The Majority Report - Can we use big data to secure a better future?

VIVIEN P.S. CHAN
K.P. CHOW
The University of Hong Kong

Abstract
With the widely adopted use of social media, it now becomes a common platform for calling supporters for civil unrest events. Despite the noble aims of these civil unrest events, sometimes these events might turn violent and disturb the daily lives of the general public. This paper aims to propose a conceptual framework regarding the study of using online social media data to predict offline civil unrest events. We propose to use time-series metrics as the prediction attributes instead of analyzing message contentsbecause the message contents on social media are usually noisy, informal and not so easy to interpret. In the case of a data set containing both civil unrest event dates and normal dates, we found that it contains many more samples from the normal dates class than from the civil unrest event dates class. Thus, creating an imbalanced class problem. We showed using accuracy as the performance metrics could be misleading as civil unrest events were the minority class. Thus, we suggest to use additional tactics to handle the imbalanced class prediction problem. We propose to use a combination of oversampling the minority class and using feature selection techniques to tackle the imbalanced class problem. The current resultsdemonstrate that use of time-series metrics to predict civil unrest events is a possible solution to the problems of handling the noise and unstructured format of social media data contents in the process of analysis and predictions. In addition, we have showed that the combination of special techniques to handle imbalanced class outperformed other classifiers without using such techniques.

Keywords
Social Media Platform, Predictive Modeling, Civil Unrest

1. The Introduction

Adoption of social media platforms, like Facebook, Twitter, discussion forums, and micro-blogging, as a platform to call for supporters to join the offline ground public events has become a phenomenon in recent years. We observed this is due to the apparent anonymous nature, low barriers to posting messages and ease of use of social media platform (Kalampokis et al., 2013).

Civil unrest as defined in this paper refers to protests, riots, or public demonstrations against the government. Even though the event organizers might initially intend to be a peaceful demonstration, such civil unrest might finally escalate into crimes or chaos. Civil unrest calling for its supporters through online social media platforms to demonstrate to the government, like the Arab Spring Movement in December 2010, and the Occupy Wall Street movement in September 2011, have become a phenomenon and spread to various forms of "Occupy" movement in different parts of the world (Juris, 2012). Yet, these peaceful demonstrations might sometimes turn into riots which might undermine public order. For

example, the London riot in August 2011, started from a peacefully march to seek justice for the death of Mark Duggan, ended up in riots which spread from London to cities and towns across England. Duggan was shot by a police officer who attempted to arrest Duggan on suspicion of planning a criminal attack. The social media platforms also played a major role in fueling the England riot crowds (Fuchs, 2012). It was reported over 1,100 were charged for offences ranging from burglary to violence during this period (BBC News UK, 2011). Thus, it becomes important for the law enforcement to forecast the intensity of the civil unrest events, so that they could be well prepared to restore the public order and protect the society.

A pertinent question to ask by both researchers and law enforcement agents would be: Can we use "collective wisdom" to predict an upcoming civil unrest events? There are several technical challenges of predicting civil unrest events based on data collected from online social media, i.e. "collective wisdom" or we call "social media intelligence" in this paper. First, a vast amount of social media data is being created every day. Such a huge amount of information available online makes it very difficult for human to identify, process and analysis. Second, the data generated on social media platforms are both noisy and unstructured. So, making analysis of text contents based on human keyword-based flagging on social media platform not an easy task because the data format is both informal and unstructured, containing misspelt words, incorrect grammars or emoticons. Third, the dynamic nature of social media also makes it hard to identify a civil unrest topic or event based on simple keyword-based searching technique by humans. For example, during the post-umbrella movement in Hong Kong from Jan to Mar 2015, the event organizers called for supporters on social media using keywords like "shopping" instead of "demonstration" when creating posts for civil unrest events. Finally, the problem of identifying what attribute(s) to be used as a predictor for an upcoming civil unrest event (Gundecha & Liu, 2012). In this paper, we propose a conceptual framework that makes use of the social media intelligence to predict the intensity of an upcoming civil unrest event organized thru online social media. Our primary contributions are:

- Suggesting an intuitive conceptual framework that predicts an upcoming civil unrest event;
- Using a real world case as a proof-of-concept for the proposed conceptual framework.

