Improving Object and Event Monitoring on Twitter Through Lexical Analysis and User Profiling

Yihong Zhang(✉), Claudia Szabo, and Quan Z. Sheng

School of Computer Science, The University of Adelaide, Adelaide, SA 5005, Australia
{yihong.zhang,claudia.szabo,michael.sheng}@adelaide.edu.au

Abstract. Personal users on Twitter frequently post observations about their immediate environment as part of the 500 million tweets posted everyday. These observations and their implicitly associated time and location data are a valuable source of information for monitoring objects and events, such as earthquake, hailstorm, and shooting incidents. However, given the informal and uncertain expressions used in personal Twitter messages, and the various type of accounts existing on Twitter, capturing personal observations of objects and events is challenging. In contrast to the existing supervised approaches, which require significant efforts for annotating examples, in this paper, we propose an unsupervised approach for filtering personal observations. Our approach employs lexical analysis, user profiling and classification components to significantly improve filtering precision. To identify personal accounts, we define and compute a mean user profile for a dataset and employ distance metrics to evaluate the similarity of the user profiles under analysis to the mean. Our extensive experiments with real Twitter data show that our approach consistently improves filtering precision of personal observations by around 22 %.

Keywords: Twitter · Microblog content classification · User profiling

1 Introduction

Micro-blogging services such as Twitter have become widely used in recent years. Twitter allows its users to create and publish short messages of maximum 140 characters, called *tweets*. Currently, around 284 million active Twitter users generate 500 million tweets every day[1], and around 80 % of Twitter users use their mobile phones to create tweets[1]. The use of mobile platforms for tweeting implies that users can report observed events and objects in their physical vicinity. For example, personal tweets have been used for tracking the movements of earthquakes and typhoons in Japan [12]; tweets about flood, hurricane, and riots have also been used for crime and disaster location [7]. In a previous work, we

[1] https://about.twitter.com/company.

W. Cellary et al. (Eds.): WISE 2016, Part II, LNCS 10042, pp. 19–34, 2016.
DOI: 10.1007/978-3-319-48743-4_2

showed that news generated based on personal observation messages on Twitter can often be hours earlier than the first news appearing in traditional media, even for the most newsworthy events such as shooting incidents [19].

Current work on microblog and short text analysis mostly relies on supervised machine learning methods [2,12], which require the manual preparation of training samples. This has several drawbacks, such as the need for significant manual effort for annotating examples, and a lack of quality guarantees of the classification solutions, when the classifier is applied to a wider pool of tweets beyond its training data. *Unsupervised methods* have the potential to address these issues. In the domain of microblogs and short texts, however, due to the complexity and uncertainty of human user data, works on unsupervised methods are still at an early stage [1,17]. The informal and unstructured text messages used on microblogs creates uncertainty for any kind of classification models, and solutions and models effective in one application often will not be as effective in another application. For example, the work in [8] has shown that the effect of user roles in Twitter rumor classification varies significantly for different rumor instances. Thus a challenge for a specific application using unsupervised methods is to find a particular model that is effective for that application and domain.

In this paper, we focus on filtering personal observations of objects and events on Twitter using an unsupervised method. To address the challenges discussed above and provide high classification accuracy, we advance a novel approach that employs lexical analysis and user profiling. The lexical analysis module filters *observation messages* based on two attributes, part-of-speech (POS) tag and message objectivity. The user profiling module separates *personal accounts* from other types of accounts based on four analyzed attributes, namely, *objectivity*, *interactivity*, *originality*, and *topic focus*. We conduct extensive experiments using real Twitter data collected for a variety of events, with significantly improved results over the existing works. Our main contributions are:

- We propose a novel unsupervised method for filtering personal observations on Twitter. Our method utilizes various natural language processing techniques in lexical analysis and user profiling.
- We propose a novel model for profiling Twitter users based on four dimensions, including *objectivity*, *interactivity*, *originality*, and *topic focus*. We also propose algorithms that can effectively distinguish personal accounts from specific-purpose accounts.
- We test our method extensively with real Twitter datasets. For controlled datasets, our method consistently improves the precision by around 22 %. We obtain even higher improvement for crowd-sourced datasets. Our method also out-performs some of the most effective supervised techniques.

2 Related Work

Twitter as a public media and news source has been studied in several works. Wu et al. [18] investigated the demographics of influential Twitter users, whom they grouped into media, organization, celebrity, and blogger. Their study concludes that bloggers are popular personal accounts that produce the most influential tweets. Kwon et al. [5] investigated supervised methods for identifying rumors on Twitter using a number of prominent features, including propagation peaks, friendship network graph, and linguistic properties based on LIWC (Linguistic Inquiry and Word Count). They found that selecting the right features is critical to classifier accuracy. Sriram et al. [14] similarly studied supervised tweet classification for five types of tweets: news, opinions, deals, events, and private messages. Although various machine learning techniques are compared, they also found selecting the right features is the key factor for classification accuracy.

