A Study on Microblog Classification Based on Information Publicness

Hongguang Zheng† Nobuhiro Kaji‡ Naoki Yoshinaga‡ and Masashi Toyoda‡

†Graduate School of Information Science and Technology, University of Tokyo

7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan

‡Institute of Industrial Science, University of Tokyo 4-6-1 Komaba, Meguro-ku, Tokyo 153-8505

E-mail: †tei@tkl.iis.u-tokyo.ac.jp, ‡{kaji, ynaga, toyoda}@ tkl.iis.u-tokyo.ac.jp

Abstract Microblog, especially Twitter today has become an important tool to propagate public information among Internet users. The content of Twitter is an extraordinarily large number of small textual messages, posted by millions of users, at random or in response to perceived events or situations. However, messages of Twitter (tweets) cover so many categories including news, spam and others that it's difficult to provide public information directly. Since the traditional search cannot meet demands of tweets of some category, we aim to classify tweets automatically into defined categories to help users search. In our paper, we focus on approaches of collecting a corpus automatically for training classifiers. We proposed two approaches that are based on typical Twitter user accounts and based on Twitter lists using label propagation respectively. Using the corpora, we built classifiers, which are able to determine news, commercial and private tweets. Experiments evaluations show our proposed techniques are effective. In our search, we worked with Japanese, but the proposed approaches can be used with any other language.

Keyword Twitter, Classification, Label propagation, Information publicness

1 Introduction

Microblogging is a broadcast medium in the form of blogging. Twitter, the most popular microblog, differs from a traditional blog in that its content is typically smaller in text size. What's more important is that Twitter exchange and share messages(tweets) in real time among Internet users. This makes it an ideal environment for the dissemination of breaking-news directly from the news source and/or geographical location of events. Some of them contain important information that is valuable for public. They include disastrous events that public are concern about such as storms, fires, traffic jams, riots, heavy rainfall, and earthquakes. They also include some big social events such as parties, baseball games, and presidential campaigns. However, tweets are very messy even on the same one subject. There is a research on distribution of tweets published in 2009. The research analyzed 2,000 tweets (originating from the US and in English) over a two-week period in 2009,8 and separated them into six categories. They found that news and commercial tweets constitute about 15% in all the tweets while private tweets (conversation and babble) constitute 78%. If you search "earthquake", that might be earthquake alarms, damages it caused somewhere or people's local reconstruction activities after earthquakes. It's hard to find real time earthquake information or activities performed locally and so on.

Therefore, we propose a new concept in our work: Information Publicness. Publicness means openness or exposure to the notice or knowledge of the community or of people at large. According to the content, we can divide tweets into two parts: tweets with publicness and tweets without publicness. In the research, we name tweets without publicness "Private tweets". And we subdivided tweets with publicness into two parts: tweets for profit and tweets for non-profit. We call them Commercial and News tweets respectively.

In our paper, we set a task that is how to classify tweets based on information publicness: news, commercial messages and private tweets.

To this end, our paper makes two main contributions:

- We proposed three categories based on information publicness. We introduced two approaches for collecting a corpus used to train a classifier. One is based on typical Twitter user accounts, while the other is based on Twitter lists using label propagation respectively.

- By using the corpora, we extracted text features and some distinctive features of Twitter. And we succeed to build effective classifiers.

The remainder of the paper proceeds as follows. In the next section, we review related work. In section 3 we discuss our approaches to form datasets. Our fundamental policy is that we first collect typical users belonging to each category and then crawl tweets from them. In section 4 we describe how to train classifiers, including section 4.4 in which we show comparison results on our two corpora. Finally, in section 6 we conclude with a brief discussion of future work.

2 Task setting

In our paper, we set a task that is how to classify tweets based on information publicness: news, commercial messages and private tweets.