2. A Proposed Cybercrime Intelligence Model

2.1 Literature Review: Civil unrest prediction and social media intelligence

In the past decade, many researchers made use of social media data for prediction of different real world phenomenon. For instance, some researchers tried to predict movie box office based on Twitter data (Asur and Huberman, 2010); some others attempted to predict stock price using information obtained from social media (Bollen et al., 2011); research of election results prediction using Twitter data (Tumasjan et al., 2010); and prediction of natural phenomena like earthquake based on Twitter data (Lampos and Cristianini, 2012).

Although there have been overwhelming interest in predicting various real world phenomenon using social media in the past decades, the study of civil unrest events prediction using social media data is still in its infancy. In a study conducted by Agarwal and Sureka (2015) on applying social media data for predicting civil unrest oriented threats, they observed a

rapid increase in interest of forecasting civil unrest by mining the social media in just the recent 3 years. We believed the reason behind such a surge in attentions from researchers is due to the Arab Spring movement which started in December 2010 and made use of Twitter to call for their supporters to protest against the government. Since the Arab Spring movement, there have been increasing number of civil unrest events which made use of the social media platforms to call for its supporters, like the various forms of Occupy movements which started from the Occupy Wall Street movement in 2011(Agarwal & Sureka, 2015).

Yu and Kak (2012) surveyed the techniques used in prediction using social media and listed two categories of predictors, namely, message characteristics and social network characteristics. Message characteristics would be used for prediction of a future event while social network characteristics would be focused on identification of groups of networks. Message characteristics measured message features and focused on either content metrics (e.g. sentiment or keywords) or time series metrics; while social network characteristics measured network structure features and using techniques like social network analysis (Yu & Kak, 2012).

In the research domain of forecasting civil unrest events, most of the researchers made use of message characteristics to predict a civil unrest event. The techniques used by these researchers relied mainly on content metrics, like, sentiment analysis or civil unrest related keyword identification or spatial-temporal keyword classification (Agarwal & Sureka, 2015). For example, Ramakrishnan et al (2014) made forecast of future civil unrest events by analyzing the presence of time, location and targeted domain related keywords in the message contents collected (Ramakrishnan et al., 2014).

Despite the usefulness of sentiment analysis on social media platforms as discussed, Gundechaand Liu (2012) pointed out that using sentiment analysis on social media contents is hard because the languages used on these platforms are usually informal and ambiguous. In addition, the performance evaluation of sentiment analysis prediction power provided a challenge for researchers due to the lack of ground truth in this area (Gundecha & Liu, 2012). For example, in our current study with data collected from social media platforms being used in Hong Kong, most of the typical post contents consisted of emoticons or Internet slang (e.g. lol = laugh out loud) or very short and spam-like text messages (e.g. "push to top"(推), "leave a name"(留名)) or just photos or foul language. So, performing content or sentiment analysis on these user-generated contents might not be good enough to predict the outcome of a civil unrest event. In addition, using content metrics means the researchers might need to define a relevant topic beforehand, for example, we need to define a collection of words related to "anti-parallel trade"event before we run the prediction model. So, it might be useful only if the researchers already knew in advance what the civil unrest event was about (Jungherr & Jürgens, 2013). Some researchers have questioned the accuracy of using sentiment or content analysis on social media data. Kalampokiset al (2013) studied the predictive power of social media and found that 65% of the social media predictive research they studied actually challenged the predictive power when using lexicon-based approaches to predict a future event using social media data. They concluded "… it seems that sentiment analysis in SM (Social Media) requires innovative approaches that could address the noisy and informal nature of SM."(Kalampokis et al., 2013).

2.2. A conceptual framework: Using social media intelligence for civil unrest event prediction

2.2.1 Reframing the problem

In order to overcome the limitations of prediction using the content metrics predictor as discussed in the previous paragraphs, we propose to use time-series metrics as the predictor for civil unrest events as it is not easily affected by the noise of the social media data, easy to interpret and efficient to handle vast amount of data. Some researchers made use of the time-series metrics, like using daily twitter rate of a specific topic to forecast movie box office before the movie was released (Asur & Huberman, 2010). Other researchers studied the detection of abnormal online or offline phenomenon by analyzing the deviation of tweet data rate from the "normal state" of online social media platform (Jungherr and Jürgens, 2013). Furthermore, researchers showed that the number of tweets typically shot up during the day of some major events, like natural disasters and breaking news, but dropped instantly afterwards (Hu et al., 2012).

