Final report

Correlation Between Major Crimes Indicators and Twitter Trends

McMaster Continuing Education
& Toronto Police Department

Team 1: Bangar P., Brzozowski M., & Khudoyarova O.

BDA 106: Capstone
July 27, 2020

Executive Summary

Business problem:

Over the last 4 years, the Toronto Police Department (TPD) has seen an increased crime rate in the City of Toronto [1]. In order to decrease the crime rate and make communities safer, the TPD must evolve and use the advancements in technology. In recent years, Twitter has become one of the most popular social media platforms in the Western world [2]. 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. The integration of publicly available social data with crime data can help police departments make their communities a safer place to live in. This report explores the possibility of correlations between Major Crime Indicators (MCI) and public opinions expressed on Twitter.

Data

In total, three datasets were explored: Twitter Trending, Twitter Tweet Archive, and TPD MCI. Over 1 terabyte (TB) of data was pre-processed to discover relationships between tweets sentiment and criminal data. Tweets and hashtags obtained from Twitter went through sentiment analysis, which provided their respective polarity of positive/ negative rating. The final datasets contained the sentiment ratings of the top 10 trending hashtags in Toronto for June and July of 2019, as well as the number of crimes committed per day split up by different variables (e.g. premise type, MCI). Sums of total crimes were divided into five tier classification: Low, Low-Medium, Medium, Medium-High, and High.

Analytic solution

Using the Multiclass Decision Forest model, average sentiment ratings and tier classification produced an average accuracy, precision and recall of 0.39. Whereas, the Multiclass Neural Network model with MCIs, average sentiment ratings and tier classification evaluated a precision of 1.0, 1.0, 0.93, 0.96, 0.92, for high, medium-high, medium, low-medium, low, tier classification respectively. Lastly, sentiment ratings, premise type and total crimes per premise per day were analysed using Neural Network, Poisson and Linear regressions. They shared a mean coefficient of determination of 0.50.

Recommendations

The following are our three highest recommendations

- 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.

Problem Description

Between the years of 2014 to 2019, the Toronto Police Department (TPD) has recorded over 200,000 incidences of crimes committed in the City of Toronto [1]. The offenses ranged from personal and commercial property crimes (e.g. *auto-theft* and *robbery*) to violent crimes (e.g. *assault* and *homicide*). Regardless of the type of offense, crimes impact not only the victims, but also relatives, friends and even entire neighbourhoods [3]. Crime motives depend on various reasons, therefore, it is hard to predict when and by whom the crime will be committed[4]. To help alleviate such hardships, the reduction of crimes is a critical step that police services must act on. This can be difficult as criminals evolve and adapt new methods in response to the latest technological progress [5].

One method of crime prevention that police services can implement is the presence and public awareness of local enforcement in areas of potential crimes [6-7]. The main issue with this method is that policing services are spread too thin, and do not have the luxury or resources to disperse officers to every single area. One solution can be selecting specific areas of the community to police, generally the ones with the highest crime rates.

Everyday, we as a community spread terabytes of information through social media [2]. Facebook, Instagram and Twitter, with a combined 5.16 billion accounts, are some of the most common mediums in the Western World [8]. Twitter alone claims 500 million public tweets everyday, which is almost 6000 tweets a second. From these public tweets, twitter aggregates the most popular topics and places them into a top trending tweet page. Such a page gives a glimpse into what people were talking about at any period of time and place. Pairing this extensive amount of information with various other sources can provide interesting discoveries.

This project proposes a way to analyze if current trending topics in the city of Toronto may have an effect on crime rate. The goal for this project is to gain insightful information by analyzing a blended dataset made up of a sentiment analysis of Twitter trends and TPD's Major Crime Indicator (MCI) data. We plan on building multiple models to look for any correlation between the datasets. While this model is descriptive and retro-looking, it can be effectively remodeled to be a forward-looking predictive model. The benefits of a project of this caliber can help reduce crime, staffing problems, public costs and promote safer communities.

Data Description

Twitter Trending: Due to cost constraints of this project, this data came from a free, unaffiliated HTML Twitter trending archive that recorded trending hashtags by location and date (https://trendogate.com). For the region of Toronto, it only had trending hashtags by the day for June and July of 2019 (Appendix A1).

Twitter Tweets: This is a much larger dataset compared to the other two and in a semi-structured form. This information was also taken from a free third-party archive source (https://archive.org). This source provides the "spritzer" version of a Twitter stream grab that contains a random sample of about 1% of all tweets for the chosen time period. This data came in multiple JSON format files, partitioned by hour, containing around 19.5 GB of data for each day. It included a lot of impractical and redundant information, as well as data in a non-English language (Appendix A2). Moreover, it contained links to images, videos, other tweets, as well as non-ascii characters.

MCI: This dataset was provided by the TPD, which contains information on crimes relating to robbery, assault, and various others from 2014 to 2019. The MCI dataset was clean, structured and contained no null values. It contains 29 columns such as location, date and crimes committed and 206,435 rows with one row for each recorded major crime committed (Appendix A3).

Due to the magnitude of data to process, only data from June and July of 2019 was used. Additionally, there were a few missing days of data. In particular, June 24th did not have twitter tweet data and July 30 & 31st did not have Hashtag data.