- News – news category contains news, notices, reports and information for public.
  - ➢ [内房線] 内房線は、強風の影響で、遅れと運休がでています。
  - ➢ 子ども手当所得制限「８６０万円以上」　民主が検討
- Commercial – commercial category contains propaganda for products, services and others including spam messages, which only aim some particular crowd of people.
  - ➢ ワンデーアキュビューモイスト超激安！！ http://bit.ly/9ykD1v
  - ➢ ＝お得なクーポン♪＝究極のウコンが登場！！
- Private – private category contains individual knowledge, experiences and opinions, which are supposed to be shared with surrounding people.
  - ➢ さっきの地震の後の地震雲。龍ですね
  - ➢ iPadの素晴らしさは、いつでもどこでもコンピュータなこと

Although three categories may not be able to cover all kinds of tweets, in our research we only focus on the three ones considering the feasibility.

To build an effective classifier, a large-scale corpus is essential. In our research, how to collect a large-scale corpus automatically is another task. And we will introduce two approaches to perform this task.

3 Related work

Although many researchers have studied document classification. they classified documents according to topics or sentiment. Recent years, tweets classification has become a popular topic due to the popularity of Twitter. Irani in [2] proposed a machine learning method to automatically identify trend-stuffing in tweets, using texts and links of tweets. Pak in [3] showed how to automatically collect a corpus for sentiment analysis and opinion mining purposes, and build a sentiment classifier, that is able to determine positive, negative and neutral sentiments for a document. Some existing works on classification of short text messages integrate messages with meta-information from Wikipedia and WordNet [4] [Hu. X., 2009]. Sakaki in [Takeshi Sakaki, 2010] showed how to detect real time events by machine learning methods. And Sankaranarayanan [Sankaranarayanan. J, 2009] introduced TweetStand to classify tweets as news and non-news. Although Sriram in [8] built a model to classify tweets to classes such as News, Events, Opinion, Deals and private messages, the dataset they used was not scalable which were labeled by human, and distinctive features of Twitter were not exploited.

4 Corpus collection

Using Twitter API[1] we collected two corpora of text posts in Japanese and we form two datasets of three classes: news, commercial, and private. To collect these three kinds of text posts on a large scale automatically, we proposed two approaches which are based on typical Twitter user account and based on Twitter list using label propagation respectively. As we emphasized in Section 1, our strategy is to collect users belonging to each categories first, and then crawl tweets from these users to form large-scale datasets. For instance, tweets of a news typical user will be regarded as news tweets and we will crawl them.

4.1 Corpus based on typical user accounts

Typical users are defined as users who post texts mostly belonging to the same one category. Such as @asashi, an official account of the Asashi Shimbun, is regarded as a

[1] https://dev.twitter.com/

news typical user for most of tweets it posts belong to the news category. In the first approach we proposed, we succeed to collect 10 typical user accounts for news and commercial categories that are showed in Table 1 and then crawled thousands of tweets from these typical user accounts by Twitter API.

Table 1: typical user accounts of news and commercial categories

News (10)	@mainichijpedit, @yomiuri_online, @YahooNewsTopics, @asahi, @livedoornews, @nikkeitter, @newsheadline, @googlenewsjp, @gnewsbot, @47news
Commercial (10)	@mixprice_com, @rakuraku360, @kaimonosuki, @ranranraku, @Chris7Brown, @yoshino1010, @yellclick, @panda_kakasi, @kadenbest, @ichichoou

But for private category, considering diversity of private tweets, an abundance of accounts belonging to private category will be preferable and hence we tried a different method. The criterion for determining a user account of the private category is whether or not user name of the account is a person’s name. To acquire large number of such accounts, we rely on Mecab[2], a morphological analyzer to judge whether a user name is a person’s name or not. Some examples of analysis results are listed in Table 2.

Table 2: morphological analysis results of “福間健二”

福間	名詞,固有名詞,人名,姓,*,*,福間,フクマ
健二	名詞,固有名詞,人名,名,*,*,健二,ケンジ

Only when each part of the user name is made up by a person’s name, we regard it as a user belonging to private category. We randomly collected 12,533 private user accounts and crawled 5 tweets from each account. Table 3 shows details of our first dataset.

Table 3:the first Tweets dataset

	News	Commercial	Private
#Tweet	38,441	50,580	62,667
#account	10	10	12,533

4.2 Corpus based on Twitter lists

In the first approach to form corpus, we only used 10 typical user accounts for news and commercial categories which may result in biased training because of insufficiency of typical user accounts. Aiming to achieve a sufficient number of typical users, in the second approach we attempted a simple iterative algorithm, label propagation on Twitter graph of users and lists to increase.

