Social Media Corporate User Identification Using Text Classification

Zhishen Yang[1], Jacek Wołkowicz[1,2], and Vlado Kešelj[1]

[1] Faculty of Computer Science
Dalhousie University, Halifax, Canada
{zyang,jacek,vlado}@cs.dal.ca
[2] LeadSift Inc.
Halifax, Canada
jacek@leadsift.com

Abstract. This paper proposes a text classification method for identifying corporate social media users. With the explosion of social media content, it is imperative to have user identification tools to classify personal accounts from corporate ones. In this paper, we use text data from Twitter to demonstrate an efficient corporate user identification method. This method uses text classification with simple but robust processing. Our experiment results show that our method is lightweight, efficient and accurate.

Keywords: Social media, Twitter, Market analysis, Text classification.

1 Introduction

Everyday, users produce vast amounts of information through various social media platforms. This makes social media a great source for data mining [5]. Social media platforms give users a place to express their attitudes and opinions. Collection of personal data is significant for companies who want to increase their products' reputation in a short time. This data is also essential for on-line marketing analysis. Kazushi et al. [4] expressed that comparing to traditional marketing and reputation analysis approaches, marketing and reputation analysis technologies using social media data are low cost, real time and high volume.

There are however several challenges with collecting data from social media platforms for marketing purposes. First, one should always avoid collecting data from users who can not be categorized as potential or target customers. Second, one should decide which kind of data is important in order to realize our marketing purposes.

With the rapid growth of social media business promotion, many corporations have social media accounts which promote their products and services. These accounts are not potential targets for social media e-commerce systems, because these accounts do not represent customers. In order to filter out these corporate users for marketing analysis, one needs a method of distinguishing these users

M. Sokolova and P. van Beek (Eds.): Canadian AI 2014, LNAI 8436, pp. 363–368, 2014.

from personal users. In this paper, we propose a text-based method to identify social media corporate users. This method uses text classification and small amount of data from social media users' profiles.

In our method, we use Twitter as our data source. Twitter is a major social media platform with over six hundred million active users. These corporate or personal users produce numerous tweets (micro-blog entries) per day, which makes Twitter a good data mining resource [5]. Meanwhile, there are no obvious methods for distinguishing between corporate and personal Twitter profiles. This is not the case for some other social media platforms, such as Facebook.

The rest of this paper is organized as follows. Section 2 outlines related work. Section 3 describes proposed method for identifying social media corporate users. Section 4 contains presentation and analysis of experiment results. Section 5 concludes this paper.

2 Related Work

There are many author profiling methods that relate to the problem we approach in this paper. Estival et al. [3] estimated author's age, gender, nationality, education level, and native language from Arabic email collection by using text attribution tool. Argamon et al. [1] introduced a text-based classification method that can estimate author's age, gender, native language, and personality from their blogs and essays. The restriction of these methods is that they use blog, email, and essays as information sources. In our circumstances, information is retrieved from social media platform which is more dynamic and inconstant. Ikeda et al.[4] proposed a text-based and community-based method for extracting demographic from Twitter users. They used their method to estimate user's gender, age, area, occupation, hobby, and marital status. Their evaluation based on a large dataset demonstrates their method has good accuracy and even worked for users who only tweet infrequently. In particular, this method is not good for building a light and efficient solution to our problem. The reason for this is that this method needs large scale data not only retrieved from users but their friends. In the case of Twitter, we want to build a solution which only requires what is readily available. Although these previous studies have high accuracy in dealing author profiling problem by extracting demographic data, but acquiring demographic information is not crucial in identifying corporate users.

There are many text classification technologies that can be applied to our method, such as Decision Trees, Support Vector Machines (SVM), Neural Networks, K-nearest Neighbours and Naïve Bayes. Bobicev et al. [2] compared performance of these text classification technologies on many Twitter data sets, they showed that Naïve Bayes performs more efficiently and nearly as effective as SVM.

3 Method

The method for social media corporate user identification consists of a training phase and a prediction phase. In the training phase, we use unsupervised text classification which produces two classifiers: name classifier and biography classifier. In the prediction phase, we use two classifiers which are produced in the training phase to score users. User's score is the criteria for identifying whether the user is corporate user.

3.1 Training Phase

Data Selection. The purpose of this step is to prepare data for unsupervised text classification.

In Twitter, every user has an unique user name and biography (brief profile description). Twitter name is a personal identifier. Normally, personal users would use personal names or nicknames, while corporate users would use corporate names or business names. A user's biography is the short text description of an individual Twitter user. Most corporate users will use words describing their business in their Twitter biography, while most individual users will talk about their personal preferences or social status.