Data Preparation

Twitter Trending: Twitter Trending data was manually copied and pasted into a spreadsheet to be cleaned. It contains only two columns: date and hashtag. Hashtags were transformed by using Microsoft Excel's replace feature, removing the hashtag symbol and unwanted spaces between words.

Twitter Tweets: In order to incorporate twitter trending data with MCI data, meaning must be first imputed onto the trending data. Sentiment analysis on too few words provides untrustworthy values; therefore, it has to be executed on the tweet bodies that contain the trending hashtags.

A script was written to open each JSON file in order to sequentially flatten it. First, to reduce the size of the data, only English tweets were pulled out of the archive, dropping the dataset from approximately 4.2 million rows to 1.2 million rows in a day. Next, only the important columns were extracted to a new dataframe in order to reduce the size of the working data from 19.5 GBs to 0.3 GBs per day. Some of the kept columns are text:bodies, user:location, & creation_date, while some of the columns dropped were user:status_id and user:screen_name. This data was further reduced by pulling out rows of tweets that only contained a hashtag found in the Twitter Trending dataset.

As our Twitter Trending dataset did not have an order as to which was the highest trending, we selected the top 10 hashtags by the number of tweets that contained a trending hashtag and Toronto location reference. This allowed us to create our own top 10 trending list.

With the ten hashtags chosen for the day, the body texts were pulled from the archive relating to the hashtags regardless of location. This was done because sentiment analysis works better on a larger dataset, and many hashtags in the top 10 list only had 4-5 Toronto based tweet text to support it.

Finally, tweet texts went through text processing, and then sentiment analysis was applied. Sentiment analysis is a text classification tool, which can classify a given text into positive, neutral and negative sentiment based on its polarity. Two dictionaries were applied onto the text body of each row in the filtered data: bag of words and labelled tweets. The sentiments of all tweet texts per hashtag were averaged and used to create the Twitter Sentiment dataframes.

MCI: Nine new columns were engineered to supplement the MCI data. Five of those columns are integer sums of each MCI classification per day, i.e. the number of assaults per day. One additional column is a boolean indicating whether the next day had less crime than the current day. Another column had the sums of crimes committed per day by premise type. Finally, the last two columns are an integer of the sum of total crimes in a day and a string of the same values represented as a tier, binned by Low, Low-Medium, Medium, Medium-High, and High crime values.

Combining the datasets

The Twitter Sentiment dataframes were appended onto the MCI data resulting in two complete MCI-Twitter datasets. The two complete sets are identical except for the values of the Twitter Sentiment ratings due to the different sentiment dictionaries applied. The sentiment ratings were arranged from highest in column 1 to lowest in column 10 for each day. The final

size for the various versions of the dataset range from 0.001 - 0.003GBs (1 to 3MBs) from an initial dataset of 1170GBs (1.17TBs).

Data Analysis

Data analysis was carried out using both Microsoft Azure Machine learning services and Python. We performed various analyses on the combined dataset, each producing different results and interesting insights. We focused on three main analyses for the report.

Simple Two Column Analysis

The first analysis featured only two columns from the entire dataset: AverageHashtagSentiment and Tier_SumofTotalCrimesPerDay. The sentiment dictionary used in this analysis was the labelled tweet dictionary. In total 61 rows, one row for each day, was put into a test-train split with a ratio of 70-30 . Various classification models were exercised on the trained data: Multiclass Neural Network, Multiclass Logistic Regression, Multiclass Decision Jungle and Multiclass Decision Forest. The most promising results came from the Multiclass Decision Forest model which produced an average accuracy, precision and recall of 0.39. While these results seem unexciting, we are certain that this model was underfitting and the addition of rows of data will further improve the results.

Crime Sum Tier Classification on Twitter Sentiment Analysis Ranks

The second analysis covered in this report features twelve columns and 7000 rows. The columns in consideration were MCI labels, top ten hashtag rank sentiments and the Tier_SumofTotalCrimesPerDay. The rows contain every crime indicated from the MCI. Again the labelled tweet Sentiment dictionary was used. Cross-validation was applied producing 10-folds on the dataset. For this model the same classification models were used as the first analysis. The Multiclass Neural Network model produced the most intriguing results with a precision of 1.0, 1.0, 0.93, 0.96, 0.92, for high, medium-high, medium, low-medium, low, respectively. Such high numbers indicate that the model was overfitting to our dataset.

Regression Analysis of Crimes Stratified by Premise Type

The third analysis of note in this report used 12 columns from the entire dataset. The first 10 columns were the sentiment values, followed by premise type and the sum of crimes per premise per day. Two datasets, each using a different sentiment dictionary, were analyzed in parallel and produced similar findings. In total 290 rows, 5 rows for each day, were cross-validated with 10 folds. Various regression models were trained on the training data: Neural Network, Bayesian Linear, Boosted decision tree, Decision Forest, Fast Forest Quantile, Poisson, and Linear regression. Neural Network regression, Poisson regression and Linear regression shared the highest mean coefficient of determination of 0.50; however, the Neural Network had the lowest standard deviation of coefficient of determination at 0.14

Conclusion

Advantages

In summary, we believe we have a footing in correlating twitter trends with MCI data. We found a minor correlation between Twitter trends and MCI recordings. In particular, we found a minor correlation between average sentiment and Tier sum total crimes per day. Additionally, a minor correlation was found between premise type, hashtag sentiments and the sum of crimes per day.

