

Master thesis

# **Classification of Microblog Posts Based on Information Publicness**

情報の公共性に基づく Microblog 記事の分類

February 8th, 2012

Supervisor

Associate Professor Masashi Toyoda

Department of Information and Communication Engineering

Graduate School of Information Science and Technology

The University of Tokyo

48-106424 Hongguang Zheng



# 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



## Acknowledgements

First of all, I would like to thank my supervisor, Associate Prof. Masashi Toyoda, for his helpful advices and criticisms during these two years. I would also like to express my gratitude to Research Associate Nobuhiro Kaiji and Naoki Yoshinaga. They provided me a lot of valuable comments and instructions on my research.

I would also like to thank members of the laboratory. They have been always supportive to me and sharing their reviews with me.

Finally, I thank my parents and my girlfriend, Xiaosha for supporting me in these years.



## Table of contents

<b>LIST OF FIGURES</b> .....	IX
<b>LIST OF TABLES</b> .....	X
<b>1 INTRODUCTION</b> .....	1
1.1 BACKGROUND.....	1
1.2 TASK SETTING.....	2
<b>2 RELATED WORK</b> .....	5
<b>3 CORPUS COLLECTION</b> .....	9
3.1 TWITTER API.....	9
3.2 OUR DATASETS.....	9
3.2.1 Corpus based on typical user accounts .....	10
3.2.2 Corpus based on Twitter lists.....	14
<b>4 TRAINING THE CLASSIFIERS</b> .....	21
4.1 FEATURES EXTRACTION.....	22
4.2 CLASSIFIERS.....	23
4.2.1 Support vector machine .....	23
4.2.2 Label propagation .....	24

4.3	EVALUATION MEASURE .....	25
5	<b>EXPERIMENT AND EVALUATION</b> .....	27
5.1	COMPARISON ON SVM AND LABEL PROPAGATION.....	27
5.2	5-FOLD CROSS VALIDATION ON CORPORA .....	27
5.2.1	5-fold cross validation on corpus1 .....	28
5.2.2	5-fold cross validation on corpus2 .....	30
5.3	CLASSIFICATION EXPERIMENTS ON TEST TWEETS .....	31
5.3.1	Test tweets .....	31
5.3.2	Performance comparison between corpus1 and corpus2 .....	34
5.3.3	Experiments exploiting over sampling method .....	36
6	<b>DISCUSSION</b> .....	39
7	<b>CONCLUSION AND FUTURE WORK</b> .....	43
	<b>REFERENCES</b> .....	45
	<b>PUBLICATIONS</b> .....	47



## List of Figures

Figure 1. Searching “ipad” on Twitter. ....	2
Figure 2. The two steps involved in calculating distances between tweets using Wikipedia. ....	7
Figure 3. Examples of tweets posted by @asahi. ....	11
Figure 4. Examples of tweets posted by @mixprice_com. ....	11
Figure 5. A list which @asahi is gathered in. ....	15
Figure 6. Snowball sampling on users and lists. ....	17
Figure 7. Cumulative distribution of typical users among all the users. ....	20
Figure 8. The framework of supervised classification. ....	21
Figure 9. Precision in each fold cross validation on corpus1. ....	29
Figure 10. Recall in each fold cross validation on corpus1. ....	29
Figure 11. Precision in each fold cross validation on corpus2. ....	31
Figure 12. Recall in each fold cross validation on corpus2. ....	31
Figure 13. Results of performance comparison on corpus1 and corpus2. ....	35
Figure 14. Results of oversampling method exploiting BOW. ....	37
Figure 15. Results of oversampling method exploiting BOW + POS +polarity. .	37
Figure 16. Results of oversampling method exploiting BOW + POS +polarity ln(friends) + ln(followers) + domain. ....	38
Figure 17. Classified tweets collected by “地震”. ....	40
Figure 18. Classified tweets collected by “AKB”. ....	41
Figure 19. Classified tweets collected by “CM”. ....	42

## List of Tables

Table 1. News and Commercial typical user accounts.....	12
Table 2. morphological analysis results of “福間健二”.....	13
Table 3. morphological analysis results of “光さん”. .....	13
Table 4. Details of the first dataset .....	14
Table 5. Seeds of News and Commercial categories.....	16
Table 6. Details of our second dataset.....	20
Table 7. Performance comparison between SVM and label propagation. ....	28
Table 8. Results of 5-fold cross validation on corpus1.....	28
Table 9. Results of 5-fold cross validation on corpus2.....	30
Table 10. An example to explain Kappa coefficient. ....	33
Table 11. Kappa value of the 3 raters.....	34
Table 12. Number of tweets in test tweets dataset. ....	34
Table 13. Results of performance comparison con corpus1 and corpus2. ....	35
Table 14. The confusion matrix of the best case.....	38

# 1 Introduction

## 1.1 Background

Microblogging is a broadcast medium in the form of blogging. Twitter, A microblog differs from a traditional blog in that its content is typically smaller in text size, which enables its users to send and read text-based posts of up to 140 characters, known as "tweets". It was created in March 2006 by Jack Dorsey and launched that July. The service rapidly gained worldwide popularity, with over 300 million users as of 2011, generating over 300 million tweets and handling over 1.6 billion search queries per day[1].

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. There is a research on distribution of tweets published in 2009<sup>1</sup>. 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.

- News – 4%
- Spam – 4%
- Self-promotion – 6%
- Pass-along value (tweets with an “RT”) – 9%
- Conversation – 38%
- Pointless babble – 40%

As we can see from this data, news and commercial tweets constitute about 15% in all the tweets while private tweets (conversation and babble) constitute 78%. Certainly news tweets contain important information that is valuable for

---

<sup>1</sup> <http://www.pearanalytics.com>



Figure 1. Searching “ipad” on Twitter.

public. They include disastrous events that public are concern about such as storms, fires, traffic jams, riots, heavy rainfall, and earthquakes. Besides news tweets, commercial tweets also include some big social events such as parties, baseball games, and presidential campaigns. And private tweets are useful for marketing such as collecting users’ reviews about a product.

