Catching Crime: Detection of Public Safety Incidents using Social Media

Vukosi Marivate
Council for Scientific and Industrial Research
South Africa
vmarivate@csir.co.za

Pelonomi Moiloa
Tohoku University
Japan
pelonomi@rii.med.tohoku.ac.jp

***Abstract*—The increasing prevalence of Social Media platform use has brought with it an explosion of new user generated public data. This data is centered around many, diverse topics. One theme of interest is how one can tap into the public safety and crime related user generated data to better understand patterns in the occurrence of crime incidents. One challenge in such data is that most of the data needs human annotation to make it usable by machines to analyse. This paper explores how different features, extracted from social media data, impact the performance of different classifiers. The classifiers are built to classify social media data as having to do with a reported crime or not. The challenge of few labelled data is discussed as well as different approaches to extracting features from the text data as well as the graph created by users interacting with each other is explored.**

Keywords*—*Social Media, Classification, Text Mining, Data Mining

I. Introduction

Social Media has become a tool not only for communication but also as a source of data emerging from different communities. We are interested in how we can leverage social media data to better understand patterns in the occurrence of public safety and crime incidents. While earlier work [1] focused on finding topics discussed in public safety and crime incident sharing/discussion or identifying vehicle descriptions through topic models [2], this paper focuses on creating classification models which can automatically label incoming social media data.

If we are able to automatically label this data, we can then start mining it for patterns [3], [4]. We will then be able to compare the outcomes of this analysis with data provided by Law Enforcement agencies and/or alternate organisations such as private security firms of non-profit organisations. In South Africa, public safety and crime data is normally shared by the South African Police Service in an aggregated form (in terms of number incidents per area). For us to be able to accurately build predictive models for crime incidents, we need to have access to incident level data. As such this paper is also proposing the use of automatic labelling as a step towards such a goal.

We first briefly discuss event detection/identification with the assistance of social media and its use in Law Enforcement. We then present our data collection followed by a discussion of the approach we will take in tackling the problem. We present our experimental results and then conclude.

II. Social Media Event Identification and Law Enforcement

Social media has been used for Law Enforcement applications such as the monitoring of terrorism [5] and the prediction of crime incidents given tagged data. Our previous work focused on extracting topics from different community discussions on public safety and crime incidents [1] and the privacy challenges when using such data [6]. In this paper we aim to build an automated labelling system for crime related incidents discussed on Social Media. This will open up opportunities for further work, especially in predictive policing models and automated flagging for follow up in a decision making situation. As such the goal of this paper is seen as providing additional information to that already available to the end-user.

There are existing works on event classification/identification with the use of Twitter data. These include detecting topics such as Earth Quakes [7] and general event extraction [8], [9]. We focus on a specific use case that spans over a long period of time, that of public safety and crime incident. We believe that the underlying patterns in how people talk about these incidents persists. For our models, we use features from text, user information as well as features extracted from the graph created when users that we collected in our dataset mentioned each other. In this paper we do not cover unique event detection, an area of active research whose goal is to build tools that automatically discover events that are occuring due to multiple social network users conversing about such events [10], [11]. We leave that to future work.

III. Data Collection

For this paper we will focus on data collected from South African public safety incident related accounts on Twitter. The community is mostly made up of individuals and non-governmental organisations that report or relay messages related to crime and public safety. This set of users were not Law Enforcement or Newspapers, but tend to be focal points of where other South African Twitter users send notices about incidents (notices of hijackings that recently occurred, motor vehicle accidents, robberies etc.). Examples of a subset of the accounts, as well as summaries of their descriptions, are:

- **CrimeAirNetwork** *(CEO - Crime Air Network Initiative*
- **gcalerts** *(Gauteng Crime Alerts ... Traffic & Accidents Updates. Community Alerts)*

- **CrimeWatchdog** *(Let us know of any illegal activity taking place in your area, from stolen cars to missing persons)*
- **TrafficNewsFeed** *(SA Traffic Updates)*
- **SAcrimefighters** *(Stand against crime and terrorism. Play your part in fighting crime by reporting suspicious or criminal activity to the authorities.)*
- **Abramjee** *(Consultant. CSI. Interpol....)*

Data was collected from Twitter from the beginning of May 2015 to the end of July 2015. We collected the data by using the Twitter Streaming API. We monitored posts coming from or being directed at these accounts (mentions). Due to using the Streaming API, we only had access to a subset of possible tweets [12]. Even with this limitation, we still can build meaningful classification models.

