Identifying vehicle descriptions in microblogging text with the aim of reducing or predicting crime

Coral Featherstone
Meraka Institute,
building 43,
627 Meiring Naude Road,
Brummeria,
Pretoria,
0001,
South Africa
Tel: +27 12 841 4925, Fax: +27 12 841 4829,
Email: cfeatherstone@csir.co.za

***Abstract*—Could Social Media, and in particular, microblogs such as Twitter, play a part in helping to track criminal movement?**

The aim of this paper is to narrow the focus of this broader problem of using social media to crowdsource information to assist in the fight against crime, to the specific problem of identifying the description of vehicles in microblog text.

As this problem has many aspects, especially in terms of data gathering and identification, an initial search is performed on preset keywords and the resulting database is tagged.

The tags are then analysed to determine which features are the most common. Topic models are then run on the data to determine if any useful keyword can be found for further searches and initial statistics are recorded as a baseline for further processing.

Our primary concern is establishing the common content of the relevant Tweets. The result could be used both for help with data collection as well as with feature selection when learning classification algorithms for data mining.

***Index Terms*—Data mining, crime prevention, social media, topic models**

A. Introduction

Microblogs are a type of blog that lets users publish very short updates. Twitter is an online microblogging service that allows people, known as Tweeters, to send and receive text messages that are limited to 140 characters called Tweets.

There are some crimes, for example cable theft, that are difficult if not impossible to police, that could benefit from independent tipoffs. Other crime prevention strategies such as reporting suspicious pedestrians or suspicious vehicles that are loitering in a neighbourhood, could be useful as a crime prevention aid, but are too trivial and numerous to be reported. While the former problem is a function of the South African crime reporting organisation “Crimeline”, the real time nature of Twitter could contribute to faster and better tracking of the reports provided the data could be sufficiently analysed for useful patterns and provided noisy, irrelevant data could be removed.

The aim of this project is to narrow the focus of this broader problem of using social media to gather information from the general public to assist in the fight against crime, to the specific problem of identifying the description of vehicles in text. One example application would be to use it to track the movements of suspicious vehicles between neighbourhoods by encouraging clients of private security companies to report suspicious vehicles to a particular user account on Twitter.

A search on Twitter, for the keyword “hijack” or ”hijacked“ or “stolen”, followed by a suburb name, indicates that people are already describing vehicles on Twitter. Vehicles are described by license plate number, make, colour, last seen location and sometimes direction of travel. An example of direction of travel is “heading south”. Examples of Tweets matching this format are shown in Table I on page 3.

There are a wider set of problems that will need to be addressed on this topic such as which techniques can be used to pick up spammers who are exploiting trending topics for the purpose of advertising. This known as Trend Stuffing[1], detecting lies in text, known as Computational Stylometry and user sentiment and location identification, but these are not the focus here.

B. Overview

As Twitter is a publicly accessible service, Tweets are considered public data provided the name of the user is not altered or obscured and the text is not altered. It is therefore of interest to researchers mining the internet for information. Twitter publishes their usage guidelines [1] online.

A number of problems need to be solved in finding and extracting the description of vehicles in text. Not only do we need to identify Tweets that contain the description we need, but we also need to further extract the features of the vehicle from the text we have identified. We could just try to use keywords to find the data but there are problems with this approach. One problem is that the keywords may not remain fixed, for example if we used a keyword list of vehicle makes, such as BMW, Mercedes or Honda we are excluding any new

[1] https://twitter.com/logo

or unknown manufacturers from our search. Another problem occurs when people mistype the text. This occurs often in microblog text and is often intentionally done when characters are dropped from a word to facilitate adherence to the 140 character limit of the service.

This problem is compounded when we consider vehicle makes which can change annually. In addition words and word stems would need to be stripped when using keywords. The keyword ”blue” is not going to isolate the description ”blueish” from the text.

Vehicle license plate identification may require its own processing to extract. A rule based system using the known structure of license plates could be used as a starting point but could be very restrictive if new or personalised number plates or international vehicles are to be spotted.

The particular nature of Twitter with its very short text messages presents its own problems. Natural language processing tools need to be able to work with very short documents where keywords may occur sparsely.

Due to the scale of the information on Twitter, this paper starts by focussing on a few the results of searching for the relevant Tweets by using a few specific keywords.

There are specific accounts known to have large quantities of Tweets containing vehicle descriptions. Three such accounts are @PigSpotter, @Afritrack_SA and @SAPoliceService although it must be noted that private Tweets appear to have a slightly different format, in that they are less consistent with the use of hashtags. Hashtags are keywords that the author of a Tweet uses to indicate the semantic content of what they are discussing. This subtle difference between hashtags can be seen from the last few Tweets in Table I .

C. Literature Review

The existing research on vehicle identification, such as the paper by Wang, Cui, Liu, Huck, Verma, Sluss & Cheng[2], appears to be mostly restricted to the task of identifying license plate numbers from photographs taken of the vehicles. Most of the discussion is centered around the optical character recognition task of extracting the letters from the images.

