Semi-supervised Domain Specified Event Extraction from Social Media

San San Nwe, Nang Saing Moon Kham
University of Computer Studies, Yangon, Myanmar
811sansannwe@gmail.com

Abstract

Social media has quickly become popular as an important means that people, organizations use to spread information of divert events for various purposes, ranging from business intelligence to nation security. However, the language used in Twitter is heavily informal, ungrammatical, short and dynamic. Automatically detecting and categorizing events using streamed data is a difficult task, due to the presence of noise and irrelevant information. Therefore, as an emerging research area, event analysis from social media, Twitter has attracted much attention since 2010 and there are many attempts to detect and categorize events from social media. This paper proposes a framework to identify the events from twitter in a semi-supervised manner for targeted domain in specific location with SVM in combination with the corpus. The demonstration shown that, with the selective use of a variety of unlabeled data, the SVM models outperform a strong state-of-the-art supervised classification model. Keywords: Social Media, Twitter, Semi-supervised, Events, SVM.

1. 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 [4].

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 [11].
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 [1].

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 [1]. 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 [3].

Moreover, Yangon city is composed of complex systems with physical, cyber, and social components. Current works on extracting and understanding city events related to education, mainly rely on technology enabled infrastructure to observe and record events. This paper also proposes a semi-supervised extraction framework to automatically detect documents containing information about educational events in Yangon location.

2. Related Works

This section describes an overview of existing approaches proposed for Event Extraction from Twitter. 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-supervised ones. There are many different applications of event extraction namely Retrospective Event Analysis (REA), Event Fast Discovery (EFD) and Future Event Forecast (FEF).

2.1. 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.

2.2. 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.

2.3. 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 happenen 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 below.
Table 1: Event Extraction Methods for Social Media

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.

Table 2: Texanomy for Domain Specific REA Applications with Semi-Supervised Methods

3. Proposed Framework Architecture

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

3.1. 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.

3.2. 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 a specific location  

\[ t_1 = \text{input tweets} \]  
\[ t_2 = \text{tweets in } t_1 \text{ whose location is within location of interest by obtaining latitude and longitude by utilizing the associated GPS location} \]  
\[ t_3 = \text{tweets in } t_2 \text{ whose location is within location of interest by obtaining latitude and longitude by utilizing geocode parameter for tweet, i.e. the registered location of tweet} \]  
\[ t_4 = t_3 \text{ whose text contains mentions of specific locations} \]  
Return \( t_4 \)

Algorithm 1: Filtering Tweets

### 3.3. Preprocessing

It is also called the denoising process. This process is needed because the time taken is more if they are present in the messages when they are sent to machine learning process.

#### 3.4. Event Message Identification System

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

##### 3.4.1. Training Phase

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

##### 3.4.1.1. 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 W (i.e, a set of words).

3.4.1.2. Labeling

All the words in the Word DB are labeled as the In CITY_EVENT_RELATED (1).

3.4.1.3. 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 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.

3.4.2. Classification Phase

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

3.4.2.1. Feature Extraction

All the documents from the filtered state are considered for token based features, dictionary based features and N-gram based features.

3.4.2.2. 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.

3.5. 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.

4. 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 3. Although the results for SVM and NaveBayes are the same, the time taken for the NaveBayes 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.

5. 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.

6. References


