Correlation Between Major Crimes Indicators and Twitter Trends

BDA 106: Capstone Final Report

Bangar P., Brzozowski M., & Khudoyarova O.

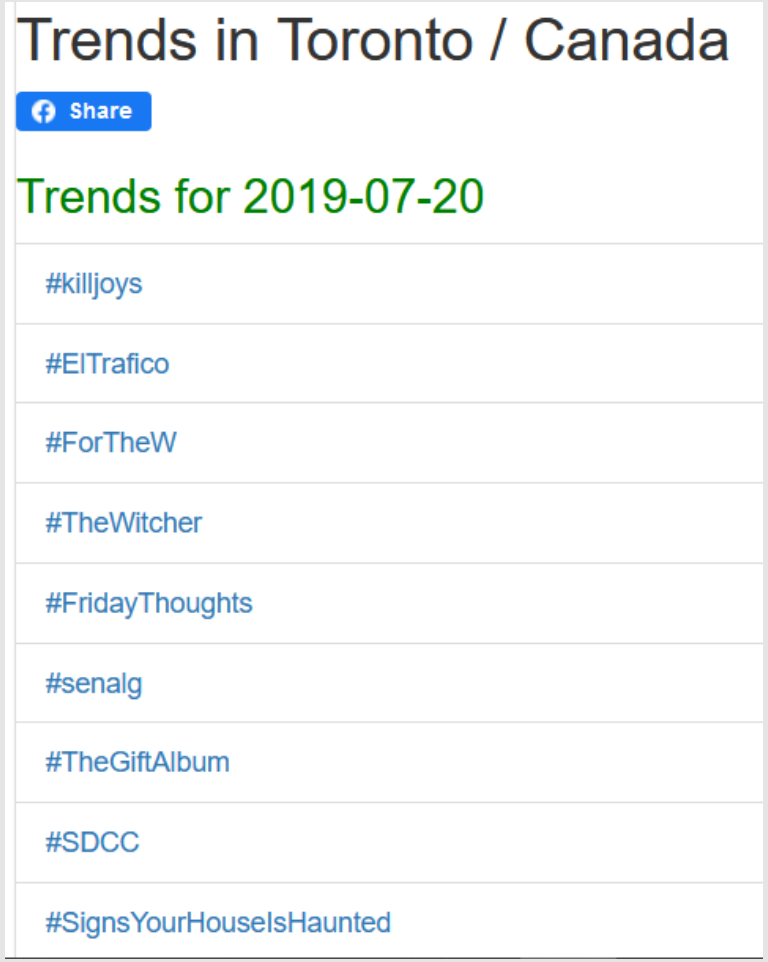
Background & Research Project

- Toronto Police Services (TPS) is dedicated to delivering police services, in partnership with their communities, to keep Toronto the best and safest place to be
- With over 500 million public tweets per day, twitter can be a great resource in understanding people's opinions on the events happening in the world.
- This report explores the possibility of correlations between Major Crime Indicators (MCI) and public opinions expressed on Twitter.

Data mining & description

Twitter Trending

- Twitter Trends are based on an algorithm that determines which topics are popular at a particular time and place.
- Trendogate pulls trends from Twitter
  - List of top 50 Trending Topics per day in Toronto
  - Unordered



Data mining & description

Twitter Tweets

- Downloaded from Archive.org
- Semi-structured Data
- 1% of all tweets
- Nested JSON format
- 19.5GB for 1 day