We decided to use only name and biography as input data for our method. Retrieving more complex information becomes infeasible in real-time applications. Name and biography is also sufficient for humans to determine whether they are dealing with a corporate account or not.

Feature Extraction. The purpose of this step is to build a feature extractor model to extract features from training data.

We applied bag of terms approach to extract both *unigrams* and *bigrams* from input data sets to generate feature sets. This model simplifies feature extraction by disregarding grammar and word order in a text, apart from their immediate vicinity.

Example:

String: Social media analysis
Unigrams: {Social, media, analysis}
Bigrams: {(Social, media), (media, analysis)}

Feature Selection. The goal of this step is to eliminate low information features. We applied feature selection on both name and biography feature sets.

We used chi-square as metric to filter out low information features to improve performance of text classification. Chi-square finds how common a *unigram* or a *bigram* is in positive training data, compared to how common it occurs in negative data.

NLTK (Natural Language Toolkit) is the tool we used in the following steps. NLTK has libraries to support text classification including chi-square.

1. We calculated frequency of each *unigram* (Overall frequency and its frequency in both positive and negative feature sets) using *FreqDist* method from NLTK. *ConditionalFreqDist* method is employed for frequency in positive and negative feature sets. These frequencies are input for next step.
2. We used *BigramAssocMeasures.chisq* method for *unigram* features and *bigram_finder.nbest* method for *bigram* features to find top N important features. In the experiments section, we show how to find number N and how this step improves the accuracy of text classification.

Unsupervised Text Classification. The final step is text classification. We used Naïve Bayes classifier due to its efficiency and performance. NLTK already provides Naïve Bayes classification function and we directly used it in this step. As mentioned earlier, we created two separate classifiers for name and biography fields.

$$Name\ Classifier = NaiveBayesClassifier.train(Name\ Training\ set)$$
$$Biography\ Classifier = NaiveBayesClassifier.train(Biography\ Training\ set)$$

3.2 Prediction Phase

The purpose of prediction phase is to generate the score of a test user based on Twitter name and biography. The score indicates how likely a given user is a corporate user. We used the same feature extractor from training phase to generate test user's feature sets without any additional feature selection.

Passing name feature set to name classifier calculates the likelihood that this user is a corporate user from Twitter name aspect. Similarly, passing biography feature set to biography classifier calculates the likelihood that this user is a corporate user from Twitter biography aspect. Then we weight those two values to give out the final user score. In the next section, we show how to weight these two values and the criterion, i.e. threshold, of judging whether a user is a corporate user.

4 Experiments

4.1 Data-Sets

To train our method, we downloaded 136 corporate users' profiles and 208 personal users' profiles. For test purposes, we collected a random and representative sample of 116 corporate and 316 personal profiles. We labelled our test data manually.

4.2 Parameters

We defined the following formula to weight name score and biography score to generate final score:

$$score_d = \alpha \times namescore_d + (1 - \alpha) \times bioscore_d \tag{1}$$

Where d is the user profile, $namescore_d$ denotes name score which is the probability produced from name classifier, $bioscore_d$ denotes biography score which is the probability produced from biography classifier and $\alpha \in [0, 1]$ denotes balance parameter to weight name score and biography score.

Next, we identified corporate users. We used $t \in [0, 1]$ as the threshold. If $score_d$ calculated from user using Formula 1 is greater than t then system identifies this user as a corporate user, otherwise as a personal user:

$$result_d = \begin{cases} corporate & \text{if} \quad score_d >= t \\ personal & \text{otherwise} \end{cases} \tag{2}$$

