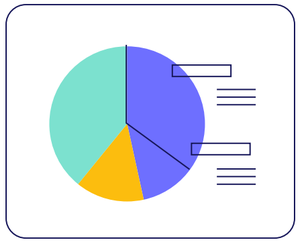
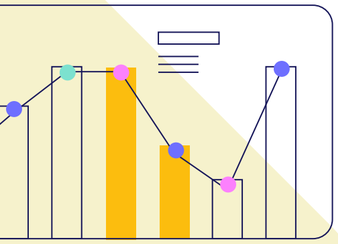


data science 402-403

Data science in social science

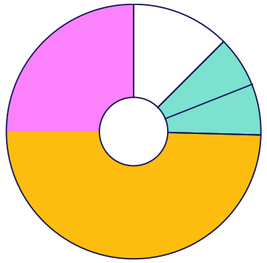
Dr. MohammadMahdi Rezapour

Mohammad Eshagh, Behnoosh Behyani, Kasra Samadi



What is social data science?

Social data science is a field that uses scientific methods, processes, algorithms, and systems to extract insights and knowledge from social data, which refers to data related to social interactions, relationships, and behaviors



The applications of social data science:

01. Predicting human behavior

02. Understanding public opinion

03. Mapping social networks

04. Identifying fake news

05. Improving education

What types of data can be used in social data science?

1. Social Media Data:

- Posts and comments
- Likes, shares, and retweets

2. Survey Data:

- Demographics
- Attitudes and opinions
- Behaviors

3. Demographic Data:

- Population characteristics
- Socioeconomic indicators
- Geographic data

What types of data can be used in social data science?

4. Online Text Data:

- News articles and blogs
- Public forums and discussion boards
- Web scraping

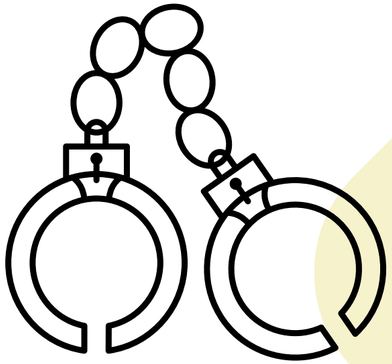
5. Transaction Data:

- Consumer behavior
- Mobility patterns
- Economic activity

01. SF-Crime Analysis & Prediction

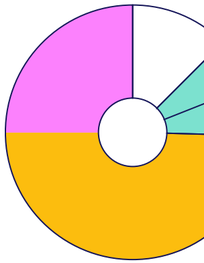
Predict crime category based on hour and address

Introduction



this project analyzes 12 years of crime reports from across all of San Francisco's neighborhoods to create a model that predicts the category of crime that occurred, given time and location.

The data ranges from 1/1/2003 to 5/13/2015 creating a training dataset with nine features and 878,049 samples



dataset features

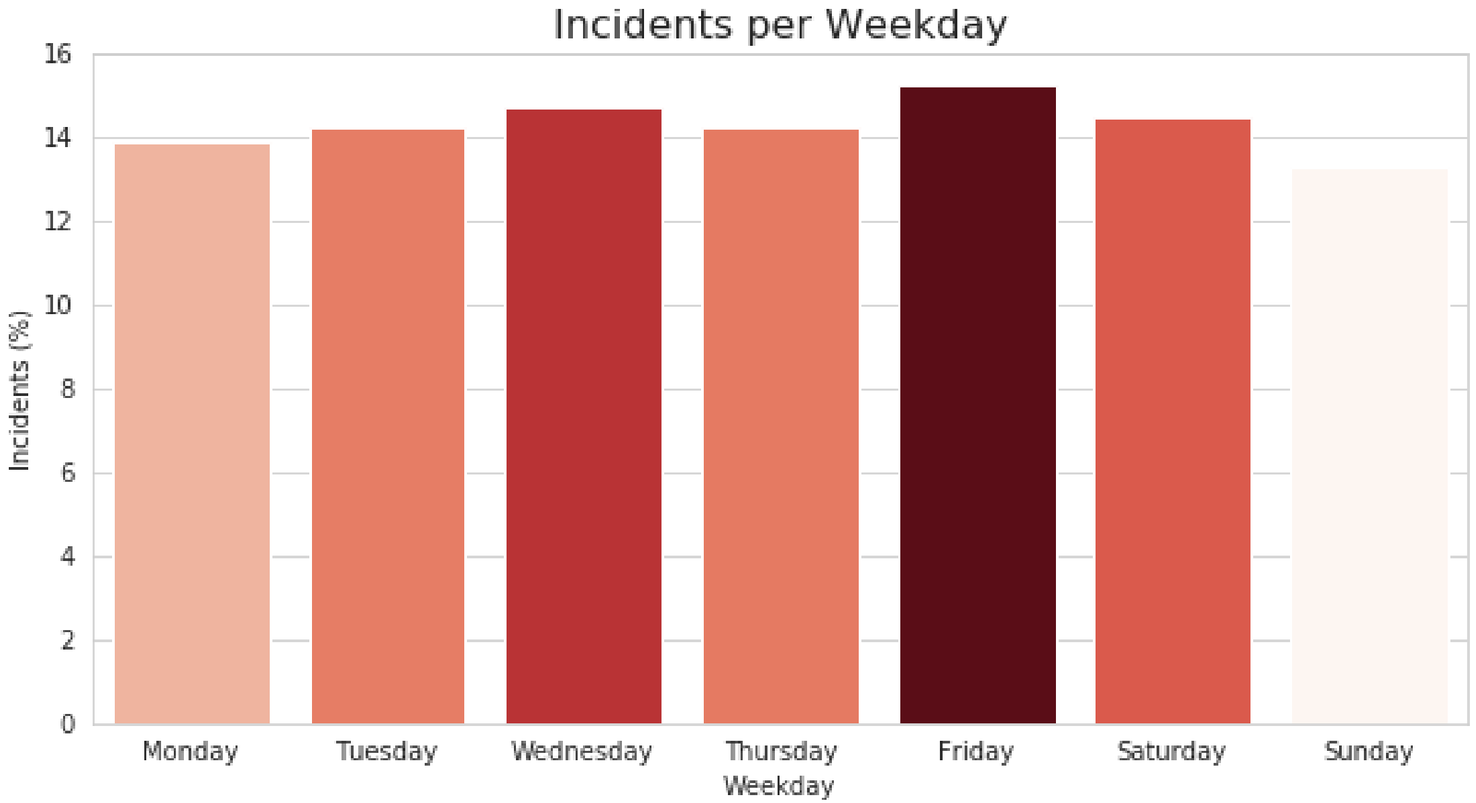
Dates	timestamp of the crime incident
Category	category of the crime incident. (This is our target variable.)
Descript	detailed description of the crime incident
DayOfWeek	the day of the week
PdDistrict	the name of the Police Department District
Resolution	The resolution of the crime incident
Address	the approximate street address of the crime incident
X	Longitude
Y	Latitude