Despite the fact that messages on Twitter are very often informal and incomplete [3], researches have used Twitter information for disaster location [6], object tracking [12], and event detection [16]. For example, Lingad et al. [7] studied locations mentioned in disaster-related messages in order to identify the position of natural disasters and affected areas. However, accurately classifying object or event-related messages is a challenging task. Sakaki et al. [12] developed a system that tracks the movement of earthquakes and typhoons based on personal reports detected on Twitter. They compared a number of features for building the event-related message classifier, with the best-performing feature set achieving a precision of 63.64 %. Li et al. [6] studied the use of tweets for detecting crime and disaster events (CDE) as they were reported on Twitter. They trained a classifier based on the words present in identified CDE tweets, and achieved a precision of 30 % and a recall of 85 %.

Due to the amount of effort required for manually annotating a large number of messages, a supervised method, however, is in many cases impractical. Unsupervised methods have the advantage of needing less manual effort. Current unsupervised methods for microblog analysis are still at an early stage of research. Carroll et al. [1] developed an unsupervised method for determining the objectivity of in Chinese microblog texts. They defined objectivity as sentiment neutrality, but the application is limited to brand and company reputation analysis. Unankard et al. [16] developed a framework for predicting election results using Twitter messages, in which message and user sentiments are calculated based on positive and negative word counts. Since observation messages are not strongly related to message sentiments, filtering personal observations, however, requires technique beyond sentiment analysis.

3 Filtering Methodology

Our method filters observations of objects and events from personal accounts, by performing the following steps. First, we identify observations from collected tweets for a specific keyword. Second, using also the collected tweets, we distinguish personal accounts from other types of accounts. A personal account is a Twitter account employed for personal use, and is assumed to be free from business or propaganda interests. Our insight is that tweets from personal accounts often contain realtime and localized observations of objects and events. Finally, from the observation tweets identified in the first step, we retain only those made from personal accounts. These personal observations of objects and events have proved useful in previous works for scenarios such as disaster location and rumor detection [5,12].

An overview of our method is shown in Fig. 1. To identify observation tweets, we run lexical analysis on tweet texts based on the par-of-speech (POS) tagging, objectivity analysis, and originality test. To identify personal accounts, we first analyze four attributes for each user, namely, *objectivity*, *interactivity*, *originality*, and *topic focus*. Then we use a clustering algorithm for classifying personal accounts based on the attribute values. We describe our method below.

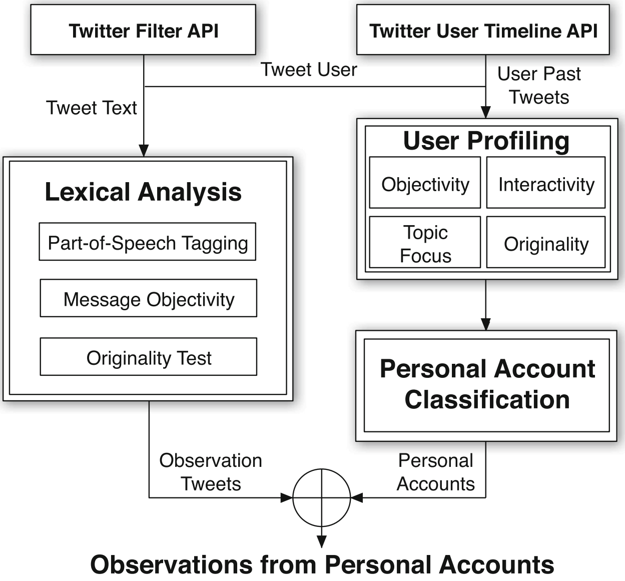

Fig. 1. Method overview

3.1 Observation Filtering

After using the Twitter Filter API[2] to obtain tweets that contain the object or event keyword such as "rainbow" or "car accident", our lexical analysis method focuses on extracting observation tweets. Not all tweets containing the keywords are observations of objects and events, since in some cases the keywords can have another semantic, context-based meaning, and the objects and events can be mentioned in general comments instead of specific observations, e.g., "I dislike car accidents". We address this by utilizing three techniques, namely, *par-of-speech* (POS) tagging, *objectivity analysis*, and *originality test*. POS tagging allows filtering of messages in which the object or event keyword is not used as a subject of observation. Objectivity analysis allows filtering of uncertain messages, such as questions and general comments. Originality test removes messages that are not originally created by the user, such as retweets or quotations.

[2] https://dev.twitter.com/streaming/public.

Filtering Based on Part-of-Speech Tagging. Our insight is that the objects and events mentioned in an observation are most likely to be nouns and gerunds, such as in “I just saw a rainbow”, or “A shooting outside my home”. On the other hand, keywords not used as nouns and gerunds often indicate that the tweet is not a specific observation. Some examples of non-observation tweets are shown in Table 1, with the role of the keyword determined by POS tagging.