However the task of extracting vehicle descriptions from text can be categorised under the broader topic of natural language processing or text data mining and in particular automatic document classification. The task of document or text classification is the task of deciding whether a piece of text belongs in a certain set.

As pointed out by Witten & Frank[3] it is known that there is no machine learning task or algorithm that is applicable to all problems and the task tends to be domain specific. For this reason each chosen topic requires investigation to ascertain the learning tool most appropriate for the topic at hand.

In this case we are interested in the machine learning task of subject or topic classification where the subject is “vehicle” and the first task is to decide whether the text is in the topic or not. This is known as binary or binomial classification. Classification in machine learning is considered a supervised learning task in that the data that the algorithm learns from, also contains an indication of which items in the training data are correct. In addition to deciding whether the text describes a vehicle, we need to then be able find an classifier, also known as a classification model, that can suitably identify the vehicle features once it has been established as belonging to the topic. A classification model, or more generally a data mining model, is a condensed representation of the information in the data, found by the classification algorithm that can be used for prediction. The model is learned from representative training documents. This concept is fairly abstract because the actual make-up of the model differs from classifier to classifier.

While there are many classification algorithms, the restriction to 140 character texts in microblogging environments, such as Twitter, presents its own problems. As pointed out by Chen, Li, Nie & Hu[4] words are often abbreviated and misspelt and slang is frequently used. Probabilistic classifiers that rely on word counts are not that useful because of the sparseness of the data. Most Tweets contain only about 12 words. While the standard smoothing techniques become essential under these circumstances[3], data mining techniques in the microblogging realm use additional techniques that are specific to this genre in order to supplement the provided information.

In traditional data mining the task of identifying and classifying data in unstructured documents, to identify entities such as location, person or vehicle, is called Named Entity Recognition, however the traditional methods used for this task, such as parts of speech tagging, perform very badly in the microblogging environment [2]. There are a number of reasons for this. Sparse data, and misspelt words are once again a major problem but the other problem is that users often use ungrammatical and half finished sentences. There is an overlap between text classification and named entity recognition and classifiers can be used to find named entities. While traditional named entity tasks output a single label it is not unheard of to have the named entity as one of the features in a larger set.

Additional detail on the practices that are used to improve data mining in the microblogging environment are presented below.

Examples of the types of topics being found by machine learning in the realm of Twitter include:- determining the gender of the author of the Tweet[5], finding Tweets about the Hiati Earthquake[6] and detecting Tweets that were typed while drunk[7].

1) Topic Models: Before discussing classifiers, I am going to briefly discuss some of the terminology.

A topic is a distribution over words in a vocabulary where each topic has all the words, however words that have been assigned a the high probability define the semantics. A topic model is a type of statistical model for discovering the abstract topics that occur in a collection of unlabelled documents. If a document is about a particular topic, then certain words can be expected to be found together. In the context of vehicles this would be the attributes discussed above such as colour and

TABLE I
EXAMPLES OF THE KIND OF TWEETS WE ARE INTERESTED IN.

Twitter account name	Unaltered Tweet text
@Afritrack_SA	ND624080 Black Opel Corsa LDV Hijacking 6pm Impala place Malvern
@Afritrack_SA	#ALERT: Gold fiat Linear 2012 Model BV81MVGP and a White BMW 2010 ZMH849GP, Both Vehicles were Taken in a House Robbery: Strubensvalley.
@Afritrack_SA	@pigspotkzn ALERT- Armed Robbery ISIPINGO, ND30710 Red TOYOTA COROLLA INVOLVED 3 Armed Males & HIJACKED NT44959 Upon Get Away. BE ON LOOKOUT
@Afritrack_SA	#ALERT: ND624080 Black Opel Corsa LDV Hijacking 6pm Impala place Malvern. Complainant was Shot in Leg. Unknown Direction.
@PigSpotter	#ATT: RT @SAR_K9_Unit: WHITE CITY GOLF JWT 296 GP Hi Jackers Rooihuiskraal The Reeds just tried robbery
@PigSpotter	#ATT 08:33:48 Black Honda Accord (2009) reg YBN778GP taken in house robbery from Newtown. 10111 if spotted
@PigSpotter	#ATT White Vw Polo TDI, reg: TRR753GP has been stolen outside a house in Melville. 10111 if spotted
@CrimeStop_RSA	RT @JPSAorg: #ATT Armed Robbery 12:00 at Inspectacar Kempton Park. BMW 325 station wagon - no plates stolen.
@jolene_deBruyn	12:00:04 Mon, May 20, 2013 - A colleague's Silver Toyota Hilux D/C - Reg: XFM938GP was Stolen (at Michael Angelo Hotel) in Sandton..
@Mlindi1	22 Oct Cc @moflavadj RT"@BraObza_Sibasa: ALEX: A charcoal black Golf 6 BJ 22 JY GP was stolen in Sandton. It was seen early this morning in Alex.
@YaseenGaffar 7 May	ALERT: Dozens of cars stolen across #Middelburg in two weeks. Mazda 323ś at risk: More here - http://tinyurl.com/cqv2uqw @ALZU_N4 @TRACN4route

make. It may be the case that there are other topics related to Twitter posts about vehicles that are not immediately obvious from the hand collected Tweets that we may wish to discover and it is these hidden words, known as latent variables, that we wish to find. Topic models can also associate words that have similar meanings. In data mining the term document is also a fairly abstract term. In this document we are referring to text however document can also refer to images or music or any other focus of the machine learning process. What is useful here is recognising that the text of one Tweet may not necessarily be the only document that we need to focus on. We can combine Tweets with other Tweets containing the same word or with other Tweets in the same conversation or added information in Retweets of the original Tweet. A completely different data source can also expand the content that the classifiers have to work with.

