Car Theft Reports: a Temporal Analysis from a Social Media Perspective

Juglar Diaz
University of Chile
Santiago, Chile
juglar.diaz@ug.uchile.cl

Barbara Poblete
University of Chile
Santiago, Chile
barbara@poblete.cl

ABSTRACT

Complex human behaviors related to crime require multiple sources of information to understand them. Social Media is a place where people share opinions and news. This allows events in the physical world like crimes to be reflected on Social Media. In this paper we study crimes from the perspective of Social Media, specifically car theft and Twitter. We use data of car theft reports from Twitter and car insurance companies in Chile to perform a temporal analysis. We found that there is an increasing correlation in recent years between the number of car theft reports in Twitter and data collected from insurance companies. We performed yearly, monthly, daily and hourly analyses. Though Twitter is an unstructured source and very noisy, it allows you to estimate the volume of thefts that are reported by the insurers. We experimented with a Moving Average to predict the tendency in the number of car theft reported to insurances using Twitter data and found that one month is the best time window for prediction.

CCS CONCEPTS

• **Information systems → Social networks**; *Spatial-temporal systems.*

KEYWORDS

Twitter, Temporal patterns, Car thefts

ACM Reference Format:
Juglar Diaz and Barbara Poblete. 2019. Car Theft Reports: a Temporal Analysis from a Social Media Perspective. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW'19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3308558.3316462

1 INTRODUCTION

Social media has been the source of information in multiple studies for understanding events in the physical world. The study of complex human activities such as criminal acts, in particular car theft, requires the analysis of multiple sources of information. In this context, social networks can contribute with valuable information that can not be found in other sources. Twitter[1] is one of the most popular social networks in Chile. In Twitter opinions and statements are shared in real time through messages with a maximum of 140 characters, called tweets. People share different types of information in Twitter including reports of car theft with the hope that it will help them to recover the stolen car. Tweets have an associated timestamp corresponding to the moment where the tweet was created. This means that temporal patterns of car theft can be extracted from information published on Twitter. Also, car insurance companies collect data about car theft reported by their clients. The objective of this work is to study temporal patterns of car theft from two perspectives: social media car theft reports, specifically Twitter and car theft insurance records (CIR).

Many have studied how to use Twitter to better understand events in the physical word [5, 6, 9, 11, 12, 16, 17]. Temporal patterns have been analyzed to study activities on urban environments [16], mobility models [17], event detection [12] and forecasting [5]. Most of the works in the intersection of crime incidents and social media data have focused on enriching event prediction models based on historical records with information extracted from Twitter. Patterns that have been extracted as Twitter indicators are Sentiment Analysis [10] and Topic Modeling [1].

In this work we use Twitter to better understand the car theft phenomenon with an emphasis on temporal patterns. With that end we study yearly, monthly, weekly, daily and hourly patterns using the two datasets. Twitter car theft reports are a sample of the whole population. Also, Twitter reports covers both insured and not insured cars. As CIR only covers those cars that are insured, the two datasets are complementary. Twitter data is freely available and produced in real time which allows analysis and predictions in the moment.

Resuming in this work we explore temporal patterns of car theft reports in Twitter and insurances. We explore frequency patterns for recent years in both datasets and study the number of daily and hourly reports. Finally we explore how we can predict the tendency of number of reports in the insurances from Twitter data. Our findings in this paper are:

- In recent years there have been an increasing correlation between the number of car theft reports in Twitter and the number of car theft reports in insurance companies.
- Twitter, despite being an unstructured source and very noisy, allows us to estimate the volume of theft that are reported in the insurance companies.
- Most of the robberies take place late in the night with a difference between weekends and weekdays. For weekdays robberies are more common in the morning while for weekends is more common the early morning.

[1] https://twitter.com/

WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA

ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308558.3316462

- The best time window estimator for the number of car theft reports for the insurers using Twitter data is a one-month time window.

It is worth mentioning that this work is complementary to that presented by [13] (In Spanish) in the sense that we use the same datasets as them but our analysis is entirely new. The rest of this paper is organized as: Section 2 describes related works aiming to better understand crime incidents. Section 3 presents the description of the experiments and finally in Section 4 we present the conclusions.

2 RELATED WORKS

In this section we describe works directly related to temporal patterns of car theft. Then we describe works that use social network data to alleviate and better understand crime incident related problems.