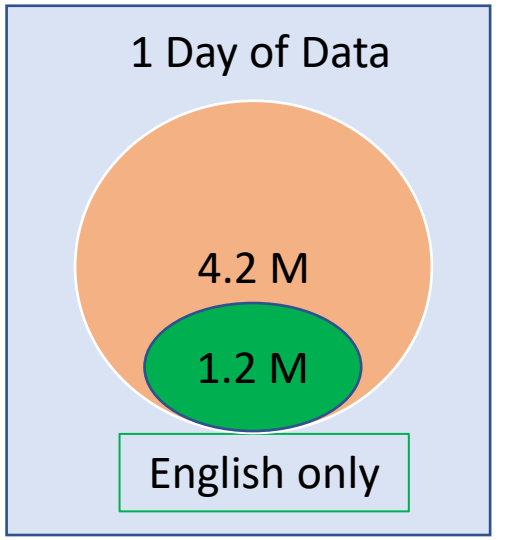
Data
mining & description

MCI

- Downloaded from the Toronto Police Services Website
- Structured Data
- 29 columns
- Major Crime details
  - Location
  - Offense
  - Date
- 206,435 rows
- 2014-2019

```
X                      float64
Y                      float64
Index_                   int64
event_unique_id         object
occurrencedate          object
reporteddate            object
premisetype             object
ucr_code                 int64
ucr_ext                  int64
offence                 object
reportedyear             int64
reportedmonth           object
reportedday              int64
reporteddayofyear        int64
reporteddayofweek       object
reportedhour             int64
occurrenceyear         float64
occurrencemonth         object
occurrenceday          float64
occurrencedayofyear    float64
occurrencedayofweek     object
occurrencehour           int64
MCI                     object
Division                object
Hood_ID                  int64
Neighbourhood           object
Long                   float64
Lat                    float64
ObjectId                 int64
dtype: object
```

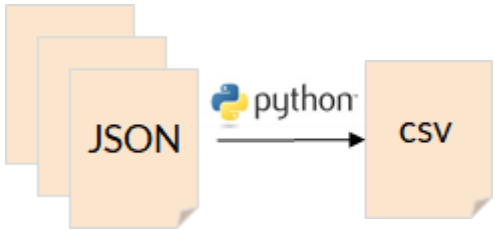
Data Cleansing



Trending Hashtag

Tweets

- Manually copied and pasted
- Hashtags transformed

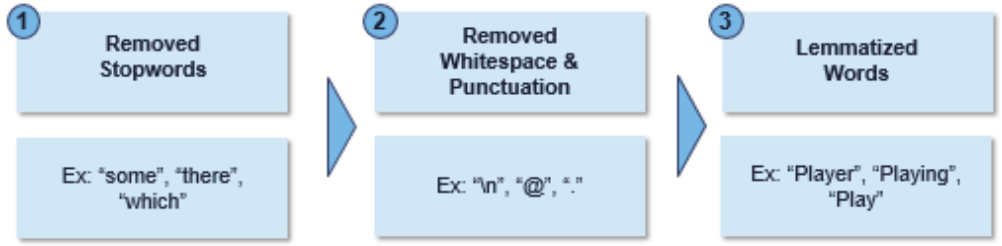


Top 10

Date	Trending Posts (Toronto Only)	Cleaned Trending Post
2019-07-20	#killjoys	kill joys
	#ElTrafico	el trafico
	#ForTheW	for the w
	#TheWitcher	witcher
	#FridayThoughts	friday thoughts
	#senalg	senalg
	#TheGiftAlbum	gift album
	#SDCC	sdcc
	#SignsYourHouseIsHaunted	signs house haunted
	#therealityhouse	reality house
	#QueerEye4	queer eye
	#VeronicaMars	veronica mars
	#FridayFeeling	friday feeling
	#thingsstupidpeoplesay	things stupid people say
	#ParksDay	park day
	#UFCSanAntonio	ufc san antonio
	#SMSociety	sm society