train dataset examples

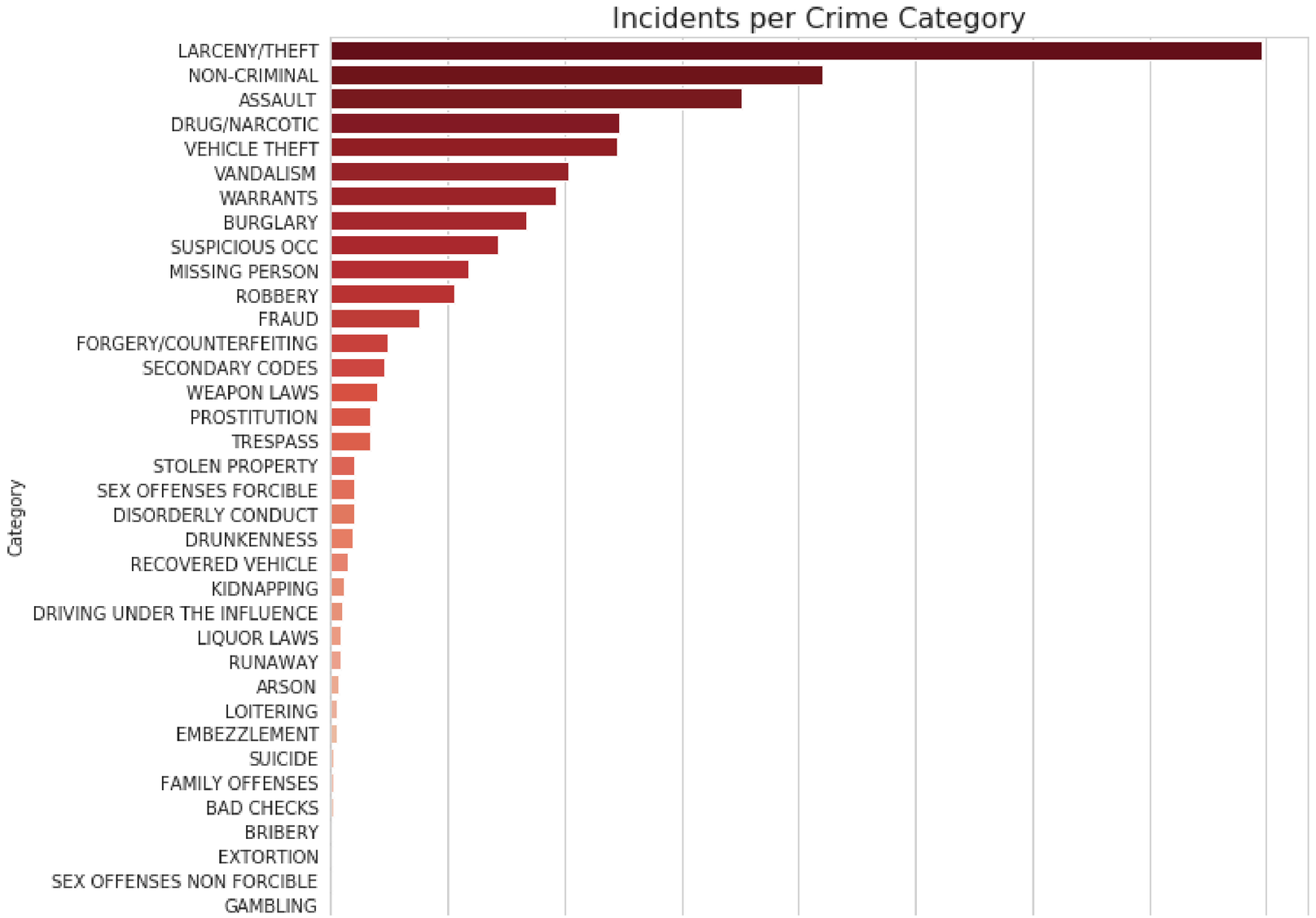
	Dates	Category	Descript	DayOfWeek	PdDistrict	Resolution	Address	X	Y
0	2015-05-13 23:53:00	WARRANTS	WARRANT ARREST	Wednesday	NORTHERN	ARREST, BOOKED	OAK ST / LAGUNA ST	-122.425892	37.774599
1	2015-05-13 23:53:00	OTHER OFFENSES	TRAFFIC VIOLATION ARREST	Wednesday	NORTHERN	ARREST, BOOKED	OAK ST / LAGUNA ST	-122.425892	37.774599
2	2015-05-13 23:33:00	OTHER OFFENSES	TRAFFIC VIOLATION ARREST	Wednesday	NORTHERN	ARREST, BOOKED	VANNESS AV / GREENWICH ST	-122.424363	37.800414
	2015-		GRAND THEFT				1500 Block		

