Automatic Crime Prediction using Events Extracted from Twitter Posts

Xiaofeng Wang, Matthew S. Gerber, and Donald E. Brown

Department of Systems and Information Engineering, University of Virginia
{xw4u,msg8u,brown}@virginia.edu

Abstract. Prior work on criminal incident prediction has relied primarily on the historical crime record and various geospatial and demographic information sources. Although promising, these models do not take into account the rich and rapidly expanding social media context that surrounds incidents of interest. This paper presents a preliminary investigation of Twitter-based criminal incident prediction. Our approach is based on the automatic semantic analysis and understanding of natural language Twitter posts, combined with dimensionality reduction via latent Dirichlet allocation and prediction via linear modeling. We tested our model on the task of predicting future hit-and-run crimes. Evaluation results indicate that the model comfortably outperforms a baseline model that predicts hit-and-run incidents uniformly across all days.

1 Introduction

Traditional crime prediction systems (e.g., the one described by Wang and Brown [14]) make extensive use of historical incident patterns as well as layers of information provided by geographic information systems (GISs) and demographic information repositories. Although crucial, these information sources do not account for the rich and rapidly expanding social media context that surrounds incidents of interest. An essential part of this context is the stream of information created by users of services such as Facebook[1] and Twitter[2]. These services allow users to instantly create, disseminate, and consume information from any location with access to the Internet. Recently, Howard et al. argued that such services played a key role in the development and perpetuation of the "Arab Spring" uprisings that took place across North Africa and the Middle East beginning in December of 2010 [10]. The authors found, among other things, evidence that social media activity of particular types preceded mass protests and other incidents.

Whereas the study conducted by Howard et al. was retrospective, this paper presents a preliminary investigation of the *predictive* power of social media information, in particular information produced by the Twitter service. We hypothesized that information extracted from the Twitter service would - if properly structured and modeled - provide indicators about the likelihood of future

[1] http://www.facebook.com
[2] http://www.twitter.com

incidents. Our investigation did not attempt to acquire and digest all "tweets" (short messages created by Twitter users); rather, we pulled tweets from the Twitter feed of a news agency covering the area of Charlottesville, Virginia. Consider the following tweets:

(1) TRAFFIC ALERT: Rt. 20 closed due to a wreck.
(2) Road closed at JPA and Shamrock due to tree falling over road.

Intuitively, these tweets provide evidence of an increased hazard level along roadways, which, in turn, might lead to an increased number of accidents or hit-and-run crimes. The goal of our investigation was to build a predictive model of criminal incidents that leverages this type of evidence. We used state-of-the-art natural language processing (NLP) techniques to extract the semantic event content of the tweets. We then identified event-based topics using unsupervised topic modeling, and used these topics to predict future occurrences of criminal incidents. With the tweet information alone, our predictive model comfortably outperformed a uniform prediction baseline on held-out data.

2 Related Work

2.1 Crime Mapping and Prediction

The task of crime prevention is constrained by scarce resources (e.g., time, patrol units, and finances). Analysts, therefore, often employ a variety of computer systems to identify and visualize areas of high crime, otherwise known as "hot-spots" [7]. Crime hot-spots indicate spatial areas of relatively high threat according to some underlying model. A common model - one promoted by Eck et al. - relies on kernel density estimation (KDE) from the criminal history record of an area [7]. KDE is an efficient method of computing a continuous surface, where the relative threat (i.e., "hotness") of an area is indicated by its color and/or vertical height. Chainey et al. investigated the use of KDE for crime prediction [6]; however, KDE as a predictive model suffers from (1) a lack of portability (crimes cannot be predicted for previously unseen regions), and (2) a lack of contextual information such as that coming from social media services. In a similar vein, Mohler et al. applied the self-exciting point process model (previously developed for earthquake prediction) as a model of crime [12]. This model, like ones based on KDE, relies on the prior occurrence of crimes in a particular area and thus cannot generalize to previously unobserved areas.