Car theft: In [4] Chen et al. analyzed car theft report data from Taiwan and discovered that the sooner the robbery is reported, the more likely it is that the car is recovered. In [3] Chen focused their analysis on the time elapsed after the incident was notified. They found that criminals prefer to commit crimes between 4:00 AM and 8:00 AM.

Social networks and Crime incidents: The main focus to enhance crime incident applications with social network data has been enriching predictive models based on historical records with information extracted from Twitter. Patterns extracted as Twitter indicators are Sentiment Analysis [10] and Topic Modeling [1]. In the work of [2] the authors use Twitter data from the ten most violent cities and the ten least violent cities in the United States. In their findings less violent cities presented a lower proportion of tweets associated with violent acts than the rest of the cities. Also, cities with the highest rates of violence presented a higher intensity of negative tweets, although some of the less violent ones also presented high levels of tweets with negative sentiments. In [5] Chen et al. proposed a crime prediction model that indicates when and where a crime will occur. Their experiments showed that the inclusion of sentiment information from Twitter improved the quality of the model. The work reported in [7] aims to create a model to predict the probability that a certain type of crime occurs in a certain place for the next day. Gerber compared two types of models, the first uses only historical records of crimes, the second includes the data from Twitter, which is processed using topic modeling. Of the 25 types of crimes considered, 19 showed improvements in the prediction when the topics of Twitter were added to the predictive model. The goal of Wang et al. in [15] is to predict whether a crime incident of hit-and-run will happen on day d considering the topics discussed on Twitter on day $d - 1$. To add the information from Twitter to the prediction model [15] follows two steps: first the tweets are processed by a Semantic Role Labeling [8] tool and then topics are extracted using LDA [1]. The model performed better than a baseline based on a uniform distribution. In [13] Vásquez et al. present an analysis similar to our proposal in the sense that they also compare car theft reported in Twitter with car theft records from insurances. Their main findings are that robberies reported on Twitter present a higher rate of recoveries and that in general cars reported on Twitter are of lower value than those reported by insurance companies.

Our work is different from the ones we described in this section in that we analyze temporal patterns of different granularities. We study yearly, monthly, daily and hourly patterns in two related datasets: car theft reports from insurance companies and Twitter car theft reports both in Santiago, Chile and in the same period of time.

3 EXPERIMENTS

Our goal is to discover new information and similarities about car theft from two different data sources. In particular, we focus on what temporal patterns we can discover from car theft reports on the social network Twitter in comparison to data of car theft reports from insurance companies in Chile.

3.1 Data

For our experiment we used two datasets of car theft reports in Chile from Jan 1, 2012 to Jan 31, 2016. One was collected from Twitter and the other from insurance companies. The Twitter dataset was collected querying to *https://twitter.com/search-home.* The query was: *[(robar | robado | robaron | robo | robada) & patente & near:Chile]* it translates to *[(steal | stolen | stolen | steal | stolen) & license plate & near:Chile]* . After deleting duplicates and data cleaning resulted in a dataset of around 7,000 records. The dataset from insurance companies consists of around 40,000 car theft reports submitted by users to car insurance companies in Chile. The timestamp is an estimation of when the incident occurred according to the client. The dataset was provided to us by the National Association of Car Insurance Companies through a collaboration agreement.[2]

3.2 Analysis of Temporal Patterns

Human activities are mostly associated to temporal patterns with different granularities as day of the week and time of day. In the morning children are taken to school and the movement of people occurs from home to the workplace. Family activities usually take place on weekends, while afternoons and evening activities are more related to leisure and nightlife in the city. These types of activities can be associated with patterns of car theft.

First we show in Figure 1 the number of robberies per month normalized according to the total (for each dataset) from 2012 to 2016. We can see that there is a great correspondence between the number of reports from Twitter and from CIR. There is a correlation of 0.42.

Later we disaggregate the graph from Figure 1 by years: 2013, 2014, 2015 and 2016 in Figure 2. By separating the data for years we can observe an increase in the correlation between the volume of thefts reported on Twitter and CIR in recent years. All this indicates that Twitter, despite being an unstructured source and very noisy, allows us to estimate the volume of car theft that are reported by the insurers.