Table 1. Non-observation tweets filtered by POS tagging, for monitoring flight delay, shooting incidents, and rainbows

Tweet text	POS
Keep praying for the typhoon to magically **delay** my flight a day	VB
Can we pretend that airplanes, in the night sky, are like **shooting** stars?	JJ
This guy got on a **rainbow** colored LV belt	JJ

VB = base form verb, JJ = adjective.

POS tagging is a technique that matches words in a text with their part-of-speech categories, such as modal, noun, verb, and adverb [13]. We use a filtering rule on top of POS tagging to effectively remove a portion of tweets that are clearly not observations. After performing POS tagging for a tweet, we accept it if the POS tag for the keyword is **NN** (Noun, singular or mass), **NNP** (proper noun, singular), and **VBG** (verb, gerund or present particle). The tweet is rejected if the keyword has other POS tags.

Filtering Based on Objectivity Analysis. Our insight is that a specific observation of an object or event usually is written in a more objective tone than a general tweet. Generally, the objectivity of a message is affected by sentimental words and uncertain words, such as “great”, “bad”, “maybe”, “anyone”. Sentimentality and uncertainty as factors for determining message objectivity has already been proposed in existing works [1,10]. We calculate tweet objectivity based on both sentimentality and uncertainty, using the following formula:

$$\begin{aligned} objectivity(t) = 1 - [senti_p(t) + 0.5 \times senti_n(t)] \\ \times(1 - \sqrt{uncertainty(t)}) \end{aligned}$$

where $senti_p$ is the positive sentiment and $senti_n$ is the negative sentiment. In our previous works, we have found that negative sentiments have a large presence in observation messages [20]. We follow this insight here and we weight down the effect of negative sentiments on reducing the objectivity in the formula. Furthermore, since uncertainty plays an important role in determining the objectivity of a message, as discovered in [10], we increase the effect of uncertainty by scaling it to a larger value.

For sentiment analysis, we employ previously proven effective methods, which employ a positive/negative words dictionary and the slang sentiment dictionary [16]. The positive and negative sentiments of a tweet text t are measured as:

$$senti_p(t) = \frac{count_p(t)}{count_w(t)}$$

$$senti_n(t) = \frac{count_n(t)}{count_w(t)}$$

where $count_p(t)$ and $count_n(t)$ are the word count for positive and negative words in t, and $count_w(t)$ is the word count of t.

For uncertainty analysis, we use a dictionary of uncertain words based on the LIWC category of hesitation words [15]. To measure the uncertainty of tweet t, we consider the number of uncertain words in the text, and whether it is a question.

$$uncertainty(t) = \begin{cases} 0.5, \text{ if } t \text{ ends with a question mark} \\ \frac{count_u(t)}{count_w(t)}, \text{ otherwise} \end{cases}$$

where $count_w(t)$ is the word count for uncertain words in t.

Originality Test. Our analysis of various datasets show that sometimes personal users may repeat some messages created by other users, which do not count as their own observations. The repeated messages not only produce redundancy, but also generate noises for analysis. Thus it is crucial to determine message originality. We proposed a set of rules to determine non-original messages based on message content, as shown in Table 2. A message satisfies any of the rules in the table is considered non-original, and will be filtered out.

Table 2. Originality test rules

Rule	Explanation
Retweet	Contains the word RT
Quotation	Contains quotation marks
Speech	Mention or capitalized word before colon
News title	All words capitalized before link
Repeat	Contains “says”, “claims”, “via”, or “according to”
News mention	Mention contains “news”, “radio”, or “breaking”
News agent	Mention contains news agent name such as “ABC” or “CNN”

Some repeated messages are easy to identify, such as retweets, which have “RT” at the beginning of the messages. Other forms of repeated messages can be more difficult to spot, such as indirect quotes, which often but not necessarily contain the word “says” or “claims”. Given the various ways a message may be repeated, the rules listed in Table 2 do not cover all non-original messages. Nevertheless, we found these rules to filter out most of the repeated messages.

Algorithm 1. Lexical Analysis on Single Tweets

```
INPUT: keyword $w$, tweet set $T$, objectivity threshold $\theta$
OUTPUT: obervation labels $O$
 1: set all $o \in O$ as $false$
 2: for each $t \in T$ do
 3:     run POS tagging for $t$
 4:     if POS tag for $w \in \{\mathbf{NN}, \mathbf{NNP}, \mathbf{VBG}\}$ then
 5:         $pp \leftarrow true$
 6:     end if
 7:     if $objectivity(t) > \theta$ then
 8:         $po \leftarrow true$
 9:     end if
10:     if $t$ fails all rules in Table 2 then
11:         $pt \leftarrow true$
12:     end if
13:     $o_t \leftarrow pp \wedge po \wedge pt$
14: end for
```

Lexical Analysis Algorithm. Algorithm 1 describes our lexical analysis method. The input is a keyword w, and a set T of tweet texts containing the keyword. The output is a set of predictions of whether each tweet text $t \in T$ is an observation, O. In line 7, we use a parameter θ to control the level of objectivity a tweet requires to meet to be considered an observation. The default value for θ is the first quartile of overall objectivity in the tweet set.