Wang and Brown proposed a different approach to crime modeling [14]. In their approach, prior criminal incidents are used as supervised training data within the predictive model. Geographic locations are characterized by a rich set of spatial and demographic features instead of the simple geographic coordinates used by KDE-based approaches. Example features include the distance to the nearest business and the number of divorced individuals in the region. This representation permits crime prediction in previously unseen places, as the correlation between, for example, burglary and business proximity can often be

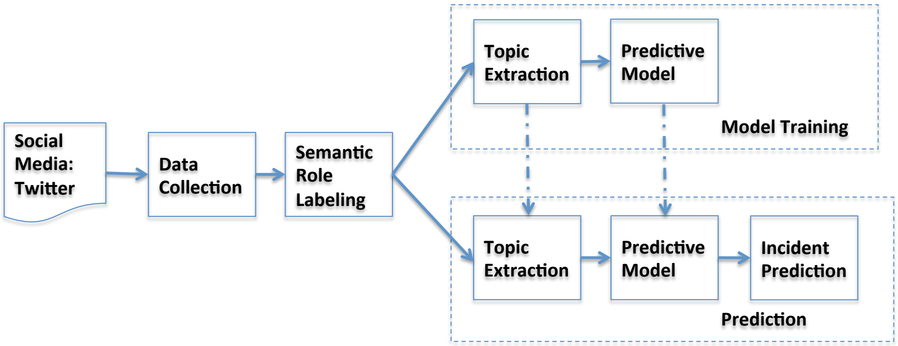


Fig. 1. Overall process of criminal incident prediction using tweets

generalized from one area of a city or country to another area. Furthermore, the feature-based representation permits the addition of information such as that extracted from social media services. Thus, the work presented in this paper should be viewed as complementary to the work reported by Wang and Brown.

2.2 Use of Social Media for Prediction

The proliferation of social media services has prompted a surge of interest in using the associated data for various predictive purposes. For example, Twitter posts have been used to predict box office results [1], election results [2], and stock market trends [5]. Popular techniques in these studies include keyword volume analysis and sentiment analysis. These methods have proven useful for the tasks mentioned above; however, a deeper semantic understanding of tweets is required to predict discrete criminal incidents, which are not mentioned ahead of time (ruling out keyword volume and sentiment analysis).

3 Data Collection and Methods

Figure 1 shows the overall operation of our Twitter-based predictive model. We first collect a corpus of tweets from Twitter. We then extract events from the main textual content of each tweet using an NLP technique known as semantic role labeling (SRL). Next, we apply latent Dirichlet allocation (LDA) to identify salient topics within the extracted events. A predictive model is then built upon these latent topics. The following sections describe these steps in detail.

3.1 Data Collection

The user base of Twitter comprises a vast community of news agencies, journalists, and casual users who post tweets from their Internet-connected devices.

Each tweet is restricted to 140 characters and can be observed by those who subscribe to the poster's Twitter feed. As of March 11, 2011, Twitter was processing approximately 140 million tweets per day, with approximately 460,000 new accounts being created daily.[3] Traditional news stations and newspapers actively use Twitter to publish breaking news in real-time. For example, CBS19[4] in Charlottesville, Virginia published 3,659 tweets during the period of February 22, 2011 through October 21, 2011 (approximately 15 per day). We collected these tweets using the public interface provided by Twitter.

In addition to Twitter data, our investigation required ground-truth criminal incident data, which we used to estimate the parameters of our predictive model and evaluate its performance. We obtained these records from local law enforcement agencies, focusing on hit-and-run incidents during the same period covered by the Twitter data. In total, we collected records for 290 hit-and-run incidents (1.2 per day).[5]

3.2 Methods

Semantic Role Labeling Our approach to Twitter-based crime prediction relies on a semantic understanding of tweets, one that goes beyond bag-of-words and sentiment representations. Such an understanding can be derived from a process known as semantic role labeling (SRL), which extracts the events mentioned in tweets, the entities involved in the events, and the roles of the entities with respect to the events. An example analysis (derived from Example 1, p. 2) is shown below:

(3) [$_{e_1:warning}$ TRAFFIC] [$_{e_1}$ ALERT]: [$_{e_2:entity}$ Rt. 20] [$_{e_2}$ closed] [$_{e_2:cause}$ due to a wreck].

