Semi-supervised Event Message Identification System for Targeted Domain

San San Nwe
Faculty of Information Science
University of Computer Studies
Yangon, Myanmar
e-mail: 811sansannwe@gmail.com

Nang Saing Moon Kham
Faculty of Information Science
University of Computer Studies
Yangon, Myanmar
e-mail: moonkhamucsy@gmail.com

***Abstract*—Social media have become increasingly popular components of our everyday lives in today's globalizing society. They provide a context where people across the world can communicate, exchange messages, share knowledge, and interact with each other regardless of the distance that separates them. This research trend, extraction of events for specific domain from these social media is emerging speedily ranging from business intelligence to nation security field. The short length of Twitter messages and frequent use of informal and ungrammatical language challenge many long standing approaches for automatically detecting and categorizing events using streamed data in Event Message Identification system. A semi-supervised approach with Support Vector Machine (SVM) in combination with the corpus to identify the events from twitter for targeted domain in specific location is proposed in this paper. The experimental results show that the proposed semi-supervised SVM model is more efficient than a strong state-of-the-art semi-supervised classification model of Logic Regression, Naïve Bayes and Decision Tree.**

Keywords-social media; twitter; semi-supervised; events; SVM

I. INTRODUCTION

Social media are web applications that allow people to share statuses, information and opinions in short messages. They provide light weight, easy and fast way of communication between us and also present a rich and timely source of information on events taking place in the world. Twitter is a very popular micro-blogging service. There are millions of people that use Twitter to share their daily stories. The topics that people usually share on Twitter range from daily stories, current events, opinions and others specific type of information [1]. Twitters user has been noticeable across areas such as education, legal proceedings, emergencies/crisis situations, survey opinion, political campaigning, protests, public relations, NASA space missions, business, fundraising and many other purposes.

The rich up-to-date sensing information allows discovering and tracking important events even earlier than news, with important applications such as public health and emergency management. Although identifying events from newspaper reports has been well studied, analyzing messages in Twitter requires more sophisticated techniques. Twitter messages are irregular, contain misspelled or non-standard acronyms, and are written in informal style. Additionally, tweets are filled with trivial events discussing daily life. Twitter's noisy nature challenges traditional text-based event detection methods and therefore specifically designed event detection approaches are needed for Twitter text analysis [2].

Social media presents a rich and timely source of information on events taking place in the world, enabling applications such as earthquake detection or identifying the location of missing persons during natural disasters. Previous work on event extraction has relied on large amounts of labeled data, or taken an open-domain approach in which general events are extracted without a specific focus. Often an information analyst might be interested in tracking a very specific type of event. However, when tracking relevant keywords, the Twitter API retrieves roughly the same total volume of data; however a much larger proportion is relevant to the education-related events of interest. But, not all tweets mentioning a relevant keyword will describe the events of interest, many system therefore leverage the seed events previously mentioned, to train a supervised extractor [3, 4].

Supervised methods have modest improvement in performance. Instead these classifier based techniques require lots of manual efforts to annotate tweets. Moreover, the event expressions in Twitter shift over time. Semi-supervised methods are designed to solve these problems, usually with the assist of news corpus or knowledge base [5].

Moreover, Current works on extracting and understanding city events related to education, mainly rely on technology enabled infrastructure to observe and record events. This paper proposes a semi-supervised event message identification system to automatically detect documents containing information about educational events in the desired location.

II. RELATED WORKS

This section describes an overview of existing approaches proposed for Event Extraction from Twitter. J. Deng et.al [6] made massively survey on the previous researches, and this paper also based on this valuable categorization and continues further analysis for semi-supervised approach. The event types can be grouped into open domain and domain specific approaches, and the existing techniques of event message identification (EMI) are also categorized into unsupervised, supervised and semi-

978-1-5386-7635-6/18/$31.00 ©2018 IEEE

supervised ones. There are many different applications of event extraction namely Retrospective Event Analysis (REA), Event Fast Discovery (EFD) and Future Event Forecast (FEF).

A. *Unsupervised, Supervised and Semi-supervised Approaches*

The techniques can be classified into supervised, unsupervised and semi-supervised approaches based on the use of labeled training data in EMI.