Filters were designed to retrieve data from Twitter to compile it with MCI data. This report and data analysis is a stepping stone for the future work with an extended amount of data instead of 1 % of Twitter data. Our results can be used to help others build upon this project.

Limitations

Data sources

For this project, we were limited to using free, publicly available data sources. The major drawback was that the top trending hashtags obtained were an unordered list per day. Paid sources may have been able to give a breakdown by the hour as well as metrics of how often a hashtag was used. Additionally, there were limitations to using tweet data from an archive that only contained 1% of all tweets from around the world as it does not provide an adequate sample of tweets from Toronto. For example, we did not have enough tweets that referenced Toronto and thus had to do sentiment analysis on any tweet that contained one of the trending hashtags.

Another limitation of this study is that there was no differentiation between regular tweets and retweets, as both were considered to be equal. Furthermore, the Trending Twitter data consists of both Trending words and Trending Hashtags posted within the Tweets. Differentiation and selection between the two subsets can help reduce similar trending tweet data.

Computing Power

Due to being limited to using our own computers, it took a while to pre-process the data. For example, it took over two hours to convert one day of twitter data to CSV format and pull out the relevant information. Proper parallel computing on a distributed file system would drastically reduce processing time.

Overfitting and Underfitting

Due to lack of variance in the data, models quickly overfit to duplicate values. MCI data records multiple columns per day, and if the Twitter Trending data from the archive are only tracked per day, the sentiment values must be repeated causing a lack of variance between days. To solve this issue, the inclusion of hourly Trending data could alleviate overfitting. In regards to underfitting with the regression models, the addition of more days of data is required.

Recommendations

First: Expand on the timeframe of data to analyze. Due to various constraints outlined above, we were only able to analyze two months worth of data with hashtags available by the day only. We recommend expanding the analysis to at least 1 year worth of data, and if possible, have trending hashtags broken down by the hour. Another option to expand the available dataset is to compare various metropolitan police records with Toronto MCI and

Sentiment Analysis. The result will reduce the underfitting of the model because of the increase in data size, plus it will prevent overfitting due to improvement of variability.

Second: Assign weight values based on hashtag ranks and sentiment values. Consider the following, Imagine the top trending hashtag corresponds to a slightly negative sentiment analysis, while the 10th ranked trending hashtag has very negative sentiment analysis. Even though the 10th rank hashtag has a very negative value, less individuals are discussing the topic, while more people are discussing the more neutral topic. A weighted function, maybe a higher order polynomial could give accurate value to the hashtag. In a similar vein, smaller weights can be assigned to retweets during the sentiment analysis stage to prevent them from overcrowding regular tweets.

Third: A non-binary sentiment analysis representing emotion on a higher dimension and high fidelity scale. As an example, pleasant: [1,0], activation [0,1], unpleasant [-1,0] deactivation [0,-1], with vectors representing various other emotions. [See Appendix C]

Additionally, we would recommend the following:

- Including a sentiment analysis dictionary for identifying and classifying emojis.
- Identifying trending topics in the tweet text in addition to hashtags.
- Overwriting tweet text with extended tweed text, when available.
- Data Warehouse and File Distributed System to store and process all data.
- Develop API - to actively pull twitter data.
- Proper flattening and filtering tools to improvise flow and speed of previous scripts.

References

[1] Public Safety Data Portal MCI 2014 to 2019, 2020
"*http://data.torontopolice.on.ca/datasets/mci-2014-to-2019"* Toronto Police Services

[2] Kayla Matthews, 2018. "Here's How Much Big Data Companies Make On The Internet"
https://bigdatashowcase.com/how-much-big-data-companies-make-on-internet/

[3] Daniel Brisson & Susan Roll, 2012. "The Effect of Neighborhood on Crime and Safety: A Review of the Evidence," Journal of Evidence-Based Social Work

[4] Patrick Sharkey and Robert J. Sampson. 2015. "Violence, Cognition, and Neighborhood Inequality in America," in *Social Neuroscience: Brain, Mind, and Society*, Russell Schutt, Matcheri S. Keshavan, and Larry J. Seidman, eds. Cambridge: Harvard University Press

[5] Taylor Simmons, 2018. "Auto theft on the rise in Toronto area, and a security expert thinks he knows why" CBC News *https://www.cbc.ca/news/canada/toronto/car-thefts-rising-1.4930890*

[6] Sarit Weisburd, 2019. "Police Presence, Rapid Response Rates, and Crime Prevention" The Mit Press Journal.

[7] Christopher Koper, 2006. "Just enough police presence: Reducing crime and disorderly behavior by optimizing patrol time in crime hot spots" Justice Quarterly

[8] Top Sites, 2020. *https://www.alexa.com/topsites*, Alexa

Data References

[1] Trends in Toronto/ Canada, 2020, *Trends for Toronto,*
https://trendogate.com/placebydate/4118/2019-06-01 TrendoGate