Two events were extracted from Example 3: (1) an *alert* event in which traffic is being brought to the reader's attention, and (2) a *close* event where a road is closed due to a wreck. Note that these events are signaled by a noun and verb, respectively. Gildea and Jurafsky documented the seminal investigation into SRL [9] and the NLP community has had a sustained interest in the technique since then (see [11] for a more recent survey and references). In our study, we used the system created by Punyakanok et al. [13] to analyze verb-based SRL structures and the system created by Gerber and Chai [8] to analyze noun-based SRL structures. In general, the SRL systems were well suited to the news tweets, since the systems were trained on news corpora. In the cited studies, the authors report F_1 scores of approximately 80% for verbal SRL and 72% for nominal SRL. The output from these systems forms the basis for event prediction, since it informs the model about current events, which (we hypothesized) might correlate with future criminal incidents.

[3] `http://blog.twitter.com/2011/03/numbers.html` (accessed November 1, 2011)

[4] `http://www.newsplex.com`

[5] `http://www.charlottesville.org/index.aspx?page=257`

Event-based Topic Extraction via Latent Dirichlet Allocation After processing the tweets with the SRL systems, we have multiple events e_i associated with each day. In topic modeling terms, each day d is associated with an abstract "document" doc_d that contains "words" $\{e_1, e_2, \ldots, e_{n_d}\}$, where n_d is the length of doc_d. These words describe what happened on day d.

As with topic modeling of actual textual documents, we hypothesized that a day's events would be related in a particular (though hidden) way. Thus, instead of using doc_d directly to predict future incidents, we further extracted topics $\{t_1, t_2, \ldots, t_k\}$ from doc_d using latent Dirichlet allocation (LDA) [4][3].[6] LDA is a probabilistic language model that can be used to explain how a collection of documents is generated from a set of hidden (or latent) topics. LDA efficiently discovers word-based topics and reduces the dimensionality of documents to lie within the k-dimensional space of topics. Given the number of topics k, LDA can estimate the topic-document distribution $\{T_{d,1}, T_{d,2}, \ldots, T_{d,k}\}$, where $T_{d,i}$ is the probability that document d is related to topic i.

We applied LDA to derive $\{T_{d,1}, T_{d,2}, \ldots, T_{d,k}\}$ for the events described in tweets on day d. Intuitively, this analysis tells us about the relationship between the k major (latent) events on day d and the observable events e_i that were reported by the news agencies. This reduces the dimensionality of doc_d and provides meaningful structured data for our predictive model, which is described next.

Predictive Model doc_d contains the events that occurred on day d. Our goal is to use doc_d to make predictions about incidents in the future. Formally, we seek a function $y_{d+1} = f(doc_d)$, where y_{d+1} is a binary random variable indicating whether an incident will occur on day d+1. For example, we can use the following generalized linear regression model (GLM):

$$log\left(\frac{Pr[y_{d+1}=1]}{1-Pr[y_{d+1}=1]}\right) = \beta_0 + \beta_1 T_{d,1} + \cdots + \beta_k T_{d,k} \tag{4}$$

Where each $T_{d,i}$ is derived via LDA. Parameters $\{\beta_0, \ldots, \beta_k\}$ can be estimated using the set of prior criminal incidents described in Section 3.1.

With both the estimated LDA model and GLM model, we can make a prediction using new tweets. To make a prediction, we first process tweets on day d' using the SRL systems described above. Then, the LDA model is used to infer the event-based topic distribution $\{T_{d',1}, T_{d',2}, \ldots, T_{d',k}\}$. Lastly, the predictive model (Equation 4) uses this distribution to predict the likelihood of an incident occurring on day $d' + 1$.