However, tweets are very messy even on the same one subject. If we 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. If we search “ipad”, there will be not only users’ reviews about ipad but many tweets promoting applications for sale (Figure 1).

## 1.2 Task setting

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 order to support user’s searching on Twitter, 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 this end, our work makes two main contributions:

- 1 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.
- 2 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 5 in which we show comparison results on our two corpora. Finally, in section 6 we conclude with a brief discussion of future work.

## 2 Related work

Although many researchers have studied document classification, they classified documents according to sentiment or topics.

McDonough in [2] described the topic identification (TID), an automatic classification of speech messages into one of a known set of possible topics. The TID task can be view as having three principal components: 1) event generation, 2) keyword event selection, and 3) topic modeling. Using data from the Switchboard corpus, the authors present experimental results for various approaches to the TID problem and compare the relative effectiveness of each. And Hsueh in [3] concerned how to segment a scenario-driven multiparty dialogue and how to label these segments automatically. They applied approaches that have been proposed for identifying topic boundaries at a coarser level to the problem of identifying agenda-based topic boundaries in scenario-based meetings and developed conditional models to classify segments into topic classes.

Recent years, tweets classification has become a popular topic due to the popularity of Internet application such as blogging, and Twitter. Bo in [4] considered the problem of classifying documents not by topic, but by overall sentiment, e.g., determining whether a review is positive or negative. Using movie reviews as data, they found that standard machine learning techniques definitively outperform human-produced baselines. However, the three machine learning methods they employed (Naive Bayes, maximum entropy classification, and support vector machines) do not perform as well on sentiment classification as on traditional topic-based categorization. Irani in [5] proposed a machine learning method to automatically identify trend-stuffing in tweets, using texts and links of tweets. Pak in [6] 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.

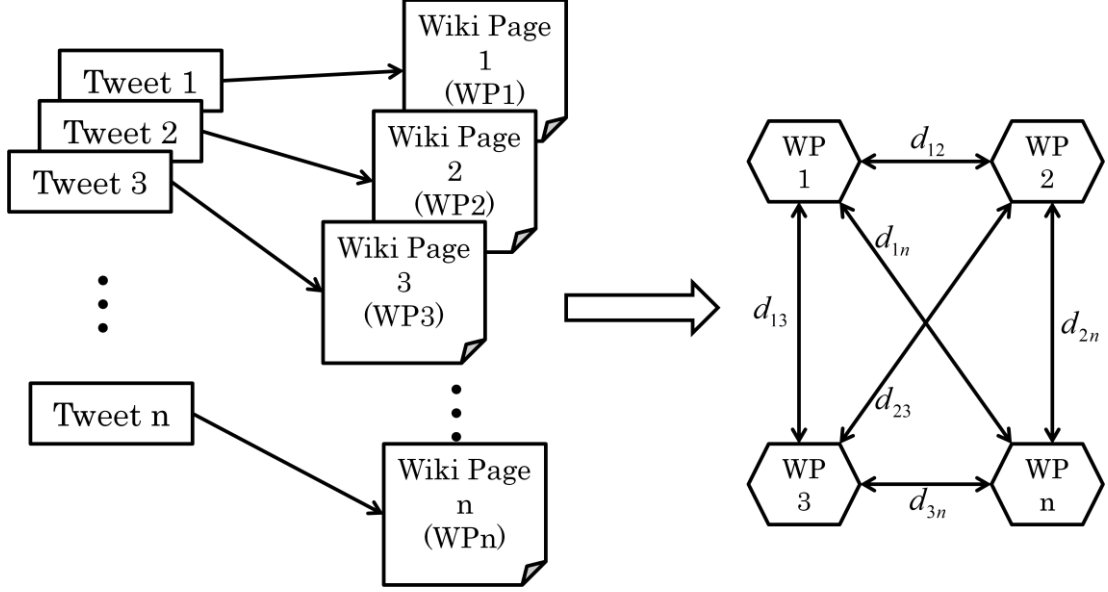
Some existing works on classification of short text messages integrate messages with meta-information from Wikipedia and WordNet [7][8]. Sakaki in [9] showed how to detect real time events by machine learning methods. Here I will focus on researches studying classifying tweets.

Sankaranarayanan [10] investigate the use of Twitter to build a news processing system, called TwitterStand, from Twitter to capture tweets that correspond to late breaking news. The reason why Twitter is that Twitter is a technology that breaks down communication barriers. It is a medium of instantaneous feedback which means that any action in the real world usually receives a near instant reaction or feedback in terms of tweets expressing opinions or reactions to the action. As a study on Twitter sentiment classification, Alec in [11] introduced a novel approach for automatically classifying the sentiment of Twitter messages. These messages are classified as either positive or negative with respect to a query term. Their approach is to use different machine learning classifiers and feature extractors. The machine learning classifiers are Naive Bayes, Maximum Entropy (MaxEnt), and Support Vector Machines (SVM). The feature extractors are unigrams, bigrams, unigrams and bigrams, and unigrams with part of speech tags.

Yegin in [12] introduced a Wikipedia-based classification technique to categorize messages streaming through Twitter. They develop this technique for calculating semantic distances between messages based on the distances between their closest Wikipedia pages. Messages are mapped onto their most similar Wikipedia pages, and the distances between pages are used as a proxy for the distances between messages.

Figure 2 shows how the system works. They used the distance between the two





**Figure 2. The two steps involved in calculating distances between tweets using Wikipedia.**

associated Wikipedia pages as an indicator of the distance. And the two steps are: finding associated wikipedia pages and calculating Distances.

Sriram’s work in [13] is closed to ours. As Twitter users may become overwhelmed by the raw data, the researchers tried to solve this problem. Their solution is to classify these short messages which shares common concepts with ours into some defined categories. The goal of their work is to automatically classify incoming tweets into different categories: News, Events, Opinions, Deals, and Private Messages. In order to train the classifier, they extracted 8 features which consist of one nominal (author) and seven binary features (presence of shortening of words and slangs, time-event phrases, opinioned words, emphasis on words, currency and percentage signs, “@username” at the beginning of the tweet, and “@username” within the tweet). However, they neglected distinctive features of Twitter. In our research, we exploited several distinctive features that only appear in social network sites. And the biggest difference between Sriram’s work and ours is that we build large-scale datasets automatically while they only crawled about 5,000 tweets and labeled them manually. In our research, we

proposed two approaches to form training datasets.