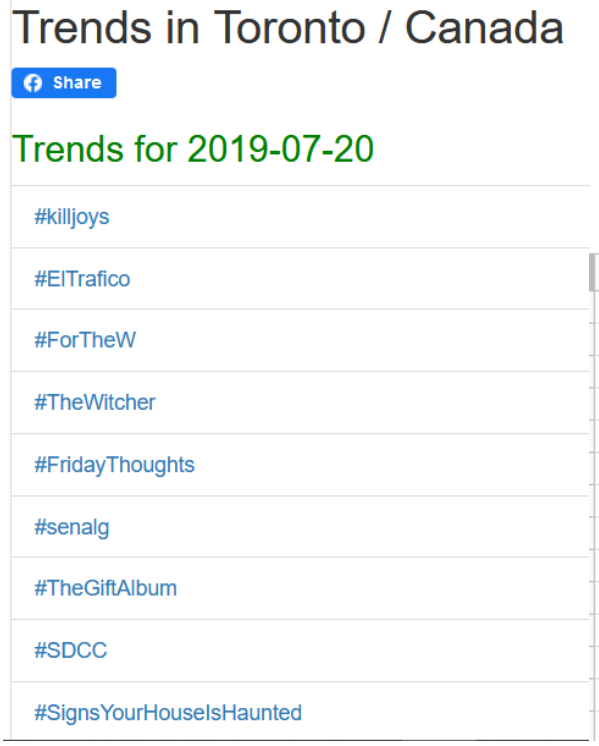
[2] Twitter Archive June-July 2019, 2020, *archiveteam-twitter-stream,*
https://archive.org/details/archiveteam-twitter-stream-2019-06 Archive.org

[3] Major Crime Indicators (MCI), 2020, *Public Safety Data Portal,*
https://data.torontopolice.on.ca/datasets/mci-2014-to-2019 Toronto Police Service

Appendix A - Data Description

Appendix A1 - Trending Hashtag Data

Images represent Trending hashtags found on TrendoGate. Left Image displays a screenshot of the data found from the webpage. Right Image corresponds to a CSV file of copied un-processed hashtags.



A	B
2019-07-27	#FlashbackFriday
2019-07-27	#QueenRadio
2019-07-27	#CFLGameday
2019-07-27	#MADDEN20
2019-07-27	#FireEmblemThreeHouses
2019-07-27	#BadFirstImpressionsIn5Words
2019-07-27	#UFC240
2019-07-27	#NHL20Beta
2019-07-27	#NeverSaidInATarantinoFilm
2019-07-27	#Fridayfeeling
2019-07-27	#intent
2019-07-27	#OITNB7
2019-07-27	#ConBravo2019
2019-07-27	#heavymontreal

Appendix A2 - Twitter Data

Figure A2-1: Image below represents a Juypter notebook. Notebook displays the columns found in the Tweet Text JSON files. Columns identified as ‘object’ contain more nested data.

```
[8]: created_at                  datetime64[ns, UTC]
     id                                      float64
     id_str                                  float64
     text                                     object
     source                                   object
     truncated                               float64
     in_reply_to_status_id                   float64
     in_reply_to_status_id_str               float64
     in_reply_to_user_id                     float64
     in_reply_to_user_id_str                 float64
     in_reply_to_screen_name                  object
     user                                     object
     geo                                      object
     coordinates                              object
     place                                    object
     contributors                            float64
     is_quote_status                         float64
     quote_count                             float64
     reply_count                             float64
     retweet_count                           float64
     favorite_count                          float64
     entities                                 object
     favorited                               float64
     retweeted                               float64
     possibly_sensitive                      float64
     filter_level                             object
     lang                                     object
     timestamp_ms                    datetime64[ns]
     display_text_range                       object
     retweeted_status                         object
     extended_entities                        object
     extended_tweet                           object
     quoted_status_id                        float64
     quoted_status_id_str                    float64
     quoted_status                            object
     quoted_status_permalink                  object
     delete                                   object
     withheld_in_countries                    object
     dtype: object
```

Figure A2-2: Image below shows the raw semi-structured data in the JSON files with non-ascii characters

Appendix A3 - MCI Data

Figure A3-1: Image below shows the MCI dataset in a Juypter notebook. Here, columns identified as ‘object’ are actually string values.

```
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
```

Appendix B - Data Analysis

Appendix B1 - Simple Two Column Analysis

Figure B1-1: Shows Azure Machine Learning Studio, running various regression models on a split dataset. The split is 70/30 with 70% going into splitting the data. Model is scored and evaluated with the test set.