A Twitter list is Twitter’s way of allowing any users to organize users they follow into groups. When click to view a list, we can see a stream of tweets from all the users included in that group, or “list”. The ground for us to use Twitter list is that Twitter user are used to organize users holding some characteristic in common into one list. Therefore we suppose the list into which typical users are gathered may contain other typical user belonging to the same category. We aim to collect 100 typical users from 10 seeds we had by the label propagation algorithm for news and commercial categories.

4.2.1 Snowball sample of Twitter lists

First, we employed snowball sampling (introduced in [Wu. S., 2011]) to collect a bunch of users that may share the same characteristics with typical users we have in common. For news and commercial categories, we choose a number u_0 of seed users from typical users. For news and commercial categories, users in Table 1 are used as seeds. But some of the seeds were not gathered in any list, we abandoned them,

Next, we selected some keywords based on their representativeness of the news and commercial categories by hand as following:

- News: news, ニュース
- Commercial: sale, commercial, shop, goods, spam, セール, ショップ, グッズ, スパム, 商品, 販売, 通販, 買い物

With seeds and keywords, we performed a snowball sample of the graph of users and lists (Figure 1).

First, we crawled all the lists in which that seed contained. Next, we chose lists l_0 whose name matched at least one of the keywords for news and commercial category. For instance, @asahi is on lists call “web service” and ”news”, but only the “news” list would be kept. We then crawled all users contained in lists l_0, and repeated these two steps to complete the crawl. In total, we crawled:

- News: 1228 users, 5267lists
- Commercial: 1688 users, 3042 lists

[2] http://mecab.sourceforge.net/

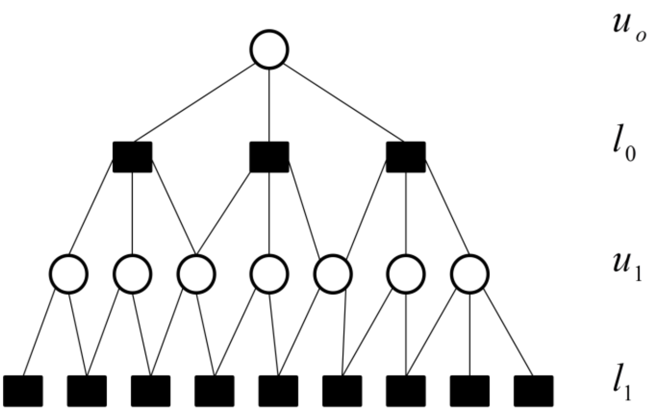

Figure 1: snowball sampling method

4.2.2 Label propagation on graph of users and lists

Although the snowball sampling is convenient, it also has some disadvantages. When users create a list, they may sometimes choose a user account without taking its representativeness of the list into account so that such. And we are not interested in such user accounts. What's more, we are not interested in user accounts that post irrelevant tweets to the category frequently. For instance, @RakutenJP, the official Twitter account of a business to customer electronic commerce site, posts commercial tweets, but many of them are noise tweets such as chats with customers, greets and so on. Here we show some examples of them.

- フォローさせていただきました。今年もよろしくお願いいたします＾＾
- 大切な買い物の思い出は深く長く残りますよね＾＾

At last, snowball sampling is also potentially biased by our particular choice of seeds and keywords.

In order to solve these problems and obtain typical users account, we exploited a simple iterative algorithm, label propagation on graph of users and lists. The goal of process is to obtain users highly related to seeds by calculating correlation weight between seeds and the other users we crawled after snowball sampling. Label propagation, by [Xiaojin Zhu, Learning from Labeled and Unlabeled Data with Label Propagation, 2002], is a semi-supervised learning method which uses a few seeds and relations between all the examples to label a large number of unlabeled examples.

Here we have users linked by lists and each user is influenced by the lists in which they are appeared. Therefore we can user label propagation algorithm to spread label distribution from a small set of seeds with the initial label information (news or commercial) through the graph. Label distributions are spread across a graph G = {V, E, W} where V is the set of n users, E is a set of link between users and lists and W is an n×n matrix of weights which we define as times every 2 users appearing in one list simultaneously.