A. Data Labelling

For this paper, we first labelled 1299 tweets, using keyword matching (the set was not constructed at random to make it easier to label). The keyword themes were specifically:

- Hijackings,
- Theft and Robberies,
- Shootings.

An example of post that would be classified as a crime incident report is

> *THEFT OF MV: KEMPTON PARK. GP. RED 2013 VW POLO COMFORTLINE CT06HWGP.*

While a benign report would be

> *Lookout for smash-n-grabbers around the Fourways area, as per @SandtonNews*

The above post is only a warning that there might be criminals in a specific area, not that an incident happened. A *smash and grab* refers to an incident where a victim's motor vehicle has its windows smashed so that the perpetrator may gain access to the interior of the car to steal/grab an item. These incidents occur most frequently when a car is stopped at a traffic light.

Two labels are allocated and simply indicate whether a tweet is referring to a crime incident or not, as such binary classification. Of the 1299 tweets labelled, 604 (46.50%) referred to crime incidents. In this paper, the goal is to build a classifier that can identify similar tweets from the larger set of unlabelled tweets.

IV. Method

In this section we discuss the approaches taken to classify our data as being related to a crime incident or not. First we will discuss extracting features from text and then proceed with the use of additional features that can be extracted from users. We then present Self-Training as a way to increase the amount of data we have for training.

TABLE I: Public Safety Community Graph Features

Feature	Mean	Std. Deviation
Number Followed	9364.66	42575.26
Number of Followers	1080.06	3872.81
Number of Favourites	1737.28	8428.24
Number of Posts	17046.04	36062.45
Number of users mentioned in post	2.06	1.81
Number of URLs in post	0.33	0.48
Number of Hashtags (#)	0.36	0.79
In Degree (mentions)	341.69	1492.20
Out Degree (mentions)	146.52	354.38
Cluster Co-efficient (mentions)	0.25	0.31
Number of Triangles (mentions)	345.16	1366.46
Closeness Centrality (mentions)	0.36	0.06
Eigenvector Centrality (mentions)	2.83e-02	3

A. Text Processing

Focusing on the text in the Twitter posts, we tokenize the data using the Twitter Tokenizer from Natural Language Toolkit(NLTK) [13]. We used a bag-of-words [14] model to attain features from the text. The bag-of-words model uses all the data collected over the 3 months. We investigated further using Term FrequencyâInverse Document Frequency (TFIDF) transforms [15] and found that for the classification of short text, the use of TFIDF does not offer much improvement compared to the use of the Bag-of-Words features. It is possible to extract more features from the text which can be an avenue of future work.

B. User, Text, Graph and Topic Features

To elaborate on the features obtained from the text, we captured user features and graph features from the social network created by messages sent between different users. The features extracted were a mixture of user features, features from the text, as well as the features from the graph created by users mentions. The approach of extracting more features to supplement short text has been attempted [14], but did not take into account features [16] from the graph as well as user features. The features extracted, as well as their descriptive statistics are shown in Table I.

Using such features, we also want to understand the impact that user information might have on how likely it is that the user is reporting an incident. This type of analysis might assist in creating a trust model for each user when it comes to the topics we are interested in. So for example, could we create a model that gives us a confidence measure on how we should trust User A on information relating to motor vehicle accidents.

Another set of features we extract is Topic Models via Latent Dirichlet Allocation (LDA) [17]. We use all of the data collected (labelled and unlabelled) to train the topic model with 50 as the number of topics. After training the topic models we can use the multinomial distribution infered for each post as a feature vector, with 50 features.

C. Classification of crime

We experimented with different classifiers to compare their performance across the different feature variations. For the results section we used Logistic Regression, Support Vector Classifiers and Random Forests. For all of the experiments we used 10 fold, cross validation.