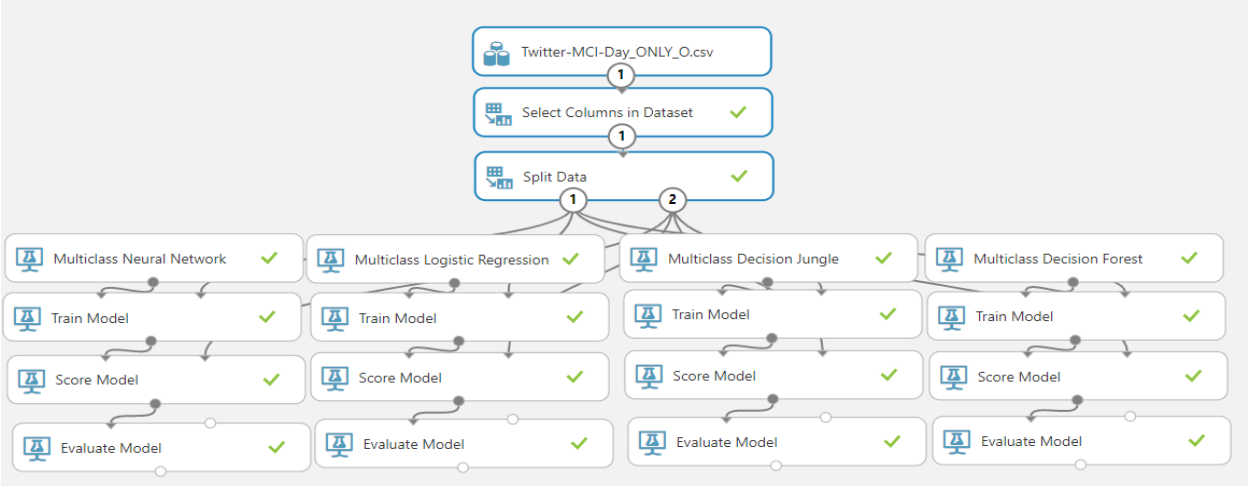


Figure B1-2: Represent Azure Machine Learning Studio, running a Multiclass Decision Forest on a split dataset. The split is 70/30 with 70% going into splitting the data. Model is scored and evaluated with the test set.

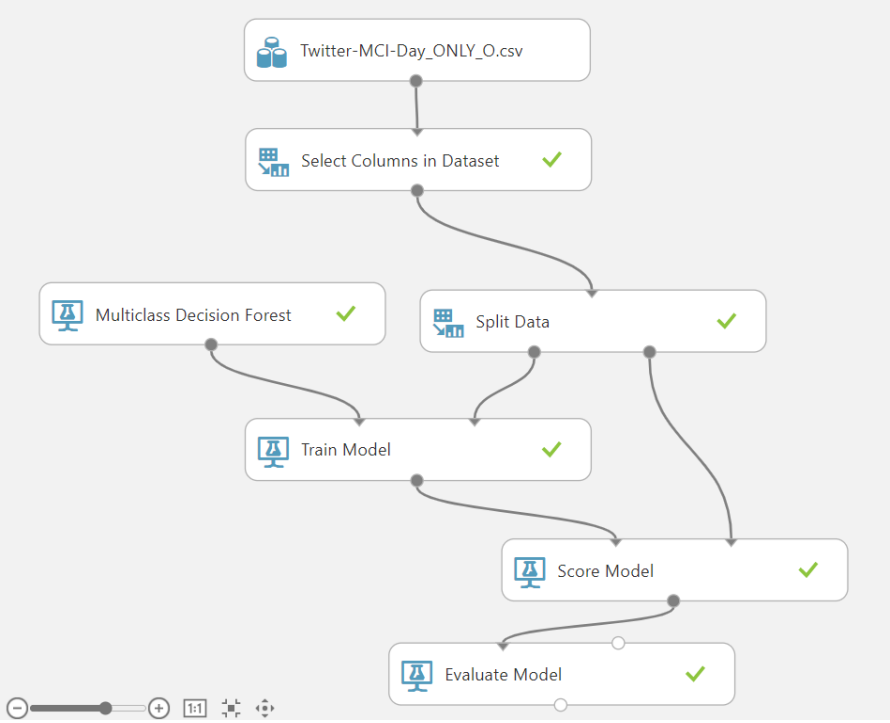
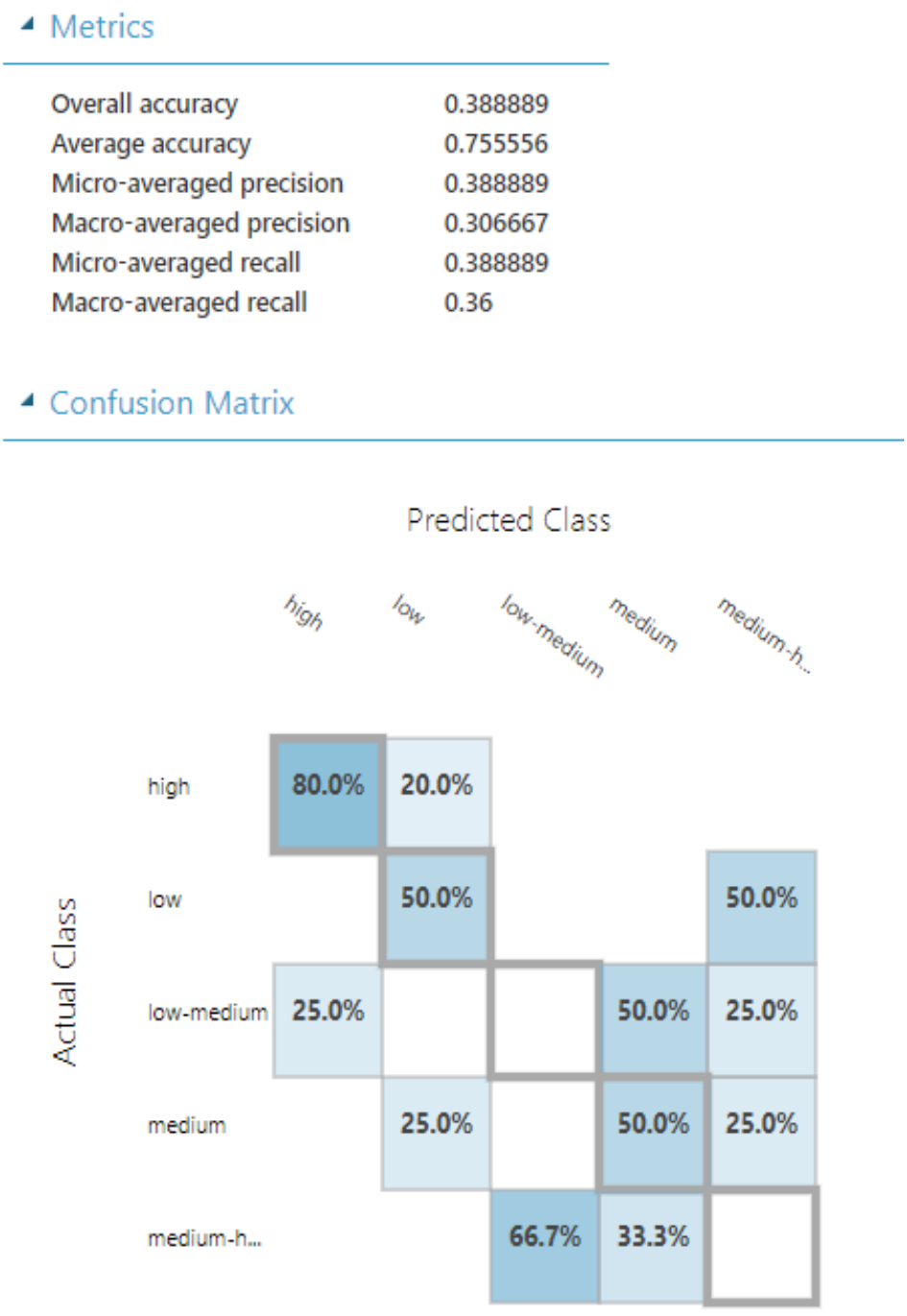