However, unlike those major events, the tweets usually peaked before the civil unrest event day and there might be several "peaks" preceding the day of a civil unrest event (Hu et al., 2013). Take for example, during the post-umbrella movement in Hong Kong from Jan to Mar 2015, there has been civil unrest event nearly every weekend. Figure 1 showed the number of posts on a commonly used discussion forum in Hong Kong peaked before a ground protest on 8 Feb 2015 when protestors demonstrated at a district in Hong Kong against the policy of the government and people from mainland China. In the figure, the number of forum topics and posts were lowest on the event date of 8th Feb 2015 (with 3,431 posts over the week). On the other hand, it was observed there were 2 peaks on 3rd and 6th Feb (with 6,223 and 7,663 posts) before the event date. Thus, by analyzing these early peaks before a civil unrest event, we might be able to shed some lights on what early attributes would be useful for the prediction of an upcoming civil unrest event.

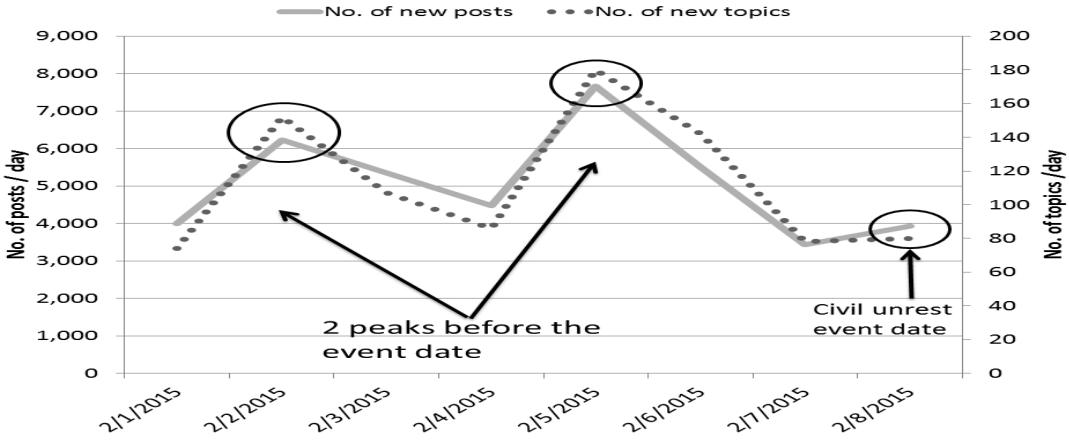

Figure 1. Distribution of a forum posts related to a civil unrest event on 8 Feb 2015 in Hong Kong.

Based on this observation using simple time-series metrics of daily post and topic volumes, we investigate the possibility of using these online social media data to predict real-world phenomenon. The question can be re-framed as:

Using the social network data, can we use the data collected N days before a civil unrest event date to predict an upcoming real-world street demonstration?

Thus, we reframe the problem into a simple classification problem to predict the target Y (civil unrest event would happen or not) as a function of the feature X (using the time-series metrics, e.g. the daily post volumes, N days before the event day).

2.2.2 Handling imbalanced classes

In the civil unrest event classification problem, we would expect the number of civil unrest event days throughout the year would normally be a rare class. In other words, one class (non-civil unrest event) is much more common than another class (civil unrest event). For example, in our experiment dataset, the ratio of civil unrest event day is about 12% of the total. When the classes are imbalanced in the dataset, guessing the more common class would probably yield very high accuracy. Thus, when we develop the model for civil unrest event prediction, we need to take into consideration how to handle this imbalanced dataset problem.

