The relevance of Social Media as it applies in South Africa to crime prediction

Coral FEATHERSTONE

Meraka Institute, building 43, 627 Meiring Naude Road, Brummeria,

Pretoria, 0001, South Africa

Tel: +27 12 8414829 , Fax: +27 12 8414829, Email: cfeatherstone@csir.co.za

Abstract: Being able to identify and predict crime trends or track criminal movement would help anyone interested in preventing criminal activity or being able to assess where crime enforcement is needed, particularly in crimes where constant policing is impossible, such as cable theft. Many neighbourhoods in South Africa have formed voluntary community policing groups, who keep in touch using SMS and two way radios. Some have adopted websites and even Twitter as a means of being more easily in touch quickly and transparently. The influential groups recognising the value and using Twitter include, Crime Line (@CrimeLineZA) and the South African Police Service (@SAPoliceService). This paper argues that existing technologies can make communication more useful in terms of data gathering, prediction and spotting broader patterns. An assessment is done to determine if South African people are already using Twitter to report crime and to find out what information they are sharing, with the goal of estabishing whether it could be useful as a source of information for the prevention of crime.

Keywords: Crime prediction, Data mining, Social Networking, Twitter, Social Media

1Introduction

Although Crime Stop, a crime reporting call centre, is very successful it operates on the level where crime is already occurring [1]. Subtler information, such as noticing a vehicle that does not belong in an area, or recognising that certain pedestrians are not local and are loitering are unlikely to be reported to anyone other than armed response companies, where the information remains local.

Social media has shown great promise in disseminating information very rapidly. Newspapers have adopted the platform and even stock brokers have stopped looking for the latest movements in the market in printed formats, preferring the speed at which the online world operates [2]. Several groups of individuals have had great success using these platforms to solve group problems. There are examples of hijacks and other crimes being foiled by Twitter [3]. There are a number of websites, forums and mobile applications, both locally, such as Mobilitate [4] and the Crime and Justice Hub [5], and internationally [6], that attempt to use social media to prevent crime.

There are Twitter accounts specifically acknowledging the need to collate crime information, such as “@crimeshouter” and “@turnitaroundsa”. The latter also SMSes crime activities as

reported on it's website [7] and claims to have a community of over 1600 users. This paper will show that the data available on social media is sufficient to be used as a tool for crime information gathering and furthermore that the information has a predictive aspect that does not appear in the existing, community based crime fighting tools.

2Methodology

The aim of the paper was to explore whether Social Media, and in particular Twitter can provide data that could be useful in the fight against crime.

The literature study in section 3 encompasses research showing that multiple topics mined from social media have a predicitive aspect to them. Being able to predict the occurance of criminal activity could be very useful as a crime prevention tool.

A brief description of the literature is provided in table 1 describing what was being predicted and which techniques were used to extract that data. Other research topics relevant to why twitter is a useful source of crime information are also suggested in table 2 as they would be relevant to further research on the topic.

In section 4 a comparison of the last year of Twitter data against the most recently released South African Police Service (SAPS) crime statistics from April 2011 - March 2012 [8] for a subset of Gauteng Suburbs is collected in order to establish whether there is any correlation between reported crimes and the crimes people discuss on Twitter. Section 4.3 does a similar Twitter search, using the same keywords, but on specific South African user accounts which are known to Tweet about crime in order to find out what crime topics people are discussing.

3Literature Review

Table 1: Studies specifically addressing the predictive nature of Social Media

Paper Title	**Description**
Predicting the future with social media [9]	Asur and Huberman point out that the content sharing and available data in social media environments is largely untapped. The paper shows how one can make fairly accurate real-world predictions using Twitter data. In this case forecasting the revenue of movies. They show that the rate of similar Tweets shows high accuracy in determining predictions. They also discuss the role of sentiment analysis, also called opinion mining, in prediction techniques.
Will Twitter Make You a Better Investor? A Look at Sentiment, User Reputation and Their Effect on the Stock Market [10]	An investigation of whether Stock market data could be predicted with social media, uses user sentiment and user reputation as part of it's Twitter data mining techniques. It also investigates whether the time of day influences the data. The preliminary research is promising.
Predicting elections with Twitter: What 140 characters reveal about political sentiment [11]	Even election results show a strong predictive trend. Tumasjan, Sprenger et al. discuss how a few heavy users can influence the data set and point out that conversations can be identified. Their results came close to traditional election polls.
Using natural-language processing to produce weather forecasts [12]	Natural language processing, used in the prediction of weather patterns both from sensor and human inputs and Multiple languages (English and French), was collated and analysed producing the same predictive results. This shows that it is conceivable to have a system combine multiple sources and languages into the result set.
Automatic crime prediction using	Wang, Gerber et al, while considering crime prediction techniques,