A similar relation is found regarding the distribution of thefts by days of the week. Figure 3 shows the number of thefts per day of the week of all the reports in both data sources. In general,

[2] Project FONDEF project ID16I10222, CONICYT.

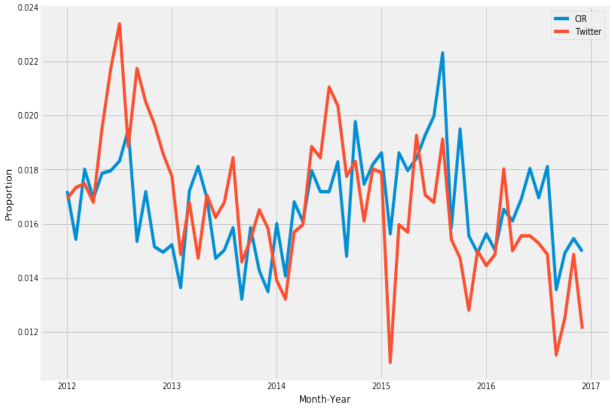

Figure 1: Number of car theft per month normalized according to the total from 2012 to 2016. Correlation 0.42.

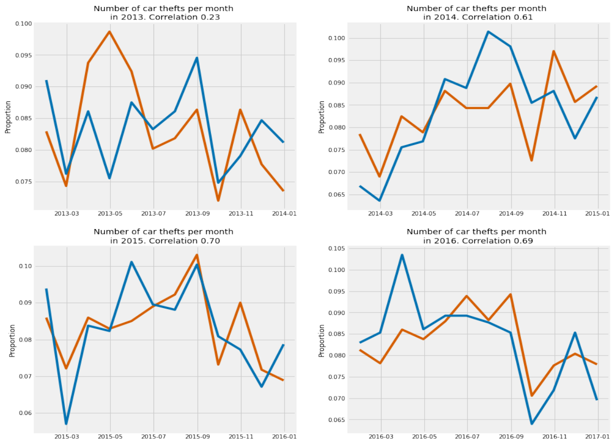

Figure 2: Number of car theft per month normalized according to the total for years 2013, 2014, 2015 and 2016.

thefts increase until mid-week, gradually decreasing towards the weekend.

Figures 4 and 5 show the normalized number of robberies by day of the week and hour of the day. The information associated with the time in both data sources is not the real moment when the robbery occurred. In the case of Twitter, it is when the tweet is created, while in the case of CIR it is the time of the robbery reported by the owner of the car and it is expected to be an approximated value.

In CIR most cases are concentrated late at night and early in the morning, while on Twitter they concentrate in the afternoon. This behavior is consistent with the fact that in CIR the timestamp is the timestamp reported by the owner, while in Twitter it is the time when the tweet is written. This may be due to the victims having more time to report the theft on Twitter during these hours. In the case of CIR on weekends there is an increase at dawn, while that on weekdays this increase is transferred to the morning hours.

From these exploratory analysis we can see that Twitter, despite being an unstructured source and very noisy, allows for an estimate

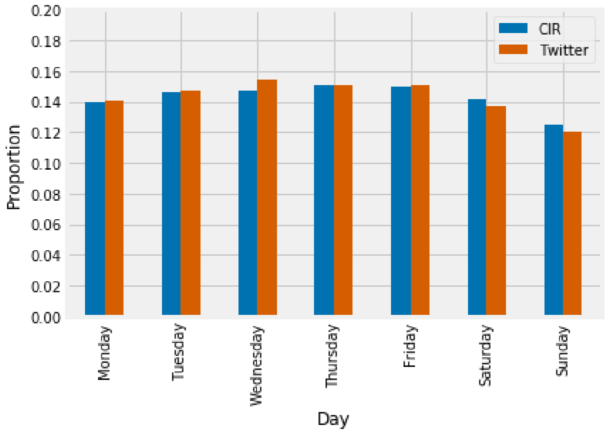

Figure 3: Number of car theft reports by day of the week.