In a review on class imbalance problem, Longadge et al. (2013) reviewed the techniques commonly used to solve this issue. These included sampling techniques, use of new learning algorithms, and feature selection. In our proposed model, we suggest to use two techniques, these are over-sampling and feature selection. First, the sampling techniques mainly are under-sampling the majority class or over-sampling the minority. In the case of civil unrest event prediction, we suggest to use over-sampling in order to preserve all useful information. Second, the goal of feature selection is to reduce the dimensionality of the feature space so as to optimize the performance of a classifier. In the case of civil unrest event prediction, the feature space might be huge as we are interested in the parameters on the social media platform X number of days before the civil unrest event. So, the use of feature selection is needed (Longadge et al., 2013). The use of new learning algorithms mainly includes a combination of sampling and feature selection techniques. We considered using currently available machine learning techniques would be a good start to develop the civil unrest event prediction.

2.2.3 Building the conceptual framework

We suggest a conceptual framework which does not rely on the content metrics, resilient to noise and easy to interpret for the prediction of an upcoming civil unrest event using social media data. The conceptual framework is depicted in Figure 2.

This conceptual framework is designed for the detection of civil unrest event when there is a detected surge of attentions online and this does not require any prior knowledge from the researchers. First, raw data is collected from online social media platforms. Second, suitable time-series attributes (e.g. hourly tweet-rate on Twitter; daily number of posts on discussion forums; etc.) will be selected as predictors of the predictive modeling stage. Third, decide the prediction window, that is how many hours, days or weeks earlier than the civil unrest event

date would be used as input. Fourth, use both oversampling the minority class and selecting a sub-set of features to tackle the imbalanced class. Fifth, a machine learning classifier would be used to predict the civil unrest event based on the time-series metrics. Finally, we suggest to evaluate the performance of the classifier using precision, recall and F1 measures instead of classification accuracy.

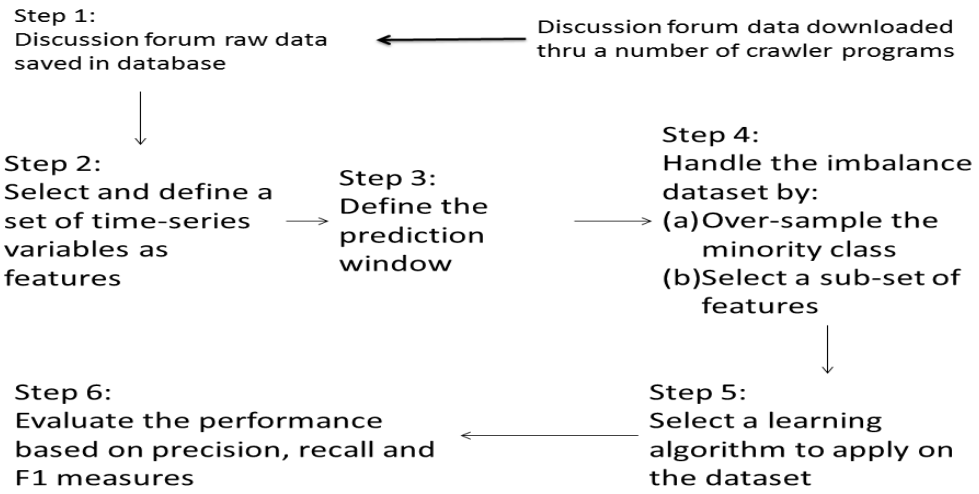

Figure 2 Conceptual framework of civil unrest event detection

3. Proof-of-concept: Case of Post-umbrella movement in Hong Kong

In our proof of concept, we made use of a data set collected from the two most commonly used discussion forums in Hong Kong during January to March 2015. This was the post-umbrella movement period in Hong Kong. The Umbrella Movement in Hong Kong started in late September 2014 by a group of scholars who formed the unit named "Occupy Central with Love and Peace". BBC News reported on 11 December 2014 that "The dismantling passed off peacefully, but many activists vowed to continue with other forms of civil disobedience." (BBC News, 2014), which marked the end of the Umbrella Movement after occupying different regions in Hong Kong since late September 2014. Protestors changed their strategies after the post-Umbrella movement. Instead of explicitly calling for their supporters to go on streets to protest, they called for supporters under the names of "shopping activities" or "recovering a region". These actions were to protest against the government's policy regarding parallel traders or mainlanders. There were different forms of ground protests in different regions of Hong Kong nearly every weekend from January to March 2015 (SCMP, 2015). As explained in the previous sections, it would make it more difficult to detect these new forms of civil unrest event if the prediction models need to rely on the use of content-metrics since it would not be possible for the researcher to know what kind of "contents" to define beforehand when the social media contents are very dynamic.