3.2 User Profiling for Personal Account Classification

Previous works have shown that news generated from personal observations on Twitter can be much faster than traditional media, and the implicitly-associated location data can be used for localizing the object or the event [12,19]. However, there are many Twitter accounts that are not for personal use, and do not have the same time and location association for their observation messages, and while they add noises to the data collected, it is usually difficult to distinguish them from personal accounts. The main issue is that all accounts on Twitter uses the same format to store data, and usually there is no effective way to judge the type of account other than looking at the content of the account posts directly. These accounts include news, business, activist and advertisement accounts. We call these latter types of accounts *specific-purpose accounts*, and show some well-known examples in Table 3.

Table 3. Examples of specific-purpose accounts

News	@cnnbrk @wsj @foxnews @huffingtonpost @bbcworld @politico
Business	@AdamDenison @GMblogs @MarriottIntl @chicagobulls @Marvel
Activist	@Greenpeace @femmajority @OU_Unheard @freedomtomarry

Our study of personal and specific-purpose accounts leads to the following observations:

- News accounts tweet about various topics in a strictly objective tone. Their tweets usually contain links to Web articles. Depending on the specialty, a media account can cover a wide range of topics.
- Business accounts contain conversations, observations, and product promotions, but the range of topic is limited to the specific business.
- Activist and advertisement accounts rarely use objective tone, and their range of topics is also limited.

A personal account, however, does not have such clear-cut characteristics as specific-purpose accounts, and usually contains a mix of information sharing, conversation with other users, and original content that covers various topics. We propose that:

Conjecture 1. A personal account has moderate levels in objectivity, interactivity, originality, and topic focus.

We use various statistics generated from Twitter data to calculate the levels of *objectivity*, *interactivity*, *originality*, and *topic focus* for Twitter users. Here we assume these user qualities are consistent over time and do not easily change. There are rare cases that the profile of a user changes drastically, for example, caused by a job change, but currently we do not consider such cases. To profile a user, first we collect a set of past tweets made by the user, H. Then we select the original tweets in H based on the rules described in Table 2, as $OH = \{oh_1, oh_2, ..., oh_l\}$, where $|OH| = l$.

The objectivity of a user is calculated based on the objectivity of each tweet in OH:

$$u_{objectivity} = \frac{\sum_{i=1}^{l} objectivity(oh_i)}{l}$$

The interactivity of a user is calculated based on the number of tweets containing mention mark "@" in H:

$$u_{interactivity} = \frac{count_{@}(H)}{|H|}$$

The originality of a user is calculated based on the fraction of original tweets in H.

$$u_{originality} = \frac{l}{|H|}$$

To calculate a user's topic focus, we count the frequency of each topic word for all topic words appearing in OH. For simplicity, we consider a topic word as a word that starts with a capital letter. The first word in a sentence is ignored. Once we have a descendingly-sorted list of topic word occurrences $\{nt_1, nt_2, ..., nt_k\}$, the topic focus of a user is calculated based on the fraction of the first quarter of the most frequent topic words:

$$u_{focus} = \frac{\sum_{i=1}^{n/4} nt_i}{\sum_{j=1}^{n} nt_j}$$

A user is thus profiled by the quadruple:

$$u = \{u_{objectivity}, u_{interactivity}, u_{originality}, u_{focus}\}$$

3.3 Personal Account Classification with Profiles

We propose an algorithm for automatically identifying personal accounts based on the user profile. First we define the difference between two user profiles u_1 and u_2 as the Euclidian distance between two profiles:

$$d(u_1, u_2) = \sqrt{\sum (u_1 - u_2)^2}$$

where

$$\begin{aligned}\sum (u_1 - u_2)^2 = &(u_{objectivity1} - u_{objectivity2})^2 \\ &+(u_{interactivity1} - u_{interactivity2})^2 \\ &+(u_{originality1} - u_{originality2})^2 \\ &+(u_{focus1} - u_{focus2})^2\end{aligned}$$

Following Conjecture 1, we see that the attributes of a personal account are usually closer to a set of mean values while a specific-purpose account usually holds more extreme values. Therefore we propose that:

Conjecture 2. Given a set of user profiles U, which contains personal account profiles P and specific-purpose account profiles S, there exists a mean profile $\bar{u}$, such that $\sum_{p \in P} d(p, \bar{u}) < \sum_{s \in S} d(s, \bar{u})$.

