Efficient User Profiling in Twitter Social Network Using Traditional Classifiers

M.A. Raghuram, K. Akshay and K. Chandrasekaran

Abstract Any discussion in social media can be fruitful if the people involved in the discussion are related to a field. In a similar way to advertise an event, it is useful to find users who are interested in the content of the event. In social networks like Twitter, which contain a large number of users, the categorization of users based on their interests will help this cause. This paper presents an efficient supervised machine learning approach which categorizes Twitter users based on three important features(Tweet-based, User-based and Time-series based) into six interest categories - Politics, Entertainment, Entrepreneurship, Journalism, Science & Technology and Healthcare. We compare the proposed feature set with different traditional classifiers like Support Vector Machines, Naive-Bayes, k-Nearest Neighbours, Decision Tree and Logistic Regression, and obtain upto 89.82% accuracy in classification. We also propose a design for a real-time system for Twitter user profiling along with a prototype implementation.

1 Introduction

Twitter is a popular micro-blogging site that allows millions of users to communicate, stay in touch, establish connections and more. Users via Twitter, can post messages called as "tweets" which are limited to 140 characters containing only text or hyperlinks. The rising popularity of social networking sites like Twitter, Facebook, LinkedIn, Tumblr etc. has produced vast resources of user-generated content. As of March 2015, Twitter reports a monthly usage of 288 million active users with more than 500 million tweets exchanged per day [2].

In this context, the problem of automatically identifying user interests [15] and

M.A. Raghuram(✉) · K. Akshay · K. Chandrasekaran
Department of Computer Science and Engineering, National Institute of Technology Karnataka, Mangalore, India
e-mail: mar.11co54@nitk.edu.in, akshaysaja44@gmail.com, kchnitk@ieee.org

S. Berretti et al. (eds.), *Intelligent Systems Technologies and Applications*,
Advances in Intelligent Systems and Computing 385,
DOI: 10.1007/978-3-319-23258-4_35

user profiling [18] has gained significant attention. Twitter's Streaming API methods provide easy and programmatic access to the vast amount of data generated in the social network [3]. This has made Twitter an active hub for user personality and profile related research. Some of the studies that have been carried out include identifying user's demographic information [17], predicting brand-related events from user's tweets [14] and tweet topic identification [19]. Researches have also been carried in finding out finely-tuned features like predicting the type of Twitter account reporting an event(individual, news organization or other) [11].

Research on Indian Twitter users is rare although Indian users constitute for more than one-third of the Twitter population [5]. In our experiments, we primarily concentrated on tweets from Indian users and obtained the handles of active Twitter users using web directory listing services like Twellow [1]. Since we make use of Tweet-based features to identify the dynamic interest of the user, we assume that the user tweets about topics that he/she is interested in. We also make use of the Weka machine learning library for various feature selection & machine learning tasks and also for implementing the real-time application in Java [4].

In this paper, we predict the user's dynamic interests based on three important characteristics - the user's static profile, tweet content [12] and simple time series features of the user's tweets. We explore the supervised learning model with several classifiers - Support Vector Machines(SVM), Naive-Bayes(NB), Decision Tree(DT), k-Nearest Neighbour(kNN) and Logistic Regression(LR) with different combinations of our proposed features. We explore the impact of principal component analysis(PCA) on our proposed feature set. We also develop a prototype to evaluate our classification scheme based on the suggested features and propose a model for implementing a real-time user profiling application.

The rest of the paper is organized as follows. In the next section, we discuss some of the related works with respect to Twitter user classification and user profiling. In Section III, we present our methodology. Section IV presents the experiments, their results and analysis. In Section V, we conclude the paper along with directions for future work.

2 Related Work

In recent years, several attempts have been made on finding out the preferences of users in the Internet. The researches primarily focused on the Internet activity of the user to determine the user's interests. Also, there are various studies on sentiment analysis and tweet analysis [8, 9] in Twitter social network alone.