- **Unsupervised approaches:** Event type is identified by event triggers or topic words. Naturally, the presence of one or more these key phrases indicates whether a tweet is event related.
- **Supervised approaches:** Supervised learning is such a machine learning task that inferring a mapping function from labeled training data. Several supervised classification algorithms have proposed for event extraction, including Naïve Bayes, Support Vector Machine (SVM), gradient boosted decision trees, Logistic Regression, and Random Forests, etc.
- **Semi-supervised approaches:** Semi-supervised techniques make use of unlabeled data for training - typically a small amount of labeled data with a large amount of unlabeled data.

TABLE I. EVENT EXTRACTION METHODS FOR SOCIAL MEDIA

Papers	*Event Types*		*EMI Methods*			*Applications*
	Open Domain	*Domain Specific*	*Unsupervised*	*Supervised*	*Semi-Supervised*	
2012, [5]	√		√			REA
2012, [7]	√			√		
2013, [2]		√			√	
2010, [8]		√		√		EFD
2013, [9]		√		√		
2015, [10]	√		√			
2014, [2]		√	√			FEF
2014, [11]		√	√			
2015, [12]		√	√			

B. *Event Message Identification Methods*

The techniques can be classified into supervised, unsupervised and semi-supervised approaches based on the use of labeled training data in Event Message Identification.

C. *Open Domain and Domain Specific Approaches*

Depending on whether the event type targeted is pre-specified or not, the event types can largely group existing approaches into two categories, i.e., open domain and domain specific EE methods.

D. *Different Applications of Event Extraction*

There are many different applications depending on the time of categorization; REA- to focus on retrieval of historical event information, EFD- to detect and alert newly happened events by listening to and monitoring incoming tweets, FEF- to identify mentions of planned events from open source indicators.

The summarization of previous research is shown in Table I.

As this framework is intended to develop for targeted domain in semi-supervised way, the following analysis table is also summarized for those related researches in many perspectives. The classifier means the techniques to identify the message whether they used the machine learning techniques or their own algorithms. The keywords they utilize to search the stream, the analysis of the common use of feature extraction methods, and the topics of the event they considered are also depicted in Table II.

TABLE II. TEXANOMY FOR DOMAIN SPECIFIC REA APPLICATIONS WITH SEMI-SUPERVISED METHODS

Papers	*Classifier*	*Keywords for Crawling*	*Keywords for Knowledge Base*	*Features*	*Event*
2014, [13]	SVM		Civil unrest, Mexico, protect, date information	Spatical and Temporal feature	Civil Unrest
2013, [14]	SVM			Textual Feature	Biomedical Event
2015, [3]	Own seed based approach	Hack, breach, ddos	Hack, breach, ddos	Textual Feature	Computer Security
2017, [15]	Neural Networks			Textual Feature	Drug
2017, [16]	Own topic based self-learning algorithms	help	earthquake, flood, hurricane, volcano eruption, tsunami, land slide, disease epidemic	Textual Feature, Spatical and Temporal feature	Disaster Event

III. PROPOSE SYSTEM ARCHITECTURE

The framework of the event message identification system for educational information is proposed as shown in Figure 1. There are five main components such as Tweets Crawling, Filtering Tweets, Pre-processing, Event Message Identification System and post processing.

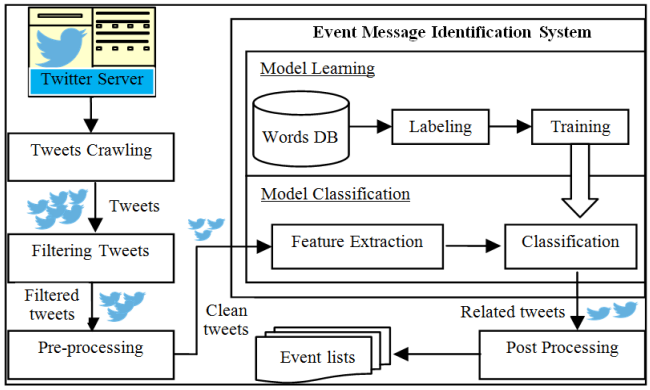

Figure 1. Proposed Framework for Targeted Event Message Identification System

A. Tweet Crawling

This is the process of crawling using Twitter Application Program Interface (API) to retrieve tweets from the server. Existing approaches mainly perform EE on individual tweets or a static set of them. This step can be referred to as Extracting Tweets from Twitter.

B. Filtering Tweets for Location

The tweets extracted from the twitter corpus are extremely large and doesn't concerned with the desired domain. Therefore the tweets are needed to filter to get the related information for education in Yangon. To be specific, our filtering algorithm can be broken down into the application of several filters which we use to continually monitor streaming data from twitter.com as in the following Algorithm.