4 Evaluation and Results

We evaluated our predictive model using Twitter data and actual hit-and-run incidents that occurred in Charlottesville, Virginia. As described in Section 3.1,

[6] We used GibbsLDA++ in all of our experiments: http://gibbslda.sourceforge.net

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
close	say	arrest	plan	report	expect	report	say	trial	come
fire	make	suspect	kill	say	remain	close	found	make	cbs
crash	hanchettjim	death	use	student	rsb	open	die	break	help
look	search	murder	ask	vote	cancel	confirm	crash	set	start
delay	confirm	shoot	plead	tell	follow	block	fall	traffic	lead
come	run	rsb	life	hear	close	wreck	want	begin	look
reopen	start	hear	sell	work	warn	follow	find	hope	check
stay	move	protest	convict	speak	make	accord	kill	bring	lawsuit
watch	begin	report	visit	head	price	move	shut	hit	arrest
driver	end	need	statement	call	list	check	come	stop	left

Table 1. Top 10 most likely words for each of the 10 topics

our data cover the period of February 22, 2011 through October 21, 2011. We studied the hit-and-run incidents per day using traditional time series methods, but discovered no trend, seasonality, or autocorrelation. Thus, without any additional information, a baseline system would assign a uniform probability of incidents to all future days. This approach constitutes our baseline model.

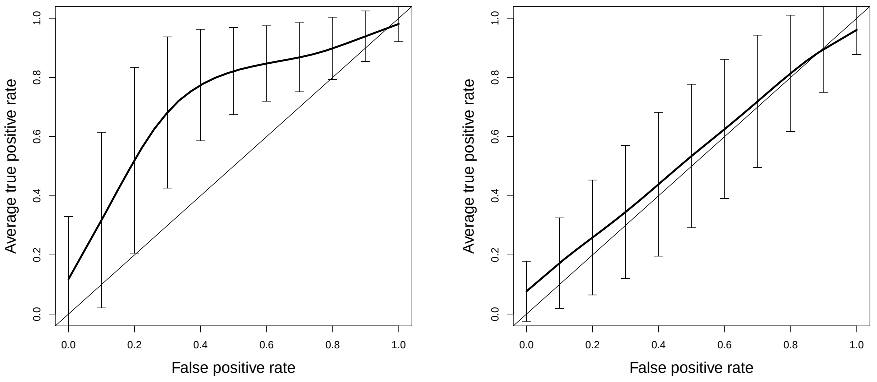
We used the data before September 17, 2011 to train the LDA and predictive models, setting k (the number of latent topics) to be 10. Table 1 presents the top 10 words for each topic. The nature of these topics is subjective; however, some structure is present. For example, topic 1 appears to be related to crashes, whereas topic 3 appears to be related to shootings and their associated criminal processes. We trained a GLM on these topics as described in Section 3.2, using stepwise selection to identify the most informative features. The resulting GLM is shown below:

$$\log\left(\frac{Pr[hit_{d+1}=1]}{1-Pr[hit_{d+1}=1]}\right) = 0.4 + 0.71T_{d,1} + 0.88T_{d,4} + 0.72T_{d,6} + 0.61T_{d,8} \quad (5)$$

In Equation 5, $Pr[hit_{d+1}=1]$ denotes the probability of at least one hit-and-run incident occurring on day $d+1$. $T_{d,\cdot}$ is the topic distribution on day d. As shown, the model emphasizes topics 1, 4, 6, and 8 in the prediction of future hit-and-run incidents.

We applied this model to predict hit-and-run incidents during the period of September 17, 2011 to October 21, 2011. Figure 2(a) shows the ROC curve of the prediction performance. Vertical bars are 95% confidence intervals derived with a bootstrap resampling procedure. The ideal ROC curve stretches toward the upper-left corner. A curve along the diagonal indicates no predictive power. As shown by Figure 2(a), the LDA/GLM model was able to predict future hit-and-run incidents; although, due to the limited amount of testing data, we observed fairly wide confidence intervals. The baseline system ROC curve lies on the diagonal, as it predicts hit-and-run incidents uniformly across all days.

Our SRL systems have a runtime complexity of $O(n^3)$, where n is the number of words and punctuation marks in the sentence. In order to justify this added complexity, we re-trained the predictive LDA/GLM model on all words in the



(a) Using events extracted from tweets

(b) Using all words from tweets

Fig. 2. ROC curves for predicting hit-and-run incidents

tweets instead of event words only. The remaining experimental conditions were held constant, resulting in the ROC curve shown in Figure 2(b). As shown in the figure, the system exhibits minimal predictive power, thus supporting the use of event extraction for incident prediction.