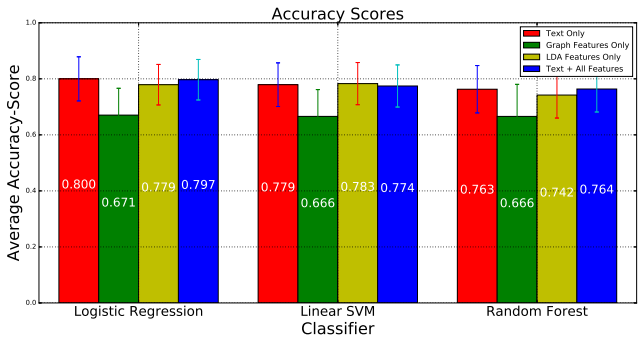


Fig. 1: Classification Results (Accuracy Score)

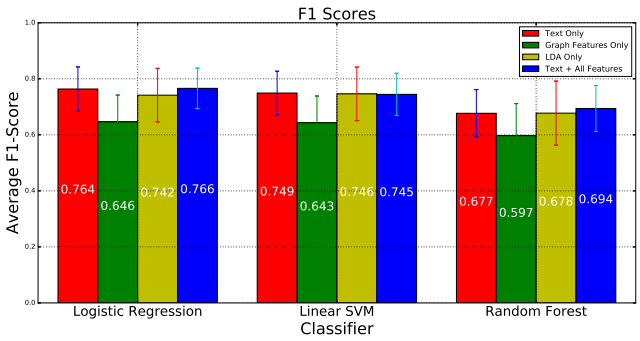


Fig. 2: Classification Results (F1 Score)

D. Semi-Supervised Learning: Self-Training

We explored two methods of self training approaches to increase the amount of information that is available to build classifiers. We briefly discuss the use of Self-Training for classification.

To increase the amount of data available for training our classifiers, we investigated the use of Self-Training [18]. The initial labelled set is composed of 1299 posts. We sample 70 % of the dataset and use it to train our initial classifiers. The initial classifiers are then used to label a set of unlabelled data from the initial set of Twitter posts. That is the initial classifier is used to predict the labels of a subset of a certain size of unlabelled data. The new labelled set is appended onto the original labelled set and a final classifier is trained on this set combination. The performance is then gauged on the held out dataset composed of 30% labelled data. In this paper we compare the dataset with only Bag-Of-Words features vs. all features that are extracted from text and user information.

V. Experiments and Results

We present the results of experiments on incident classification using our data. We have covered the descriptions of the methods earlier and the algorithms themselves are widely used. Our goal is to understand how different features impact the performance of classifiers for incident labelling.

A. Experiments with labelled data

We first present the results when using different feature sets to train different classifiers. The most significant results, in relation to accuracy score, are shown in Figure 1. The accuracy score gives an indication of how many labels in the test set were predicted correctly by the classifier. The F1 score is used to further gauge the success of the prediction capabilities of each classifier. This is due to the fact that accuracy is not holistically representative of a classifiers success when dealing with skewed data (we obtain 54 % accuracy when labelling all data in the set as non-crimes). The F1 score gives an indication of the relationship between classifier precision and recall as it is the weighted average of the two. A value of one indicates perfect precision and perfect recall [19]. The F1 scores are seen in Figure 2.

With only the Bag-Of-Words features (red), the highest accuracy score is achieved by Logistic Regression. This indicates it is able to correctly predict the classification class of a test subject better than the other classifiers. The highest F1 score is attained by Logistic Regression as well. This indicates that the Logistic Regression classifier is able to deal better with the slightly skewed distributed nature of the data at hand. At the same time all the other models have similar performance. When evaluating the performance when using User and Graph features (green), we see that the performance degrades slightly. This result is still not as bad as just guessing that all the data may be classified as not a crime which would yield an accuracy of 0.54 and an F1 score of 0.63. Using LDA features (yellow) is better than user or graph features, this should be expected as LDA features (50 topics) are a summary of the text features. The final comparison is the combination of the Bag-Of-Words features, the user and graph features and LDA features (blue). We see that the performance of the models based on F1 score and accuracy score increases from that of using the extra features on their own but does not beat simply using words. One would expect that we would get slightly better performance to the simple Bag-Of-Words model, as such this warrants further investigation in future.

It's important to note that the LDA features are also useful as they use all of the data for training, and not only the labelled training data. As such we should expect that on other unseen data, the predicted labels will likely be more accurate than using features that incorporate more generalized data. As such we can see the LDA as first a clustering algorithm that we then fit a classification model on top of the soft clusters/membership of new data to the clusters/topics.

B. Feature importance for User and Graph features

Even though the user and graph feature results are not as successful as the Bag-Of-Words features, we analysed the magnitude of weights given to each feature in the Logistic Regression classifier to find out if there were any standout features that dominated the classifier. The largest positive feature was the number of previous posts the user had, while the largest negative feature was the number of posts the user had favourited before. The former does give an indication that large accounts that attract a lot content tend to help filter content and as such would likely share information about real incidents.