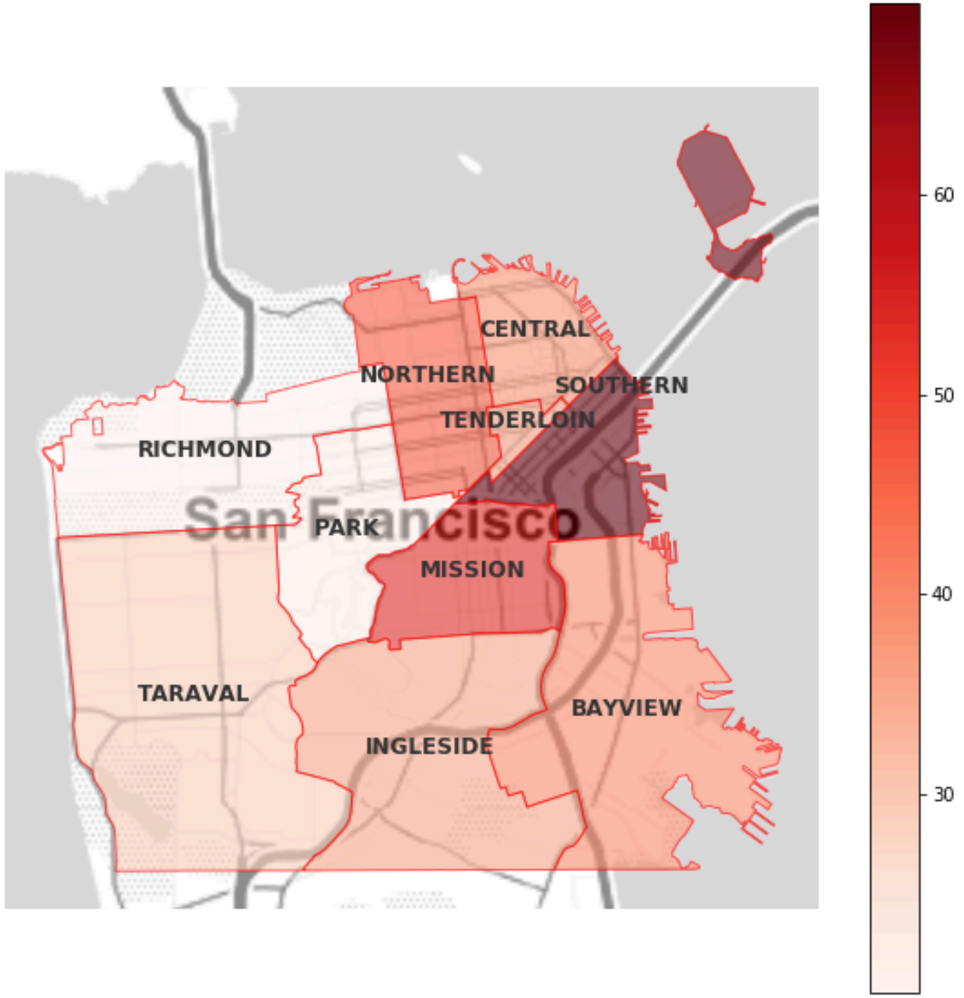
test dataset examples

Id	Dates	DayOfWeek	PdDistrict	Address	X	Y
0	5/10/2015 23:59	Sunday	BAYVIEW	2000 Block	-122.4	37.73505
1	5/10/2015 23:51	Sunday	BAYVIEW	3RD ST / R	-122.392	37.73243
2	5/10/2015 23:50	Sunday	NORTHERI	2000 Block	-122.426	37.79221
3	5/10/2015 23:45	Sunday	INGLESIDE	4700 Block	-122.437	37.72141
4	5/10/2015 23:45	Sunday	INGLESIDE	4700 Block	-122.437	37.72141
5	5/10/2015 23:40	Sunday	TARAVAL	BROAD ST	-122.459	37.71317
6	5/10/2015 23:30	Sunday	INGLESIDE	100 Block	-122.426	37.73935
7	5/10/2015 23:30	Sunday	INGLESIDE	200 Block	-122.413	37.73975
8	5/10/2015 23:10	Sunday	MISSION	2900 Block	-122.419	37.76516