Figure B1-3: Image corresponds to the evaluation of the test set on the trained model. Azure Machine Learning Studio is running a Multiclass Decision Forest model. The split is 70/30 with 70% going into splitting the data. Accuracy and Recall are the metrics with an included confusion matrix.



Appendix B2 - Crime Sum Tier Classification on Twitter Sentiment Analysis Ranks

Figure B2-1: Shows an Azure Machine Learning Studio, running one Multiclass Neural Network model with cross-validation that performs 10 folds.

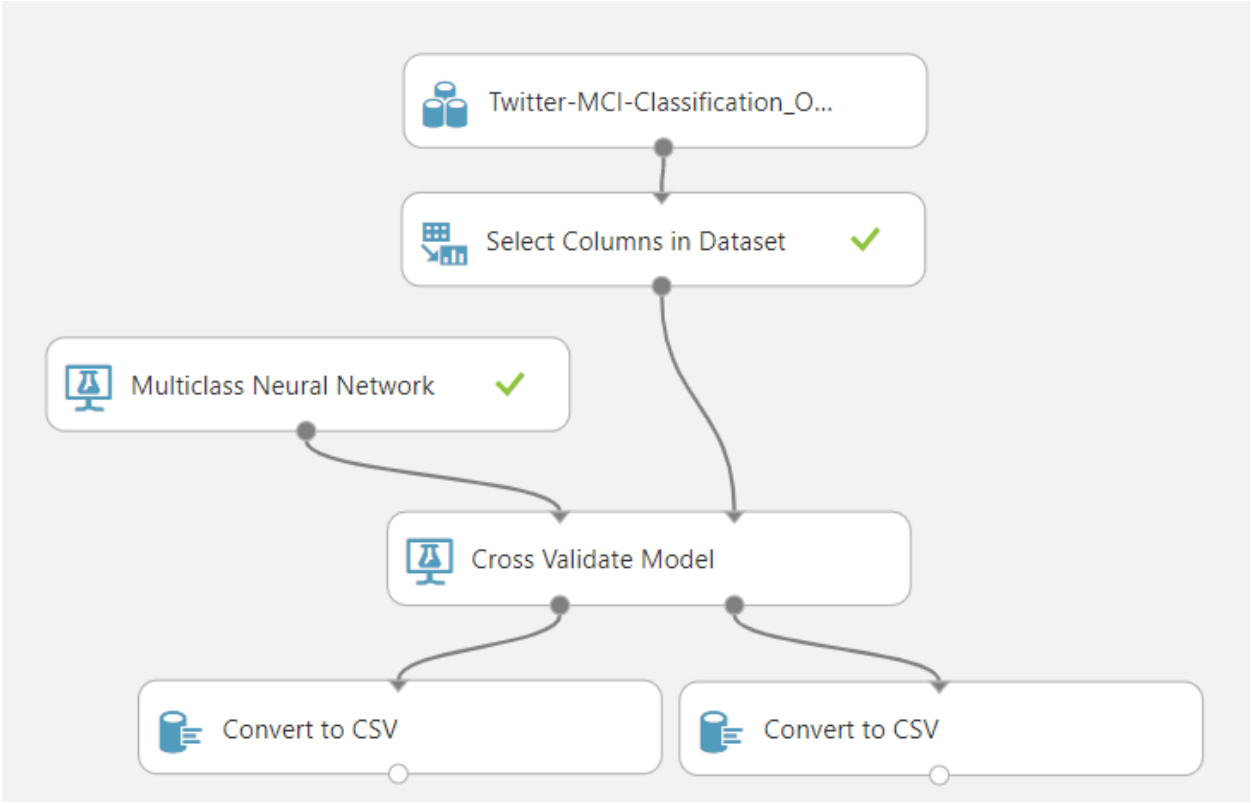


Table B2-1: Shows Azure Machine Learning Studio, executing various regression models on a split dataset. Model is scored and evaluated with the 10-fold cross-validation. Results in 5 outcomes one for each tier.