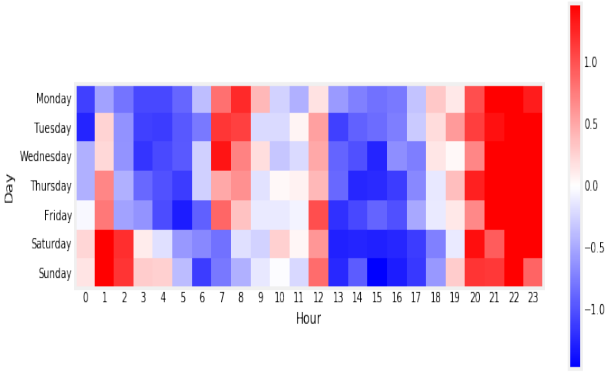

Figure 4: Normalized number of car theft reports by day of the week and hour of the day in CIR.

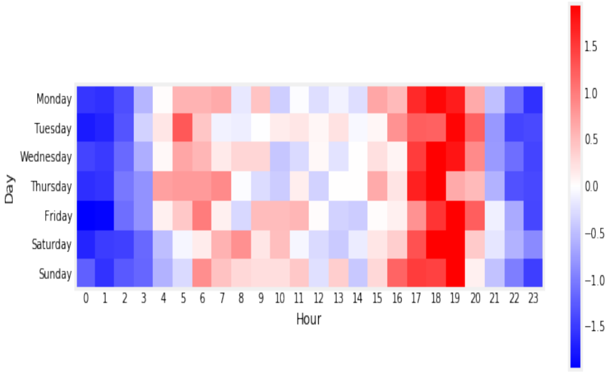

Figure 5: Normalized number of car theft reports by day of the week and hour of the day in Twitter.

of the volume of thefts that are reported to the insurers. Finally we study the best time window to make predictions about the tendency of car theft reports in the CIR using Twitter data. For that purpose we use a Moving Average (MA) from Twitter to predict the tendency of the car theft reports in CIR. We use the Pearson correlation coefficient as evaluation metric. In Figure 6 we show MA for 1,2,3,4 and 5 months. We can see that the best correlation between the MA and CIR is using the last month as predictor. In Figure 7 we show a finer granularity of the prediction by predicting the number of reports weekly. We can see that despite testing with fewer (1,2,3) weeks than a month, still the mean of one month of reports in Twitter is the best predictor of the tendency for the number of car theft reports in CIR.

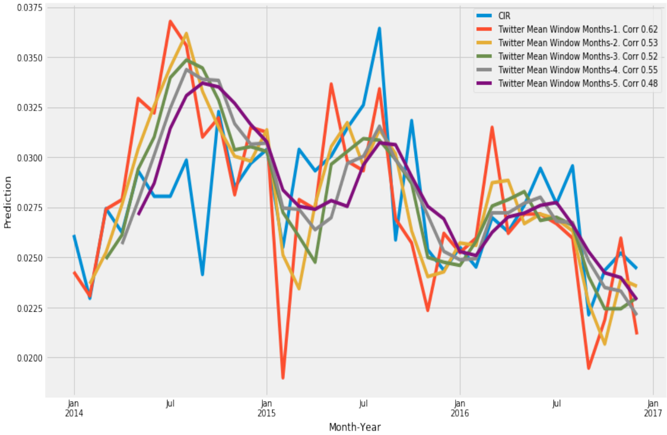

Figure 6: Monthly Moving Average from Twitter to predict the tendency of the car theft reports in CIR.

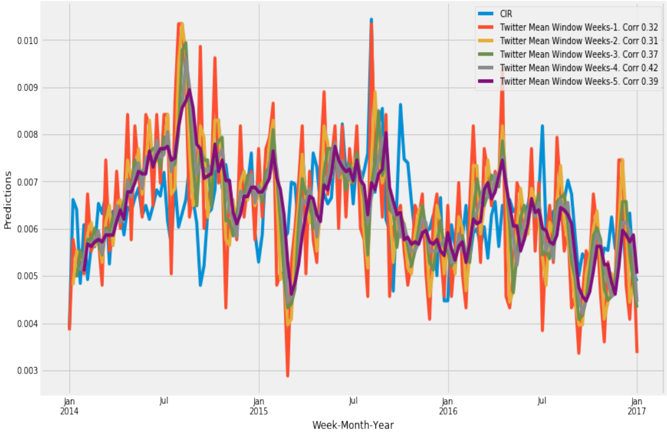

Figure 7: Weekly Moving Average from Twitter to predict the tendency of the car theft reports in CIR.