3.1 Experiment

We collected the data from the two most commonly used discussion forums in Hong Kong, namely, Hong Kong Golden (www.hkgolden.com) and Hong Kong Galden (www.hkgalden.com) from January to March 2015 just after the Occupy Central Movement in December 2014. The major reason is because many demonstrations were called out through these two discussion forums and they are the most popular platforms used by Hong Kong users to discuss political issues. The posts are collected from one sub-categories ("Current Affairs") of the discussion forums. Hong Kong Golden has a long historical standing among Hong Kong users and it was set up in the year 2000. On the other hand, Hong Kong Galden, is setup in 2013 and has a very similar structure as Hong Kong Golden, but also very popular among users who are either de-registered by Hong Kong Golden or select to join Hong Kong Galden who claims itself as the most "liberal" discussion forum in Hong Kong.

	Jan	**Feb**	**Mar**
Total number of topics	3,273	3,896	5,912
Total number of posts	130,400	135,806	199,118
Total number topic authors	1,026	1,154	1,506
Total number post authors	13,484	13,221	14,694

Table 1 Data Collected from Jan to Mar 2015

Table 1 showed the summary of data collected during Jan to Mar 2015. In the current data set collected during Jan-Mar 2015, a total of 13,081 topics and 465,324 posts were downloaded from the discussion forum. Figure 3 showed the total number of daily posts and topics during Jan to Mar 2015. Most street demonstrations became violent during weekends between 8th Feb to 15th Mar (except the weekend on 22nd Feb which was the long weekend holidays of Lunar New Year).

To prepare the data for running the prediction model, several steps were involved. First, pre-process the raw data by applying normalization to the total number of daily posts and topics for both forums. In this way, the volume of posts from different discussion forums would become comparable on the same scale. Second, a 6-day lag before the civil unrest event date prediction window was selected. Since there were street demonstrations nearly every weekend, it seemed logical to choose one week's data for prediction. Third, we could also include other time-series metrics as the predictor attributes, like, posting rate (i.e. number of posts/hour, number of topics/hour, number of post authors, number of topic authors, etc.). In this experiment, we have selected four time-series attributes during the 6-day time window to form the feature set as shown in Table 2 below. Thus, for each record, the feature set consisted of 24 features.

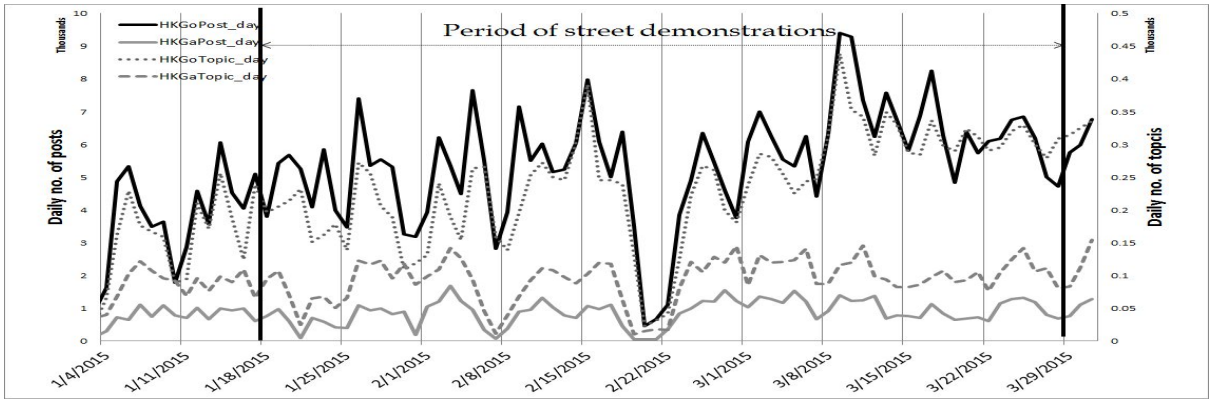

Figure 3 Distribution of daily posts and topics on the sub-forum "Current Affairs" of Hong Kong Golden and Hong Kong Galden during Jan-Mar 2015