De Choudhury et al. [11] categorized Twitter accounts reporting worldwide events into journalists, organizations and ordinary users, based on the analysis of their Twitter time-line. The Twitter data of 1850 users was collected and the supervised machine learning model was explored. Pennacchiotti and Popescu [16] tried to build a machine learning model to determine the political affiliation of a Twitter user. The research was aimed to find whether a user is democratic or republic and also tells whether a user is attracted towards particular brand of business.

Table 1 Distribution of Twitter User Profiles collected

Interest Class	User Instances
Politics	97
Entertainment	142
Entrepreneurship	108
Journalism	72
Science & Technology	75
Healthcare	96
Total Instances	**593**

Table 2 Specific Twitter accounts collected for each class

Interest Class	Twitter account handles
Politics	@narendramodi,@NandanNilekani, @ShashiTharoor, @arunjaitley, @SushmaSwaraj
Entertainment	@iHrithik, @arrahman, @SrBachchan, @iamsrk, @AnilKapoor
Entrepreneurship	@dharmesh, @FareedZakaria, @hnshah, @kiranshaw, @Aishwarya_N
Journalism	@ndtv, @zeenews, @sardesairajdeep, @BDUTT, @timesofindia
Science & Technology	@elakdawalla, @Atul_Gawande, @sanjayguptaCNN, @IndianScience, @pallavbagla
Healthcare	@GEHealthIndia, @drsanjaygupta, @3MHealthcare_In, @fortis_hospital, @INDHEALTHCARE

Some researches focused on identifying the subjects that a user was interested in. Thongsuk et al. [19] used the topic model for identifying businesses of particular interest which match the interests of a user. The Latent Dirichlet Allocation(LDA) algorithm was used to construct the topic model. User-based features were used to identify businesses which might interest the user. Medvet and Bartoli [14] used precompiled topic description and unsupervised machine learning algorithm to detect popular themes, events and associated sentiment polarity. They focused on creating a customized user profile for Twitter users. The Time-series based features were introduced by Yang et al. [21] on a large corpus of Twitter data containing information from political and sports domains. They classify users on periodicity of activities and the patterns in which users express their opinions and also applied their proposed methods to both binary and multi-class classification of sports and political interests of Twitter users. Siswanto et al. [18] used the supervised learning model and lexical features and tried to determine users interest based on bio data and collection of users tweets. They propose two approaches for dynamic interest prediction of a Twitter user- The first method was based on classification of user's tweets using multi-label classification method and the other approach used specific accounts for classification.

A participant-based event summarization approach was proposed in Chakrabarti and Punera [10]. The proposed approach studies Twitter event streams at the participant level. Their work finds the sub-events that are associated with each participant involved in that event. They use a mixture model, combining burstiness and cohesiveness found in tweet properties related to the event which was used to generate event summaries. They collect the event related information from the tweets and re-tweets of the participants involved in the Twitter event and with time generate the

Table 3 Most commonly used words for each class of users

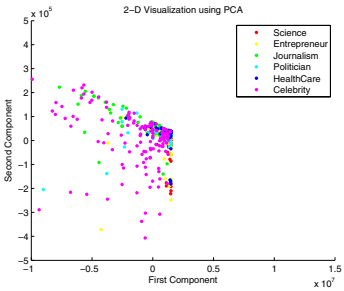
Interest Class	Frequent Terms(Stemmed words)
Politics	'india', 'peopl', 'pm', 'nation', 'wish', 'state', 'govern', 'parti', ,'congress', 'develop'
Entertainment	'love', 'thank', 'like', 'watch', 'great', 'happi', 'show', 'night', 'amaz'
Entrepreneurship	'busi', 'time', 'come', 'help', 'design', 'work', 'need', 'know'
Journalism	'polic', 'year', 'report', 'live', 'kill', 'world', 'video', 'week', 'right', 'court', 'presid', 'offici', 'attack', 'women', 'march'
Science & Technology	'scienc', 'think', 'research', 'data', 'find', 'interest', 'brain', 'post', 'human'
Healthcare	'drug', 'medic', 'care', 'risk', 'heart', 'learn', 'share', 'check', 'prevent', 'futur', 'food'

summaries about the event and sub-events related to it. Kouloumpis et al. [12] evaluated the usefulness of the linguistic features which capture data about the informal languages used by the users in activities like blogging. Sentiment detection based on linguistic studies of messages exchanged over Twitter can also be viewed as a supervised machine learning model.