While it is difficult to prove Conjecture 2, we find it generally true in our analysis, as we will show with our experiments. Given a set of user profiles U, and a mean profile $\bar{u}$, we can separate from U a subset C that is more likely to contain personal accounts, by selecting profiles that have shorter distance to $\bar{u}$.

We devise an iterative algorithm for finding the mean profile $\bar{u}$. Intuitively, we can use the mean attribute values of all profiles in U. However, the extreme attribute values of the specific-purpose account profiles can bias the mean significantly, making it inaccurate for deciding personal accounts. In Algorithm 2, we use an iterative approach and a cluster size threshold δ for selecting a cluster of $|U| \times \delta$ profiles that are close to an unbiased $\bar{u}$. Starting from an initial mean profile $\bar{u}_0$, the algorithm alters between cluster updating (line 2, 6) and mean updating (line 4 and 5). In the cluster updating step, a number of profiles close to the mean are selected. In the mean updating step, a new mean is calculated based on the selected profiles. If there are extreme values that cause a bias in the cluster, the mean will move away from the bias, and replace the extreme value profiles with more average profiles in the cluster. The output of the algorithm, F, is a set of personal account predictions.

Algorithm 2. Predicting Personal Accounts

```
INPUT: user profiles $U$, mean profile $\bar{u}_0$, selected cluster size $\delta$
OUTPUT: $F$
set all $f \in F$ as $false$
$C \leftarrow |U| \times \delta$ profiles closest to $\bar{u}_0$
while $C \neq C'$ do
    $C' \leftarrow C$
    $\bar{u} \leftarrow$ mean attribute values of profiles in $C$
    $C \leftarrow |U| \times \delta$ profiles closest to $\bar{u}$
end while
for each $u \in U$ do
    if $u \in C$ then
        $f_u \leftarrow true$
    end if
end for
```

While Algorithm 2 generally finds a good mean profile that separates personal accounts and specific-purpose accounts. However, depending on the choice of the initial mean $\bar{u}_0$, the algorithm sometimes produces undesirable results. To address this issue, we derive a particle swarm optimization (PSO) algorithm for finding the optimal $\bar{u}_0$.

PSO is an optimization technique that takes a population of solutions, and iteratively improves the quality of the solutions by moving them toward the best solution in each iteration [4]. A solution in our PSO algorithm is an initial mean $\bar{u}_0$ to be given to Algorithm 2. A PSO algorithm requires the definition of the quality measure and the solution movement. To define the quality of a solution, we rely on our initial observation that personal accounts exhibit higher variance than any types of specific-purpose accounts. Therefore we propose that:

Conjecture 3. Given two user profile clusters C_1 and C_2, if the profiles in C_1 are more diverse than C_2, than C_1 is more likely to contain personal accounts.

We use pairwise profile differences to calculate the diversity of profiles in a cluster, $C = \{c_1, c_2, ..., c_k\}$,

$$div(C) = \frac{2 \times \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} d(c_i, c_j)}{k \cdot (k-1)}$$

For the solution movement in PSO, we set a moving speed v so in each iteration, a solution p moves towards the best solution p_b as:

$$p \leftarrow p + (p_b - p) \cdot v \tag{1}$$

Our PSO algorithm is shown as Algorithm 3. It starts with a number of random solutions (line 1) and for each solution, a profile cluster is generated using Algorithm 2 (line 2 to 4). Then iteratively, the PSO algorithm moves the best solution towards an optimal solution by comparing the cluster diversity with each solution (line 5 to 13).

Algorithm 3. PSO for Finding Optimal $\bar{u}_0$

```
INPUT: user profiles $U$, selected cluster size $\delta$, number of particles $n$, speed $v$
OUTPUT: $p_b$
 1: randomly choose $n$ solutions $P$ in the solution space
 2: for each $p \in P$ do
 3:     generate a cluster $C_p$ using Algorithm 2
 4: end for
 5: $p_b \leftarrow p$ with highest $div(C_p)$
 6: while $p_b \neq p'_b$ do
 7:     $p'_b \leftarrow p_b$
 8:     for each $p \in P$ do
 9:         $p \leftarrow p + (p_b - p) \cdot v$
10:         generate a cluster $C_p$ using Algorithm 2
11:     end for
12:     $p_b \leftarrow p$ with highest $div(C_p)$
13: end while
```

The optimal initial mean produced by Algorithm 3 can then be used in Algorithm 2 for selecting the cluster of personal account profiles. Although Algorithm 3 requires two more parameters, during our experiments we find the effect of changing n and v negligible for any $n > 1,000$ and $v < 0.2$, as the solution already reaches optimal values. Therefore we can confidently set n and v to fixed values. The only parameter that still affects the classification result is the cluster size parameter δ, which controls the portion of profiles in the data to be selected as personal account profiles.

