Machine Learning and Crime Prediction from Location-based Sentiment Analysis of Twitter Data

Uchi Uchibeke (uiu040)

University of Saskatchewan, Canada

Machine Learning and Crime Prediction from Location-based Sentiment Analysis of Twitter Data

Police departments have established different strategies to optimally allocate officers to different subdivisions of their departments. They do this to provide the best service to their community and keep the community and people in it safe. They aim to make the best of their scarce resources for this purpose. Chen, Cho, and Jang (2015) suggest that automatic crime prediction is a key technique used by police departments to better serve the community by identifying optimal patrol strategies (p. 63). Technological innovation and the availability of real-time and on-demand user-generated data have made current crime prediction methods used by various police departments suboptimal. Gaber (2014) proposed the hypothesis that the location, timing, and content of tweets are informative for predicting future events like criminal activity in a location (p. 2). Gaber answered the question: can we use the tweets posted by residents in a major U.S city to predict criminal activity? In his study and research paper, he answered affirmatively to that question.

Gaber (2014) achieved this by doing three things. (1) quantify the crime prediction gains achieved by adding Twitter-derived information to a standard crime prediction approach based on kernel density estimation (KDE), (2) identify existing text processing tools and associated parameterizations that can be employed effectively in the analysis of tweets for crime prediction, and (3) identify performance bottlenecks that most affect the Twitter-based crime prediction approach. The objective of the current research and paper is to answer the question: can we use the tweets posted by residents in a major Canadian city to predict criminal activity in that city? The current study aims to replicate the Gaber (2014) study.

Problem Statement

The approach of the current study is to employ the process in Gaber (2014) combined with the approach of Chen, Cho, and Jang (2015). This will be achieved by completing four objectives: (1) analyze textual content in Twitter data and perform sentiment analysis to score the emotion, using Ekmans' Atlas of Emotions, for various Canadian cities (2) identify if there is a correlation between the kind emotion found in tweets posted in a Canadian city and the crime rate in that city (3) identify the time of the day that crimes occur and their relationship with tweets posted at the time with the goal of predicting crime incidents (4) set the stage for further work on building and evaluating Twitter-derived data from Canadian cities.

Literature Review

Emotions and Crime over the Life Course

Giordano's et al (2007) work, *Emotions and Crime over the Life Course: A Neo - Meadian Perspective on Criminal Continuity and Change*, highlighted ways in which emotions influence long-term patterns of criminal involvement. They found the following:

> Anger identity is associated with variations in adult crime, net of the actor's current level of social bonding or early delinquency. Depression also distinguishes stable desisters from persisters and those who evidenced an unstable pattern. Significant associations between these aspects of emotional identity (particularly anger identity) and violence as well as drug/alcohol problems are also potentially important, as these adult problem areas often figure directly or indirectly into further adult legal difficulties. (p. 47)

This shows that emotions influence long-term patterns of criminal involvement. What does this mean for research in crime prediction? Because real-time geotagged data from users can be mined online, this means that we can improve the efficiency of crime prediction by combining emotion data from users that live in a city.

Method

Data Collection

There were two primary sources of data: (1) geotagged twitter data posted by users and collected in real-time using the official twitter data collection service, and (2) offense data from the Toronto Police Service, grouped by major crime indicators (MCI). Data sources from other Canadian cities were investigated but found to be inconsistent or too small to be statistically significant. Given the timeframe of this project, the Saskatoon police service was not contacted to provide offense data for the city of Saskatoon. In Saskatchewan, the only information that was found online were graphs of crimes and their rate in different Saskatoon neighborhoods. Raw data was not available online. To that end, other cities were explored and Toronto was the best choice because they provide crime data grouped by major crime indicators in a format that is analytically meaningful. Additionally, most tweets collected were found to be geotagged with Toronto as the location.

Tools

Twitter Service: Data collection was done using the official twitter data collection service. As a Developers, I signed up for a data access account and I was assigned a key that allowed me to collect user post in real-time.

Pandas: Pandas was used for data collection, filtering, process and analysis. Pandas offer many mathematics and statistics function that allow complex data process and analysis

Python: All the tools were imported into the python programming language environment. The sentiment analysis algorithm, data filtering and saving were all implemented in python.