As far as know, classifying tweets based on information publicness is not done by other people yet and it's a useful work for user's searching, so in our work we try to propose some effective method to train such classifiers.

## 3 Corpus collection

### 3.1 Twitter API

In our research, we make use of Twitter API<sup>2</sup> which is developed by Twitter. The Twitter micro-blogging service includes three APIs: the Rest API, the Search API and Streaming API.

The Twitter REST API methods allow developers to access core Twitter data. This includes update timelines, status data, and user information. With the REST API, you can read your timeline and direct messages, tweet and retweet from it, follow and unfollow other users, send direct messages and so on. In the other hand, the Search API methods give developers methods to interact with Twitter Search and trends data. It duplicates what the Twitter search page does. It can do everything that the Twitter advanced search can do, and nearly the Twitter monitoring tools depend on this API. The last API, Streaming API released into production in 2011. It allows high-throughput near-realtime access to various subsets of public and protected Twitter data.

The API presently supports the following data formats: XML, JSON, and the RSS and Atom syndication formats, with some methods only accepting a subset of these formats. So through the REST API provided by Twitter, users can programmatically perform almost any task that can be performed via Twitter's web interface. For non-whitelisted users, Twitter restricts a user to 150 requests per hour. For authorized user, Twitter restricts 350 requests per hour. Furthermore, searches are limited to returning 1500 posts for a given request.

### 3.2 Our datasets

---

<sup>2</sup> <https://dev.twitter.com/>

Using Twitter API 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 before, our strategy is to collect users belonging to each categories first, and then crawl tweets from these users to form large-scale datasets.

### 3.2.1 Corpus based on typical user accounts

First, we give the definition of a typical user. Typical users are defined as users who post texts mostly belonging to the same one category. Let's take some examples for each category.

1. @asahi – an official account of the Asahi Shimbun, is regarded as a News typical user for most of tweets it posts belong to the News category (in Figure 3).
2. @mixprice\_com - an official account of “Mixprice.com”, is for conveying messages to market products of the company (in Figure 4).
3. For Private category, because a user can be regarded as a typical private user if it only posts private tweets and ordinary people usually post private tweets, we have so many private user accounts. The method to acquire such accounts will be introduced later.



Figure 3. Examples of tweets posted by @asahi.



Figure 4. Examples of tweets posted by @mixprice\_com.

In the first approach we proposed, we succeed to collect 10 typical user accounts for News and Commercial category which are showed in Table 1. We then crawled tweets from these typical accounts by Twitter API on a large scale.

**Table 1. News and Commercial typical user accounts.**

News	@mainichijpedit, @yomiuri_onlie, @YahooNewsTopics, @asahi, @livedoornews, @nikkeitter, @newsheadline, @googlenewsjp, @gnewsbot, @47news
Commercial	@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. Here we focus on users' profile information. 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. According to the Twitter document, a user has two names: the screen name and the user name. Let's take @ Yomiuri\_Online for example.

- @ Yomiuri\_Online
  - Screen name: Yomiuri\_Online
  - User name: 読売新聞 YOL

Considering that people using person's name in their profile information would hardly post News and Commercial tweets, if the user name of a user is constituted by person's name, we regard the user as a Private typical user. To acquire large number of such accounts quickly, we rely on Mecab<sup>3</sup>, a morphological analyzer to judge whether a user name is a person's name or not. Morphological analysis is the process of breaking down morphologically complex words into their constituent morphemes (word meaning parts). For instance, the words “携帯電話” is composed of two meaning units, the base “電話”, and the modifier “携帯”, which conveys the meaning of an agent (person or thing) that

---

<sup>3</sup> <http://mecab.sourceforge.net/>

**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

**Table 3. morphological analysis results of “光さん”.**

光	名詞,固有名詞,人名,名,*,*,光,ヒカリ,ヒカリ
さん	名詞,接尾,人名,*,*,*,さん,サン,サン

does whatever is implied in the base. Therefore the words “携帯電話” mean a mobile phone. With Mecab, we can obtain parts of speech of the words. For example:

- 携帯電話
  - 携帯            名詞,サ変接続,\*,\*,\*,携帯,ケイタイ,ケイタイ
  - 電話            名詞,サ変接続,\*,\*,\*,電話,デンワ,デンワ

Following our criterion, if among the parts of speech “人名” appears and the user name are only constituted by such words, we say that the user is a typical Private user. Some examples of analysis results are listed in Table 2 and Table 3.

We randomly collected 12,533 private user accounts and crawled 5 tweets from each account through Twitter API. Table 4 Shows details of our first dataset.

**Table 4. Details of the first dataset**

	News	Commercial	Private
#Tweets	38,441	50,580	62,667
#accounts	10	10	12,533

### 3.2.2 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 the number of typical users of News and Commercial category.

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”. Lists help you organize people in ways that make sense to you, and help improve the signal to noise ratio of Twitter. You can call them whatever you like by naming the lists, and can add or remove people at any point. Examples of lists include: News, Earthquake Experts, Celebrity and so on. From homepages of Twitter users, we can figure out details of lists the users are gathered in. Let’s take @asahi for an example (Figure 5):





Figure 5. A list which @asahi is gathered in.

The List named “ニュース” in Figure 5 is created by a user “ゆきち”. When we pay attention to other users in the list, we found that they are all related to news. Besides this list, we can see this characteristic from other lists. This characteristic supports our approach. 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 lists into which typical users are gathered may contain other typical users belonging to the same category.

For the second dataset, we aim to collect 100 typical users from 10 seeds we had by the label propagation algorithm for news and commercial categories.