Paper Title	Description
events extracted from Twitter posts [13]	focused specifically on predicting hit and run incidents. They points out that prior work on criminal incident prediction has relied primarily on the historical crime and demographic information. They use statistical methods, specifically, latent Dirichlet allocation, also known as semantic analysis to determine Tweet semantics. It does this via recognition of nearby words, which build up topic sets that can then be used to extract data from new text. The result showed that forecasts were accurate enough to warrant further research.
Social Media Analysis and Geospatial Crime Report Clustering for Crime Prediction & Prevention [14]	Saraf, P.; Milo, M.W.;et al show how using crime rate localization, the Geospatial aspects in social media as well as temporal data can be used to demonstrate and predict trends. Actual crime report data is mixed with social media data in order to match crime trends. The combination of structured and unstructured data is interesting

Table 2: Other studies that have relevance to techniques and adoption of such a tool

Paper Title	Description
Communities of practice [15]	Studies of communities of practise have been around for a long time and pre date the internet. These are also known as learning networks. Recognising that community of practise, have already formed around crime prevention, means that all the applicable research also applies here. The Internet has made it easier for these communities to get together and share information
Sentiment Analysis and Opinion Mining: A Survey [16]	In comparing sentiment classifiers, this study points out that the performance of the chosen tool is dependent on the domains or topic. Some investigation to determine the best way to mine for topics specific to the domain of crime and criminal activity would be relevant.
Locating the Source of Diffusion in Large-Scale Networks [17]	It is shown that even though we cannot measure every node in a complex network, it is still possible to estimate the original source of information by studying only a random subset of the nodes. They tested their algorithm by calculating the source of a cholera outbreak in the KwaZulu-Natal, but point out that the technique would extend to any complex network, including social media, meaning that you could locate the original source of information on the topic of a Tweet

4Data Analysis

4.1 Twitter data versus crime statistics comparison

In order to compare the last year of Twitter data against the most recently released South African Police crime statistics access to old Twitter data was required. Unfortunately Twitter does not allow realtime access to more than 7 days of data for any given search term when searching multiple user timelines. The Google Custom Search API, which allows searches against specific websites gave programmatic access to the required data.

Six suburbs, namely Hillbrow, Diepsloot, Douglasdale, Midrand, Olievenhoutbosch, Sandton and Wierdabrug were chosen from the SAPS list. They were chosen randomly, but include both low costs housing and middle class suburbs. Five topics were chosen, from the crime categories, namely hijacking, shoplifting, burglary, vehicle theft and murder. The crime statistics distinguish between car-jacking and truck-jacking. For the purposes of this study close categories where combined under a common category, in this case, hijacking. There was no geospatial or place information available searching Twitter from Google. To work

around this the search was made with a combination of the suburbs name and a few synonyms of the keyword chosen.

Table 3 shows a condensed section of the SAPS crime statistics for April 2011 to March 2012 for Diepsloot showing some of the keywords that were considered.

Table 3: a condensed section of the SAPS crime statistics for April 2011 to March 2012 for Diepsloot

Crime in Diepsloot (GP) for April 2011 to March 2012			
Murder	52	Theft of motor vehicle and motorcycle	69
Burglary at non-residential premises	80	Theft out of or from motor vehicle	146
Burglary at residential premises	436	Shoplifting	60

4.2 Twitter data versus crime statistics results

Figure 1 shows the SAPS percentages graphed against the twitter percentages after the Google search was performed.

Figure 1: Comparison of twitter counts against SAPS counts

The percentages were calculated as crime category count divided by total count. For example, referring again to Figure 1, for Diepsloot there were (52 + 80 + 436 + 60 + 69 + 7 + 1= 705) crimes reported for the chosen period so the percentage for burglary is calculated as (residential + non-residential burglary)/(total over all keywords). So the burglary percentage is 73% of the reported crime. The Twitter search count percentages were obtained from the search results in the same way.

Table 4 shows a summary of the results matching the search terms count for the various searches while tables 5 and 6 show some of the results typical of each search. Table 7 shows results that matched the search but not the semantics.

Table 4: Tweet count summary from twitter via google search

Tweet count summary By Neighbourhood (over all crimes)			Summary By Crime Topic (over all neighbourhoods)	
Diepsloot	73		Hijacking	7
Hillbrow	47		Shoplifting	66
Midrand	5		Burglary	209
Olievenhoubosch	0		Theft	7
Sandton	898		Murder	754