Algorithm 4. Filter Observations from Personal Accounts

```
INPUT: keyword $w$, messages $M$, objectivity threshold $\theta$, selected cluster size $\delta$
OUTPUT: $R$
set all $r \in R$ as $false$
$T \leftarrow$ tweet text from $M$
$U \leftarrow$ user profiles from $M$
$O \leftarrow$ run Algorithm 1 with $w$, $T$, $\theta$
$p_b \leftarrow$ run Algorithm 3 with $U$, $\delta$
$F \leftarrow$ run Algorithm 2 with $U$, $p_b$, $\delta$
for each $m \in M$ that has text $t$ and user profile $u$ do
    if $o_t \wedge f_u$ then
        $r_m \leftarrow true$
    end if
end for
```

3.4 Overall Algorithm

Algorithm 4 identifies observations from personal accounts. Given the input of a keyword w and a set of tweets M, and the control parameter θ and δ, the output is a set of predictions, R, of whether each respective tweet is an observation of the object or event of interest from personal accounts.

4 Experimental Analysis

We tested the effectiveness of our method for filtering personal observations on Twitter with two real Twitter datasets, comprising of a controlled dataset and a crowd-sourced dataset. In this section, we present the setup, measurement, baseline methods, and results of our experiments in detail.

4.1 Experiment Setup

We implemented the algorithms presented in the previous section in Java. The experiments were run on a MacBook Pro laptop computer, with 2.3 GHz Intel Core i7 CPU and 8 GB 1600 MHz DDR3 memory. We deployed an existing implementation for POS tagging. After comparing several existing POS tagging implementations including OpenNLP and LingPipe, we chose StanfordNLP POS module to run our POS tagging because it is relatively fast and provides a high tagging accuracy of around 95 % [9].

For parameter θ in Algorithm 1, we chose the first quartile of overall objectivity in the dataset for all experiments, which generally provides good results. For parameter δ, we compared three different values, including 0.7, 0.8, and 0.9. To ensure the consistency of the experiments, instead of randomly choosing initial values for the particles in Algorithm 3, we chose combinations of evenly distributed values for the four attributes as the initial values, i.e., 0.2, 0.4, 0.6, 0.8, and 1. Our analysis shows that randomly initialized particles provide similar results. For user profiling, up to 1,000 recent tweets were collected for each user using Twitter Timeline API.

4.2 Baseline Methods and Comparison Metrics

We compared our approach with three baseline filtering strategies, namely, *Accept All*, *Sakaki* filter, and *Sriram*. Accept All takes all tweets in the dataset as the positive for personal observations. Sakaki classifier, proposed by Sakaki et al. in [12], is a supervised method that deploys a Support Vector Machine (SVM) classifier with linear kernel built on manually annotated training data. Among the three feature sets proposed in [12], we implemented the reportedly most effective set, Feature Set A, which is based on word counts and keyword positions. We deployed an existing SVM implementation in an R language package called e1071[3]. We used a weighting function according to class imbalance to ensure optimal performance of the classifier. The performance of the Sakaki classifier was measured using the three-fold cross validation. One drawback of the

[3] https://cran.r-project.org/package=e1071.

Sakaki classifier is that it requires the presence of a keyword. The user profiling in our approach, though, does not have this requirement.

The Sriram classifier, proposed by Sriram et al. in [14], is also a supervised method that is based on eight features and the Naive Bayes model. The eight features include author name, use of slang, time phrase, opinionated words, and word emphasis, presences of currency signs, percentage signs, mention sign at the beginning and the middle of the message. The evaluation is based on the five-fold cross validation. The Sriram classifier is shown to be effective in classifying tweets into categories such as news, opinions, deals and events, but has not been tested in other applications.

All datasets for evaluation were manually annotated according to whether each tweet is a personal observation of an object or event of interest, which were considered ground truth in our experiments. The output of the filtering methods were compared with the manual annotations. If a filtering output is positive in manual annotations, it is considered a *true positive*. We use *precision*, *recall* and $f - value$ as the measurements of filtering accuracy, where given the set of positive filtering results P and the set of true positives in the dataset TP, The $precision = \frac{|P \bigcap TP|}{|P|}$, $recall = \frac{|P \bigcap TP|}{|TP|}$, and f-value $= 2 \cdot \frac{precision \cdot recall}{precision + recall}$.

4.3 Effectiveness on Controlled Datasets

We first tested our method on two controlled datasets. We collected a dataset of around 5,000 tweets containing the keyword *hailstorm* during August, 2015, and a dataset of around 5,000 tweets containing the keyword *car accident* during September, 2015. After removing retweets and tweets containing links, we manually labelled the remaining tweets as positive or negative examples, according to whether the message is about a direct observation of a hailstorm or a car accident. The resulted *hailstorm* dataset contains 675 tweets, with 251 positive examples and 424 negative examples. The labelled *accident* dataset contains 954 tweets, with 347 positive examples and 607 negative examples.