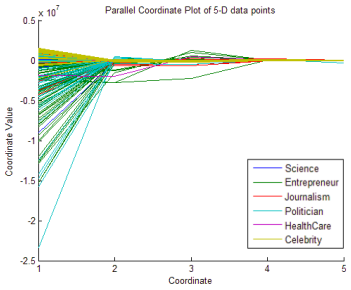
An et al. [7] carried out a novel study on the media and information broadcasting landscape in Twitter. They use the Twitter data and the user accounts followed by people to reveal the relationship between different the classes of media and its diversity in content. Their work aimed at explaining how news related to a particular field spreads in a social network. This study helps to find the behaviour of users who read news and how publishers interact with their readers. They also explain why Twitter users follow multiple news sources.

3 Methodology

The aim of this work is to propose a generic and scalable model for automatically classifying Twitter users based on their dynamic interests given the user attributes and tweet history for a large set of Twitter users. Our approach consists of the following sequence of steps - building a large corpus of Twitter data, extracting proposed features and comparing the performance of traditional classifiers along with the effect of principal component analysis and finally develop a real-time application for Twitter user categorization.



(a) Using PCA to visualize data set in 2D space



(b) Parallel Coordinates Plot of 5-dimensional feature space obtained after dimensionality reduction using PCA

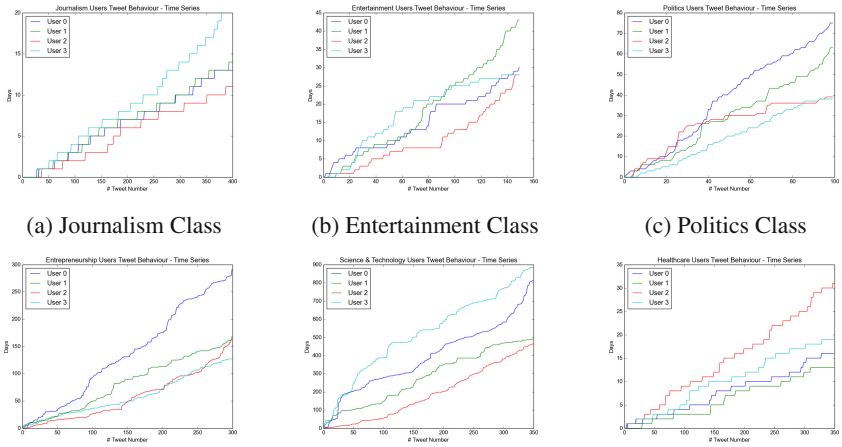
3.1 *Data Collection*

The labelled data set of Twitter users was obtained in a semi-automatic manner wherein we first manually gather Twitter account handles for a specific class over the Internet and then use the Twitter4J Java library for extracting the Twitter feed of the user. We first focused on Indian Twitter handles that were available in the Internet(services like Twellow) and then filled the remaining instance slots with users across the globe so as to balance the data-set. For every Twitter User we collect up to last 500 tweets made by the user and pre-process the tweet data. This includes replacing URL, number with standard texts, removing special symbols/emoticons and stop words and finally performing word stemming using the Porter stemmer.

The Table 1 gives an overview of the data collected for each type of dynamic interest class. Table 2 shows some of the Twitter accounts used for collecting the data. These Twitter handles are manually labelled after inspecting the Twitter account and the tweets posted from the account. Also, only those Twitter handles that could be clearly identified with respect to a single interest class were used. Figure 1a and Figure 1b are obtained after applying principal component method for dimensionality reduction [20]. They are helpful in visualizing our data set. The first component after PCA accounts for more than 99% of the total variance which suggests that most of the data points are along a hyper-plane of higher dimensions. Figure 1a suggests a tendency of close relation among the data points although clustering of individual interest classes is not clearly visible.