Time-series feature	Number of days before a particular date					
	1-day	2 days	3 days	4 days	5 days	6 days
Daily new posts	P_L1	P_L2	P_L3	P_L4	P_L5	P_L6
Daily new topics	T_L1	T_L2	T_L3	T_L4	T_L5	T_L6
Daily unique post authors	PA_L1	PA_L2	PA_L3	PA_L4	PA_L5	PA_L6
Daily unique authors	TA_L1	TA_L2	TA_L3	TA_L4	TA_L5	TA_L6

Table 2 List of features for each record

Finally, we manually label the days from Jan to Mar 2015 into dates of no civil unrest event and dates of civil unrest event. Table 3 listed the type of street demonstrations during this period. These civil unrest event dates were labeled as "0" and other non-civil unrest event dates were labeled as "1". The percentage of civil unrest event date to non-civil unrest event date ratio in the date set is 20:146, that means the percentage of civil unrest event date was 12.05%.

Date	Title of demonstration posted on forum	Demonstration
17 Jan	Ultimate Shopping activity ("「數百人終極鳩鳴」活動")	Peaceful[a]
31 Jan	Voluntary surrender to Hong Kong Police (網上號召自發到灣仔警察總部自首)	Peaceful
8 Feb	Recover TuenMun district (「光復屯門」活動)	Not peaceful[b]
15 Feb	Shopping discount at Shatin (「星期日沙田好似有大減價」活動)	Not peaceful
1 Mar	Recover Yuen Long district (「光復元朗」活動)	Not peaceful
8 Mar	Recover SheungShui district (「光復上水」活動)	Not peaceful
15 Mar	Be a parallel trader at Hong Kong Governors' Office (往禮賓府「齊做水貨客，在禮賓府邊賞花，邊分貨」活動)	Not peaceful
22 Mar	Anti-parallel trader at SheungShui district (上水反水貨客集會)	Peaceful

Date	Title of demonstration posted on forum	Demonstration
29 Mar	Cleaning SheungShui district (「還原上水清潔大行動」)	Peaceful

[a] Demonstrators came out on the street and demonstrated peacefully.

[b] Demonstrators came out on the street and ended with conflict with police & some of the demonstrators got arrested.

Table 3 List of street demonstration during Jan-Mar 2015

In this experiment, we have chosen random forest as the base learning algorithm. The random forest algorithm is a combination of decision trees classifiers which make predictions independently on a randomly selected samples. The combined trees formed the forest based on majority voting and resulted in minimized error. The random forest algorithm is relatively robust to outliers and noise, thus suitable to be applied on the social media data (Breiman, 2001). Then, as discussed in Section 2.2.2, we applied over-sampling of minority class; selected a sub-set of features; and combined over-sampling and feature selection to compare the performances of difference learning algorithms. Thus, we have a total of four learning algorithms coded as "Random Forest" (RF), "Random Forest-Oversample" (RF-O), "Random Forest-Feature selection" (RF-F), and "Random Forest-Oversampling-Feature Selection" (RF-O-F). We have used 30% of the data set as testing data set.

3.2 Results

The metrics we used in measuring the classifier's performance include, accuracy, precision, recall, F1 measure, and the area under the precision-recall curve ("AUCPR"). While accuracy measures both true positives and true negatives, the remaining metrics mainly measure the true positive results.