### 3.2.2.1 Snowball sample of Twitter lists

First, we employed snowball sampling (introduced in [14]) to collect a bunch of users that may share the same characteristics with the 20 typical users (News

**Table 5. Seeds of News and Commercial categories.**

News	@mainichijpedit, @yomiuri_onlie, @YahooNewsTopics, @asahi, @livedoornews, @nikkeitter, @newsheadline, @googlenewsjp, @gnewsbot, @47news
Commercial	@kaimonosuki, @Chris7Brown, @kadenbest, @ichichoou

and Commercial categories) in common.

For News and Commercial categories, we picked up a number  $u_0$  of seed users from typical users we have (Table 1). Because some users are not gathered in any list, we abandoned such users. The users used as seeds are as follow in Table 5 .

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

As we mentioned before, we can obtain information of lists in which a particular user is gathered in, and lists which a particular user created. With seeds and keywords, we performed a snowball sample of the graph of users and lists (Figure 6) to get a bunch users related to the seeds.

First, we crawled all the lists in which seeds contained. Next, we chose lists  $l_0$  whose name matches at least one of the keywords for News (if News seed) or Commercial (if Commercial seed) category. For instance, @asahi is on lists called “web service” and “news”, but only the “news” list will be kept for next process. We then crawled all users contained in the lists  $l_0$ . As soon as we got the users  $u_1$ , we repeated the first step to complete the snowball sampling.

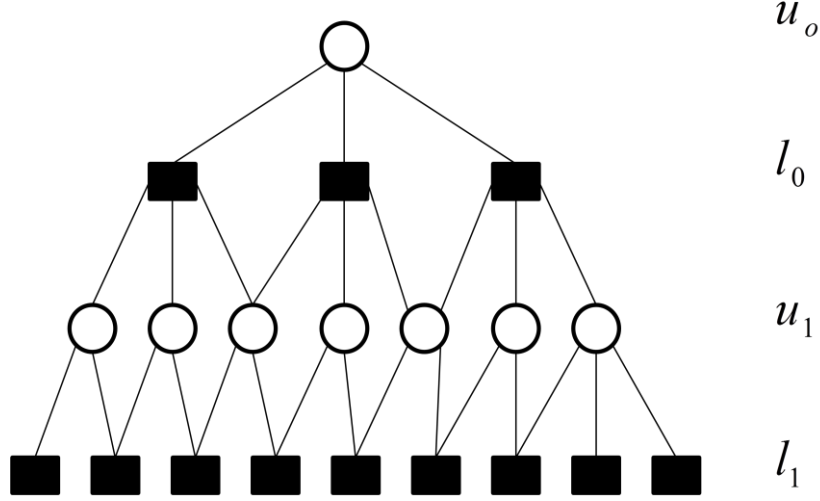


Figure 6. Snowball sampling on users and lists.

In total, we crawled:

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

### 3.2.2.2 Label propagation on graph of users and lists

Although the snowball sampling is convenient for crawling users, 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. For instance, @asahi is gathered in a list named “@hrk107/ニュース” and there are other 6 users in the list. However, some users don’t share the same characteristic with @asahi as a news account, such as the user “@atcosmenet” is gathered in the list which posts commercial tweets on cosmetic. Moreover, 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 ^^
- つぶやきありがとうございます ^^ 私もヘルスケアを意識しなくては . . .

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 user accounts, 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, introduced in [15], 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 use 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  $n \times n$  matrix of weights which we define as times every 2 users appearing in the same list simultaneously.

The algorithm proceeds as follow:

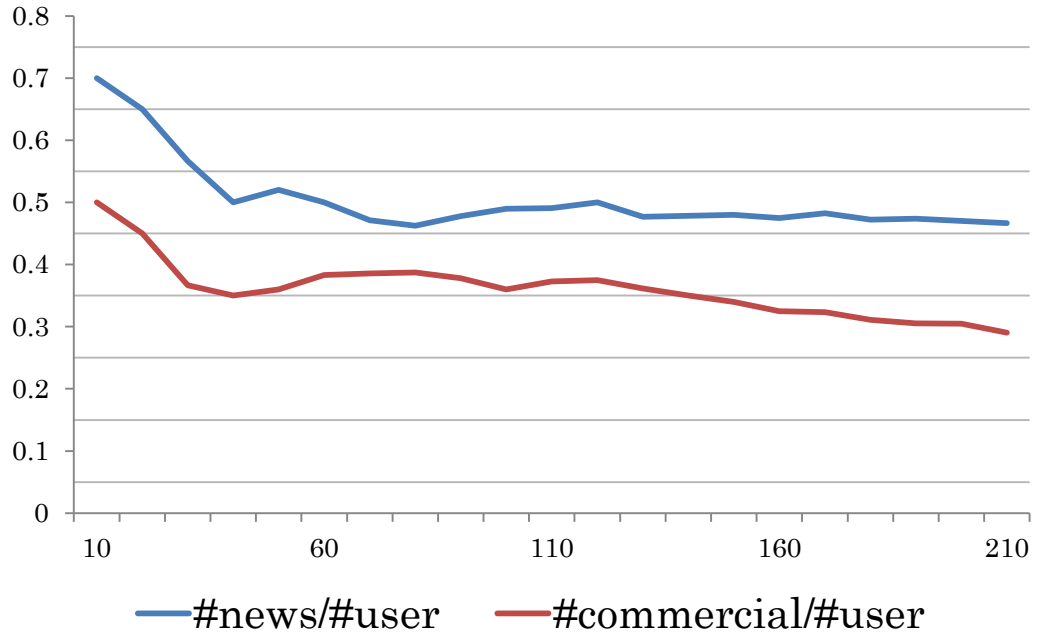
1. For news and commercial categories, separately assign a  $n \times n$  matrix  $T$  with times every 2 users appearing in the same list simultaneously, where  $n$  is the number of users. And then we assign another  $n \times c$  matrix  $Y$  with the initial assignment labels, where  $c$  is 2 (for News category, there are two kinds of labels: News or not News / for Commercial category, there are two kinds of

labels: 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. In Figure 7, cumulative distribution of typical users among all the users for the two categories, shows the effectiveness of label propagation.