3.2 *Feature Selection*

We propose three types of features for efficient classification of Twitter users: User-based features(Static profile), Tweet-based features and simple Time-series based features.



(a) Journalism Class (b) Entertainment Class (c) Politics Class

(d) Entrepreneurship Class (e) Science & Technology Class (f) Healthcare Class

Fig. 2 Tweet Behaviour Plots for each category of users

1. **User-based features**, representing the user's static profile i.e. user information that is very unlikely to undergo rapid changes. This includes user's gender, location, timezone and friends, followers and favourites count. These kind of features are also helpful in spam detection, for example Twitter's spam and abuse policy considers users having small amount of followers compared to the number of people followed by the user to be spam [13]. We consider the reputation score of the user, which is defined by Equation 1,

$$Reputation(user) = \frac{F_o(user)}{F_o(user) + F_r(user)} \tag{1}$$

where F_o and F_r represent the number of followers and friends respectively.

2. **Tweet-based features**, are extracted from the tweets posted by the user. *TF-IDF(Term Frequency - Inverse Document Frequency)* is a commonly used feature in data mining. It is used to provide a weighting method to the usual term frequency of specific dictionary words. Specifically, the weight assigned to each term is determined as shown in Equation 2,

$$Weight(t) = -\log \frac{df(t)}{U} \tag{2}$$

where t represents a term, df(t) is the document frequency - no. of users whose tweet contains the term t, U is the total number of users and weight(t) represents the corresponding weight that is multiplied to the term frequency. Other Tweet-based features include, No. of Hash-tags, No. of replies/mentions, No. of Sensitive tweets and No. of Hyperlinks/URLs etc. In Table 3, we show the most

commonly occurring words in the tweets of the users of each class. Note that these words are the stemmed words obtained using the Porter Stemmer which is used in text normalization.

3. **Time Series features** are used to represent the temporal behaviour of a user's tweets. In our experiments, we use simple statistical features of the user's time series like average, maximum/minimum, standard deviation and derivative of the time series. This is different from the features considered in [21], where class-specific keywords were used to measure the temporal behaviour of different types of users. In Figure 2, we show the tweet behaviour of different types of users. We record the time of tweeting of at-least last hundred tweets and use it to plot the graphs. We can observe that users belonging to a specific interest class tend to have almost similar tweeting patterns. Also, some class of users can be easily characterized with respect to the time series of its users. For example, the users of Journalism class have a step function pattern of tweeting which is due to the periodic news-related information posted by these type of users.

3.3 Classification Methods

We briefly summarize the classification principles of the traditional classification algorithms that were used for training/testing our data set.

1. **Support Vector Machines** is a popular classification technique which tries to determine a large-margin hyperplane that can act as decision boundary. It is therefore, a non-probabilistic linear binary classifier, which performs an implicit mapping from the input to high-dimension feature space for identifying a clear margin. The polynomial kernel(Equation 3) is commonly used for measuring the similarity between feature vectors and computing the cost function.

$$K(x, y) = (\sum_{i=1}^{N} x_i y_i + c)^d \tag{3}$$

2. **Naive Bayes** classifier is one among the many different classifiers that are based on the Bayes Theorem and is useful particularly when the input feature space is of high dimensionality. Given a set of features $X = f_1, f_2, ...f_d$ extracted from a user and a set of classification categories $c_1, c_2, ...c_k$, the Naive Bayes classifier assigns that class c_i that has the maximum posterior probability i.e. $c_i = c_j | P(C_j|X)$ is maximum
3. **Decision Tree** algorithm works on the principle of information gain i.e. at each node of the decision tree, the attribute that splits the data-set effectively is chosen. This process is then repeated on the smaller data-sets obtained by splitting the original data-set. The decision tree classifier can work efficiently with independent features and also even in the presence of outliers but does not scale suitably with large number of features compared to other classifiers like SVM or Naive Bayes.