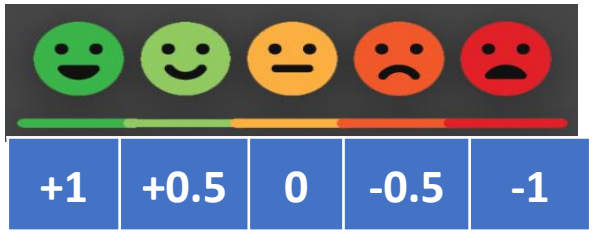
Rank	Hashtag	Body
0	SDCC	RT @ComixVillain: He has it almost right. They have been trying to tell new stories....but much of it is unreadable. I mean, there are pan...
0	SDCC	RT @danielwarrenart: Batman sketch cover at #SDCC https://t.co/GW3leloKLr
0	SDCC	RT @MCU_Direct: Here's is the officially-released slate of upcoming Phase Four movies and series set to release in the next two years! #S
0	SDCC	RT @MarvelStudios: Just announced in Hall H at #SDCC, Marvel Studios' THE FALCON AND THE WINTER SOLDIER, an original series
0	SDCC	RT @starwars: 5 things we learned from the Star Wars fashion collaborations panel at #SDCC: https://t.co/gjLBJf18sy https://t.co/9tqD2M0
0	SDCC	RT @SFXmagazine: Black Panther 2
0	SDCC	RT @Fandomopolis: Lena: I don't want to kill Supergirl. I just want to inflict on her the same pain she inflicted on me. #SupergirlSDCC
0	SDCC	RT @MarvelStudios: Just announced in Hall H at #SDCC, Marvel Studios' SHANG-CHI AND THE LEGEND OF THE TEN RINGS, with Sir
0	SDCC	RT @MarvelStudios: Just announced in Hall H at #SDCC, Marvel Studios' BLACK WIDOW with Scarlett Johansson, David Harbour, Florer
0	SDCC	RT @_RyanGajewski: Tessa Thompson on Valkyrie: "First of all, as king, she needs to find her queen." #MarvelSDCC https://t.co/XR1l2ws
0	SDCC	RT @scarlettsgreys: THAT'S MY GIRL! #SDCC2019 https://t.co/n2aCpvtUel
0	SDCC	RT @MarvelStudios: Just announced in Hall H at #SDCC, Marvel Studios' WHAT IF...?, the first animated series in the MCU, with Jeffrey