Sklean: Sklean is a machine learning framework that enables planning, developing and training of machine learning models

Numpy: Numpy provides complex mathematics functions for data analytics and processing

Process

More than 150,000 tweets were collected between November 21st and November 30th, 2017. Five (5) concurrent data collection scripts were set up to collect tweets posted from Saskatoon, Toronto, Vancouver, Calgary and Canada. The collection happened every 5 seconds for the 5 groups and the tweets were grouped by the day they were collected. For each day, they were grouped by the hour were collected. Tweets with less than 50 English characters were discarded during the collection process.

The next process was preprocessing the tweets. This involved removing non-English characters, stop words, hashtags, URLs and phone numbers and other punctuations that will confuse the prediction algorithm. Tweets with less than 50 characters were then discarded, again.

Sentiment Analysis

To use the tweets for prediction, sentiment analysis was used to determine if a tweet was positive or negative. Ekman's' Atlas of Emotions was first used to group the tweet. Using a pre-trained machine learning model, the tweets were grouped into the groups developed by Paul Ekman, the Ekmans' Atlas of Emotions. For the present study, tweets were grouped into one of Anger, Disgust, Fear, Joy, Sadness, or Surprise. Tweets in the Anger, Disgust, and Sadness group

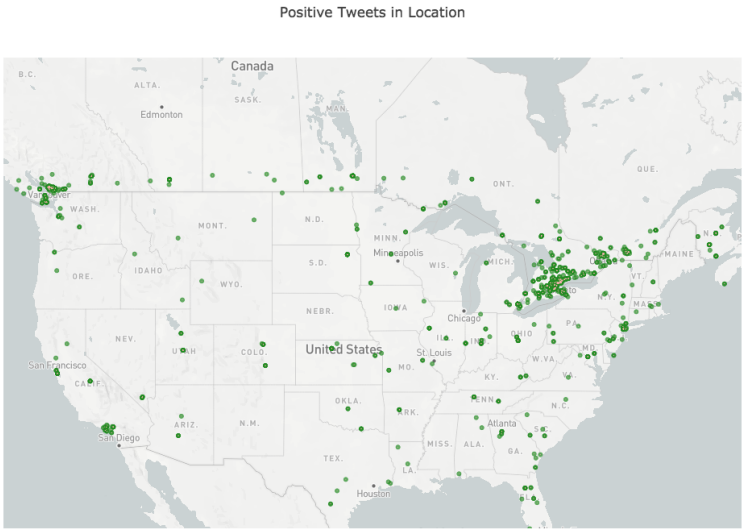
were then labeled as negative tweets and the remaining as positive tweets. Two columns, positive and negative was added to the data set and this was used for analysis.

		followers	lat	lon	negative	positive
Emotion						
Anger	**count**	315.00	315.00	315.00	315.00	315.00
	mean	1368.62	38.83	-72.54	1.00	0.00
	std	2730.93	19.04	40.76	0.00	0.00
	min	1.00	-23.56	-150.05	1.00	0.00
	25%	229.00	43.53	-95.91	1.00	0.00
	50%	504.00	43.66	-79.39	1.00	0.00
	75%	1287.50	50.07	-74.83	1.00	0.00
	max	34224.00	62.84	73.53	1.00	0.00
Disgust	**count**	121.00	121.00	121.00	121.00	121.00
	mean	1163.32	37.51	-62.66	1.00	0.00
	std	1512.18	21.52	49.60	0.00	0.00
	min	3.00	0.00	-125.10	1.00	0.00
	25%	201.00	39.32	-111.68	1.00	0.00
	50%	565.00	43.66	-79.38	1.00	0.00
	75%	1431.00	51.05	0.00	1.00	0.00
	max	7185.00	63.03	15.02	1.00	0.00
Fear	**count**	720.00	720.00	720.00	720.00	720.00
	mean	3343.52	39.12	-71.65	0.00	1.00
	std	20053.18	18.74	45.32	0.00	0.00
	min	0.00	-28.07	-150.05	0.00	1.00
	25%	194.75	43.65	-97.61	0.00	1.00
	50%	619.00	43.66	-79.39	0.00	1.00
	75%	1639.00	49.26	-74.01	0.00	1.00
	max	350033.00	67.27	153.36	0.00	1.00
Joy	**count**	4588.00	4588.00	4588.00	4588.00	4588.00
	mean	5089.53	40.71	-76.14	0.00	1.00
	std	69673.25	17.43	42.63	0.00	0.00
	min	0.00	-44.06	-178.22	0.00	1.00
	25%	234.75	43.65	-113.21	0.00	1.00
	50%	653.50	43.66	-79.39	0.00	1.00
	75%	2202.00	49.27	-75.69	0.00	1.00
	max	3134323.00	64.23	174.92	0.00	1.00
Sadness	**count**	1461.00	1461.00	1461.00	1461.00	1461.00
	mean	2674.60	39.14	-72.30	1.00	0.00
	std	16201.90	18.59	44.61	0.00	0.00
	min	1.00	-38.08	-154.22	1.00	0.00
	25%	223.00	43.59	-97.61	1.00	0.00
	50%	578.00	43.66	-79.39	1.00	0.00
	75%	1702.00	49.26	-73.55	1.00	0.00
	max	404149.00	69.21	148.20	1.00	0.00
Surprise	**count**	4836.00	4836.00	4836.00	4836.00	4836.00
	mean	4099.49	41.02	-75.94	0.00	1.00
	std	66167.16	16.24	39.31	0.00	0.00
	min	0.00	-41.27	-178.22	0.00	1.00
	25%	222.00	43.65	-95.91	0.00	1.00
	50%	418.00	43.66	-79.39	0.00	1.00
	75%	1348.00	49.26	-79.38	0.00	1.00
	max	3134323.00	72.70	173.28	0.00	1.00