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 6 shows the details.



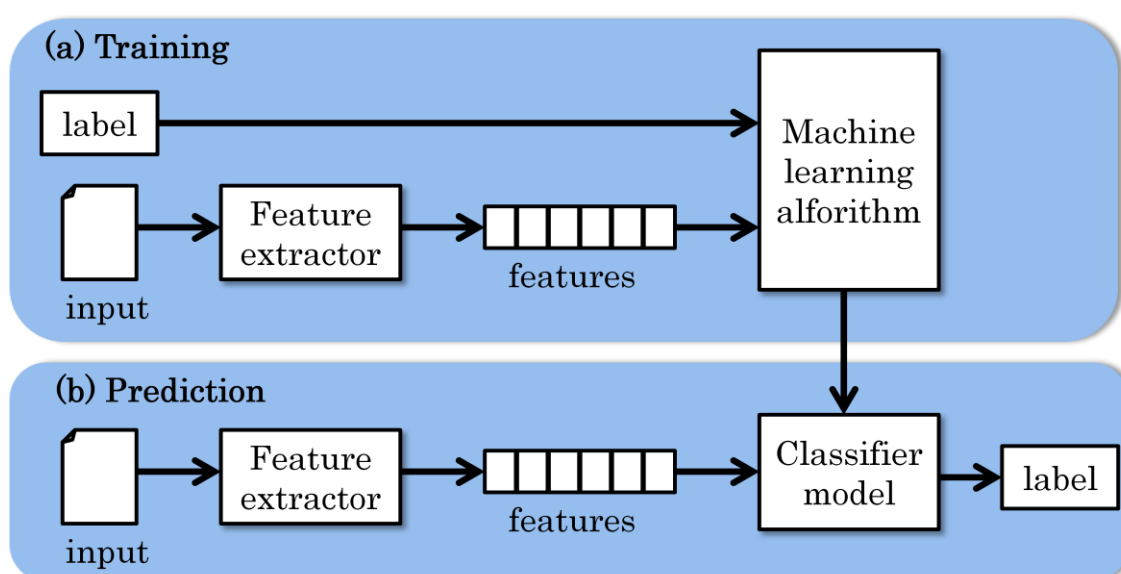
**Figure 7. Cumulative distribution of typical users among all the users.**

**Table 6. Details of our second dataset.**

	News	Commercial	Private
#Tweets	23,683	20,068	23,263
#accounts	100	98	196

## 4 Training the classifiers

Classification is the task to choose a correct label for a given input. A classifier is called supervised if it is built based on training corpora containing the correct label for each input. The framework used by supervised classification is shown in Figure 8 [16].



**Figure 8. The framework of supervised classification.**

Supervised classification has two parts:

- During training, a feature extractor is used to convert each input value to a feature set. This process, which captures the basic information about each input that should be used to classify it, is called feature extraction. Pairs of feature sets and labels are fed into the machine learning algorithm to generate a model. Therefore during this process, whether we are able to extract appropriate features or not will be the key.
- During prediction, the same feature extractor is used to convert unlabeled inputs to feature sets. These feature sets are then fed into the model, which generates predicted labels according to the training sets.

## 4.1 Features 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.

At first, we performed some preprocess before extracting features in order to clean the data.

- 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. In our work, this process helps us prevent from repetition of retweeted tweets.
- Removing stopwords – by using Mecab, we removed words including particle, aux, symbol, noun-pronoun, noun-affix , exclamation.

And then, we extracted features following document classification:

- 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.
- Parts of speech - In grammar, a part of speech is a linguistic category of words (or more precisely lexical items), which is generally defined by the syntactic or morphological behavior of the lexical item. Common linguistic categories include noun and verb, among others. Using Mecab, we can acquire information on parts of speech of each word, split into several layers. For instance, parts of speech of “東京” will be “名詞,固有名詞,地域,一般”. And we use the former two parts of speech as features such as “名詞” and “固有名詞”.
- Polarity items – polarity items are defined as negative polarity items or



positive polarity items. Here we used a Japanese polarity dictionary exploited by researchers in Nara Institute of Science and Technology<sup>4</sup>.

Since our work is performed on Twitter, a social Microblog, we extracted some specific features for our work.

- 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 value of number of friends and followers are used as features.
- Url domain – Twitter allows users to share url links in tweets, which are converted into shortened ones. We managed to reverse them back to original ones and extracted domain of them as a feature. For instance: now we have a shorten url “t.asahi.com/5foy”, and then we succeed to reverse it into “http://www.asahi.com/science/update/0201/SEB201202010014.html”. At last we extracted domain from it “www.asahi.com”.

## 4.2 Classifiers

We trained two kinds of classifiers: support vector machine(SVM) and label propagation.

### 4.2.1 Support vector machine

SVM is a concept in statistics for a set of related supervised learning methods that recognize patterns. The standard SVM takes a set of input data (training data and test data) and predicts, for each given input (test data) , which of two or more possible classes forms the input, making the SVM a non-probabilistic

---

<sup>4</sup> <http://cl.naist.jp/index.php>

classifier. Given a set of training examples with features, each marked as belonging to one of the categories, an SVM training algorithm builds a model that assigns new examples (test examples) into one category or others. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on.

Support vector machine is a popular classification technique [17]. We use liblinear<sup>5</sup>, a library for large linear classification. Our input data are sets of vectors and each element in the vectors represents a feature.

- BOW, parts of speech, polarity item and url domain – if the feature presents, the value is 1, otherwise it is 0.
- User property information – the value of friend and follower features are logarithm value of number of friends and followers.

#### 4.2.2 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 News, commercial and Private seeds respectively, initial labels will be  $(1, 0, 0)$ ,  $(0, 1, 0)$  or  $(0, 0, 1)$  while initial labels of other test tweets 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.

---

<sup>5</sup> <http://www.csie.ntu.edu.tw/~cjlin/liblinear/>

### 4.3 Evaluation measure

We use four indicators to evaluate the performance of the model.