Model: Multi-class Neural Network

Fold Number	Average Log Loss for Class "high"	Precision for Class "high"	Recall for Class "high"
Mean	0.001378235	1	1
Standard Deviation	0.000286132	0	0
Fold Number	Average Log Loss for Class "low"	Precision for Class "low"	Recall for Class "low"
Mean	0.082997699	0.992110123	0.941872436
Standard Deviation	0.019669	0.018001849	0.01418652
Fold Number	Average Log Loss for Class "low-medium"	Precision for Class "low-medium"	Recall for Class "low-medium"
Mean	0.103263502	0.980156768	0.926307306
Standard Deviation	0.030659607	0.041466201	0.026938255
Fold Number	Average Log Loss for Class "medium"	Precision for Class "medium"	Recall for Class "medium"
Mean	0.084329085	0.898249821	0.987637635
Standard Deviation	0.014841577	0.042503768	0.025318285
Fold Number	Average Log Loss for Class "medium-high"	Precision for Class "medium-high"	Recall for Class "medium-high"
Mean	0.002992208	1	1
Standard Deviation	0.000760206	0	0

Appendix B3 - Regression Analysis of Crimes Stratified by Premise Type

Figure B3-1: The image below shows the different regression models being trained and cross-validated with 10 folds in Azure Machine Learning Studio on the dataset that was stratified by premise type.