The algorithm proceeds as follow:

1. For news and commercial categories, separately assign an n*n matrix T with times every 2 users appearing in one list simultaneously, where n is the number of users. And then we assign another n*c matrix Y with the initial assignment labels, where c is 2 (news or not news / commercial or not commercial). For seeds, initial labels will be (1, 0) while other will be (0.5, 0.5).
2. Propagate labels for all users by computing Y = TY
3. Row-normalize Y such that each row adds up to 1
4. Reset labels of seeds to be original values (1, 0).
5. Repeat 2-5 until Y converges.

Through label propagation proceeds, we got the converged Y matrix with values of correlation between seeds and other users. The bigger such values are, the higher the possibility to be a typical user will be. So at last, we select typical users up from the users in the matrix. We check tweets such users post and pick up them only most of their tweets belong to the supposed category. We stopped selection until we obtained 100 typical users separately belonging to news and commercial categories. Figure 2, cumulative distribution of typical users among all the users for the two categories, shows the effectiveness of label propagation.

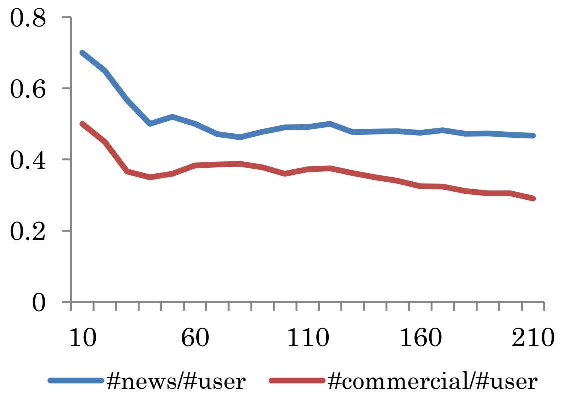

Figure 2: Cumulative distribution of typical users among all the users.

We succeed to acquire 100 news user accounts and 98 commercial user accounts. For private category, we selected 196 users from the first corpus. At last we crawled up to approximately 200 tweets from each typical user account to form our second corpus. Table 4 shows the details.

Table 4: the second tweets dataset

	News	Commercial	Private
#Tweet	23,683	20,068	23,263
#account	100	98	196

5 Training the classifiers

5.1 Feature extraction

Since we had formed corpora, we extracted features from them to train 3-class classifier. Following work on document classification, we extracted ordinary text features as well. Besides them, we tailored some are that are specific to the task.

First, we perform some preprocess:

- Filter – we removed cited text from a retweet. A retweet is supposed to help user quickly share other users' tweet with their followers, adding comments or not.
- Removing stopwords – we removed words including particle, aux, symbol, noun-pronoun, noun-affix , exclamation.

Features we extracted following document classification are as followed:

- Constructing bag of words (BOW) model – BOW model is a simplifying assumption user in natural language processing in which a text is represented as an unordered collection of words, disregarding grammar. We first performed morphological analysis on each word by Mecab, and then represented tweets by words analyzed.

Features that are specific to the task are as followed:

- User property information – since Twitter is a social microblog, it has a feature allowing users to subscribe to other users' tweets as a follower. According to rules of Twitter, if you follow someone, he(she) will be regarded as your friend while if you are followed by someone, he(she) will be regarded as a follower. We extracted information on friends and followers of each user and logarithm of #friend and #follower are used as features.
- Url domain – Twitter allows users to share url links in their tweets, but shortened ones. We managed to reverse them back to original ones and extracted domain of them as a feature.

5.2 Classifiers

We tested two different classifiers: support vector machine and label propagation.

5.2.1 Support vector machine

Support vector machine is a popular classification technique [11]. We use liblinear[3], a library for large linear classification. Our input data are sets of vectors and each element in the vector represents a feature. If the feature is present, the value is 1, otherwise it is 0.