4. **K-Nearest neighbours** is a simple classification method that does not require training / model fitting. It assumes that points that are close in the feature space are more likely to belong to the same class. Here, K represents the number of nearest neighbours to be considered for classifying the user. The most common mechanism for aggregating the k-points is to use a voting scheme where the class with highest votes is assigned as the predicted class. There are different measures that are used to determine the distance between two points, the most common being the Euclidean distance $D(x, y) = \sqrt{(x - y)^2}$
5. **Logistic regression** is a probabilistic statistical classification scheme which uses probabilistic scores to infer the relation between the classification variable(dependent) and the set of features(independent variables). The logistic function $\sigma(t)$ which is used to determine the probability score, is defined as follows

$$\sigma(t) = \frac{1}{1 + e^{-t}} \tag{4}$$

3.4 Real-Time Twitter User Classification

We propose a simple algorithm for developing a real-time user classification system that periodically updates itself. The system naturally adapts itself to changing tweet behaviour over time, of users of different categories and hence improves in its accuracy of classification. Such a system can also be easily integrated with other applications such as recommender systems to dynamically provide user interests for better performance in the integrated applications.

4 Experiments and Analysis

In all our experiments, we use the 10-fold cross validation for testing and measuring the accuracy of classification. The feature space contains the normalized term frequencies of thousand nine hundred most commonly used words in the English dictionary in their stemmed form, eight User-based features and ten Time-series based features giving a total of 1918 attributes. The Table 4 shows the result of the testing phase with the different combinations of classifier and feature set. The highest accuracy of classification is achieved when the SVM classifier and combining User, Tweet and Time-series based features. The total number of instances are 593 out of which 528 were classified correctly and 65 were classified incorrectly yielding the best accuracy of 89.04% without performing Principal Component Analysis. From Table 5 and Table 6, we can observe that the classifier is having trouble distinguishing between the fields Journalism and Entrepreneur. The True Positive rate for Journalism class is much less compared to other classes which means that actual instances of Journalism class are not being identified correctly and instead are misclassified as Entrepreneur class, thereby increasing the False Positive rate of Entrepreneur class. This type of error occurs mainly because of Twitter users who tweet about business

Algorithm 1. Real-time Twitter User Classification

```
Data: New Twitter User v
Result: Interest Class Label for User v
Model M = Load existing classification model;
/* Loads a set of Twitter users U where each user is
   labelled from a set of classes C and extracted set of
   features F                                          */
if M is old then
    t = 25;
    NewUsers = Read t new users from Twitter Stream;
    for each user u in NewUsers do
        f = ExtractFeatures(u) ;
        label = M.classify(f) ;
        Model M = merge (u,f,label) with model M ;
    end
    Store new classification model M;
end
f = ExtractFeatures(v) ;
label = M.classify(v) ;
return label
```

Table 4 Performance comparison - Testing results of different classifiers with different feature set

Classifier	Tweet-based features	User,Tweet-based features	User,Tweet & Time-series based features
Support Vector Machine	81.80 %	84.68 %	89.04 %
Naive Bayes	70.71 %	76.50 %	84.14 %
J48 Decision Tree	61.11 %	65.87 %	63.91 %
K-Nearest Neighbours	36.90 %	36.90 %	64.41 %
Logistic Regression	79.16 %	81.12 %	83.98 %

related news which can be similar to tweets by Entrepreneurs(For example, users emarketer, theeconomist are examples of biz news accounts).

Another important observation that can be made from Table 5 is with respect to the Entertainment class which has the perfect True Positive rate and Recall which means that any user who belongs to the Entertainment class will always be identified. The Entrepreneurship class has the lowest False Positive rate which is because it is the class that is most commonly misclassified into. This suggests that users interested in business are also likely to be interested in other topics as well.