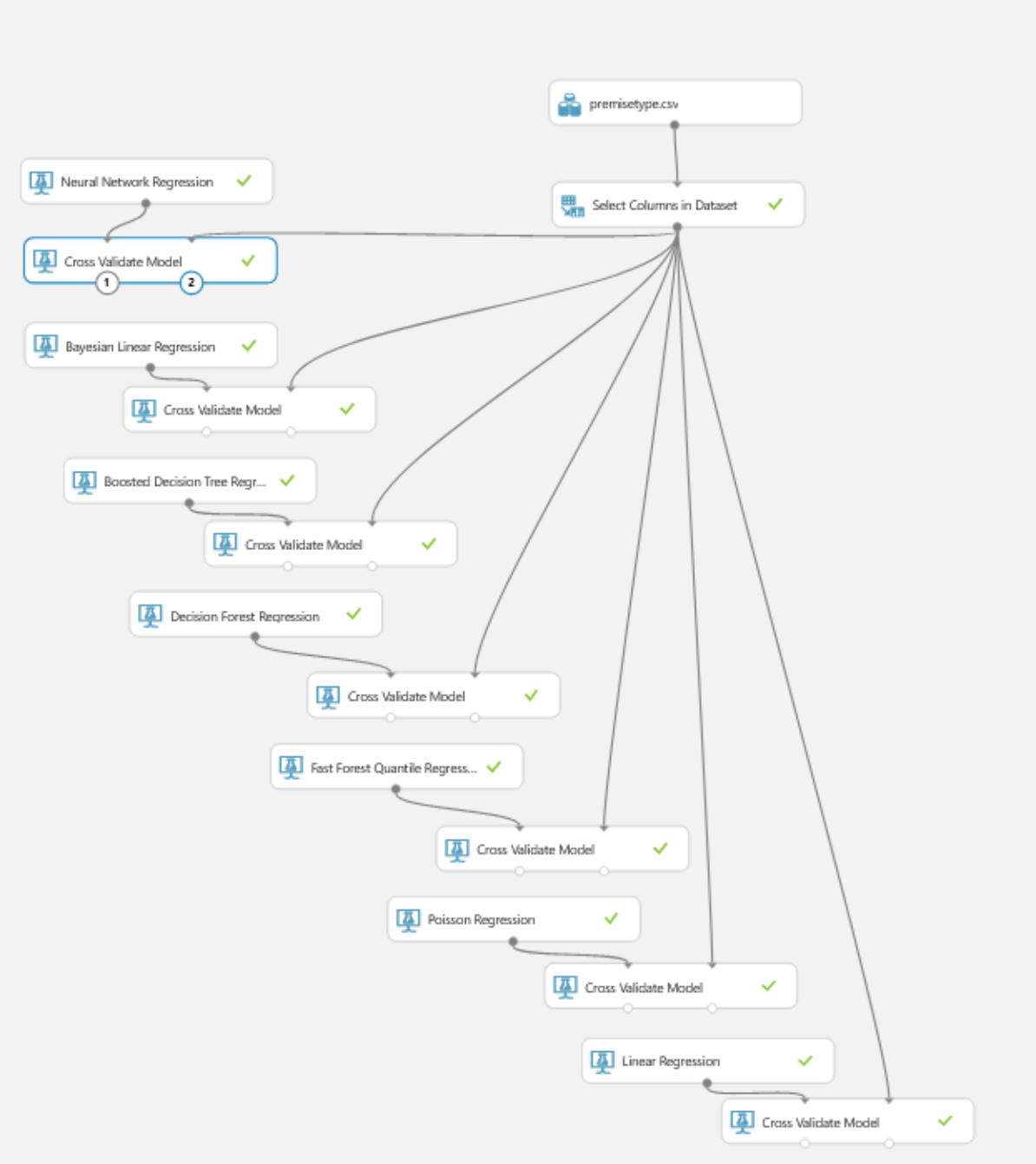
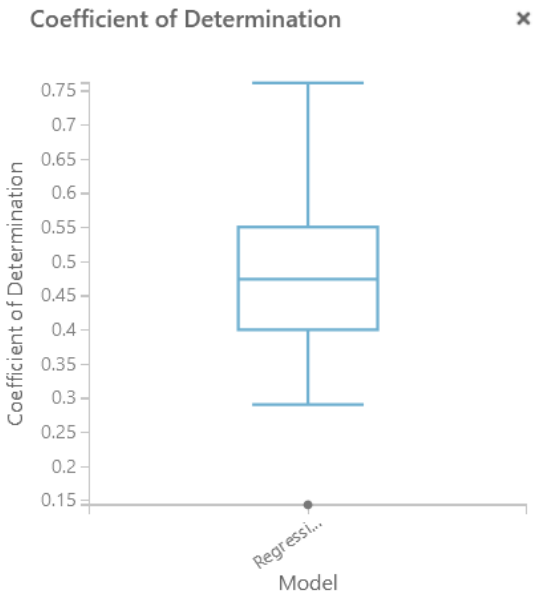
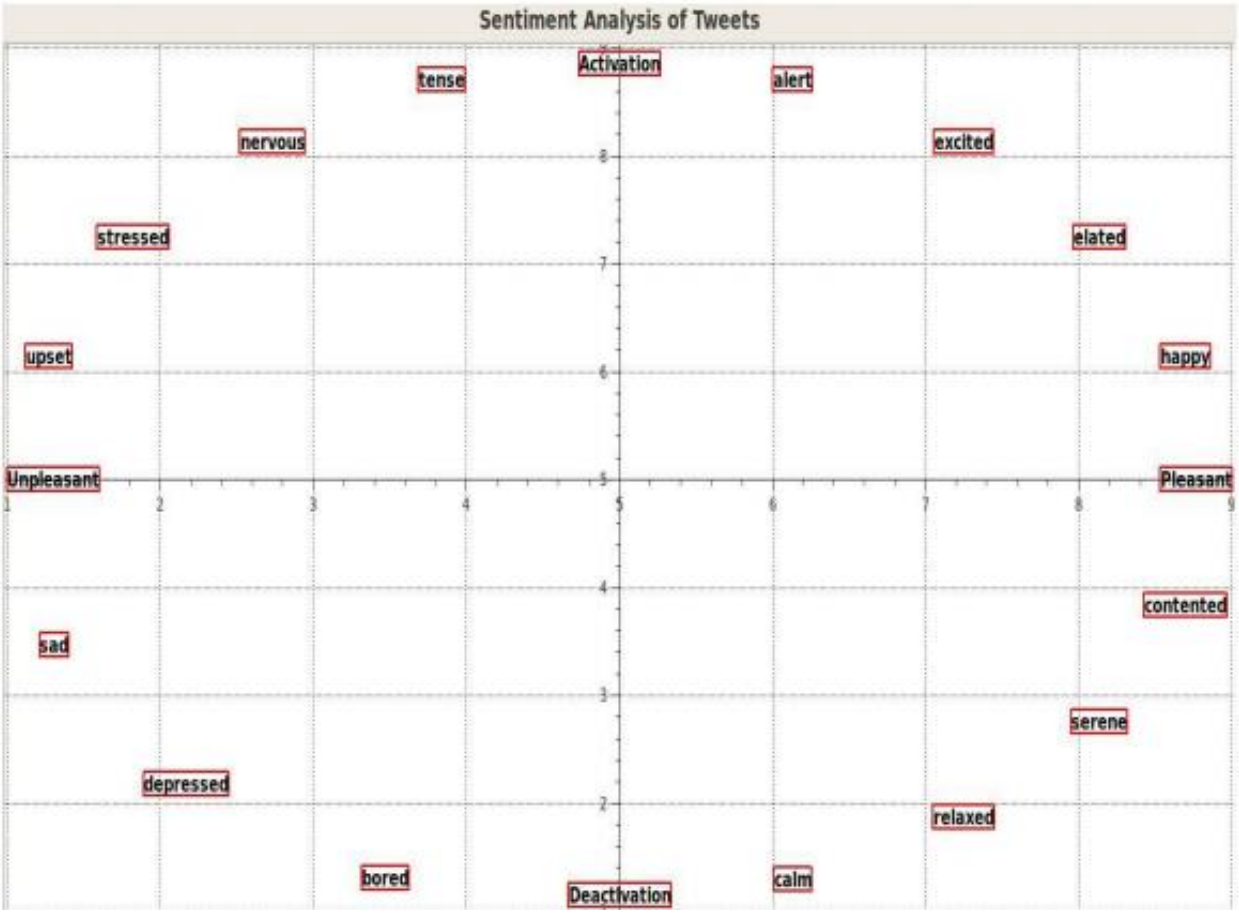


Figure B3-2: The Image below is a boxplot of the coefficients of determination from the 10 folds of the Neural Network regression model.



Appendix C - Conclusion

Figure C1-1: The image below is one possible model for doing multi-faceted sentiment analysis in order to get a better understanding of peoples' emotions.



Bolla, Raja Ashok, "Crime pattern detection using online social media" (2014). Masters Theses. 7321. https://scholarsmine.mst.edu/masters_theses/7321