5.2.2 Label propagation

As we introduced, label propagation is a semi-supervised learning method that can be trained to classify tweets. The proceeds of classifying tweets are almost consistent with classifying users but three differences exist as following:

- In the graph G = {V, E, W} where V is the set of m tweets, W here represents an $m \times m$ matrix of weights which we define as number of words every two tweets share in common.
- We assign $m \times c$ matrix Y with the initial assignment labels, where c is 3 (news, commercial and private). For seeds, initial labels will be (1, 0, 0), (0, 1, 0) or (0, 0, 1) while other will be (0.33, 0.33, 0.33).
- In converged matrix Y, for each tweet the biggest one among the three label values determines which category the tweets should belong to.

5.3 Methodology

We use four indicators to evaluate the performance of the model built based on training tweets data.

Accuracy: It represents how many the label of test data are predicted by the model correctly.

$$\text{Accuracy} = \frac{\text{N(correct classified tweets)}}{\text{N(all tweets)}}$$

Precision: In the field of information retrieval, it is the fraction of retrieved documents that are relevant to the search.

[3] http://www.csie.ntu.edu.tw/~cjlin/liblinear/

$$\text{Precision} = \frac{|\{\text{relevant tweets}\} \cap \{\text{retrieved tweets}\}|}{|\{\text{retrieved tweets}\}|}$$

Recall: in Information Retrieval it is the fraction of the documents that are relevant to the query that are successfully retrieved.

$$\text{Recall} = \frac{|\{\text{relevant tweets}\} \cap \{\text{retrieved tweets}\}|}{\text{relevant tweets}}$$

F-measure: A measure that combines precision and recall is the harmonic mean of precision and recall.

$$F = 2 \times \text{precision} \times \text{recall}/(\text{precision} + \text{recall})$$

This is also known as the F_1 measure due to the evenly weighted recall and precision.

5.4 Experiment and evaluation

In the research, we performed a classification experiments on test tweets to confirm the effectiveness of features and a experiments exploiting over sampling method to dissolve the gap of quality between manual corpus and automatic corpus.

First, we performed a classification experiment on manually labeled tweets. We crawled test tweets from Twitter with 10 hot keywords in 2011: AKB, 授業 (lecture), CM (commercial message), 地震 (earthquake), 福島 (fukushima), NHK, ワンピース (onepiece), ラーメン (noodle), サッカー (soccer), 電車 (train). These keywords belong to different genres and we chose them randomly. Then we crawled about 150 tweets randomly by each keyword. The test tweets are all labeled by 3 Japanese master students. But before the labeling work, we had to confirm whether they have good agreement on the criterion of deciding the category of a tweet or not. We prepared a small dataset for them to label and computed Kappa value of them. Kappa value is a statistical measure of inter-rater agreement for categorical items. The mean value is 0.67 which means there is a good agreement among the raters.

Since the three raters have good agreement, it's acceptable for them to label our test tweets showed in the following table.

Table 5. Number of tweets in test tweets dataset.

Tweets	News	Commercial	Private
1605	248	39	1318

We exploited many kinds of combinations of the features and we did comparison experiments on corpus1 and corpus2. The features we exploited are following:

- I – BOW
- II – BOW + parts of speech
- III – BOW + parts of speech + polarity
- IV – BOW + parts of speech + polarity + ln(friends) + ln(followers) + domain

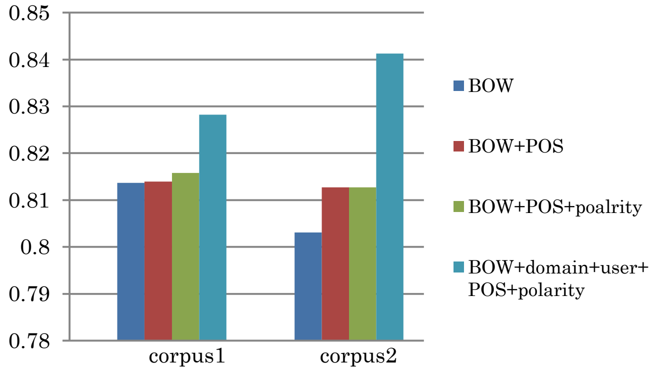

Figure 3. Accuracy of classification on test tweets.

From the figure, we can observe that the distinctive features of Twitter are effective.