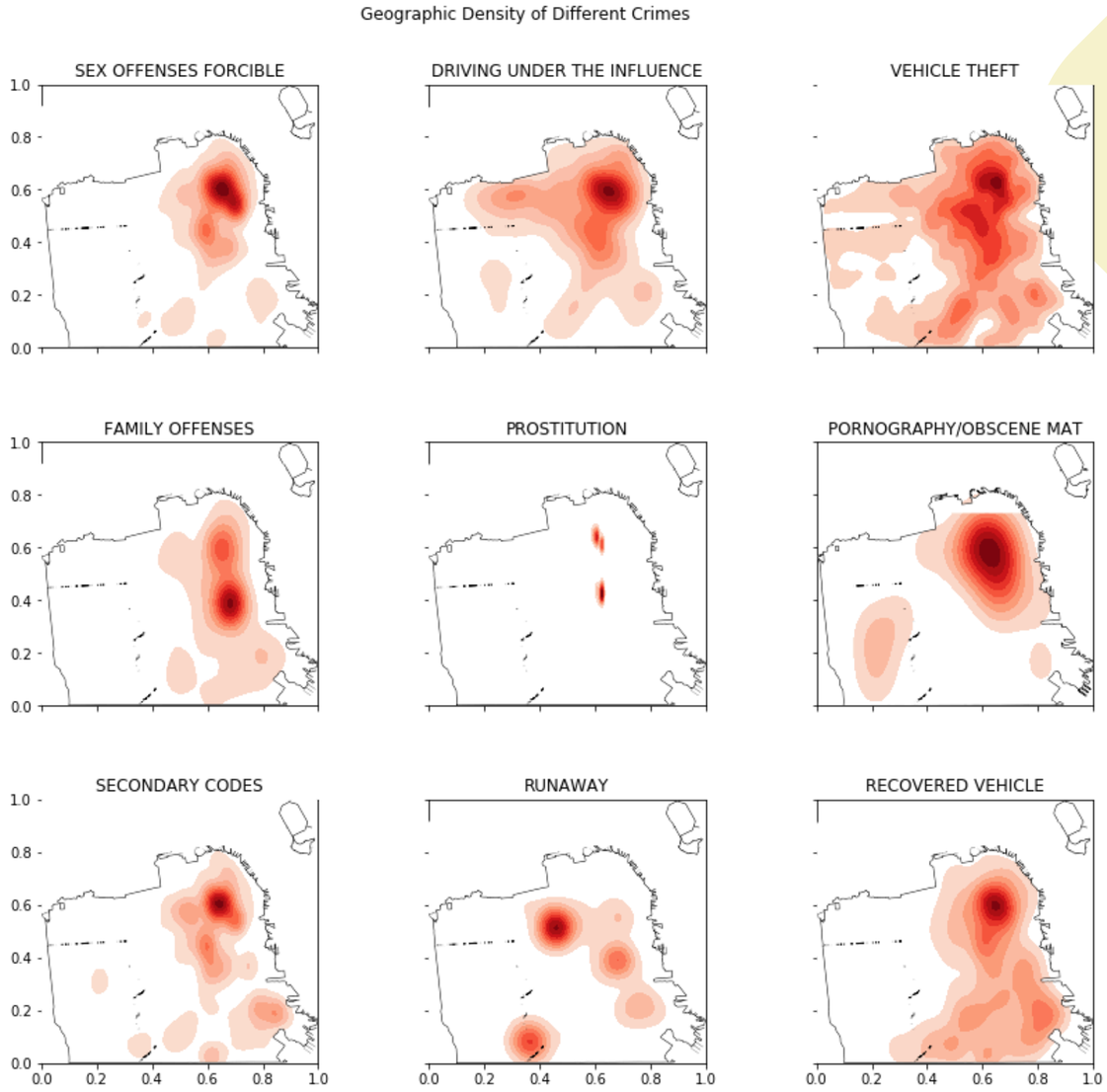
here is no significant deviation of incidents frequency throughout the week. Thus we do not expect this variable to play a significant role in the prediction.


Incidents per Crime Category
Category
LARCENY/THEFT
NON-CRIMINAL
ASSAULT
DRUG/NARCOTIC
VEHICLE THEFT
VANDALISM
WARRANTS
BURGLARY
SUSPICIOUS OCC
MISSING PERSON
ROBBERY
FRAUD
FORGERY/COUNTERFEITING
SECONDARY CODES
WEAPON LAWS
PROSTITUTION
TRESPASS
STOLEN PROPERTY
SEX OFFENSES FORCIBLE
DISORDERLY CONDUCT
DRUNKENNESS
RECOVERED VEHICLE
KIDNAPPING
DRIVING UNDER THE INFLUENCE
LIQUOR LAWS
RUNAWAY
ARSON
LOITERING
EMBEZZLEMENT
SUICIDE
FAMILY OFFENSES
BAD CHECKS
BRIBERY
EXTORTION
SEX OFFENSES NON FORCIBLE
GAMBLING



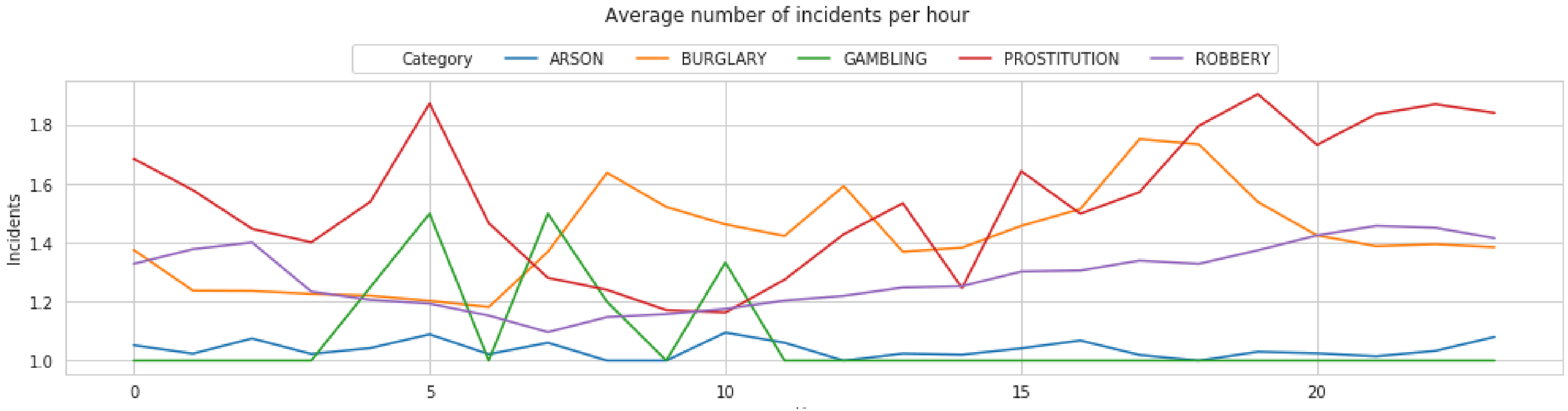
The frequency of incidents registration by different districts

