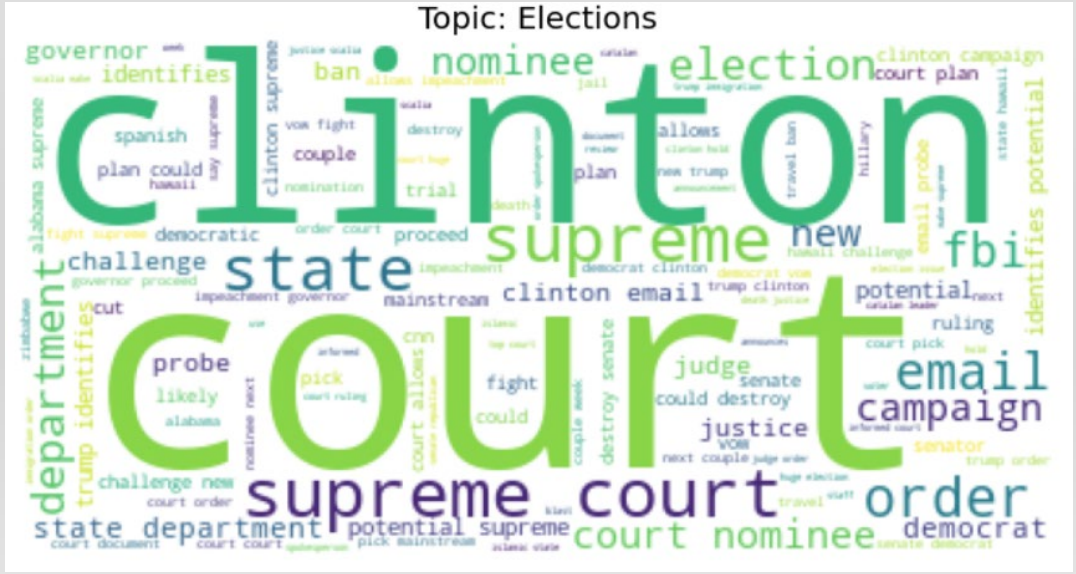




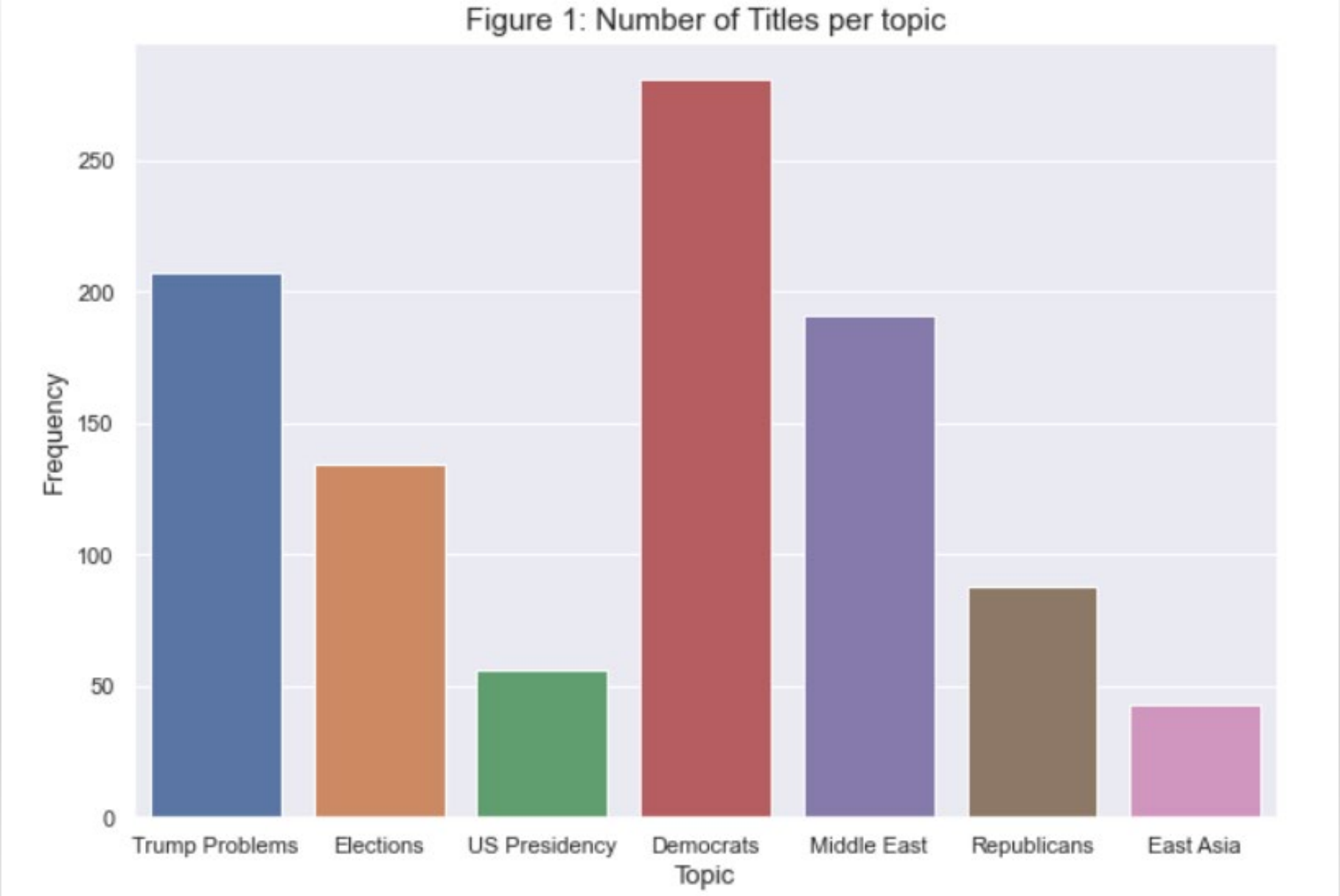
TOPIC MODELING



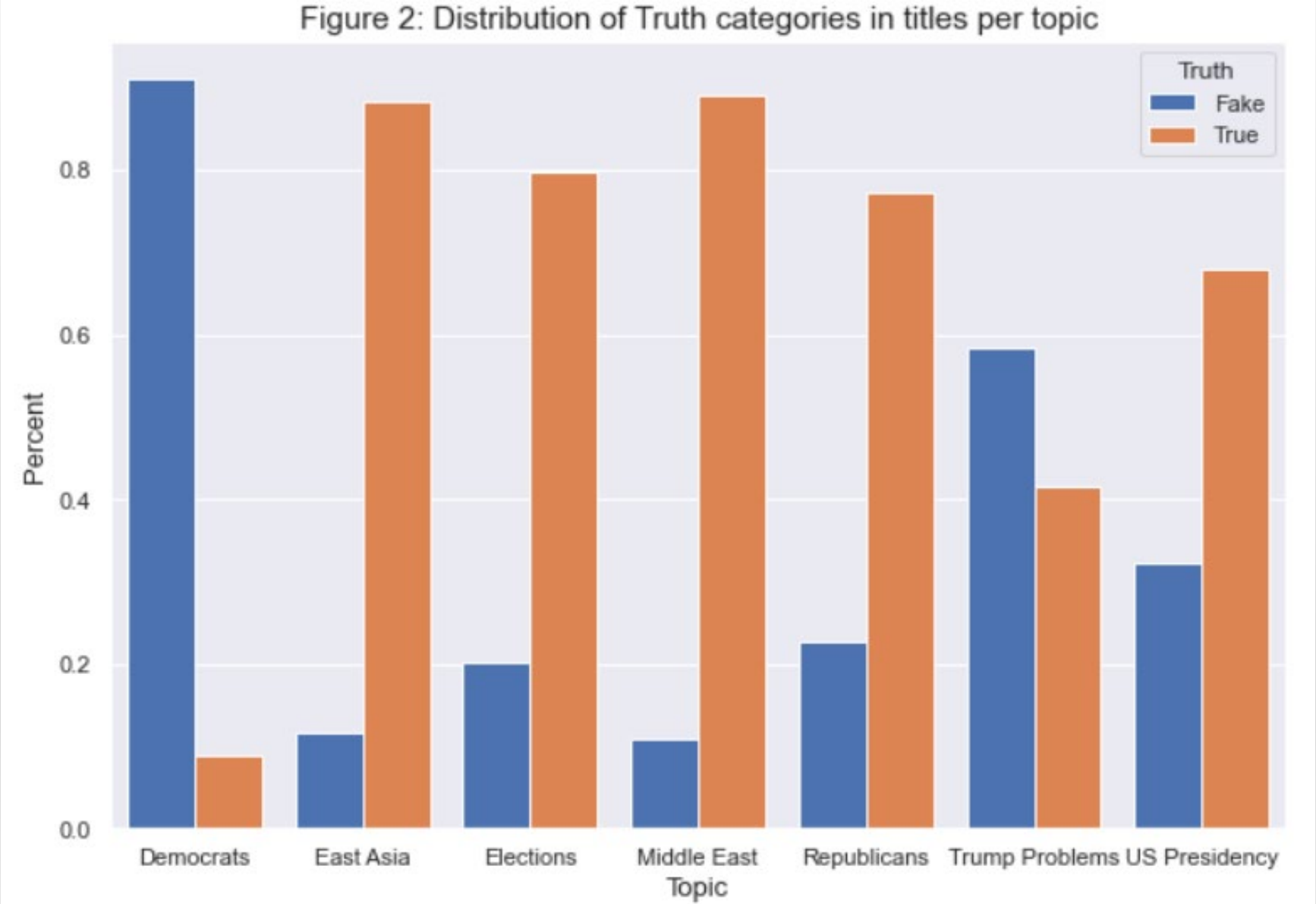




TOPIC MODELING



TOPIC MODELING



RECOMMENDATION

- There are multiple algorithms could be used in topic modeling
- There is an overfitting when performing a classification model on the dataset because of lack of the rows

THANK YOU