We tested the filtering methods on the two datasets. Accuracy results for the baseline methods, lexical analysis-only filtering (LX), and lexical analysis combined with personal account filtering using three δ values, PA($\delta = 0.9$), PA($\delta = 0.8$), and PA($\delta = 0.5$), are presented in Table 4.

As shown in the table, the Accept All strategy captured all the positives in the annotations and had the maximum recall of 1. All other methods improved the precision by sacrificing the recall to some degree. Personal account filtering with δ set to 0.9 achieved the highest overall performance, indicated by the highest f-value. Using lexical analysis only and PA with $\delta = 0.9$ and $delta = 0.8$ all performed better than the Sakaki classifier and the Sriram classifier, the latter of which provided almost no filtering effect in the hailstorm dataset. Setting δ to a lower value improved the precision but also lowered the recall. When setting δ to 0.5, PA achieved the highest precision, while still held a relatively high f-value. The performance of all methods were consistent across two datasets, with LX improving precision from the Accept All strategy by around 15 % and PA($\delta = 0.9$) further improved it by 5 %–7 %.

Table 4. Filtering accuracy for hailstorm and car accident datasets

	Accept all	Sakaki	Sriram	LX	PA($\delta = 0.9$)	PA($\delta = 0.8$)	PA($\delta = 0.5$)
`Hailstorm dataset`							
Precision	0.37	0.43	0.37	0.53	0.62	0.64	**0.70**
Recall	**1**	0.70	0.98	0.80	0.76	0.71	0.46
f-value	0.54	0.53	0.54	0.63	**0.68**	0.67	0.56
`Car accident dataset`							
Precision	0.38	0.50	0.44	0.53	0.58	0.59	**0.60**
Recall	**1**	0.73	0.84	0.76	0.74	0.69	0.43
f-value	0.55	0.60	0.57	0.63	**0.65**	0.64	0.50

4.4 Effectiveness on Crowd-Sourced Dataset

We tested our approach on a publicly available dataset produced by Castillo et al. [11], and is available online[4]. The dataset contains around 20,000 tweets related to crisis events, such as the Colorado wildfires and the Pablo typhoon in 2012, and the Australia bushfire and New York train crash in 2013. These crisis tweets were manually annotated by hired workers on Crowdflower, a crowdsourcing platform[5]. The tweets were labelled according to their relevance to the crisis event, and the type of information they provide into four categories, namely, *related and informative*, *related but not informative*, *not related*, and *not applicable*. The seven information types include Eyewitness, Government, NGOs, Business, Media, Outsiders, and Not applicable.

We consider that the Eyewitness-type tweets in the dataset are personal observations, while other types of tweets are not. Hence we expect our approach to filter Eyewitness tweets from other tweets. With this goal, we re-organized the dataset. First we selected two categories of related tweets from the dataset. Then we selected five information types of tweets: Eyewitness, Government, NGOs, Business and Media. We then produced a list of labels, with Eyewitness tweets as positives, and other types of tweets negatives. We also removed retweets from the data. Our labelled dataset had 3,646 tweets with 528 positives.

Since the tweets do not contain a specific keyword, we did not run POS and objectivity analysis. The Sakaki classifier is not applicable without a keyword. As such we ran the originality test in the lexical analysis and the personal account classification, and compared only to the Sriram classifier (Table 5).

The results are similar to previous experiments, where PA($\delta = 0.9$) achieved the highest f-value and PA($\delta = 0.5$) achieved the highest precision. The lexical analysis was particularly effective for this dataset, improving the precision by around 38 %, mainly because the dataset includes a large portion of news messages, which failed the originality test. After the lexical analysis, PA($\delta = 0.9$) further improved the precision by 12 %. Both LX and PA($\delta = 0.9$) significantly outperformed the Sriram classifier.

[4] http://crisislex.org/.

[5] http://www.crowdflower.com/.

Table 5. Filtering accuracy for the crisis dataset

	Accept all	Sriram	LX	PA($\delta = 0.9$)	PA($\delta = 0.8$)	PA($\delta = 0.5$)
Precision	0.14	0.32	0.52	0.64	0.64	**0.65**
Recall	**1**	0.52	0.47	0.50	0.48	0.27
f-value	0.24	0.40	0.47	**0.56**	0.55	0.38

5 Conclusion

Personal observations of objects and events published on micro-blogging platforms such as Twitter are an invaluable information source, and can be utilized in applications such as natural disaster tracking and crime monitoring. However, given the various ways users post messages and the large variety of account types, information about a particular object or event is usually noisy and misleading. Thus it is critical to develop a novel approach that filters out noises before the information can be further utilized. Current filtering approaches based on supervised machine learning techniques require large manual efforts and thus are impractical in many scenarios.