2) Classifiers: There are many existing binary classification algorithms, but in the small text environment Support Vector Machines, Naive Bayes and Latent Dirichlet Allocation (LDA) and their variants appear frequently.

Naive Bayes and Latent Dirichlet Allocation are both generative models, that is the probability of all words, including those that are latent, are taken into account (as opposed to what is known as the discriminative model which only conditions on observed data). The term latent refers to hidden data that can be calculated but is never observed. Latent Dirichlet Allocation is one of simplest topic models. It computes hidden models given observations of words by assuming independence and then using the dirichlet distribution, which is the multivariate version of the beta distribution, to infer both the distributions of words in each topic as well as the distribution of topics in each document.

Naive Bayes is a probabilistic method that uses Bayes' theorem and the assumption of independence between the attributes to assign topic membership. Despite this simplistic and unrealistic assumption, Naive Bayes continues to perform remarkably well against competing methods in all aspects including speed, accuracy and simplicity of implementation. It manages to score quite well when compared against much more sophisticated methods[5] [7] [6]. Witten & Frank [3] mention that because simple methods often work very well that it is well worth adopting their usage first before trying more advanced techniques but once again points out that different datasets may lend themselves to different approaches. He also mentions that Naive Bayes as a classifier is good at handling documents with sub attributes. He uses, as an example, documents that are news items that have the subclasses of overseas news, financial news or sport.

Support vector machines were originally designed for binary classification of linearly separable attributes and were extended to support non separable data. [5] claim that Support Vector Machines are quite slow and [6] and [7] point out that the results from Support Vector Machines don't offer much benefit over classification by keywords as done by the Naive Bayes method, which is a much simpler technique. [6] points out that binary Support Vector Machine classifiers vary significantly from one category to another, with some categories being much harder to classify than others

3) Preprocessing: Text is commonly cleaned up before processing, by eliminating stop words, such as "a" and "the", or using feature selection, which are methods to remove extraneous information considered to be redundant. The aim is to increase the semantic content of the text by eliminating anything that doesn't contribute towards the the intended. Clearly, while this improves accuracy, it once again increases the sparseness of already sparse data. One needs to differentiate between "feature selection" which reduces the words in the text and another text data mining preprocessing method called "feature extraction which is also used to reduce the dimensionality of the data, but does so by creating a new but smaller data set, thereby increasing the semantics. Most of the time, choosing the representative words instead of all words as features can improve the classification[3] [7], with some attributes or keywords doing most of the work.

Some preprocessors for microblog datasets also try to correct the abbreviated and informal words that occur so frequently because of the 140 character restriction. This can be done with a smaller dictionary of commonly occurring words or by cross referencing against other data sources such as wikipedia[2] or urban dictionary[3]. In data mining n-grams are also frequently used to fix spelling.

Designing classifiers over different topic sets is very much the same, regardless of topic, however how feature selection is done can alter the results[7].The importance of data preprocessing, and in particular attribute selection is not only critical to performances but also at a later stage to data visualisation[8].

This suggests that identifying which features most accurately pinpoint Tweets describing vehicles should be a starting point and indeed feature selection is regarded to be just as essential as the choice of algorithm.

4) Terminology Specific to Twitter: Before explaining some of the techniques used to supplement data mining in the microblogging realm, I will briefly explain some of the terminology specific to these environments. A Twitter account (also know as a timeline) is the collection of Tweets from a particular user listed from the most recent message. While many Twitter accounts are public, and can therefore be seen and read by anyone, a Twitter user can indicate their specific interest in a particular user's Twitter messages by subscribing to their account. This allows Twitter client applications to present to this user just the content that they are interested in. This subscription is referred to as "Following" another user. The term "Retweet" is use to to describe the process of an author of a Twitter account repeating another authors contribution to the users who "Follow" them. Hastags are not restricted to real words and are often concatenated phrases. Hashtags are indicated by prepending the '#' character to the keyword. Tweeters can reply to other users by including that users account name in their Tweet prepended by the "@" character.