Figure 1. Summary of emotion types

The density of the negative sentiments and that of the positive sentiments were plotted on a graph to show how negative and positive a city is, respectively.

Data Analysis



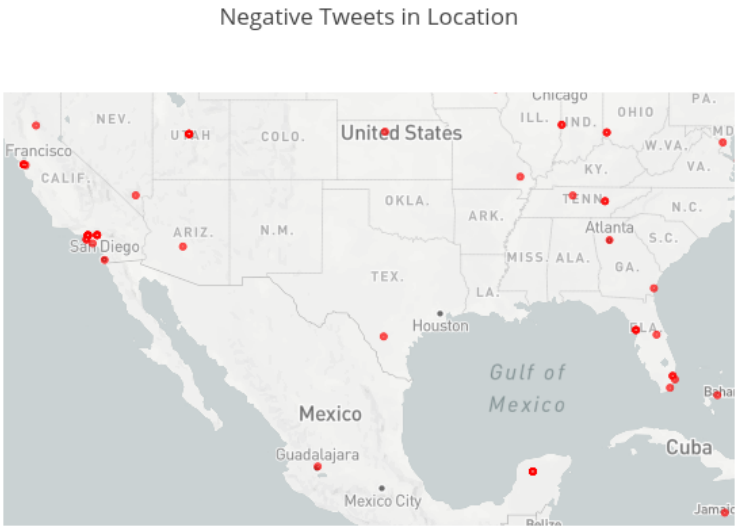


Figure 2.a Positive Emotions | Figure 2.b Negative Emotions

Multiple analysis method was applied to the data set. One of the methods was Kernel Density estimation (KDE). This did not yield any meaningful result because of overfitting. Additionally, there was limited time to tune the estimates to get meaningful statistics.