Learning algorithm	Accuracy	Precision	Recall	F1 measure	AUCPR
Random Forest (RF)	0.86*	0.74	0.86*	0.80*	0.90
Random Forest-Oversample (RF-O)	0.84	0.74	0.84	0.79	0.93
Random Forest-Feature selection (RF-F)	0.86*	0.74	0.86*	0.80*	0.79
*Random Forest-Oversampling-Feature Selection (RF-O-F)***	0.78	0.84*	0.78	0.80*	0.96*

Note: *Highest score of the performance metrics

** "Random Forest-Oversampling-Feature Selection" (RF-O-F) showed the highest score in Precision, F1 measure and area under the precision-recall curve (AUCPR)

Error! Reference source not found. Performance metrics of 4 learning algorithms

Table 4 summarizes the performances of the four learning algorithms. On the accuracy, recall and F1 measure metrics, "Random Forest" and "Random Forest-Feature Selection" got the highest score (accuracy = 0.86, recall = 0.86, F1 measure = 0.80). However, "Random Forest-Oversampling-Feature Selection" showed the highest value for the AUCPR (AUC = 0.96), precision (precision = 0.84) and F1 measure (F1-measure = 0.80). Among the four

learning algorithms, "Random Forest-Oversampling" got the lowest score across all performance metrics.

The confusion matrix of the 4 learning algorithms in Table 5 showed that only "Random Forest-Oversample-Feature Selection" could predict the actual civil unrest event while the remaining three could not predict any actual civil unrest event.

		Predicted civil unrest event	Predicted non-civil unrest event
RF	Actual civil unrest event	0	7
	Actual non-civil unrest event	0	43
RF-O	Actual civil unrest event	0	7
	Actual non-civil unrest event	1	42
RF-F	Actual civil unrest event	0	7
	Actual non-civil unrest event	0	43
RF-O-F	Actual civil unrest event	4	3
	Actual non-civil unrest event	8	35

Table 5 Confusion matrix of 4 learning algorithms

4. Discussion and conclusions

To predict civil unrest event from online social media data, we propose to use time series metrics as predictors by detecting the sudden surge of daily posts as compared to normal online pattern before a civil unrest event date. In our proof of concept experiment, we showed how simple it could be to predict real world civil unrest event by just measuring the number of daily posts on Hong Kong Golden and Hong Kong Galden. In addition, using the techniques of over-sampling minority class and selecting a subset of features seem to be effective to handle imbalanced class when applying random forest learning algorithm.

The current results showed that "Random Forest-Oversampling-Feature Selection" has the highest precision (value = 0.84), F1 measure (value = 0.80) and highest AUCPR (value = 0.96). The results also showed that "Random Forest-Oversampling-Feature Selection" has the lowest accuracy (accuracy = 0.78). The current confusion matrix results showed that when dealing with a very skewed dataset, a classifier would easily get high accuracy if it always predicts no civil unrest event. As discussed in Section 2.2.2, accuracy might not be an appropriate measure as it would be easy to get very high prediction accuracy on the majority class (i.e. non-civil unrest event day) even when the prediction for minority class is missed. When we predict whether civil unrest would happen at a future date, the accuracy of predicting the minority class (civil unrest event) would be more important. Thus, using precision metrics would be a more appropriate measure. In a civil unrest event prediction, it would be critical for the decision maker to know when there is actual civil unrest event. Thus, with the highest precision score, "Random Forest-Oversampling-Feature Selection" seems to be more preferred to the other three learning algorithms when handling imbalanced class in the case of civil unrest prediction.

By using random forest as the learning algorithm, we were able to explain and formalize a real world problem. This conceptual framework adds to the literature about the prediction of offline civil unrest event using online social media data.

5. Research limitations and direction for future research

Similar to all other research using online social media data, there are still some limitations in our current work. First, the online social media data might be very volatile. Unlike other online data (e.g. those on websites or blogs), online social media data could be deleted within hours after posting. In particular for the civil unrest event, the organizer might post their online callings for just a few hours and removed them so that there would be no traces for the law enforcement agencies. Second, despite the use of time-series metrics, there still exists other "noises" related to the posting volume of the online social media data. For example, it is well known that there exist a lot of spam messages on the social media data which might create a false surge on the volume of daily data.

Future directions in this area might include tackling the problems and limitations mentioned. In our current work, we have only used the very basic version of machine learning techniques, which is random forest. In addition, we have only made use of a simple time series metric, i.e. the daily number of posts and topics. Future research might include expanding the repertoire of time series metric, including, hourly rate of posting, number of users with high social influence index participating in the discussions, just to name a few. Another direction of research is to include link analysis to detect any possible relation and networks among the participants. In addition, apart from detection of civil unrest events, we suggest that it is possible to apply the proposed conceptual framework in other cybercrimes. For example, identification and estimation of the risk of cyber-attacks.