5) Techniques Specific to microblogging: The techniques used to supplement microblogging data mining include some of the following:-

When urls are present in the text, the classifier loads the page that is referred to by that url and uses its content to add features to the model. Often the importance of the content of the Tweet can be determined and increased by counting how many times it has been Retweeted and also by noting the frequency produced by the time between Retweets. The honesty, validity and intent of the Tweeter can be established by acknowledging the age since they started Tweeting using a particular user account. This is valid for two reasons. Firstly, a person who is using a Twitter account maliciously usually closes it themselves shortly after starting to Tweet, so that they can avoid being traced. Secondly Twitter has the facility to allow users to report accounts that they think are being abused and those accounts get rapidly shut down.

Noting that the hashtags used indicate metadata or semantics are normally strongly related when they co-occur in a single Tweet also helps to increase the keywords in topic sets by adding the word or phrase from the second hashtag to any feature sets containing the keyword from the first. There are implied conversations when users reply to each others content. These replies can go on for several Tweets and since the content of these chained Tweets is generally about the same topic the topics of these conversations can be combined so that they all contribute to the same semantic. Additional techniques used to increase the semantics include comparing the words in the Tweet to the content on other sites such as Wikipedia and even Google Books. Wikipedia's disambiguation pages are particularly useful. Google Sets was an online tool that generated lists of similar items when provided with a few examples. Keywords could be run through Google Sets to find additional related keywords for the topics the search is trying to classify. Unfortunately Google sets is no longer available as an online tool, however the functionality is available as a feature on the spreadsheets on the Google Drive [4] website. The lexical database Wordnet[5] is often used. Theoretically of course, any thesaurus would suffice. In addition to this unusual hidden data, the follower, followee and total Tweet count are also very relevant to classification[9] since these indicate different types of users as well as whether their Tweets are likely to have a high Retweet rate. A user account that has many followers and Tweets many times a day is likely to be a high profile site. News accounts would be an example in this category. @PigSpotter is an example of this kind of user with over 52 000 Tweets and over 181 500 followers, but the account is only following 16 users. The user, @YaseenGaffar, also show in Table I is an example of a more low profile user with only 384 followers. It is also worth noting that older accounts will have accumulated more followers. Followers between users form a directed relation. It is interesting to note that this kind of hidden information can enable the classification of spammers.

6) Data collection and labelling: A very important consideration in machine learning is a concept known as overfitting or overtraining. If a classifier is trained on a set of data that is too similar to the data it needs to discover it may happen that it builds up a model that gets very good at correctly classifying the training set, while getting less and less accurate at recognising or predicting new data that it has to classify. This also happens where the training was performed for too long and often occurs in the case of sparse data. Latent Dirichlet Allocation became popular because it does well at avoiding this problem. One solution to overfitting is to include

[2] en.wikipedia.org

[3] http://www.urbandictionary.com

[4] https://drive.google.com

[5] http://wordnet.princeton.edu/

in the training data, documents from another source. For vehicle description in text, sites like Junkmail[6] or AutoTrader[7] could be considered for the vehicle descriptions, although the latter is not likely to have any negative classifications.

Since Twitter has a limit on the amount of Tweets that can be downloaded in an hour, most researchers are collecting the Tweets over a period of time into a database. A larger continuous stream of data, called a firehose can be purchased. Either way, the resulting data can be very large.

Most classifiers tend to be supervised learning processes the training data needs to be labelled after collection. One of the problems inherent in a site like Twitter, is that there are vast quantities of unlabelled data available for machine learning, but virtually none of it is labelled. This task, which is manually done by a human, can be extremely time consuming and too few labels produce inaccurate results from the classifiers[8]. The problem of hand labelling training data on such large collections, where there is an abundance of unlabelled data and a shortage of labelled data, can be reduced by combining the learning with unlabelled data[10] [10] [8]. This process is called semi-supervised learning or sometimes bootstrapping. An unsupervised process can also start with a title word or seed word, such as ”automobile”, for each category it is looking for[8]. The most confident matches in the first run over the seed data is then added to the training set and the training is rerun until a low error rate is achieved.

7) Software: WEKA (Waikato Environment for Knowledge Analysis)[8] [6] and MALLET (MAchine Learning for Language Toolkit)[9][5] [6] are existing Java based software packages that appear frequently in the data mining research. WEKA does data preprocessing, classification and helps with attribute selection. It also analyses the resulting classifier, measures its performance and has data visualisation facilities. It has Naive Bayes support and the facility to implement additional algorithms.

MALLET provides document classification, classifier performance evaluation, topic modelling, information extraction and feature selection as well as Nave Bayes and many other algorithms for natural language processing.

Both tools provide Java Programming Language Application Programming Interfaces (APIs) so that they can be programatically accessed from other software.

Stanford University has a number of natural language programs[10] written in Java including the Stanford Topic Modelling Toolbox (TMT)[11] as well as the Stanford Named Entity Recognizer (NER)[12].

While this is not an extensive software list, it provides a starting point and some of these applications may be more useful than others.