Table 5: Searching for: sandton OR Buccleuch OR Wendywood hijack, OR hijacked

Title:Twitter / sandtontimes: Anyone with information to suspects in the murder of a couple in Sandton this weekend can message a tip off to Crimeline on 32111.
Title:Twitter / Schickrose10: RT @AlertZaAfrica: sapsJHB 5 suspects arrested in Meadowlands attempted robbery. 4 unlicensed firearms recovered. Yaris is believed to be stolen in Sandton.
Title:Twitter / TuksFMNews: One man has been injured after an armed business robbery in Sandown Estate

in Sandton Today. Reply; Retweeted …
Title:Afritrack SA™ (Afritrack_SA) on Twitter:#ALERT: ND624080 Black Opel Corsa LDV Hijacking 6pm Impala place Malvern . ... #ALERT: BMW 335i White (HIJACKED) South Coast Road by 6 Armed Males Driving Red ... silver Subaru station wagon SBL007GP Hijacked in midrand ...
Title:Graeme McCormack (drivability) on Twitter:Silver Ford Fiesta reg WYN057GP , just been hijacked in the Wendywood area. Pls inform police If ...
Title:Lauren Dada (Lau009) on Twitter:#ATT Just hijacked by gun point, BMW, charcoal colour reg:VRL 853GP Tracking unknown. Hijacked in Griffen ... #ATT: Missing person + vehicle.RZS817GP …
Title:Liezl E Oosthuizen (LiezlEOosthuize) on Twitter Vehicle hijacked at Bromley spar. Brand new Fortuna with .... #ALERT: SAVO GP VW Polo TDI Black (STOLEN) Parkmore Sandton. Suspects and Direction .

Table 6: Searching for: wierdabrug OR Rooihuiskraal burglary OR robbery

Title:Twitter / insanemedic: @PigSpotter: #ATT: RT @SAR_K9_Unit: WHITE CITY GOLF JWT 296 GP Hi Jackers Rooihuiskraal The Reeds just tried robbery. 10111 if ...
Title:Bedfordview (BedfordviewArea) on Twitter A woman (70 or older) died following a house robbery in the Bedfordview SAPS .... Stolen Rooihuiskraal Centurion mazda 323 sting reg; LTK 766 GP green r/r ..

Table 7: The type of Tweets that matched the search but did not match the semantics

Title:fahd shanawaz (fahdshanawaz) on Twitter Rheecha Ratnam @rheecharatnam1. @ ShashiTharoor 23 Indians on ship hijacked in Nigeria by pirates ...
Title:Twitter / MixwellDigital: Daylight Robbery: iPad 2 costs R13 000 at the Phone Shop in Sandton!

4.3 Individual User account search

The same keywords were searched on Twitter, using Twitter4J, on accounts known to be South African in order to get a feel for the results not restricted to those specifically mentioning the suburb. The user accounts chosen were those that specifically encourage the reporting of crime and suspicious activity. The timespan of the individual user search is shorter than the previous searches and spans just over two months, from 28 August – 7 November 2012.

4.4 Individual User account results

A summary of keyword matches for the 6 accounts searched is shown in table 8. The following two tables, 9 and 10, give an indication of the data.

Table 8: Twitter4J result counts for individual user's timelines

	SAPoliceService	CrimeLineZA	PigSpotter	CrimeStop_RSA	crimeshouter	turnitaroundsa
burglary	1	0	0	0	0	48
hijack	5	6	3	17	6	5
murder	35	18	1	12	6	14
shop-lifting	0	0	0	0	0	0
theft	16	12	0	2	5	25

Table 9: Results from Turn It Around @turnitaroundsaCrime , location='South Africa'

Motor vehicle theft - Corner Of Blake Road And Collingwood Road, Observatory - 2012/08/22 at 20:26 http://t.co/AMVMjJs8

Attempted burglary - Chemnen Avenue, Weltevreden Park - 2012/08/23 at 10:00 http://t.co/OSOwN14a
Attempted burglary - Mahogany Street, Noordwyk - 2012/08/27 at 03:27 http://t.co/0ZtNqmR9
Motor vehicle theft - Wessels Road, Rivonia - 2012/09/02 at 08:30 http://t.co/m4OPQWkC

Table 10: Results from Crime Stop @CrimeStop_RSA, location='South Africa, Gauteng, Benoni',

RT @K9_UNIT_GP: RT @llrottie: look out: white bantum Bakkie possible black bonnet or tailgate - just hijacked in ... http://t.co/hzqNScfg
RT @PigSpotter: #ATT 3x Silver VW Polo Vivo cars, that I know of, have been hijacked/stolen in the last 2 days. Please be cautious.
RT @SouthAfrica_SOS: A woman who was hijacked last month by 3 men, spotted one of the men in a bar having a drink ... http://t.co/IhMShHL5

5Discussion

An initial look at the chart in Figure 1 looked promising, however a count of the matched data proved too low for any accurate conclusions to be drawn. It appears to show higher Tweet rate for emotive categories such as hijacking and burglary. Olievenhoutbosch was removed from the graphed results because there was no data.