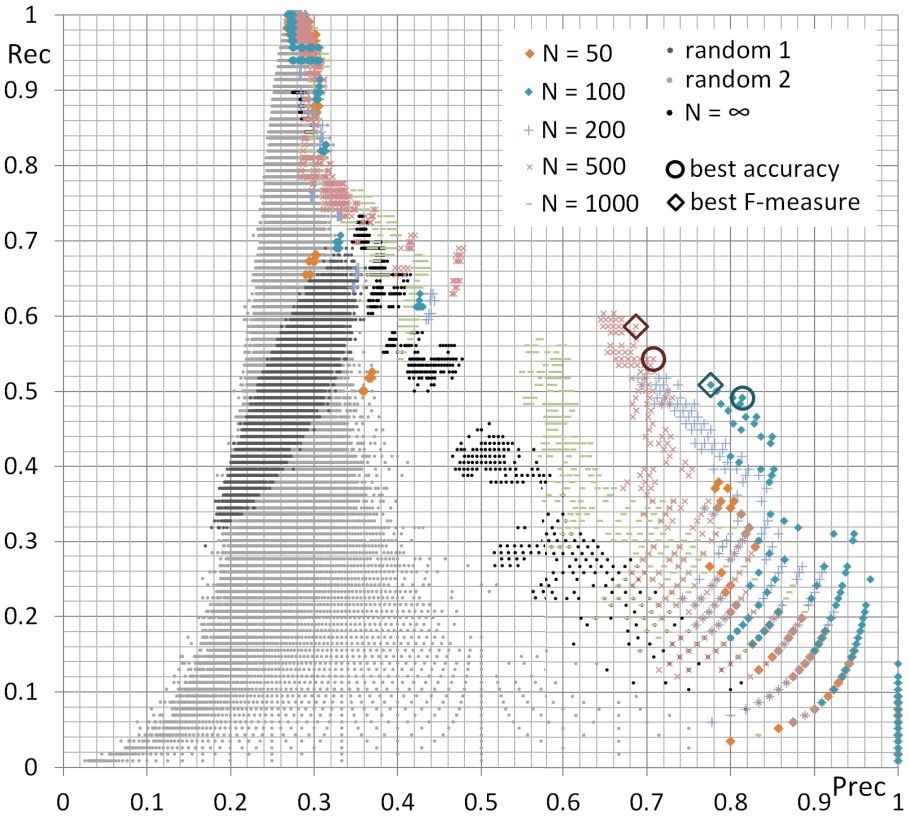

Fig. 1. Precision (P) — Recall (R) graph for our method for various N features classifiers comparing to random classifiers. Each cloud of points represent a system setup with varying α and t parameters, hence different precision and recall values. We indicated points of highest accuracy and F-measure in the graph for 100 and 500 features classifier.

4.3 Experiment Result

We built confusion matrix for all *alpha*, t values with step of 0.01 and various N number of features. Having that, we calculated accuracy, precision, recall and F-measure performance. The highest accuracy of 83% was achieved for $\alpha = 0.27$, $t = 0.58$ and $N = 100$. The highest F-measure (0.64) was obtained for $\alpha = 0.57$, $t = 0.53$ and $N = 500$.

We compared the system to two random classifiers, the first one (`random 1`) was returning a random result from the set $\{corporate, personal\}$ with even probability. The second one (`random 2`) was returning a random score given Formula 1 where $namescore_d$ and $bioscore_d$ were drawn randomly. We also used the full featureset ($N = \infty$) as our baseline. Figure 1 shows precision-recall graph indicating all N features classifiers of the system comparing to those two random classifiers. Clearly, `random 2` much closer resembles the actual behaviour of our system, indicating as well, that even our baseline system outperforms both random classifiers. It also shows that by applying feature selection in our method, accuracy improves nearly 6% and F-measure performance improves nearly 13% compare to full features classifiers. Our method doesn't achieve much better results for high recall values, but if one wants to retrieve just a small sample from a bigger data set (low recall requirements), our model can do it with high precision, which is relevant to the search engine type of scenarios.

5 Conclusion

In this research paper, we presented a method for social media corporate user identification that is lightweight, robust and efficient. We combined analysis of user name and description showing that both are important to achieve good results. From our experiment result, we show that feature selection improved performances of our method and this method has good performance while using small volume of unbalanced training data.

References

1. Argamon, S., Koppel, M., Pennebaker, J., Schler, J.: Automatically profiling the author of an anonymous text. Communications of the ACM 52(2), 119–123 (2009)
2. Bobicev, V., Sokolova, M., Jafer, Y., Schramm, D.: Learning sentiments from tweets with personal health information. In: Kosseim, L., Inkpen, D. (eds.) Canadian AI 2012. LNCS, vol. 7310, pp. 37–48. Springer, Heidelberg (2012)
3. Estival, D., Gaustad, T., Pham, S., Radford, W., Hutchinson, B.: Tat: an author profiling tool with application to arabic emails. In: Proceedings of the Australasian Language Technology Workshop, pp. 21–30 (2007)
4. Ikeda, K., Hattori, G., Ono, C., Asoh, H., Higashino, T.: Twitter user profiling based on text and community mining for market analysis. Knowledge-Based Systems 51, 35–47 (2013)
5. Khan, F., Bashir, S., Qamar, U.: Tom: Twitter opinion mining framework using hybrid classification scheme. Decision Support Systems (2013)