Text pre-processing



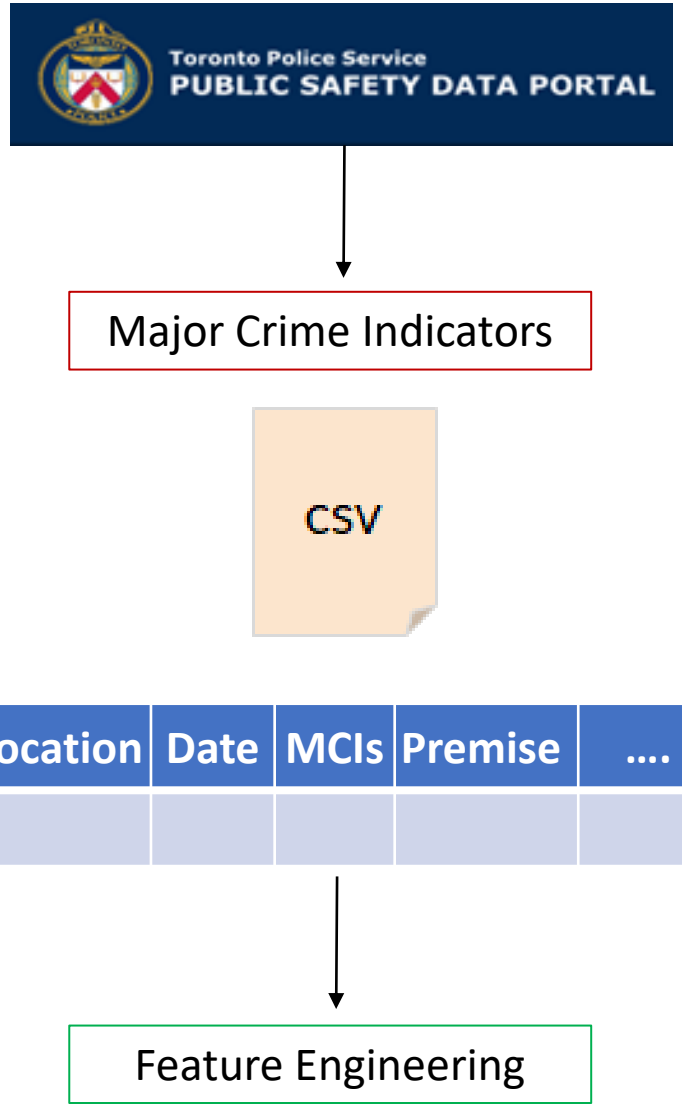
Sentiment Analysis

- Text Blob
- Labelled Tweets



	Rank	Hashtag	Body	Sentiment	AVG_Sentiment
0	0	SDCC	RT @ComixVillain: He has it almost right. They...	-1	-0.179081
1	0	SDCC	RT @danielwarrenart: Batman sketch cover at #S...	1	-0.179081
2	0	SDCC	RT @MCU_Direct: Here's is the officially-relea...	1	-0.179081
3	0	SDCC	RT @MarvelStudios: Just announced in Hall H at...	-1	-0.179081
4	0	SDCC	RT @starwars: 5 things we learned from the Sta...	1	-0.179081
...	...	...	...	...	...
21099	9	TheGiftAlbum	RT @TheGift_Franco: Y'all spent 7 years callin...	-1	-0.437340
21100	9	TheGiftAlbum	RT @USATODAY: Fans are loving Beyoncé's new so...	1	-0.437340
21101	9	TheGiftAlbum	RT @TheGift_Franco: Y'all spent 7 years callin...	-1	-0.437340
21102	9	TheGiftAlbum	RT @TheGift_Franco: Y'all spent 7 years callin...	-1	-0.437340
21103	9	TheGiftAlbum	RT @Sakpo007: Yesssssss!!!!\n\nWe have a UM IT...	1	-0.437340

21104 rows × 5 columns



Feature Engineering resulted in nine new columns needed for analysis.

Final dataset for analysis:

- 2 months of data (June-July 2019)
- MCI + sentiment analyzed Twitter Tweets
- Important factor – Align them based on date

Data Analysis

- Analysis were carried out using both Microsoft Azure Machine Learning Services and Python Methods.
- Main Three Analysis Performed
  - Simple Two Column Analysis
  - Crime Sum Tier Classification
  - Crimes Stratified by Premise Type