Algorithm: Filtering Tweets

```
Input  : the tweets from real world twitter corpus within
         a specific time frame
Output: a few dozen posts relevant to target events for
         specific location
t1 = input tweets
 t2 = tweets in t1 whose location is within location of
      interest by obtaining latitude and longitude by
          utilizing the associated GPS location
 t3 =  tweets in t1 whose location is within location of
      interest by obtaining latitude and longitude by
      utilizing geocode parameter for tweet, i.e, the
              registered location of tweet
t4 = tweets in t1 whose text contains mentions of
     specific locations
t5= t2 + t3 + t4
Return t5
```

C. Preprocessing

It is also called the denoising process. It include Removing hash tag and link url, Processing stop word removal, Stemming, Normalizing to lowercase, Removing duplicate tweets and Removing tweets shorter than three words. This process is needed because the time taken is more if they are present in the tweet messages when they are sent directly to machine learning process.

D. Event Message Identification System

It consists of two phase, training phase and classification phase.

1) Training Phase

The training phase include the creation of words Database, Labeling and training processes.

a) Words DB: The first step in learning phase is the creation of bag of words model that uses a dictionary of trigger words to detect and characterize events; these words are manually labeled by experts and decision makers. It consists of words that are directly related to education and words which partially "characterize" to education. It is represent as WDB (i.e, a set of words).

b) Labeling: All the words in the Word DB are labeled as the In EDUCATION_EVENT_RELATED (1).

c) Training: In the training phase, a one-class classification model for identifying documents of class +1 (which are education reporting news), as against any other kind of document (class = -1) is trained. The training model is in the form of a word set W, consisting of words which characterize only the class (+1). The interest is not included in characterizing class -1, and in that sense this is a one-class classification problem. The model W is trained in the form of a small labeled seed set D, where each document in D is labeled with class = +1 (i.e., each document is a known to be related to education), and a small set W0 of known seed words which partially "characterize" class +1 (i.e., are related to education). The model is trained by using multi layer perception.

2) Classification Phase

This phase consists of two main parts, Feature Extraction and classification.

a) Feature Extraction: All the documents from the filtered state are considered for token based features, dictionary based features and N-gram based features.

b) Classification: This is the process of classifying of tweets in an incoming stream as EDUCATION-RELATED (+1) or NOT-EDUCATION-RELATED (-1). To classify a tweet into a positive class or a negative class, this paper use a support vector machine (SVM), which is a widely used machine-learning algorithm. By preparing a training set at the previous state, the framework can produce a model to classify tweets automatically into positive and negative categories.

E. Post Processing

Since, this system is domain specific event extraction, there is no need to categorize events, but need to summarize the information for the targeted event. The goal is to generate the event of the form in 3 tuple: Date-the submitting date of the message, User- the screen name of the post and Message- the original tweets that is related to educational events in Yangon.

IV. EXPERIMENTAL RESULTS

The experiment is built on tweets of targeted event by using the 1000 messages in a specified time frame. The investigation on the different EMI methods is performed. The number of event produced by the unsupervised method is nearly the same as the keyword search method of thousand messages. The baseline evaluation of different machine learning methods for supervised methods is done as shown in Figure 2. Although the results for SVM and Naïve Bayes are the same, the time taken for the Naïve Bayes is larger than the SVM. Therefore this paper chooses the SVM depend on the experiment results. Because of these supervised methods require lots of manual efforts to annotate tweets and take longer time to classify, this system tends to propose semi-supervised manner as mention as above with the help of Words DB.