- each crime has a different density on the rest of the city
- the location (coordinates / Police District) will be a significant factor for the analysis and the forecasting



- It is evident that different crimes have different frequency during different times of the day.

- the time parameters will have a significant role also.

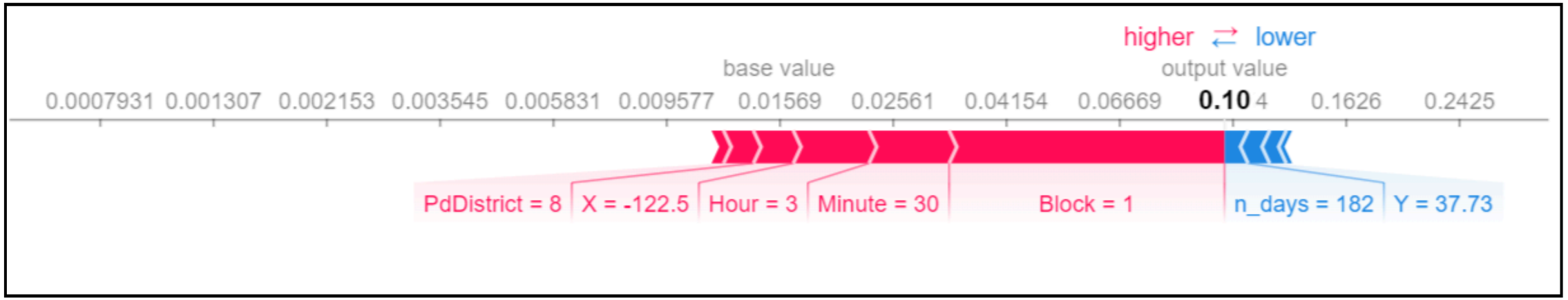


train dataset after finding better features

	Category	DayOfWeek	PdDistrict	X	Y	Hour	n_days	Day	Month	Year	Minute	Block
0	WARRANTS	2	NORTHERN	-122.425892	37.774599	23	4510	13	5	2015	53	False
1	OTHER OFFENSES	2	NORTHERN	-122.425892	37.774599	23	4510	13	5	2015	53	False
2	OTHER OFFENSES	2	NORTHERN	-122.424363	37.800414	23	4510	13	5	2015	33	False
3	LARCENY/THEFT	2	NORTHERN	-122.426995	37.800873	23	4510	13	5	2015	30	True
4	LARCENY/THEFT	2	PARK	-122.438738	37.771541	23	4510	13	5	2015	30	True

sample test for test model validation

	DayOfWeek	PdDistrict	X	Y	n_days	Day	Month	Year	Hour	Minute	Block
Id											
846262	2	8	-122.484968	37.732351	182	2	7	2003	3	30	True



02. Quora Insincere Questions Classification

Detect toxic content to improve online conversations

Insincere questions

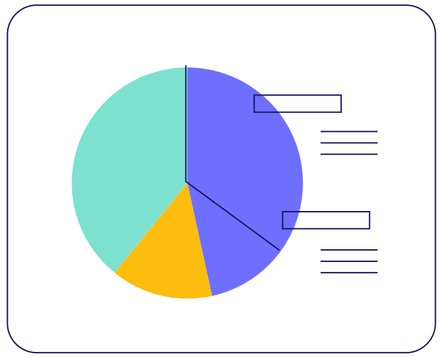
Some characteristics that can signify that a question is insincere:

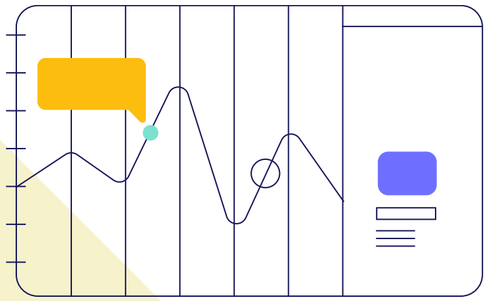
- Has a non-neutral tone
- Is disparaging or inflammatory
- Isn't grounded in reality

Insincere questions examples

- Has the United States become the largest dictatorship in the world?
- Which babies are more sweeter to their parents? Dark skin babies or light skin babies?
- If both Honey Singh and Justin Bieber fall from the 5th floor, who will survive?
- Could the leader of Iran be dead many years ago and the leader of today's Iran is actually a fake leader?