4 CONCLUSIONS

In this work we studied the car theft phenomenon from a social media perspective. We studied temporal patterns and found that there is an increasing correlation in recent years between the number of car theft reports in Twitter and data collected from insurance companies. Twitter, despite being an unstructured source and very noisy, allows us to estimate of the volume of thefts that are reported to the insurers. We experimented with a Moving Average to predict tendency in the number of reports using Twitter data and found that one month is the best predictor.

As future work we are interested in predicting not only the tendency of the number of reports, but the exact number of reports. Also, we would like to enrich the analysis with general information from Twitter and study finer granularities of prediction like predicting the number of thefts by branches.

ACKNOWLEDGMENTS

The work of Juglar Diaz was supported by CONICYT-PCHA/Doctorado Nacional/2016-21160142.

REFERENCES

[1] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3, Jan (2003), 993–1022.
[2] Raja Ashok Bolla. 2014. Crime pattern detection using online social media. (2014).
[3] Patrick S Chen. 2008. Discovering Investigation Clues through Mining Criminal Databases. In *Intelligence and Security Informatics*. Springer, 173–198.
[4] Patrick S Chen, KC Chang, Tai-Ping Hsing, and Shihchieh Chou. 2006. Mining criminal databases to finding investigation cluesâĂŤby example of stolen automobiles database. In *Intelligence and Security Informatics*. Springer, 91–102.
[5] Xinyu Chen, Youngwoon Cho, and Suk Young Jang. 2015. Crime prediction using Twitter sentiment and weather. In *Systems and Information Engineering Design Symposium (SIEDS), 2015*. IEEE, 63–68.
[6] Jessica Elan Chung and Eni Mustafaraj. 2011. Can collective sentiment expressed on twitter predict political elections?. In *AAAI*, Vol. 11. 1770–1771.
[7] Matthew S Gerber. 2014. Predicting crime using Twitter and kernel density estimation. *Decision Support Systems* 61 (2014), 115–125.
[8] Daniel Gildea and Daniel Jurafsky. 2002. Automatic labeling of semantic roles. *Computational linguistics* 28, 3 (2002), 245–288.
[9] Niels Buus Lassen, Rene Madsen, and Ravi Vatrapu. 2014. Predicting iphone sales from iphone tweets. In *Enterprise Distributed Object Computing Conference (EDOC), 2014 IEEE 18th International*. IEEE, 81–90.
[10] Bing Liu. 2012. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies* 5, 1 (2012), 1–167.
[11] Anshul Mittal and Arpit Goel. 2012. Stock prediction using twitter sentiment analysis. *Standford University, CS229 (2011 http://cs229. stanford. edu/proj2011/GoelMittal-StockMarketPredictionUsingTwitterSentimentAnalysis. pdf)* 15 (2012).
[12] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. 2010. Earthquake shakes Twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*. ACM, 851–860.
[13] Alejandro Vásquez, Rodrigo Joannon, and Richard Weber. 2018. Análisis de redes sociales para mejorar la identificación de patrones de robo de vehıculos. *Revista Ingenierıa de Sistemas Volumen XXXII* (2018).
[14] Xiaofeng Wang, Donald E Brown, and Matthew S Gerber. 2012. Spatio-temporal modeling of criminal incidents using geographic, demographic, and Twitter-derived information. In *Intelligence and Security Informatics (ISI), 2012 IEEE International Conference on*. IEEE, 36–41.
[15] Xiaofeng Wang, Matthew S Gerber, and Donald E Brown. 2012. Automatic crime prediction using events extracted from twitter posts. In *International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction*. Springer, 231–238.
[16] Chao Zhang, Keyang Zhang, Quan Yuan, Haoruo Peng, Yu Zheng, Tim Hanratty, Shaowen Wang, and Jiawei Han. 2017. Regions, periods, activities: Uncovering urban dynamics via cross-modal representation learning. In *Proceedings of the 26th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 361–370.
[17] Chao Zhang, Keyang Zhang, Quan Yuan, Luming Zhang, Tim Hanratty, and Jiawei Han. 2016. Gmove: Group-level mobility modeling using geo-tagged social media. In *KDD: proceedings. International Conference on Knowledge Discovery & Data Mining*, Vol. 2016. NIH Public Access, 1305.