TABLE II
THE INITIAL KEYWORDS USED TO SEARCH TWITTER FOR THE BASE DATA

theft OR (stolen AND vehicle) OR (vehicle AND theft)
murder OR murdered OR killed
shoplifting OR loot OR pinch OR snatch
burglary OR breakin OR robbery OR break-in OR holdup OR housebreaking
hijack OR hijacked vehicle OR HIJACKING OR carjack OR truckjack

TABLE III
THE DATA TAGGING ATTRIBUTES FOR THE TWEETS ABOUT VEHICLES

vehicle	if the Tweet mentions a vehicle
license plate	if a license plate number is provided
location	Tweet contains a place, road name or suburb
make	Tweet contains the make of the vehicle (eg. MERCEDES-BENZ)
colour	the vehicles colour is provided
model	Tweet contains the make of the vehicle (eg. M-CLASS ML500)
direction	the direction of travel is indicated in the Tweet

D. Methodology

It is clear that being able to extract the data from Twitter is a task that needs to be done before any classifiers, topic models, or any other machine learning tasks can be performed. Having established long term goals we focus here on an initial investigation into the data collected by the use of keywords that appear to accompany the Tweets we are interested in.

Twitter is queried for results matching the search terms presented in table II and the resulting data is stored in a database. The chosen search terms came out of an earlier paper[11] that identified that crime is being discussed on Twitter and that highly emotive crimes, such as hijacking were particulary prevalent. The data is collected from February to April of 2013. The data is then hand tagged according to the attributes in table III. This is done to establish percentages of occurrence of the attributes known to exist in Tweets about vehicles. The result was a set of 10 000 Tweets from the predefined search, each with a boolean indicator of whether it was describing a vehicle and boolean indicators of whether it contained each of the other attributes.

Interesting observations from the hand labelling exercise are pointed out in section -E1. The percentages of the resulting data is discussed in -E2. MALLET was used to generate topic models for each set of attributes. Those results are presented in -E3.

E. Results

1) Interesting data and observations: As can be seen in table IV, duplicate removal is made difficult by shortened urls of same story by different users, emphasising the need to remove duplicates, as necessary, by some other mechanism such as a Retweet id. The text content is almost identical

[6] http://www.junkmail.co.za
[7] http://www.autotrader.co.za
[8] http://www.cs.waikato.ac.nz/ ml/weka/index.html
[9] http://mallet.cs.umass.edu
[10] http://nlp.stanford.edu/software/index.shtml
[11] http://nlp.stanford.edu/software/tmt/tmt-0.4
[12] http://nlp.stanford.edu/software/CRF-NER.shtml

TABLE IV
DIFFICULTIES WITH DUPLICATE REMOVAL CAUSED BY SHORTENED URLS

Twitter account name	Unaltered Tweet text
aladinzuko	Trailer Park Tragedy: Woman Popped For Brutally Stabbing 80-Year-Old Woman And Her Dog During Robbery: Woman ... http://t.co/SyidNwjEEz
abdulqoodir	Trailer Park Tragedy: Woman Popped For Brutally Stabbing 80-Year-Old Woman And Her Dog During Robbery: Woman ... http://t.co/6AnIAsIQgr
funsaah	Trailer Park Tragedy: Woman Popped For Brutally Stabbing 80-Year-Old Woman And Her Dog During Robbery: Woman ... http://t.co/znAWV99eAE

TABLE VII
STATISTICS FROM THE HAND TAGGED DATA FOR 10 000 TWEETS

label	count	Percentage of 10000	Percentage of vehicle Tweets
vehicle	203	2.03	n/a
license plate	84	0.84 %	41.37%
location	102	1.02%	50.24%
make	95	0.95%	46.8%
colour	88	0.88%	43.35%
model	62	0.62%	30.54 %
direction	8	0.08 %	3.94 %

between these Tweets but the shortened url to the full content differs.

While clearly Twitter translates the search terms and supplies relevant results, returned non English results would need special handling by any machine learning tool, should they be established as relevant.

An example of this is shown in table V which contains Tweets that demonstrate how following urls in Tweets increases semantic content since the keyword doesn't exist in the translation of the Tweet text. The article at the url contains the keyword.

Twitter also searches the account name for keywords as could be seen from the non English Tweet returned from the user with account name ”unity_breakin“.

The game, ”Grand theft auto“, resulted in many undesirable Tweets, both in English and in other languages.

The search did turn up user accounts dedicated to crime prevention.

2) Data distribution: As was to be expected the data returned from such a wide search is fairly sparse. One of our goals will be to increase that match rate by identifying the relevant keywords and features.

A summary of these statistics is shown in table VII.

Out of the 10 000 tagged Tweets only 203 described vehicles. Of those, location, vehicle colour and license plate number are the most reported properties. Surprisingly, location is given for half of those.

As was pointed out in the overview, there are user accounts with much higher hit rates for this data, but his particular analyses did not target particular user accounts.