Dataset





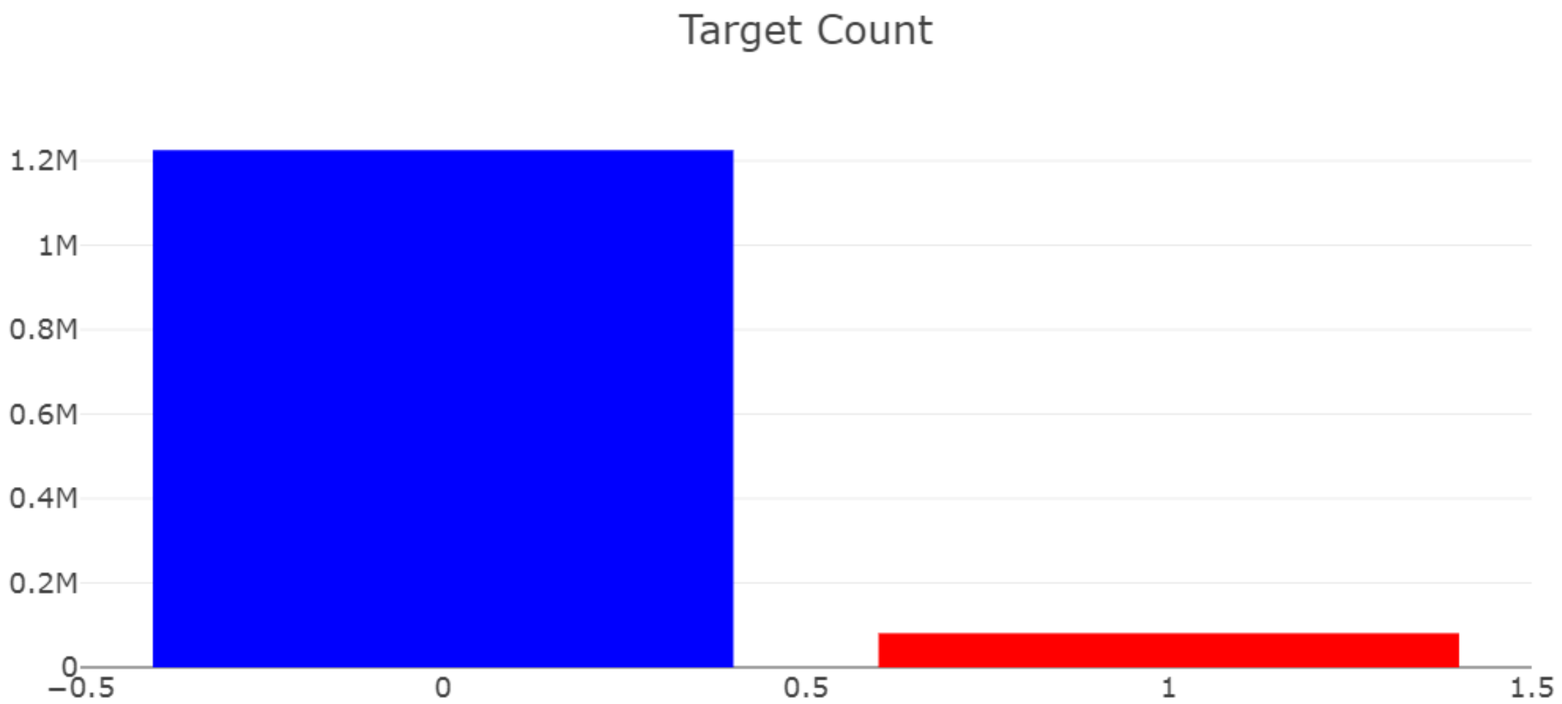
File descriptions

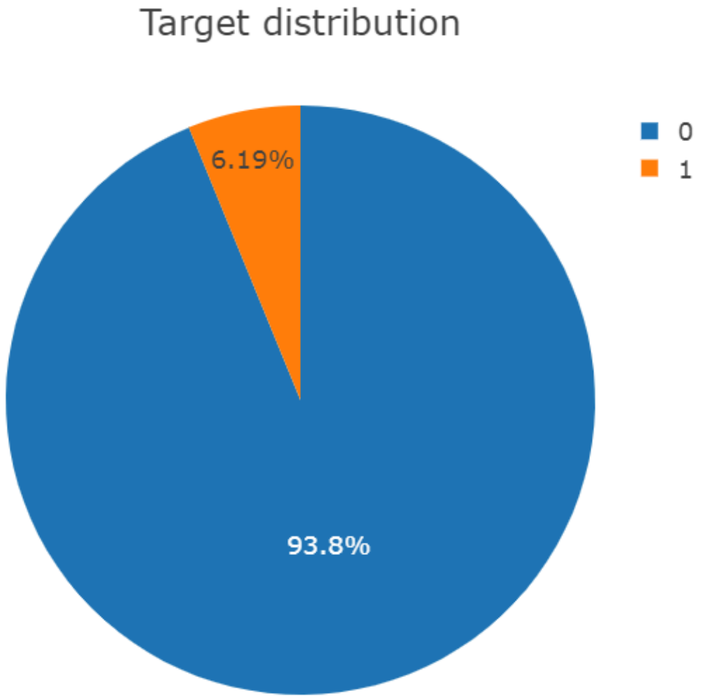
- train.csv
- test.csv
- sample_submission.csv