It is quite common for people to Tweet the license plates and description of stolen vehicles. The appearance of Tweets from news sources is quite high, however there are also many Tweets from individuals. The occurrences that matched the keywords but didn't match the intended semantic content was not very high. Although the Tweets themselves often do not contain geospatial information, the user account itself contains an accurate location. Expanding the keywords used for searching as well as the suburbs searched and making use of the conversational aspect of retweets, would all increase the results. The counts are skewed because word stems were not accounted for so while hijacking had a count of 7, hijacked, and hijack are common in the results.

Because tracking of hijacked vehicles seems to be such a common occurrence in the data, the tracking of criminal movement is probably a good target for further investigation. It is interesting to note that the user timelines dedicated to reporting suspicious activity and encouraging people to report what they see have much higher keyword matches than the others. In particular the kind of Tweets on (@turnitaroundsa) would be very useful to analyse because most of them are from people reporting specifics of criminal activity.

Other Advantages of this approach are:-

- No specific registration is needed since users are already using these social mechanisms
- The system is more real time than existing systems
- Emotion, timeline, frequency and other partially hidden mechanisms are available for increasing the accuracy and urgency of the data

6Conclusions

This paper aimed to show that people are already using social media to discuss criminal activity and this can be seen in the data. The predictiveness of Twitter has already been established, as indicated in the case studies. It is clear that the quantity of the data could be vastly improved with ongoing data collection from the intended source and more specific use of user location. Including the suburb in the search obviously narrows the collected data.

The comparison presented has shown that there is sufficient usage in the Social Media space to warrant further investigation. Ongoing collection of data into a database would more accurately mirror how such a tool would function and would be key to getting a deeper understanding of the usefulness and future directions.

Future research could focus on how to extract the most relevant data. As seen in the existing research, retweets and the rate of Tweets on similar subject matter are good predictors of accurate information. Research on how to most affectively group users in close approximation would be useful. Firstly since location is so important in the type of information being investigated, but also because of the feedback mechanisms. A person in Pretoria is not going to be interested in following the Tweets of someone in Cape Town.

A combination of statistical searching, Crowdsourcing, modelling, natural language processing and other methods could be combined to build up and report on criminal activity in a vastly superior and predictive way.

The data gathering in this paper fell short of what is needed for accurate conclusions to be drawn, however it did show that South Africans are discussing crime on Twitter.

7References

[1] J. F. De Beer. Crime stop: Message from the divisional commissioner. *2012(6),* 2010.

[2] G. Nevill, "JSE killing off advertising cash cow for business media," vol. 2012, .

[3] K. van Schie, "Twitter alert foils hijack," vol. 2012, April 9 2012 at 08:09am, 2012.

[4] Anonymous "Mobilitate: Community Policing Forums," vol. 2012, 2012.

[5] Anonymous "Crime and Justice Hub: Information and analysis sharing for a safer and just society. A civil society initiative by the Crime and Justice Programme (CJP) of the Institute for Security Studies (ISS)," vol. 2012, 2012.

[6] Anonymous "Community Policing & Neighborhood Crime Statistics," vol. 2012, 2012.

[7] Anonymous "Turn It Around - Fight crime in South Africa " vol. 2012, 2012.

[8] Anonymous South african police service, crime statistics for gauteng for april 2011 - march 2012. *2012(11/19/2012),* 2012.

[9] S. Asur and B. A. Huberman. Predicting the future with social media. *Arxiv Preprint arXiv:1003.5699* 2010.

[10] E. D. Brown. Will twitter make you a better investor? A look at sentiment, user reputation and their effect on the stock market. 2012.

[11] A. Tumasjan, T. O. Sprenger, P. G. Sandner and I. M. Welpe. Predicting elections with twitter: What 140 characters reveal about political sentiment. Presented at Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media. 2010, .

[12] E. Goldberg, N. Driedger and R. I. Kittredge. Using natural-language processing to produce weather forecasts. *IEEE Expert 9(2)*, pp. 45-53. 1994.

[13] X. Wang, M. Gerber and D. Brown. Automatic crime prediction using events extracted from twitter posts. *Social Computing, Behavioral-Cultural Modeling and Prediction* pp. 231-238. 2012.

[14] P. Saraf, M. W. Milo, S. C. Richards, T. Bhattacharjee and L. Malambo. Social media analysis and geospatial crime report clustering for crime prediction & prevention. pp. http://filebox.vt.edu/users/mmilo/CS5525/Proposal_milo_saraf_richards_moonga_bhattacharjee.pdf.

[15] E. Wenger. Communities of practice. *Learning Meaning and Identity* 1998.

[16] G. Vinodhini and R. Chandrasekaran. Sentiment analysis and opinion mining: A survey. *International Journal 2(6)*, 2012.

[17] P. C. Pinto, P. Thiran and M. Vetterli. Locating the source of diffusion in large-scale networks *Phys. Rev. Lett. 109(6)*, pp. 68702. 2012. Available: http://prl.aps.org/.