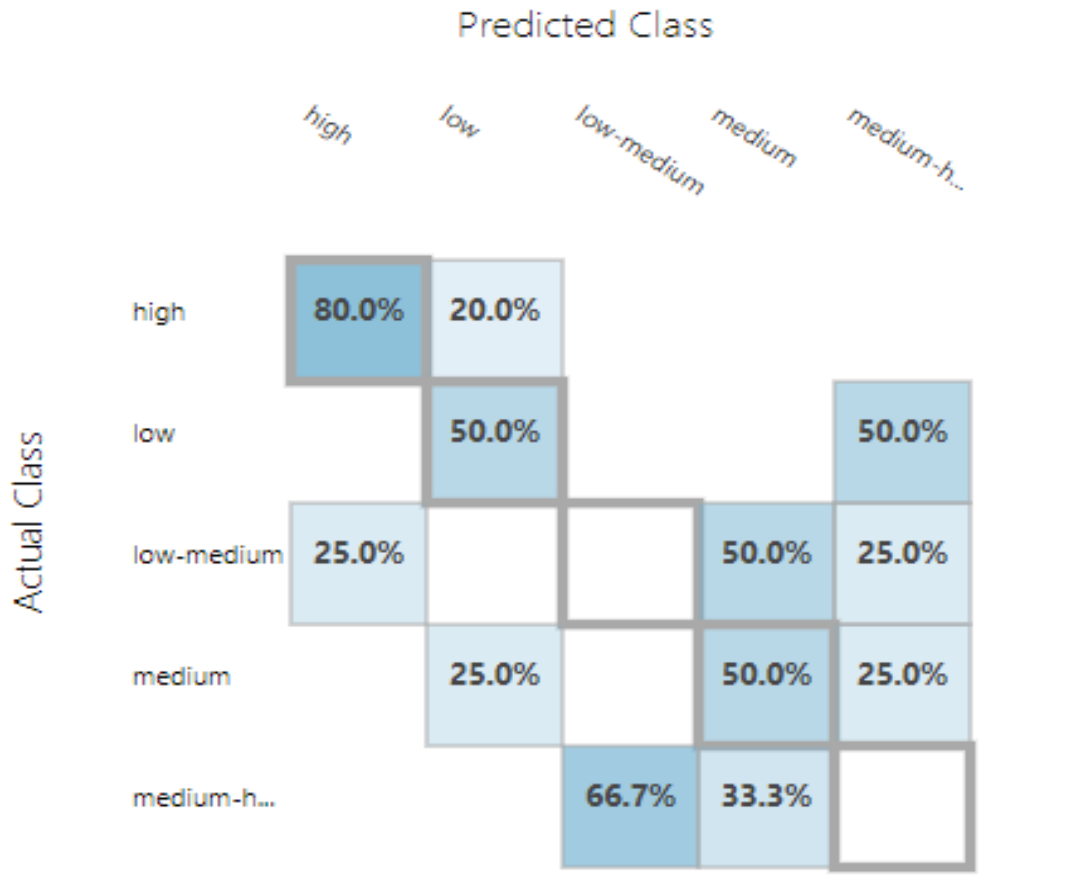
Simple Two Column Analysis

- Two columns considered
  - Total of 61 rows, one row for each day.
- Sentiment dictionary: labelled tweet
- Feature: **AverageHashtagSentiment**
- Target: **Tier_SumofTotalCrimesPerDay**
  - Divided into 5 equal bins depending on the sum of crimes per day
  - High, Medium-High, Medium, Low-Medium, Low
- Multiclass Decision Forest Model
  - Accuracy, Prediction, Recall produced a value of 0.39.

Metrics

Overall accuracy	0.388889
Average accuracy	0.755556
Micro-averaged precision	0.388889
Macro-averaged precision	0.306667
Micro-averaged recall	0.388889
Macro-averaged recall	0.36