Data fields

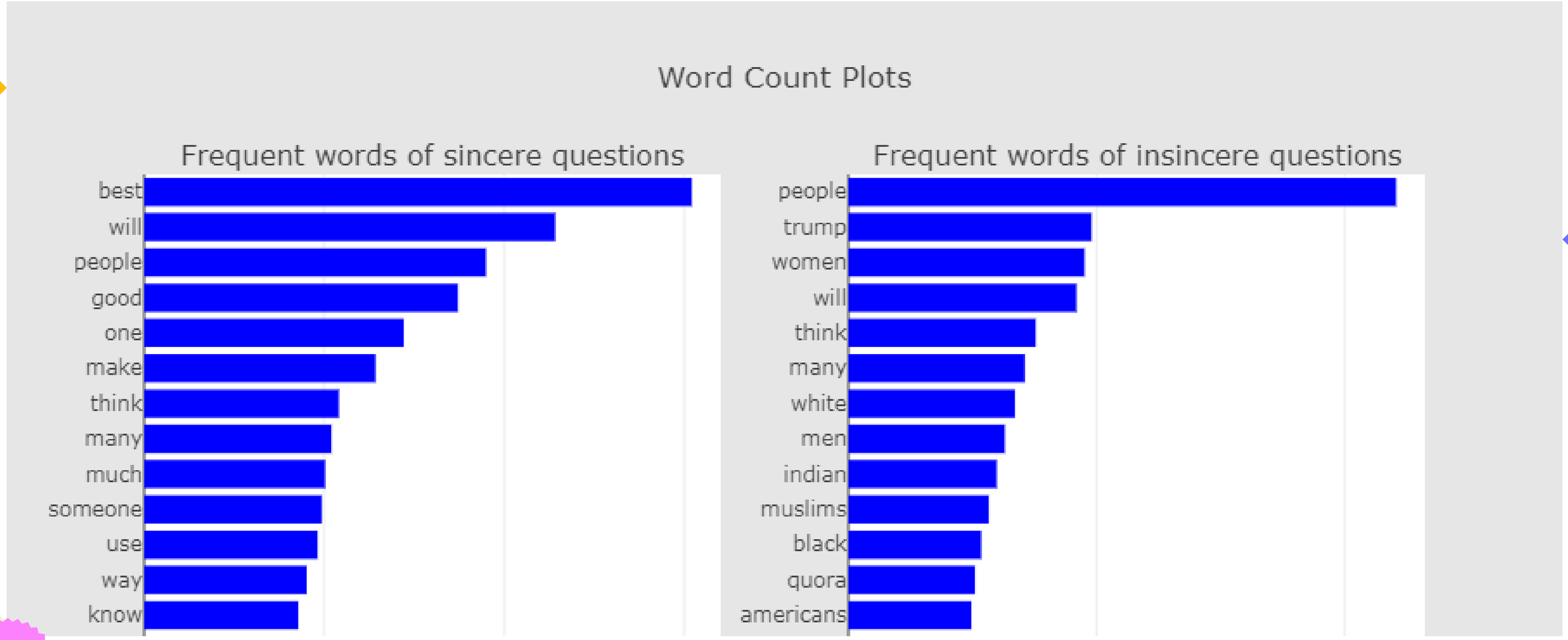
- qid
- question_text
- target

Training data target distribution:

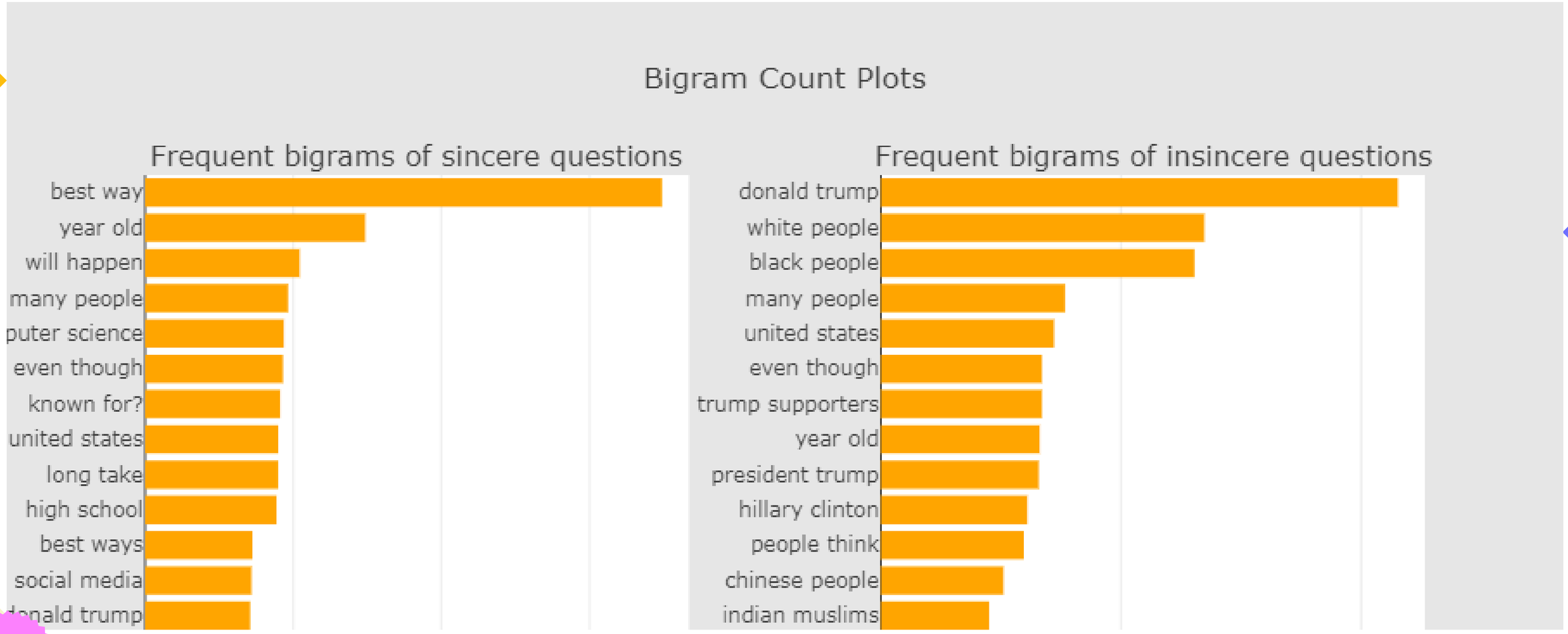




Training data word count:



Training data bigram count:



Solution

Preprocessing

- Exclude filter of punctuations
- Apply misspell corrections

Solution

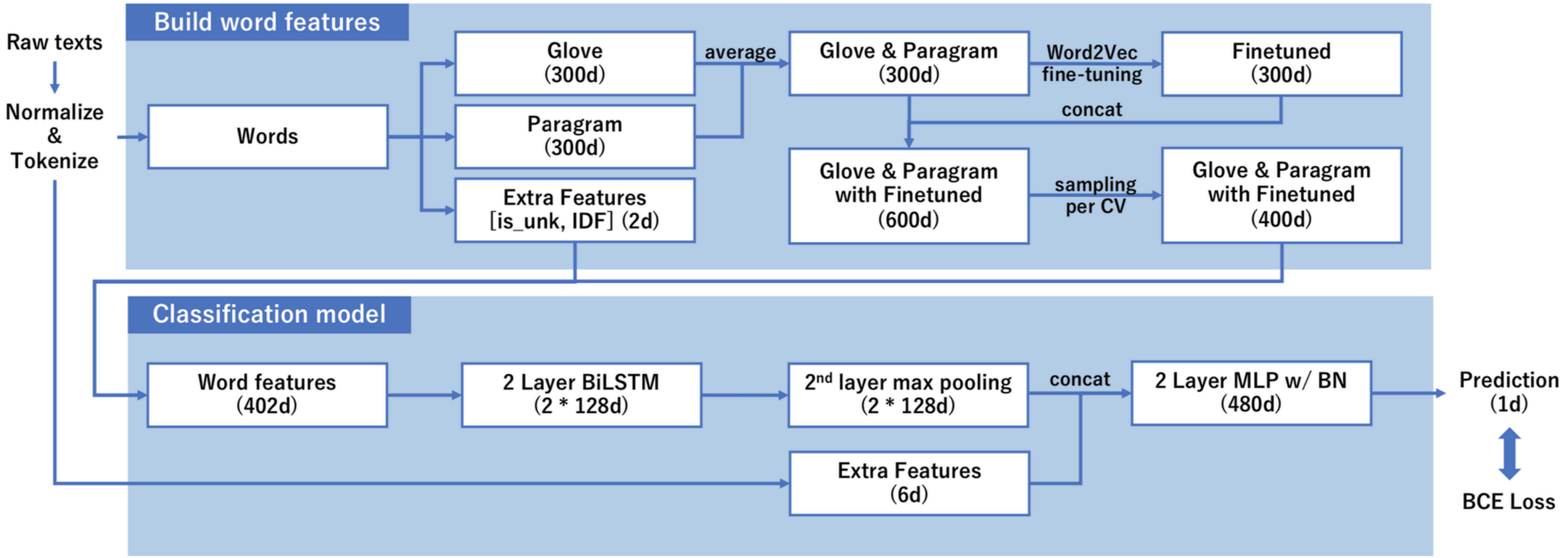
Embedding

Some words in the Quora dataset are common but do not have pretrained vectors available in Glove and Paragram

Adjusted or improved the word embeddings by training them further on the Quora dataset using a method called Word2Vec

Solution

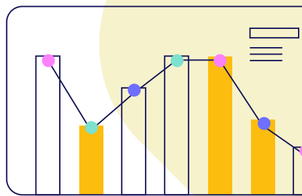
Model architecture



03. Analysis of tweets and social media content of prominent political figures

Analysis of tweets and social media content of prominent political figures

- Influence on public opinion
- Political impact
- These Contents are historical artifact



Trump’s Twitter History

We will already be familiar with the political career of Donald Trump, and will no doubt understand the level of influence his twitter account, his tweets attracted the attention of the world. Analyzing this dataset allows us to imagine how political leaders of the future will use their rising technological powers to rule the populos of the future.

Trump’s Twitter History

1- Import Packages

2- Extracting Hashtags and Mentions

3- Sentiment Analysis

4- Tokenize and get keywords using spacy

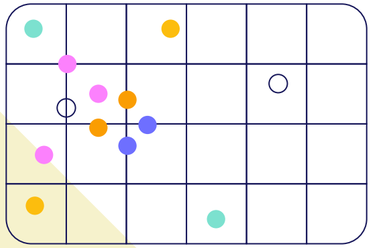
5- Create Word Cloud

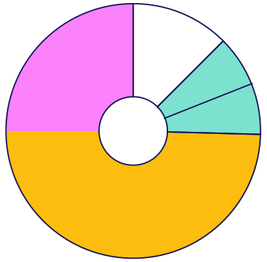
6- TF_IDF

Resources

https://gemini.google.com/
https://chat.openai.com/
https://www.kaggle.com/code/yannisp/sf-crime-analysis-prediction/notebook
https://www.kaggle.com/competitions/quora-insincere-questions-classification
https://www.kaggle.com/code/jonathanmoore2/trump-tweets-1-data-cleaning

THANKS FOR YOUR ATTENTION!





have any question?