Table 5 Detailed classification accuracy - SVM classifier + Tweet,User & Time series features

Class	TP Rate	FP Rate	Precision	Recall	F-Measure
Politics	0.969	0.002	0.989	0.969	0.979
Entertainment	1.000	0.036	0.901	1.000	0.948
Entrepreneurship	0.880	0.052	0.792	0.880	0.833
Journalism	0.681	0.013	0.875	0.681	0.766
Science & Technology	0.787	0.014	0.894	0.787	0.837
Healthcare	0.896	0.018	0.905	0.896	0.901

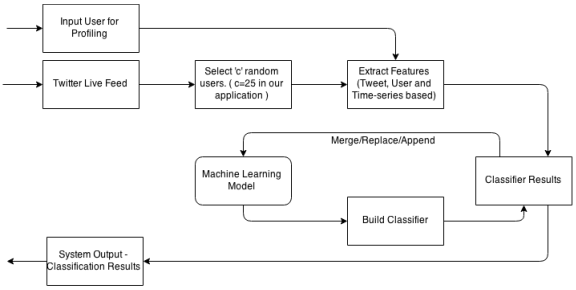
Table 6 Confusion Matrix for SVM classifier + Tweet,User & Time series features

Class	A	B	C	D	E	F
A - Politics	94	3	0	0	0	0
B - Entertainment	0	145	0	0	0	0
C - Entrepreneurship	0	6	95	5	1	1
D - Journalism	1	3	14	49	3	2
E - Science & Technology	0	3	6	1	59	6
F - Healthcare	0	1	5	1	3	86

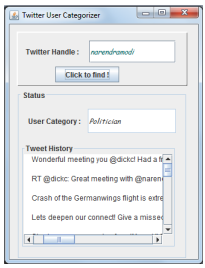
Table 7 Classification accuracy for different classifiers with different feature set size obtained after using PCA

Classifier	500 features	250 features	100 features	50 features
Support Vector Machine	51.09%	78.58%	86.68%	89.71%
Naive Bayes	84.65%	84.65%	84.65%	84.82%
J48 Decision Tree	74.04%	74.87%	75.37%	76.89%
K-Nearest Neighbours	31.70%	26.81%	53.45%	77.23%
Logistic Regression	87.18%	87.68%	88.19%	88.87%

Principal Component Analysis(PCA) is a common technique in machine learning that is used to reduce the dimensionality of the feature space(can also be used to increase the dimension but is rarely done). It is based on the principle of maximum variability along each component of the new set of dimensions. The Table 7 shows the classification accuracy with different classifiers with decreasing feature space dimension obtained using the PCA tool in Weka. The results show that the highest result for the classification task is achieved by reducing the 1918 features to 50 features and using the SVM classifier. Also, general the trend across the table is increasing accuracy of classification with shrinking feature dimensions.



(a) Design of Real-time Twitter User Classification System



(b) Real-time Twitter User Classification Application

4.1 Real-Time Application

The Twitter user profiling application can be made more accurate and scalable by deploying it as a real-time system. We propose a simple design for a real-time Twitter user classification system along with a prototype implementation using the Weka library for machine learning in Java and python scripts for feature extraction. The application periodically collects information about a fixed number of random users from the Twitter streaming API which provides a live feed of tweets. These set of users are classified using an existing machine learning model and then these users are incorporated into the model either by replacement, addition or any other insertion scheme(We have implemented the addition scheme in our application). At any point of time, the request for user classification may be received and the machine learning model available at that time is used to classify the user. The source code and data set used for developing the application are available online [6].

5 Conclusion and Future Work

In this study we present an efficient method of categorizing Twitter users based on their interests using Tweet-based, User-based and Time-series based features. We also compared several approaches for improving the performance of the classification task along with the effect of Principal Component Analysis on our feature space. The best approach categorizes Twitter users into six interest categories namely Politics, Journalism, Entrepreneurship, Entertainment, Science & Technology and Healthcare with 89.82% accuracy. We also propose an algorithm for developing a real-time Twitter user classification system. The application we developed replaces 25 new sample users in the old model periodically and can easily scale to cater to large number of users. For future work, we plan to propose a new framework that uses category-specific keywords and also incorporates the user's social network to improve the accuracy of classification. The Multi-label approach to user interest classification is also an area that can be focused to improve performance. We also plan to increase

the number of interest classes to cover a wide range of dynamic user profiles without degrading the performance of the classification task.