In the next experiment, we exploited over sampling to dissolve the gap of quality between datasets. As we explained, we formed two automatic corpora of comparatively low quality. And in the training process is dominated by the large automatic corpora so we enlarge the manual corpus by over sampling. First, we perform 5-fold cross validation on the test tweets. Then in the over sampling process, we increased the size of test tweets to 5, 10, 15, 20 and 25 times size, then performing 5-fold cross validation on test tweets adding corpus1 and corpus2 respectively. I will show the results of case when we exploit the IV features.

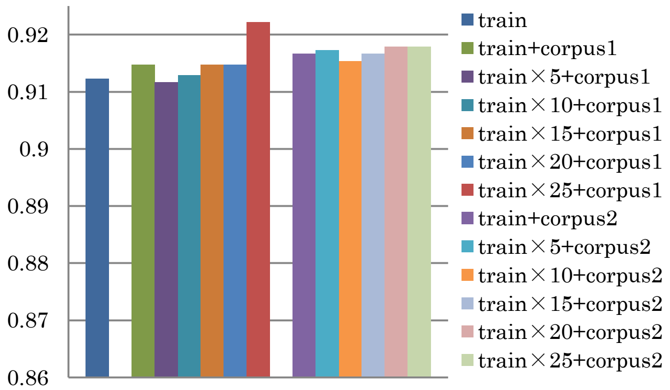

Figure 4. Results of over sampling method exploiting IV features.

As we can see from the figure that, as we amplify the size of the test tweets, the value of accuracy in each figure has gone up which means over sampling the original test tweets can help improve efficient of the SVM classifiers. The objective to dissolve the gap of quality between manual corpus and automatic corpora by over sampling is carried out. In the next Table, I will show the confusion matrix of the best case (in Table 6, train×25+corpus1) to show the details of the classified tweets. We can see from the table, about 70%of the News and 37.5% of the Commercial tweets are correctly classified which shows our methods have good performance.

Table 6. The confusion matrix of the best case.

Correct Predict	News	Commercial	Private
News	35	1	5
Commercial	0	3	0
Private	15	4	259
Amount	50	8	264

6 Conclusion

In our research, we proposed three categories based on information publicness. We focus on approaches of collecting a corpus automatically for training classifiers. We proposed two approaches that are based on typical Twitter user accounts and based on Twitter lists using label propagation respectively. Using the corpora, we built classifiers, which are able to determine news, commercial and private tweets. Experiments evaluations show our proposed techniques are effective.

For the future work, we can develop our study to classify tweets into more specific categories. For instance, the Private category can be divided into two categories: experience of people and thoughts (opinions) of people. As we all know, detecting what the people are thinking and their will is useful for marketing, social investigation and so on. Also, we can consider methods to improve efficiency of classification on imbalanced data.

7 References

1. Andreas Kaplan. The early bird catches the news: Nine things you should know about micro-blogging. : Business Horizons, 2011.

2. D. Irani Webb, C. Pu, and K. LiS. Study of trend-stuffing on twitter through text classification. : Proceedings of 7th Annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference, 2010.

3. Alexander Pak, Paroubek Patrick. Twitter as a Corpus for Sentiment Analysis and Opinion Mining. : Proceedings of the Seventh conference on International Language Resources and Evaluation LREC, 2010.

4. Banerjee Ramanthan, K., and Gupta, AS.,. Clustering short text using Wikipedia. : Proc. SIGIR, 2007.

5. HuSun, N., Zhang, C., and Chua, T.-SX.,. Exploiting internal and external semantics for the clustering of short texts using world knowledge. : Proc. CIKM, 2009.

6. Takeshi Sakaki, Makoto Okazaki, Yutaka Matsuo. Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors. : In Proceedings of the Nineteenth International WWW Conference, 2010.

7. Sankaranarayanan Samet, H., Teitler, B. E.,J.,. TwitterStand: news in tweets. : In Proc. ACM GIS, 2009.

8. B. SriramFuhry, and M.DemirbasD. Short text classification intwitter to improve information filtering. : Proceedings of 33rd International ACM SIGIR Conference, 2010.