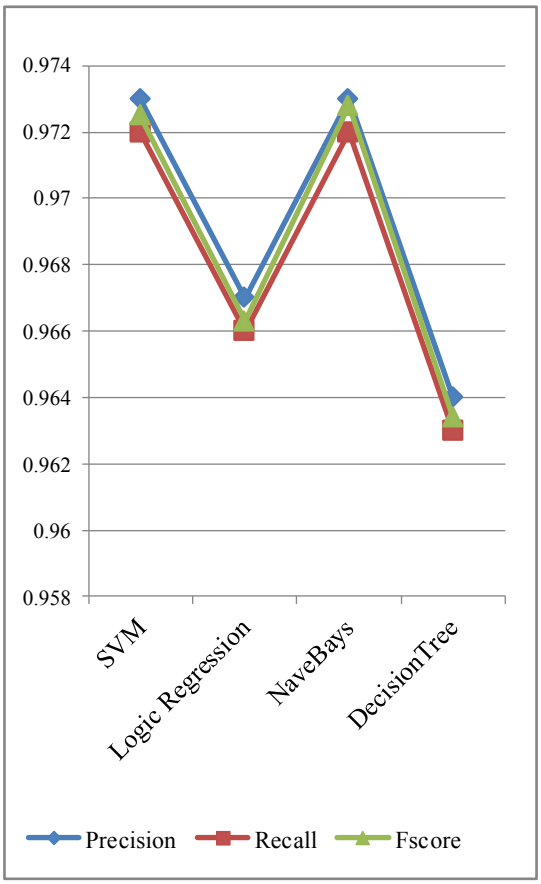

Figure 2. Accuracy Results for baseline supervised methods

V. CONCLUSIONS

Motivated by the wide variety of event extraction which might be of interest to track, the system to find automatically the information about event from Social Media is needed. A number of approaches were investigated to address this challenge and this lead to a novel framework to identify relevant events. This paper has proposed with a semi-supervised SVM-based framework for classification of adverse educational events in tweets for Yangon city. This system could facilitate search for social event and aid users in exploring and discovering social events on a larger scale.

REFERENCES

[1] F.A. Elsafouary, "Monitoring urban traffic management using twitter message", Enschede, The Netherlands, 2013.

[2] R. Compton, C. Lee, J. Xu, L. Artieda-Moncada, T.-C. Lu, L. De Silva, and M. Macy, "Using publicly visible social media to build detailed forecasts of civil unrest," Security Informatics, vol. 3, no. 1, pp. 1–10, 2014.

[3] A. Ritter et. al., "Weakly Supervised Extraction of Computer Security Events from Twitter", International World Wide Web Conference Committee (IW3C2), WWW 2015, May 18–22, 2015, Florence, Italy, ACM 978-1-4503-3469-3/15/05.

[4] T. Hua, F. Chen, L. Zhao, C.-T. Lu, and N. Ramakrishnan, "Sted: semisupervised targeted-interest event detectionin in twitter," in Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2013, pp. 1466–1469.

[5] D. Metzler, C. Cai and E. Hovy, "Structured Event Retrieval over Microblog Archives", 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Canada, pages 646–655K.

[6] J. Deng, F. Qiao, H. Li, X.Zhang and H.Wang, "An Overview of Event Extraction from Twitter", 2015 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery.

[7] 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.

[8] 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.

[9] R. Li, K. H. Lei, R. Khadiwala, and K.-C. Chang, "Tedas: A twitterbased event detection and analysis system," in Data engineering (icde),2012 ieee 28th international conference on. IEEE, 2012, pp. 1273–1276.

[10] Y. Wang, D. Fink, and E. Agichtein, "Seeft: Planned social event discovery and attribute extraction by fusing twitter and web content," in Ninth International AAAI Conference on Web and Social Media, 2015.

[11] N. Ramakrishnan, P. Butler, S. Muthiah, N. Self, R. Khandpur, P. Saraf, W. Wang, J. Cadena, A. Vullikanti, G. Korkmaz et al., "'beating the news' with embers: forecasting civil unrest using open source indicators," in Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2014, pp. 1799–1808.

[12] S. Muthiah, B. Huang, J. Arredondo, D. Mares, L. Getoor, G. Katz, and N. Ramakrishnan, "Planned protest modeling in news and social media," 2015

[13] J. Xu, T.-C. Lu, R. Compton, and D. Allen, "Civil unrest prediction: A tumblr-based exploration," in Social Computing, Behavioral-Cultural Modeling and Prediction. Springer, 2014, pp. 403–411.

[14] J. Wang, Q. Xu, H. Lin, Z. Yang and Y. Li, "Semi-supervised method for biomedical eventbExtraction", IEEE International Conference on Bioinformatics and Biomedicine 2012 Philadelphia, PA, USA. 4-7, 2013.

[15] K. Lee et al., "Adverse Drug Event Detection in Tweets with Semi-Supervised Convolutional Neural Networks", International World Wide Web Conference Committee (IW3C2), WWW 2017, April 3–7, 2017, Perth, Australia, ACM 978-1-4503-4913-0/17/04.

[16] G.K. Palshikar, M. Apte and D. Pandita, "Weakly Supervised Classi_cation of Tweets for Disaster Management", April 2017, TCS Research, Tata Consultancy Services Limited, India.