**Accuracy:** It represents how many the label of test tweets are assigned by the model correctly.

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

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

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

**Recall:** In the field of 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.

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



## 5 Experiment and evaluation

Our experiments consist of next variants.

- Performance comparison between SVM and label propagation
- 5-fold cross validation of the two corpora
- Classification experiments on test tweets
  - Performance comparison among features
  - Experiments exploiting over sampling method

### 5.1 Comparison on SVM and label propagation

At first, we simply compared the performance between SVM and label propagation. We performed it on corpus2, using the BOW feature. The results are presented in Table 7. As we see from the table, SVM performs much better than label propagation method almost by 7%. Therefore, we determined to train SVM classifier for our following experiments other than label propagation.

### 5.2 5-fold cross validation on corpora

Next, we examined the performance of our two corpora by 5-fold cross validation which is a technique for assessing how the results of a statistical analysis will generalize to an independent dataset. We set two restrictions on 5-fold cross validation:

1. Users of training and test tweets are not reduplicate. For instance, if in some fold cross validation @asahi appears in users of training tweets, there won't be any tweet posted by @asahi in test tweets.
2. Time period of test tweets comes after training tweets. The reason for this restriction is that we try to train classifiers to predict labels of later tweets.

**Table 7. Performance comparison between SVM and label propagation.**

	Accuracy	Precision	Recall	F1
SVM	0.828	0.817	0.817	0.817
LP	0.758	0.745	0.718	0.731

**Table 8. Results of 5-fold cross validation on corpus1.**

	Accuracy	Precision	Recall	F1
I	0.859	0.784	0.841	0.812
II	<b>0.901</b>	<b>0.893</b>	<b>0.891</b>	<b>0.892</b>
III	0.895	0.907	0.884	0.895

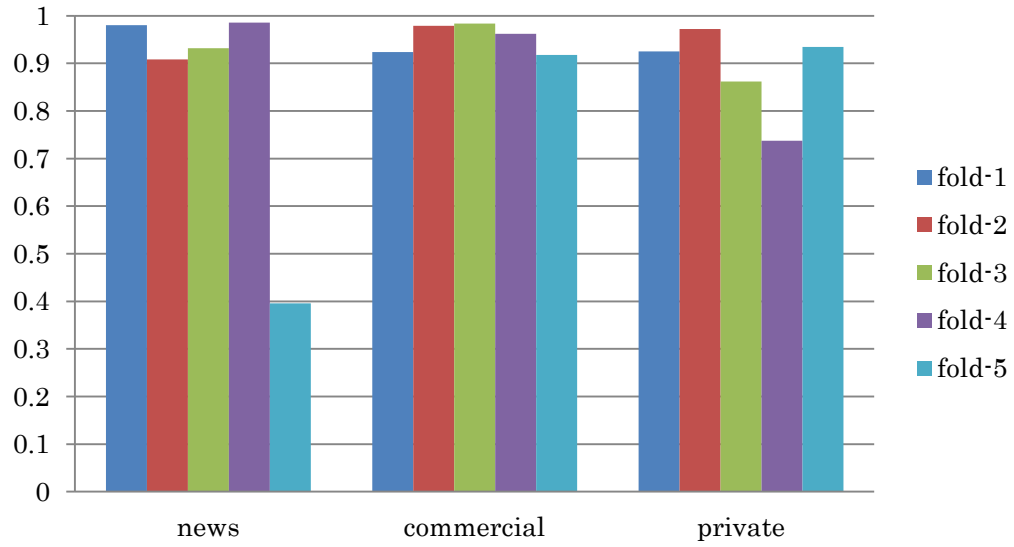
### 5.2.1 5-fold cross validation on corpus1

The results of experiments on corpus1 are showed in Table 8. In the table,

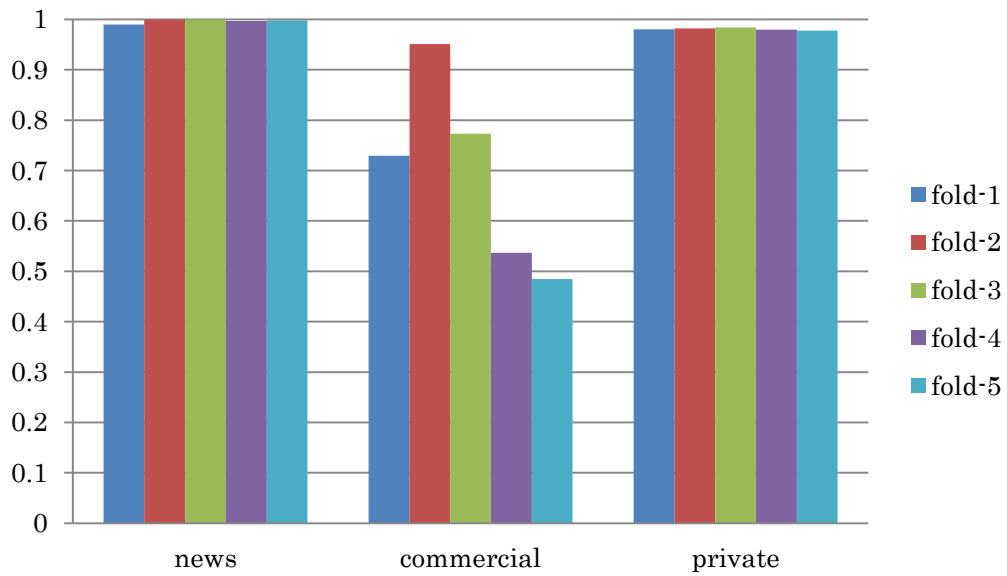
- I – BOW
- II – BOW +  $\ln(\text{friends})$  +  $\ln(\text{followers})$
- III – BOW +  $\ln(\text{friends})$  +  $\ln(\text{followers})$  + domain.

Values showed in the table are mean value of the 5-fold cross validation.

As we see from the table, we acquire high accuracy that is more than 0.85 and the F-measure value are over 0.80 as well. And we figured out that classifier using features of BOW and user information performs best among the three classifiers. When we did deeper analysis on each fold cross validation, we found recall value of Commercial category is not stable and low compared with other two categories in Figure 10.