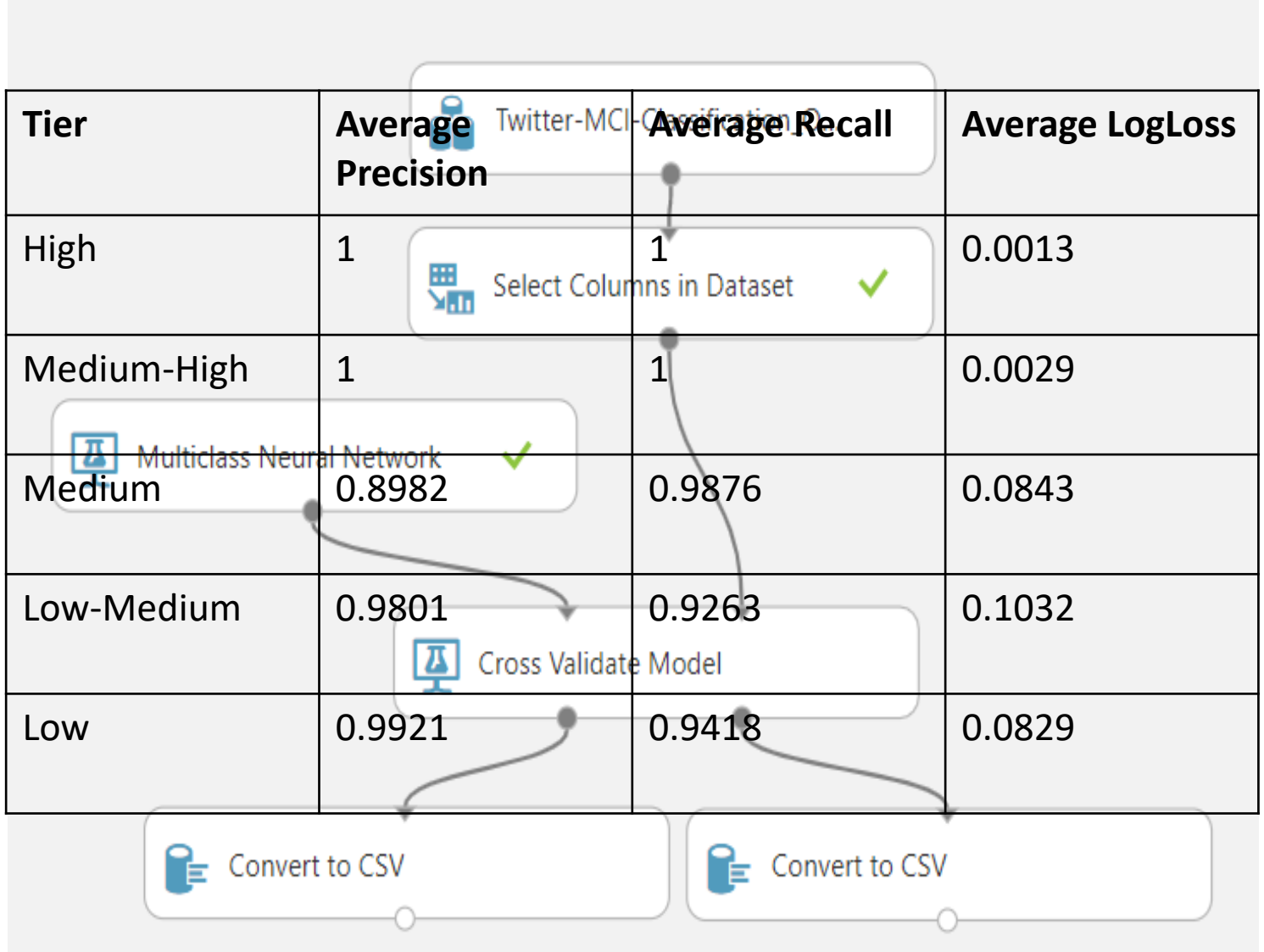
Confusion Matrix



Crime Sum Tier Classification

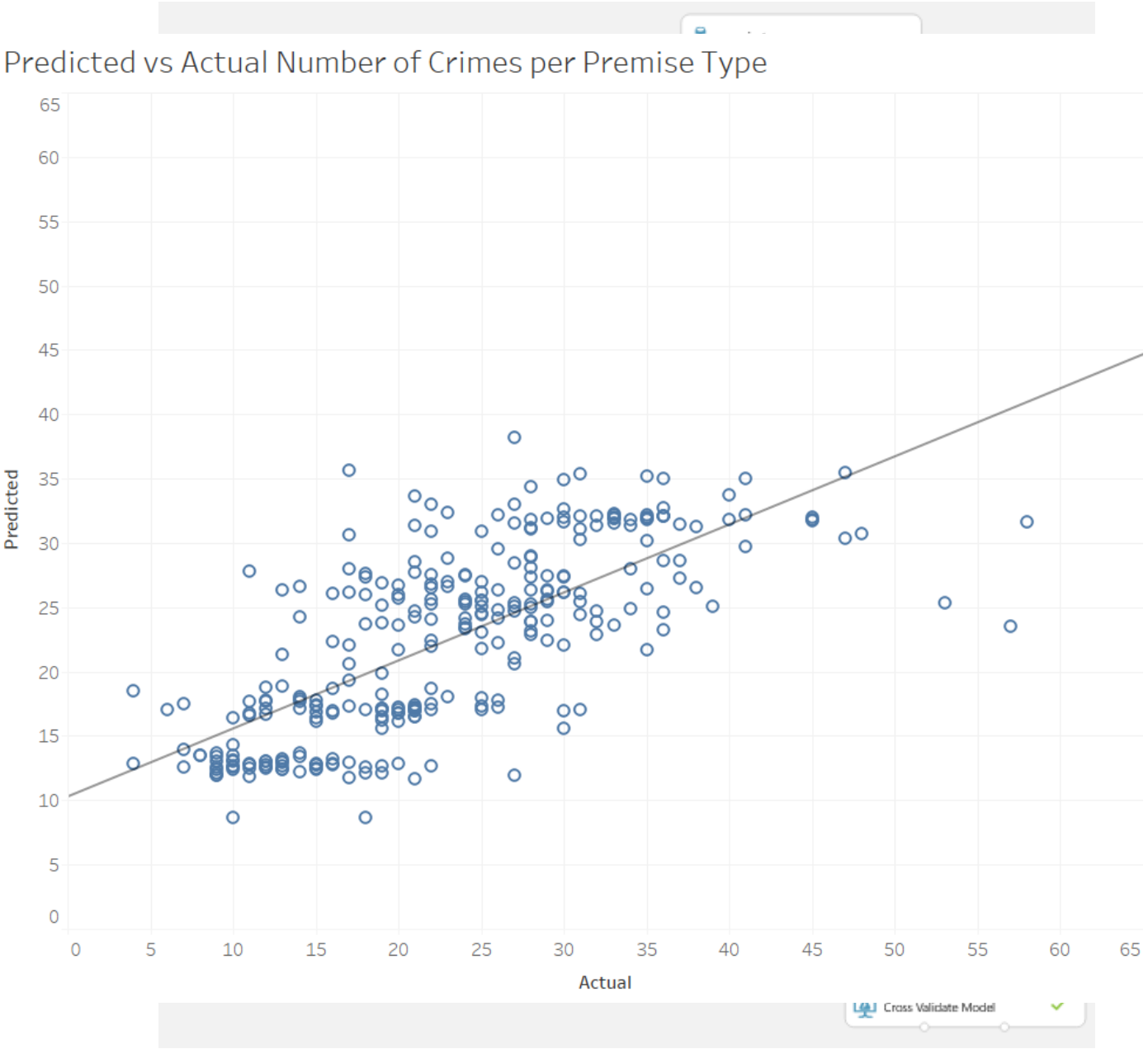
- Twelve columns from considered
  - Total of 7000 rows, one row for each crime.
- Sentiment dictionary: labelled tweet
- Features:
  - **Top 10 Hashtag Sentiment**
  - **MCI labels**
- Target: **Tier_SumofTotalCrimesPerDay**
- Multiclass Neural Network Model
  - Overfit
  - Lack of variance with trending data.

Tier	Average Precision	Average Recall	Average LogLoss
High	1	1	0.0013
Medium-High	1	1	0.0029
Medium	0.8982	0.9876	0.0843
Low-Medium	0.9801	0.9263	0.1032
Low	0.9921	0.9418	0.0829



Regression Analysis of Crimes Stratified by Premise Type

- Features: **Top 10 Hashtag Sentiments, PremiseType**
- Target: **Sum of crimes per premise type per day**
- Neural Network Regression
  - Average coefficients of determination after 10 folds: 0.5
  - Standard deviation of coefficients of determination: 0.14



Implications & Recommendations

- We found a minor correlation between Twitter trends and major crime occurrence; however, more data is needed to confirm our findings.
- Expand on the timeframe of data to analyze from two months to a year. Use trending hashtags by the hour if available.
- Assign weight values based on hashtag ranks and sentiment values. Weights can also be assigned to retweets during the sentiment analysis stage to prevent them from overcrowding regular tweets.
- A non-binary sentiment analysis representing emotion on a higher dimension and high-fidelity scale.

Questions?

Special thanks to Guido and Tom