5 Conclusions and Future Work

This paper has presented a preliminary investigation into the use of social media for criminal incident prediction. Although our data source (Twitter) and the prediction domain (criminal incidents) are not novel, we are not aware of any prior work that brings these topics together. Our approach is based on the automatic semantic analysis and understanding of natural language tweets, combined with dimensionality reduction via latent Dirichlet allocation and prediction via linear modeling. Evaluation results demonstrate the model's ability to forecast hit-and-run crimes using only the information contained in the training set of tweets. Given the widespread use of social media services such as Twitter, our results indicate a fruitful line of future research.

There are many ways in which this work can be extended. In our semantic analysis step, we assumed that the events contained in a tweet occurred on the day that the tweet was posted; however, tweets often describe events that occurred days, weeks, or even years ago using overt linguistic expressions (e.g., "Last year's storm..."). This information needs to be taken into account when predicting future incidents, since events in the distant past might lose their influence. Regarding the GLM model, we used a simple stepwise selection method to choose features for prediction. A better alternative might be to apply the penalized GLM (PGLM) to select the most predictive features. Lastly, our approach does not leverage the massive amount of information produced by the

Twitter service. Large-scale analysis of tweets might provide insights that are not apparent within a single feed.

Acknowledgments

We would like to thank the anonymous reviewers for their comments on a previous version of this paper. The work presented in this paper was funded by United States Army Grant W911NF-10-2-0051.

References

1. Asur, S., Huberman, B.: Predicting the future with social media. In: 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology. pp. 492–499. IEEE (2010)
2. Bermingham, A., Smeaton, A.: On using twitter to monitor political sentiment and predict election results. In: Proceedings of the Workshop on Sentiment Analysis where AI meets Psychology (SAAIP 2011). pp. 2–10. Asian Federation of Natural Language Processing, Chiang Mai, Thailand (November 2011)
3. Blei, D., Carin, L., Dunson, D.: Probabilistic topic models. Signal Processing Magazine, IEEE 27(6), 55–65 (2010)
4. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. 3, 993–1022 (2003)
5. Bollen, J., Mao, H., Zeng, X.: Twitter mood predicts the stock market. Journal of Computational Science (2011)
6. Chainey, S., Tompson, L., Uhlig, S.: The utility of hotspot mapping for predicting spatial patterns of crime. Security Journal 21, 428 (2008)
7. Eck, J., Chainey, S., Cameron, J., Leitner, M., Wilson, R.: Mapping crime: Understanding hot spots (2005)
8. Gerber, M., Chai, J., Meyers, A.: The role of implicit argumentation in nominal SRL. In: Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics. pp. 146–154. Association for Computational Linguistics, Boulder, Colorado (June 2009)
9. Gildea, D., Jurafsky, D.: Automatic labeling of semantic roles. Computational Linguistics 28, 245–288 (2002)
10. Howard, P.N., Duffy, A., Freelon, D., Hussain, M., Mari, W., Mazaid, M.: Opening closed regimes: What was the role of social media during the arab spring? Tech. rep., Project on Information Technology and Political Islam, University of Washington, Seattle (January 2011)
11. Màrquez, L., Carreras, X., Litkowski, K.C., Stevenson, S.: Semantic role labeling: an introduction to the special issue. Comput. Linguist. 34(2), 145–159 (2008)
12. Mohler, G.O., Short, M.B., Brantingham, P.J., Schoenberg, F.P., Tita, G.E.: Self-exciting point process modeling of crime. Journal of the American Statistical Association 106(493), 100–108 (March 2011)
13. Punyakanok, V., Roth, D., Yih, W.t.: The importance of syntactic parsing and inference in semantic role labeling. Comput. Linguist. 34(2), 257–287 (2008)
14. Wang, X., Brown, D.E.: The spatio-temporal generalized additive model for criminal incidents. ISI (2011)