In this paper, we propose an unsupervised message filtering approach that consists of a lexical analysis module, which examines the message, and a personal account classification module, which examines the message history of the user and determines if the user account is a personal account. We tested our approach extensively on real Twitter datasets. For the controlled dataset, our method consistently improves the precision by around 22 %, with the lexical analysis module improves it by 15 %, and personal account classification further improves it by 7 %. We see even higher improvement in a crowd-sourced dataset, increasing the precision from 14 % to 65 %. Compared with the Sakaki classifier and the Sriram classifier, our approach was able to achieve more than 10 % higher accuracy. In the future, we will continue to investigate unsupervised methods for further filtering accuracy improvement by incorporating location and name-entity analysis.

References

1. Carroll, T.Z.J.: Unsupervised classification of sentiment and objectivity in Chinese text. In: Third International Joint Conference on Natural Language Processing, p. 304 (2008)
2. Castillo, C., Mendoza, M., Poblete, B.: Information credibility on Twitter. In: Proceedings of the 20th International World Wide Web Conference, pp. 675–684 (2011)
3. Chung, D.S., Nah, S.: Media credibility and journalistic role conceptions: views on citizen and professional journalists among citizen contributors. J. Mass Media Ethics **28**(4), 271–288 (2013)
4. Kennedy, J.: Particle swarm optimization. In: Sammut, C., Webb, G.I. (eds.) Encyclopedia of Machine Learning, pp. 760–766. Springer, Heidelberg (2010)

5. Kwon, S., Cha, M., Jung, K., Chen, W., Wang, Y.: Prominent features of rumor propagation in online social media. In: Proceedings of 13th International Conference on Data Mining, pp. 1103–1108 (2013)
6. Li, R., Lei, K.H., Khadiwala, R., Chang, K.-C.: TEDAS: a Twitter-based event detection and analysis system. In: Proceedings of 28th International Conference on Data Engineering, pp. 1273–1276 (2012)
7. Lingad, J., Karimi, S., Yin, J.: Location extraction from disaster-related microblogs. In: Proceedings of the 22nd International World Wide Web Conference Companion, pp. 1017–1020 (2013)
8. Maddock, J., Starbird, K., Al-Hassani, H., Sandoval, D.E., Orand, M., Mason, R.M.: Characterizing online rumoring behavior using multi-dimensional signatures. In: Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work and Social Computing, pp. 228–241 (2015)
9. Manning, C.D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S.J., McClosky, D.: The stanford CoreNLP natural language processing toolkit. In: Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pp. 55–60 (2014)
10. Mukherjee, S., Weikum, G., Danescu-Niculescu-Mizil, C.: People on drugs: credibility of user statements in health communities. In: Proceedings of the 20th ACM International Conference on Knowledge Discovery and Data Mining, pp. 65–74 (2014)
11. Olteanu, A., Castillo, C., Diaz, F., Vieweg, S.: CrisisLex: a lexicon for collecting and filtering microblogged communications in crises. In: Proceedings of the 8th International AAAI Conference on Weblogs and Social Media, pp. 376–385 (2014)
12. Sakaki, T., Okazaki, M., Matsuo, Y.: Tweet analysis for real-time event detection and earthquake reporting system development. IEEE Trans. Knowl. Data Eng. **25**(4), 919–931 (2013)
13. Santorini, B.: Part-of-speech tagging guidelines for the penn treebank project (3rd revision). Technical report MS-CIS-90-47, University of Pennsylvania Department of Computer and Information Science Technical (1990)
14. Sriram, B., Fuhry, D., Demir, E., Ferhatosmanoglu, H., Demirbas, M.: 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, pp. 841–842 (2010)
15. Tausczik, Y.R., Pennebaker, J.W.: The psychological meaning of words: LIWC and computerized text analysis methods. J. Lang. Soc. Psychol. **29**(1), 24–54 (2010)
16. Unankard, S., Li, X., Sharaf, M., Zhong, J., Li, X.: Predicting elections from social networks based on sub-event detection and sentiment analysis. In: Benatallah, B., Bestavros, A., Manolopoulos, Y., Vakali, A., Zhang, Y. (eds.) WISE 2014. LNCS, vol. 8787, pp. 1–16. Springer, Heidelberg (2014). doi:10.1007/978-3-319-11746-1_1
17. Unankard, S., Li, X., Sharaf, M.A.: Emerging event detection in social networks with location sensitivity. World Wide Web Journal (2015, in press)
18. Wu, S., Hofman, J.M., Mason, W.A., Watts, D.J.: Who says what to whom on Twitter. In: Proceedings of the 20th International World Wide Web Conference, pp. 705–714 (2011)
19. Zhang, Y., Szabo, C., Sheng, Q.Z.: Sense and focus: towards effective location inference and event detection on Twitter. In: The Proceedings of the 16th International Conference on Web Information Systems Engineering (2015)
20. Zhang, Y., Szabo, C., Sheng, Q.Z., Fang, X.S.: Classifying perspectives on Twitter: immediate observation, affection, and speculation. In: The Proceedings of the 16th International Conference on Web Information Systems Engineering (2015)