**Figure 9. Precision in each fold cross validation on corpus1.**



**Figure 10. Recall in each fold cross validation on corpus1.**

For instance, recall values of fold-4 and fold-5 in Commercial category are only about 0.50. Meanwhile we figured out that many Commercial tweets are easily incorrectly assigned to Private and News category respectively which results in low precision value of fold-4 in Private and fold-5 in News category.

### 5.2.2 5-fold cross validation on corpus2

**Table 9. Results of 5-fold cross validation on corpus2.**

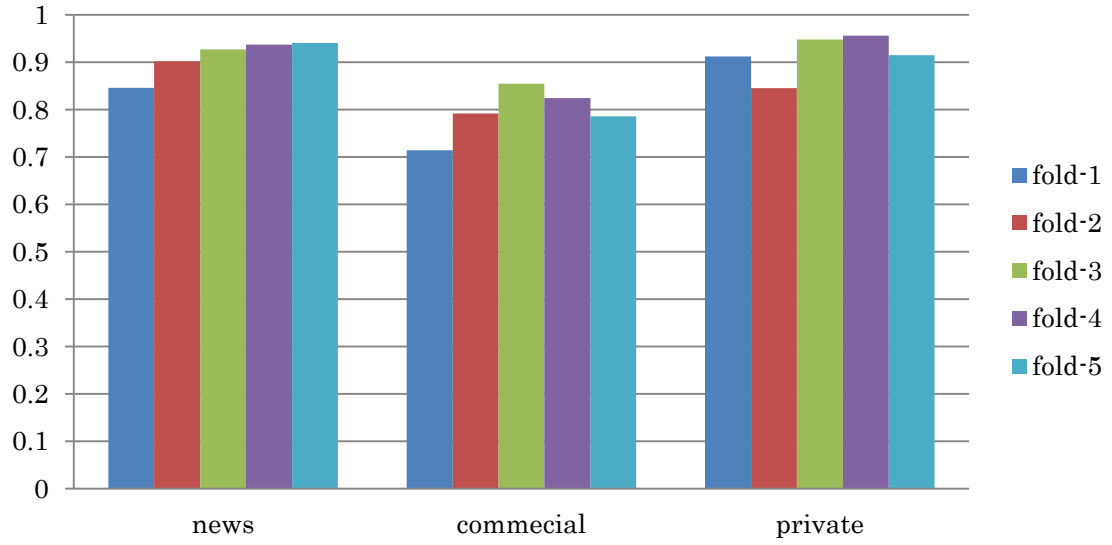
	Accuracy	Precision	Recall	F1
I	0.828	0.817	0.817	0.817
<b>II</b>	<b>0.875</b>	<b>0.873</b>	<b>0.873</b>	<b>0.873</b>
III	0.827	0.831	0.842	0.836

We did the same experiment as we did on corpus1 (Table 9).

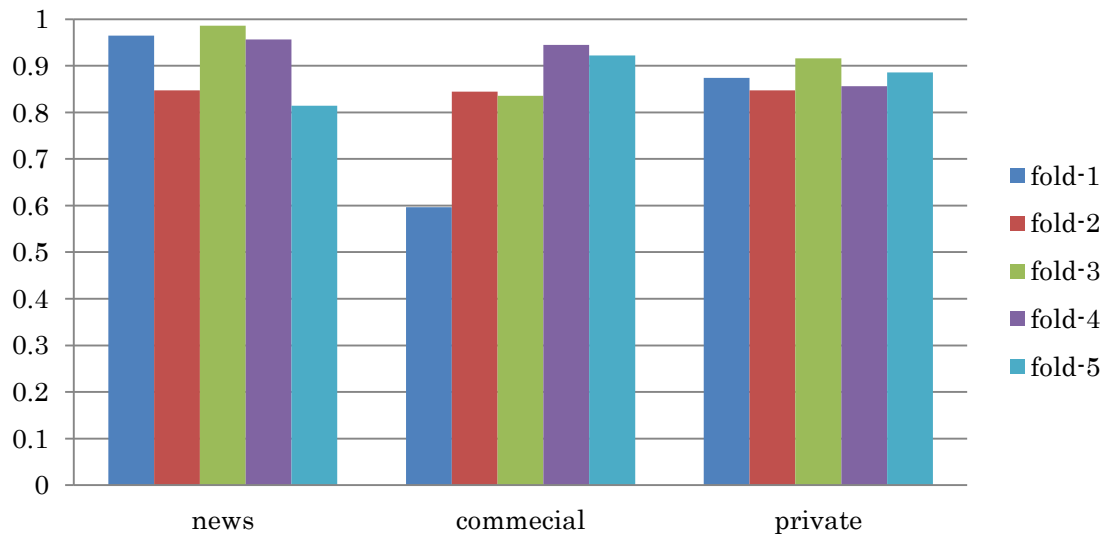
- I – BOW
- II – BOW +  $\ln(\text{friends}) + \ln(\text{followers})$
- III – BOW +  $\ln(\text{friends}) + \ln(\text{followers}) + \text{domain}$ .

As we see from the table, we acquire high accuracy and F-measure value as well like experiment on corpus1. And we also figured out that classifier using features of BOW and user information performs best among the three classifiers. When we focused on precision and recall value of each fold cross validation, we can see that recall value of Commercial category is stable (Figure 11 and Figure 12).





**Figure 11. Precision in each fold cross validation on corpus2.**



**Figure 12. Recall in each fold cross validation on corpus2.**

### 5.3 Classification experiments on test tweets

In the following experiments, we will show results of classification on test tweets that are labeled manually.

#### 5.3.1 Test tweets

We crawled test tweets from Twitter with 10 hot keywords in 2011, which are released by NEC Biglobe<sup>6</sup>: 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.

Since we have collected the test tweets, labeling them will be our next work. 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.