References

1. Twellow. https://www.twellow.com/splash/ (accessed March 10, 2015)
2. Twitter. https://about.twitter.com/company (accessed March 10, 2015)
3. Twitter Streaming APIs. https://dev.twitter.com/streaming/overview (accessed March 10, 2015)
4. Weka 3: Data Mining Software in Java. http://www.cs.waikato.ac.nz/ml/weka/ (accessed March 16, 2015)
5. India to have third-largest Twitter population by 2014: eMarketer (2014). http://indianexpress.com/article/india/politics/india-to-have-third-largest-twitter-population-by-2014-emarketer (accessed March 10, 2015)
6. Github - Twitter User Categorization (2015). https://github.com/AKSHAYH/twitterusercategorization (accessed March 19, 2015)
7. An, J., Cha, M., Gummadi, P.K., Crowcroft, J.: Media landscape in twitter: a world of new conventions and political diversity. In: ICWSM (2011)
8. Bifet, A., Frank, E.: Sentiment knowledge discovery in twitter streaming data. In: Pfahringer, B., Holmes, G., Hoffmann, A. (eds.) DS 2010. LNCS, vol. 6332, pp. 1–15. Springer, Heidelberg (2010)
9. Bollen, J., Mao, H., Zeng, X.: Twitter mood predicts the stock market. Journal of Computational Science **2**(1), 1–8 (2011)
10. Chakrabarti, D., Punera, K.: Event summarization using tweets. In: ICWSM 2011, pp. 66–73 (2011)
11. De Choudhury, M., Diakopoulos, N., Naaman, M.: Unfolding the event landscape on twitter: classification and exploration of user categories. In: Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work, pp. 241–244. ACM (2012)
12. Kouloumpis, E., Wilson, T., Moore, J.: Twitter sentiment analysis: the good the bad and the omg! In: ICWSM 2011, pp. 538–541 (2011)
13. McCord, M., Chuah, M.: Spam detection on twitter using traditional classifiers. In: Calero, J.M.A., Yang, L.T., Mármol, F.G., García Villalba, L.J., Li, A.X., Wang, Y. (eds.) ATC 2011. LNCS, vol. 6906, pp. 175–186. Springer, Heidelberg (2011)
14. Medvet, E., Bartoli, A.: Brand-related events detection, classification and summarization on twitter. In: 2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology (WI-IAT), vol. 1, pp. 297–302. IEEE (2012)
15. Michelson, M., Macskassy, S.A.: Discovering users' topics of interest on twitter: a first look. In: Proceedings of the Fourth Workshop on Analytics for Noisy Unstructured Text Data, pp. 73–80. ACM (2010)
16. Pennacchiotti, M., Popescu, A.-M.: Democrats, republicans and starbucks afficionados: user classification in twitter. In: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 430–438. ACM (2011)
17. Siswanto, E., Khodra, M.L.: Predicting latent attributes of twitter user by employing lexical features. In: 2013 International Conference on Information Technology and Electrical Engineering (ICITEE), pp. 176–180. IEEE (2013)

18. Siswanto, E., Khodra, M.L., Dewi, E., Joni, L.: Prediction of interest for dynamic profile of twitter user. In: 2014 International Conference of Advanced Informatics: Concept, Theory and Application (ICAICTA), pp. 266–271. IEEE (2014)
19. Thongsuk, C., Haruechaiyasak, C., Saelee, S.: Multi-classification of business types on twitter based on topic model. In: 2011 8th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON), pp. 508–511. IEEE (2011)
20. Van der Maaten, L.J.P., Postma, E.O., van den Herik, H.J.: Matlab toolbox for dimensionality reduction. MICC, Maastricht University (2007)
21. Yang, T., Lee, D., Yan, S.: Steeler nation, 12th man, and boo birds: classifying twitter user interests using time series. In: 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pp. 684–691. IEEE (2013)