TABLE VIII
TOPIC MODEL SETS OVER ALL TWEETS LABELLED AS VEHICLE TWEETS

id	distribution	words in topic with counts
0	0.050	vehicle (8) recovered (6) kzn/dbn (6) umlazi (5) taken (5)
1	0.050	car (14) classic (5) prevent (3) secure (3) ways (3)
2	0.050	family (13) woman (11) car (9) members (7) along (7)
3	0.050	kingsway (5) silver (5) blk (4) toti (4) hyundai (4)
4	0.050	vehicle (11) police (10) crash (5) suspects (4) pursuit (3)
5	0.050	door (3) armed (3) sleeping (2) worth (2) mall (2)
6	0.050	leave (3) garden (2) vandalism (2) sweets (1) steal (1)
7	0.050	man (4) snatched (3) case (3) charges (3) report (3)
8	0.050	reported (8) car (6) block (6) auto (5) driving (3)
9	0.050	gas (12) city (11) car (8) station (8) police (8)
10	0.050	news (4) truck (3) surveillance (2) street (2) gunpoint (2)
11	0.050	car (5) window (2) north (2) more (2) sheldon (2)
12	0.050	vehicle (10) black (6) males (4) passat (2) durban (2)
13	0.050	armed (6) park (4) carjacking (3) during (3) saps (3)
14	0.050	vehicle (11) arrested (7) carjacking (4) armed (4) yesterday (3)
15	0.050	people (4) van (3) motor (2) two (2) over (2)
16	0.050	reg (6) view (3) because (2) spotted (2) rkt (2)
17	0.050	last (7) update (6) night (3) golf (3) silver (3)
18	0.050	raleigh (3) cars (2) crime (2) male (2) lol (2)
19	0.050	vehicle (12) nur (9) corolla (9) blue (9) driver (9)

3) Topic Models: Because of the small data set the topic model size was kept small at 20. MALLET's[13] Topic Modelling, which is an implementation of the LDA algorithm, was used to generate the 20 topics for all vehicle labelled Tweets, model matching Tweets, make matching Tweets, colour matching Tweets and license matching Tweets. There weren't enough location containing Tweets to make a coherent Topic Model.

The results are shown against the word counts in table VIII, XI, X, IX and IX. The first column is just a unique identifier for the topic, the second column is the dirichlet distribution between topics. The last column contains the keywords that make up the topic along with how frequently they occurred within the topic.

I. CONCLUSION

Clearly there is still much work to do to extract the information required to establish the content of microblog posts about vehicles.

Reducing the sparseness of the labelled data should be a starting point since there is not enough data at this point to accurately do full feature extraction or to derive useful topic models on the data.

Future research could achieve this through pulling Tweets off the accounts known to contain large quantity of this

[13] http://mallet.cs.umass.edu/

TABLE V
TRANSLATION AND URL FOLLOWING FOR INCREASED SEMANTIC CONTENT

Twitter account name	Search term	Unaltered Tweet	Tweet translation	Snippet translation of article at url
schlagzeilen1	burglary OR breakin OR robbery OR break-in OR holdup OR housebreaking	#actuell Schon ber 50 Flle LKW-Diebe schlagen auch in Deutschland zu: Sie kommen in der Dunkelheit... http://t.co/b86lZkdnYW	#actuell For over 50 cases - Truck thieves strike in Germany: they come in the dark	Truck Robbery - Police Dortmund has already set up on the orders of the state criminal office a special commission.
elliotmbyrne	burglary OR breakin OR robbery OR break-in OR holdup OR housebreaking	@leytonorientfc http://t.co/reSc1JwVMJ YAY!		Promising young professional footballer, 19, used his OWN Smart Car as armed robbery getaway vehicle

TABLE VI
THE SEARCH DID TURN UP USER ACCOUNTS DEDICATED TO CRIME PREVENTION

Twitter account name	Search term	Unaltered Tweet
CPLCSindh	theft OR (stolen AND vehicle) OR (vehicle AND theft)	Todays Crime Stats: 4 Wheelers: Snatched 1 , Theft 11 2 Wheelers: Snatched 5, Theft 21 Mobile phones: Snatched 16, Theft 24
GAngell (ReTweet of SgtSimmonds)	burglary OR breakin OR robbery OR break-in OR holdup OR housebreaking	RT @SgtSimmonds: Crime in #Tandridge 6th-7th April: House burglary = 0 Theft of motor vehicle = 0 Theft from motor vehicle = 0 Criminal damage = 1