C. Self-Training results

Here we present the results on our Semi-supervised learning approach with Self-Training. Figures 3 and 4 present the F1 and accuracy score for different levels of data for Self-Training.

We show 3 different settings for our experiment. First we have a baseline where we have not added any extra data through Self-Training. We then show the result with all the extra data added with Self-Training (green). Here we see that for most models the average performance (both accuracy and F1 score) decreases with self training. We also show the maximum performance achieved with a subset of the full unlabelled set (blue). For some of the classifiers the maximum obtained is the best performance as it is slightly better than the baseline performance.

VI. Discussion and Conclusion

In this paper we investigated the use of machine learning to classify social media text in order to be able to identify potential crime incidents that are being discussed online. With such a system we may be able to create an extra sensor for organisations to pick up on crime patterns and be able to work towards better public safety. We are able to build models that can classify using text features as well as a mixture of text features, features from the graph and user information on Twitter as well as features generated from topic models. Extracting more user features that can assist in establishing what characteristics within the communication network correlate with high likelihood of the information shared being useful is an avenue for further work. We still would like to explore other methods of semi-supervised learning as a way to expand the dataset. We plan to expand the labelled instances using crowdsource labelling as well in conjunction with semi-supervised learning methods to guide what is shown to users. Further, we aim to expand our labelling to multiple labels instead of binary. We would like to pick out events from newspapers, possible eyewitnesses and threaded conversations (a group of people discussing an incident together) from the data. Here crowdsourcing will play a major role.

Acknowledgement

We would like to acknowledge the assistance of Nyalleng Moorosi and Patrick Monamo for their assistance in shaping some of the work discussed in this paper.

References

[1] V. N. Marivate, "Extracting south african safety and security incident patterns from social media," in *Pattern Recognition Association of South Africa and Robotics and Mechatronics International Conference (PRASA-RobMech), 2015*. IEEE, 2015, pp. 106–111.

[2] C. Featherstone, "Identifying vehicle descriptions in microblogging text with the aim of reducing or predicting crime," in *2013 International Conference on Adaptive Science and Technology*. IEEE, 2013, pp. 1–8.

[3] P. J. Brantingham and G. Tita, "Offender mobility and crime pattern formation from first principles," *Artificial crime analysis systems: using computer simulations and geographic information systems*, pp. 193–208, 2008.

[4] S. Chaturapruek, J. Breslau, D. Yazdi, T. Kolokolnikov, and S. G. McCalla, "Crime modeling with lévy flights," *SIAM Journal on Applied Mathematics*, vol. 73, no. 4, pp. 1703–1720, 2013.

[5] G. Dean, P. Bell, and J. Newman, "The dark side of social media: review of online terrorism," *Pakistan Journal of Criminology*, vol. 3, no. 3, pp. 103–122, 2012.

[6] N. Moorosi and V. Marivate, "Privacy in mining crime data from social media: A south african perspective," in *2015 Second International Conference on Information Security and Cyber Forensics (InfoSec)*. IEEE, 2015, pp. 171–175.

[7] T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake shakes twitter users: real-time event detection by social sensors," in *Proceedings of the 19th international conference on World wide web*. ACM, 2010, pp. 851–860.

[8] A. Ritter, O. Etzioni, S. Clark *et al.*, "Open domain event extraction from twitter," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2012, pp. 1104–1112.

[9] H. Becker, M. Naaman, and L. Gravano, "Beyond trending topics: Real-world event identification on twitter," *Fifth International AAAI Conference on Weblogs and Social Media*, 2011.

[10] X. Dong, D. Mavroeidis, F. Calabrese, and P. Frossard, "Multiscale event detection in social media," *Data Mining and Knowledge Discovery*, vol. 29, no. 5, pp. 1374–1405, 2015.

[11] J. Chae, D. Thom, H. Bosch, Y. Jang, R. Maciejewski, D. S. Ebert, and T. Ertl, "Spatiotemporal social media analytics for abnormal event detection and examination using seasonal-trend decomposition," in *Visual Analytics Science and Technology (VAST), 2012 IEEE Conference on*. IEEE, 2012, pp. 143–152.

[12] F. Morstatter, J. Pfeffer, H. Liu, and K. M. Carley, "Is the sample good enough? comparing data from twitter's streaming api with twitter's firehose," *arXiv preprint arXiv:1306.5204*, 2013.