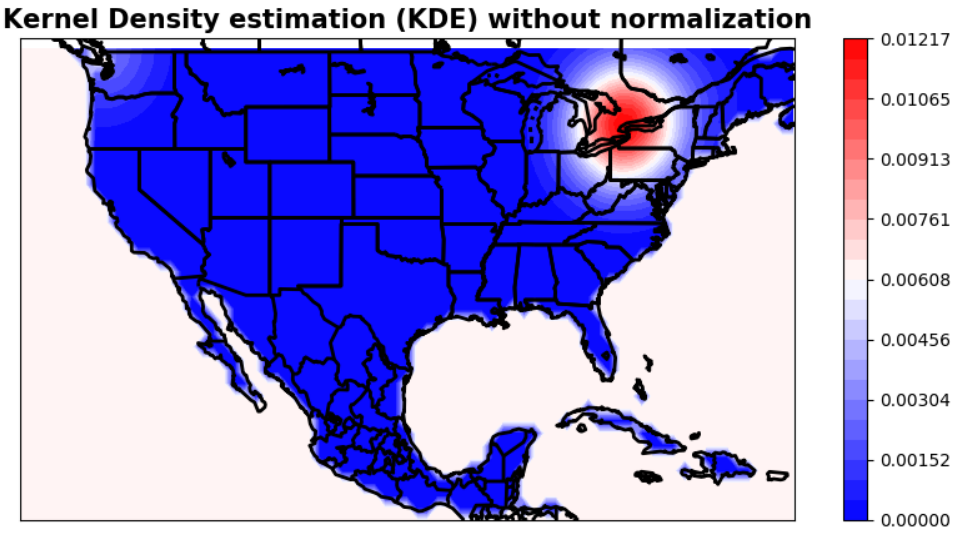


Figure 3. Kernel Density Estimation (KDE)

Results

The present study successfully harnessed hundreds of thousands of tweets and analyzed textual content in the data using sentiment analysis to score the emotion, based on Ekman's' Atlas of Emotions. The goal was to find a trend for various emotions in Canadian cities. The raw data and result from the analysis will be useful and readily available for researchers who are hoping to do a more extensive research with twitter data posted in Saskatoon, Toronto, Vancouver and other Canadian cities. The processed data and analytics can also be refined for a more extensive study. More importantly, the method used in the data collection and analysis can be applied to harvesting data from multiple online sources and geotagged with the different location.

The present study also used pre-trained emotion models to predict the emotion of a tweet text. The model and process can also be applied to more extensive research projects for sentiment analysis. The result from a prediction of crime rate in a city is, however, inconclusive. Given the limited time and short-term data collected, applying a prediction model to detect crime from tweets and the time they occurred will hardly accurate.

Future Work

Prior to this project, large real-time geotagged online messages from Saskatoon has never been collected. Specifically, geotagged messages with the GPS coordinates of the user posting it has not been readily available. The data collected in this study can be used for other studies and the method employed in the current study can be used to collect data from other locations, for specific demographics, and with certain keywords and emotions. This has the potential of empowering law enforcement to prevent crime before they occur.

The analysis of the data in the current study is certainly suboptimal. Future work and investigation need to be applied to the current data and crime prediction process described to enable us to get meaningful insights that will allow law enforcement to continue to stay ahead of the game.

References

Chen, X., Cho, Y., & Jang, S. (2015). Crime prediction using Twitter sentiment and weather. *2015 Systems And Information Engineering Design Symposium*. http://dx.doi.org/10.1109/sieds.2015.7117012

Data Set of Toronto Offences. (2017). *Data.torontopolice.on.ca*. Retrieved 1 December 2017, from http://data.torontopolice.on.ca/datasets/

Feature Extraction for Sentiment Classification on Twitter Data. (2016). *International Journal Of Science And Research (IJSR)*, *5*(2), 2183-2189. http://dx.doi.org/10.21275/v5i2.nov161677

Gerber, M. (2014). Predicting crime using Twitter and kernel density estimation. *Decision Support Systems*, *61*, 115-125. http://dx.doi.org/10.1016/j.dss.2014.02.003

Giordano, P., Schroeder, R., & Cernkovich, S. (2007). Emotions and Crime over the Life Course: A Neo-Meadian Perspective on Criminal Continuity and Change. *American Journal Of Sociology*, *112*(6), 1603-1661. http://dx.doi.org/10.1086/512710

kandluis/crime-prediction. (2017). *GitHub*. Retrieved 1 December 2017, from https://github.com/kandluis/crime-prediction

Shendruk, A., & Treble, P. (2017). *Canada's most dangerous cities 2016: How safe is your city?*. *Macleans.ca*. Retrieved 1 December 2017, from http://www.macleans.ca/news/canada/canada-most-dangerous-cities-2016-safe-your-city/

Text analysis of Trump's tweets confirms he writes only the (angrier) Android half. (2017). *Variance Explained*. Retrieved 1 December 2017, from http://varianceexplained.org/r/trump-tweets/

Yang, C., & Ng, T. (2007). Terrorism and Crime Related Weblog Social Network: Link, Content Analysis and Information Visualization. *2007 IEEE Intelligence And Security Informatics*. http://dx.doi.org/10.1109/isi.2007.379533