TABLE IX
TOPIC MODEL SETS OVER ALL TWEETS LABELLED AS VEHICLE AND MODEL TWEETS

id	distribution	words in topic with counts
0	0.050	ormonde (2)almera (2)spotted (1)rkt (1)view (1)
1	0.050	color (1)pencil (1)camry (1)toyota (1)
2	0.050	station (8)city (8)buick (4)gas (2)http://t.co/aat (1)
3	0.050	red (2)contact (1)
4	0.050	used (2)http://t.co/c (1)res/hills (1)crv (1)week (1)
5	0.050	http://t.co/urmeqibkyg (1)
6	0.050	gas (6)vehicle (6)lasabre (4)man (3)another (3)
7	0.050	benonicpf (1)detail (1)unable (1)polo (1)
8	0.050	reg/no (4)polo (3)golf (3)citi (3)grey (3)
9	0.050	vehicle (4)honda (1)reorted (1)
10	0.050	passat (2)durban (2)greyville (2)males (2)black (2)
11	0.050	thought (1)yeah (1)bank (1)ago (1)few (1)
12	0.050	last (5)silver (3)middelburg (2)night (2)trailers (2)
13	0.050	reg (3)nasrec (2)jbmalebe (2)here (1)spotted (1)
14	0.050	gun (3)park (3)athlone (2)mel (1)silver (1)
15	0.050	http://t.co/kyu (1)new (1)npn (1)inanda (1)recovered (1)
16	0.050	vehicle (5)found.thanks (1)kindly (1)bdg (1)light (1)
17	0.050	months (1)laugh (1)camry (1)apvjle (1)cchjsrne (1)
18	0.050	kingsway (5)hyundai (4)toti (4)nissan (3)armed (3)
19	0.050	saps (3)tida (3)bms (3)blk (3)athone (1)

TABLE X
TOPIC MODEL SETS OVER ALL TWEETS LABELLED AS VEHICLE AND LICENSE PLATE CONTAINING TWEETS

id	distribution	words in topic with counts
0	0.050	silver (5)hyundai (2)tonight (1)afritrackinfo (1)monaseem (1)
1	0.050	toyota (3)contact (2)pencil (1)please (1)found (1)
2	0.050	reg (2)new (2)sacrimefighters (1)http://t.co/kyu (1)npn (1)
3	0.050	last (4)night (2)trailers (2)side (2)red (2)
4	0.050	toti (5)tida (3)nissan (3)armed (3)hyundai (3)
5	0.050	light (5)ago (4)contact (4)nur (4)blue (4)
6	0.050	cell (2)reg (2)fvgp (1)crescent (1)sherwood (1)
7	0.050	found.thanks (1)bdg (1)camry (1)gidi_traffic (1)
8	0.050	polo (3)reg/no (3)grey (3)vehicle (3)golf (2)
9	0.050	reg (3)vehicle (3)tweet (2)spots (2)corolla (2)
10	0.050	http://t.co/m (1)heij (1)suspected (1)bmw (1)alertza-africa (1)
11	0.050	jeep (2)black (2)eky (2)please (1)found (1)
12	0.050	taken (4)recovered (4)kzn/dbn (4)road (3)thompson (3)
13	0.050	middelburg (2)tipper (2)urgent (2)update (2)silver (2)
14	0.050	http://t.co/cesn (1)syndicate (1)shape (1)look (1)
15	0.050	driver (7)reg (7)corolla (7)vehicle (7)stonebridge (6)
16	0.050	white (3)vehicle (3)spotted (2)rkt (2)nasrec (2)
17	0.050	black (4)males (3)vehicle (3)passat (2)of_the_south (2)
18	0.050	kingsway (5)saps (3)blk (3)gun (3)bms (3)
19	0.050	benonicpf (1)detail (1)inanda (1)

content, purely to get to a point where the data can be sufficiently analysed for further manipulation. The topic model keywords with high word counts could also add to the search hit rate. As discussed in the literature review, topic models over larger data and classifiers would be a good starting point, with some of the data mining software able to help with feature selection.

The statistics have pointed out that vehicle model is reported less while location is a frequently supplied attribute, followed closely by make and colour. Clearly vehicle make and colour is an easier feature to supplement the training set with when attempting to classify the data via machine learning since it they are not an exhaustive list. Location on the other hand can

TABLE XI
TOPIC MODEL SETS OVER ALL TWEETS LABELLED AS VEHICLE THAT ALSO PROVIDES ITS COLOUR

id	distribution	words in topic with counts
0	0.050	derwent (1)drive (1)
1	0.050	nur (9)blue (9)reg (7)female (5)tonight (1)
2	0.050	males (3)passat (2)durban (2)greyville (2)brother (2)
3	0.050	grey (4)golf (3)citi (3)info (3)correct (3)
4	0.050	kingsway (5)toti (5)hyundai (4)silver (4)blk (4)
5	0.050	toyota (3)contact (3)please (2)found (2)jeep (2)
6	0.050	spotted (2)rkt (2)nasrec (2)view (2)ormonde (2)
7	0.050	last (5)silver (4)trailers (2)side (2)hwl (2)
8	0.050	reg (2)crescent (1)afritrackinfo (1)plates (1)monaseem (1)
9	0.050	armed (2)http://t.co/cesn (1)suspected (1)new (1)cchjsrne (1)
10	0.050	vehicle (4)road (3)thompson (3)taken (3)rhy (3)
11	0.050	khumalo (3)detective (3)contact (3)please (2)reg (2)
12	0.050	found.thanks (1)kindly (1)color (1)donkorleone (1)avd (1)
13	0.050	vehicle (6)polo (4)reg/no (3)bayley (1)farrarmere (1)
14	0.050	black (2)heij (1)syndicate (1)shape (1)look (1)
15	0.050	driver (7)stonebridge (6)corolla (6)hijackers (5)vehicle (5)
16	0.050	black (4)vehicle (3)of_the_south (2)
17	0.050	gun (2)pigspotkzn (1)charmskil (1)mel (1)fled (1)
18	0.050	middelburg (2)tipper (2)nissan (2)look-out (2)night (1)
19	0.050	light (6)vehicle (5)ago (4)corolla (2)benonicpf (1)

be a street name, a suburb or even a province, but would be an important attribute if the final goal of predicting the movement of criminals were to be achieved.