[13] E. Loper and S. Bird, "Nltk: The natural language toolkit," in *Proceedings of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics-Volume 1*. Association for Computational Linguistics, 2002, pp. 63–70.

[14] B. Sriram, D. Fuhry, E. Demir, H. Ferhatosmanoglu, and M. Demirbas, "Short text classification in twitter to improve information filtering," in *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2010, pp. 841–842.

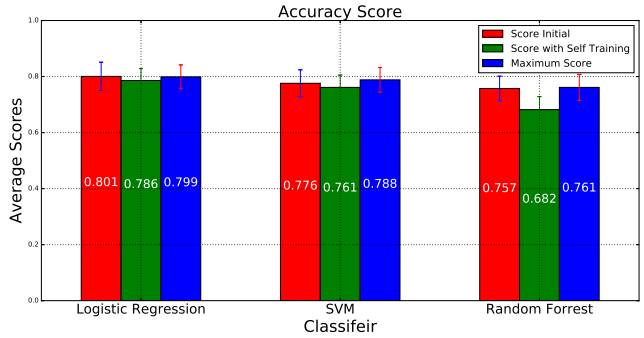
[15] S. Bird, E. Klein, and E. Loper, *Natural language processing with Python*. " O'Reilly Media, Inc.", 2009.

[16] A. Rajaraman, J. D. Ullman, J. D. Ullman, and J. D. Ullman, *Mining of massive datasets*. Cambridge University Press Cambridge, 2012, vol. 1.

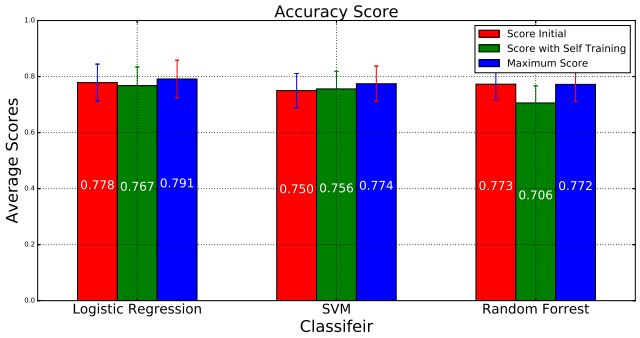
[17] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *the Journal of machine Learning research*, vol. 3, pp. 993–1022, 2003.

[18] O. Chapelle, B. Schölkopf, A. Zien *et al.*, "Semi-supervised learning," 2006.

[19] C. Sammut and G. I. Webb, *Encyclopedia of Machine Learning*. Springer New York, 2011.

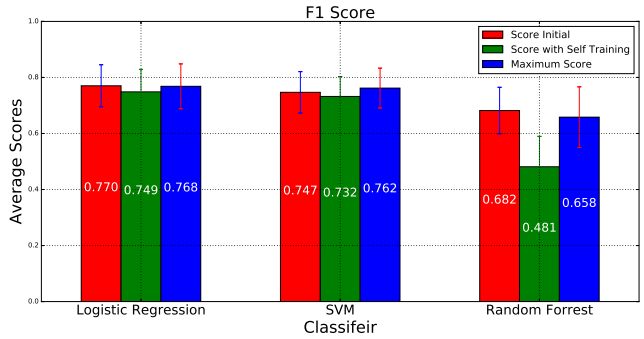


(a) Self-Training accuracy score (BOW + all features)

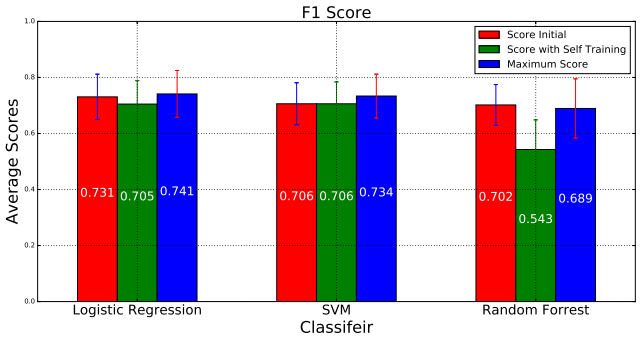


(b) Self-Training accuracy score (BOW only)

Fig. 3: Accuracy for labelling Self-Training results



(a) Self-Training F1 score (BOW + all features)



(b) Self-Training Accuracy (BOW only)

Fig. 4: F1 Score for labelling Self-Training Results