Finally, in the current research, we observe a rapid transition from private email communications to public postings on social media and these social media posts might include personal information. Thus, privacy might need to be protected during the process of collection, use and sharing of these collected data for research purposes. Research related to the ethical use of social media data or privacy-enhanced data collection methodologies would be an important area for researchers.

References

Agarwal, S. and Sureka, A. (2015) Applying Social Media Intelligence for Predicting and Identifying On-line Radicalization and Civil Unrest Oriented Threats. *CoRR,abs/1511.06858*.

Asur, S. and Huberman, B. A. (2010) Predicting the Future with Social Media. In Paper Presented to the Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Volume 01.

BBC News, (2014). *Hong Kong protests: Arrests as Admiralty site is cleared*. (11 December 2014) BBC News, [online]. Available at: http://www.bbc.com/news/world-asia-china-30426346 [Accessed 19 April 2016].

BBC News UK, (2011). *England's week of riots*. (15 August 2011) BBC News UK, [online]. Available at: http://www.bbc.co.uk/news/uk-14532532 [Accessed 19 April 2016].

Bollen, J., Mao, H. and Zeng, X.-J. (2011) Twitter mood predicts the stock market. *J. Comput. Science,2*(1), pp. 1-8.

Breiman, L. (2001) Random Forests.inSchapire, R. E., (ed.) *Machine Learning*,The Netherlands: Kluwer Academic Publishers. pp. 5-32.

Fuchs, C. (2012) Social media, riots, and revolutions.*Capital & Class,36*(383-391).

Gundecha, P. and Liu, H. (2012) Mining Social Media: A Brief Introduction.*2012 Tutorials in Operations Research*, pp. 1-17.

Hu, M., Liu, S., Wei, F., Wu, Y., Stasko, J. and Ma, K.-L.(2012) Breaking news on twitter. In Paper Presented to the Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Austin, Texas, USA.

Hua, T., Lu, C. T., Ramakrishnan, N., Chen, F., Arredondo, J., Mares, D. and Summers, K. (2013) Analyzing Civil Unrest through Social Media. *Computer,46*(12), pp. 80-84.

Jungherr, A. and Jürgens, P. (2013) Forecasting the pulse: How deviations from regular patterns in online data can identify offline phenomena. *Internet Research,23*(5), pp. 589-607.

Juris, J. S. (2012) Reflections on #Occupy Everywhere: Social media, public space, and emerging logics of aggregation. *American Ethnologist,39*(2), pp. 259-279.

Kalampokis, E., Tambouris, E. and Tarabanis, K. (2013) Understanding the predictive power of social media.*Internet Research,23*(5), pp. 544-559.

Lampos, V. and Cristianini, N. (2012) Nowcasting Events from the Social Web with Statistical Learning.*ACM Trans. Intell. Syst. Technol.,3*(4), pp. 1-22.

Longadge, R., Dongre, S. and Malik, L. (2013) Class Imbalance Problem in Data Mining Review.*International Journal of Computer Science and Network (IJCSN),2*(1).

Ramakrishnan, N., Butler, P., Muthiah, S., Self, N., Khandpur, R., Saraf, P., Wang, W., Cadena, J., Vullikanti, A., Korkmaz, G., Kuhlman, C., Marathe, A., Zhao, L., Hua, T., Chen, F., Lu, C. T., Huang, B., Srinivasan, A., Trinh, K., Getoor, L., Katz, G., Doyle, A., Ackermann, C., Zavorin, I., Ford, J., Summers, K., Fayed, Y., Arredondo, J., Gupta, D. and Mares, D. (2014) 'Beating the news' with EMBERS: forecasting civil unrest using open source indicators. In Paper Presented to the Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, New York, New York, USA.

SCMP, (2015) *Topics: Parallel Trading*. (2015) South China Morning Post [online].Available at: http://www.scmp.com/topics/parallel-trading [Accessed 3 June 2015].

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

Yu, S. and Kak, S. (2012) A Survey of Prediction Using Social Media.*CoRR,abs/1203.1647*.