For this reason it may still be reasonable, when enough data has been gathered to investigate incorporating another source of data with the data used for training.

REFERENCES

[1] D. Irani, S. Webb, C. Pu, and K. Li, "Study of trend-stuffing on twitter through text classification," in *Collaboration, Electronic messaging, Anti-Abuse and Spam Conference (CEAS)*, 2010.

[2] S. Wang, L. Cui, D. Liu, R. Huck, P. Verma, J. J. Sluss, and S. Cheng, "Vehicle identification via sparse representation," *Intelligent Transportation Systems, IEEE Transactions on*, vol. 13, no. 2, pp. 955–962, 2012, iD: 1.

[3] I. H. Witten and E. Frank, *Data Mining : Practical Machine Learning Tools and Techniques*. Burlington, MA, USA: Morgan Kaufmann, 200506 2005, iD: 10127947.

[4] Y. Chen, Z. Li, L. Nie, and X. Hu, "A semi-supervised bayesian network model for microblog topic classification," in *Proceedings of COLING*, 2012.

[5] J. D. Burger, J. Henderson, G. Kim, and G. Zarrella, "Discriminating gender on twitter," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2011, pp. 1301–1309.

[6] C. Caragea, N. McNeese, A. Jaiswal, G. Traylor, H.-W. Kim, P. Mitra, D. Wu, A. H. Tapia, L. Giles, and B. J. Jansen, "Classifying text messages for the haiti earthquake," in *Proceedings of the 8th International Conference on Information Systems for Crisis Response and Management (ISCRAM2011)*, 2011.

[7] H. Mao, X. Shuai, and A. Kapadia, "Loose tweets: an analysis of privacy leaks on twitter," in *Proceedings of the 10th annual ACM workshop on Privacy in the electronic society*. ACM, 2011, pp. 1–12.

TABLE XII
TOPIC MODEL SETS OVER ALL TWEETS LABELLED AS VEHICLE THAT CONTAIN ITS MAKE

id	distribution	words in topic with counts
0	0.050	armed (2)driving (2)months (1)mvo (1)http://t.co/zebkn (1)
1	0.050	woman (2)afritrack (1)sanantonio (1)volkswagen (1)drive-thru (1)
2	0.050	kingsway (5)toti (4)saps (3)tida (3)gun (3)
3	0.050	last (5)tipper (2)look-out (2)vaquero (1)avd (1)
4	0.050	reg (9)blue (8)driver (7)corolla (7)stonebridge (6)
5	0.050	polo (3)reg/no (3)vehicle (3)golf (2)citi (2)
6	0.050	light (6)ago (5)nur (5)vehicle (5)please (4)
7	0.050	kzn/dbn (5)taken (4)recovered (4)toti (2)new (2)
8	0.050	black (6)vehicle (4)males (3)durban (2)greyville (2)
9	0.050	toyota (3)found (2)jeep (2)highlander (2)eky (2)
10	0.050	vehicle (4)road (3)rhy (3)merc (3)umlazi (3)
11	0.050	gas (8)city (8)vehicle (4)man (3)another (3)
12	0.050	silver (8)nissan (5)hyundai (5)blk (4)park (3)
13	0.050	benonicpf (1)detail (1)durban_fishing (1)back (1)anti (1)
14	0.050	passat (2)brother (2)pigspotter (2)silver (2)trafficsa (1)
15	0.050	tonight (1)fvgp (1)sherwood (1)pigspotkzn (1)charmskil (1)
16	0.050	bmw (2)yif (1)http://t.co/ndz (1)metallic (1)chrome (1)
17	0.050	camry (2)thought (1)yeah (1)bank (1)few (1)
18	0.050	station (7)lasabre (4)buick (4)fortwayne (1)http://t.co/ly (1)
19	0.050	middelburg (2)night (2)trailers (2)side (2)red (2)

[8] Y. Ko and J. Seo, "Text classification from unlabeled documents with bootstrapping and feature projection techniques," *Information Processing and Management*, vol. 45, no. 1, pp. 70–83, 2009.

[9] B. Suh, L. Hong, P. Pirolli, and E. H. Chi, "Want to be retweeted? large scale analytics on factors impacting retweet in twitter network," in *Social Computing (SocialCom), 2010 IEEE Second International Conference on*. IEEE, 2010, pp. 177–184.

[10] K. Nigam, A. K. McCallum, S. Thrun, and T. Mitchell, "Text classification from labeled and unlabeled documents using em," *Machine Learning*, vol. 39, no. 2-3, pp. 103–134, 2000.

[11] C. Featherstone, "The relevance of social media as it applies in south africa to crime prediction," 2013.