At first, we prepared a small dataset  $D_0$  of tweets which contained about 250 tweets collected by the 10 hot keywords. We set several demands for them to label:

- Follow definition of each category.
- Assign only one label (News, Commercial or Private) for one tweet.
- When labeling, first judge whether the tweets are Commercial or not, and then News, and at last Private. We made such order because of the difficulty of judging each category.
- Refer to the profile of users and pages of url link written in the tweets.

After they labeled the small set of tweets, here we calculated the Kappa value to evaluate their rate of concordance. Kappa value is a statistical measure of inter-rater agreement for categorical items. It is generally thought to be a more robust measure than simple percent agreement calculation since k value takes into account the agreement occurring by chance.

Suppose two raters (A and B) are asked to classify objects into categories 1 and 2. The table below contains the number of objects  $p_{ij}$  labeled  $i$  by rater B while

---

<sup>6</sup> [http://tr.twipple.jp/2011\\_top.html](http://tr.twipple.jp/2011_top.html)

**Table 10. An example to explain Kappa coefficient.**

	1	2	Total
1	$p_{11}$	$p_{12}$	$p_{1*}$
2	$p_{21}$	$p_{22}$	$p_{2*}$
Total	$p_{*1}$	$p_{*2}$	$P$

labeled  $j$  by rater A (Table 10).

To compute Kappa, you first need to calculate the observed level of agreement.

$$p_0 = p_{11} + p_{22}$$

This value needs to be compared to the value that you would expect if the two raters were totally independent,

$$p_e = p_{*1} \times p_{1*} + p_{*2} \times p_{2*}$$

The value of Kappa is defined as:

$$K = \frac{p_0 - p_e}{1 - p_e}$$

And there is a acknowledged interpretation of Kappa value: Poor agreement = less than 0.20; Fair agreement = 0.20 to 0.40; Moderate agreement = 0.40 to 0.60; Good agreement = 0.60 to 0.80; Excellent agreement = 0.80 to 1.00.

In our work, there are 3 people to label the tweets, so we calculated Kappa values of each two people and calculated mean value of them (Table 11). The mean value is 0.67 which means there is a good agreement among the raters.

**Table 11. Kappa value of the 3 raters.**

	A	B	C
A	1	0.77	0.55
B	0.77	1	0.68
C	0.55	0.68	1

**Table 12. Number of tweets in test tweets dataset.**

Tweets	News	Commercial	Private
1605	248	39	1318

Since the three raters have good agreement, it's acceptable for them to label our test tweets which contains about 1605 tweets showed in Table 12.

### 5.3.2 Performance comparison between corpus1 and corpus2

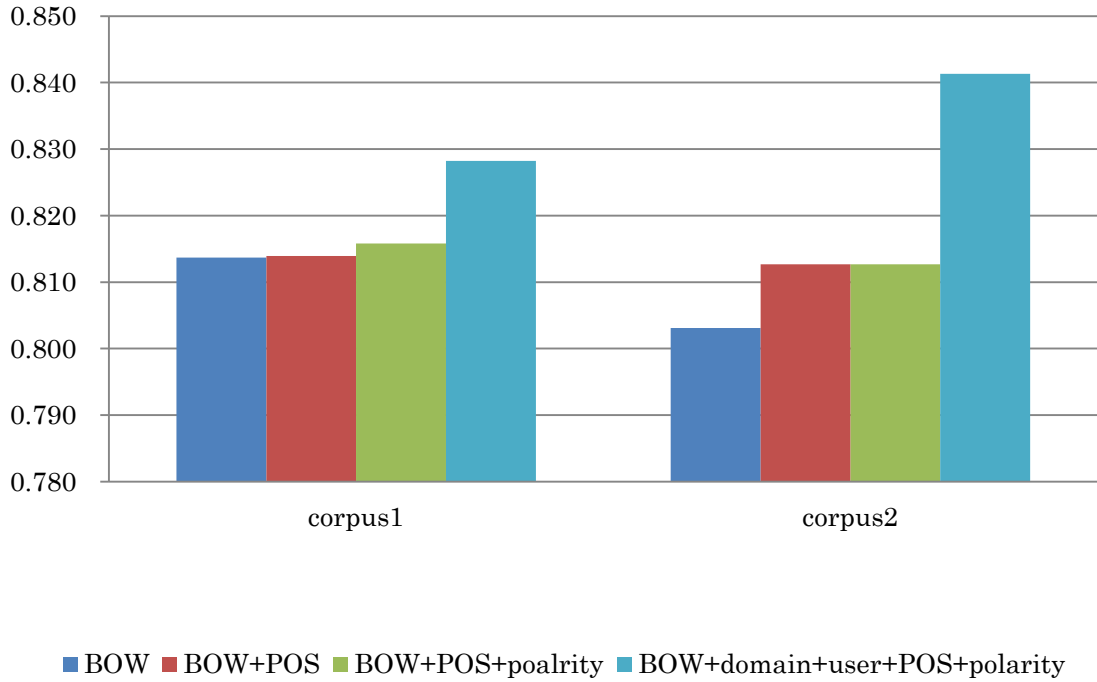
In the following experiments, we exploited many kinds of combinations of the features and we did comparison experiments on corpus1 and corpus2. The features we exploited are following:

- I – BOW
- II – BOW + parts of speech
- III – BOW + parts of speech + polarity
- IV – BOW + parts of speech + polarity +  $\ln(\text{friends})$  +  $\ln(\text{followers})$  + domain

Our experiment results are showed in the Table 13 and Figure 13.

**Table 13. Results of performance comparison con corpus1 and corpus2.**

Accuracy	I	II	III	IV	Mean
Corpus1	0.814	0.814	0.816	0.828	0.822
Corpus2	0.803	0.813	0.813	<b>0.841</b>	0.817



**Figure 13. Results of performance comparison on corpus1 and corpus2.**

Here we use accuracy to evaluate our experiments. In Table 13 and Figure 13, we list each value of accuracy under each combination on corpus1 and corpus2. As we can see from the results:

1. Our two classifiers trained by corpus1 and corpus2 both acquired high performance of an average accuracy value more than 0.80.
2. As we exploited more and more complicated features, the value of accuracy did go up which means that not only the linguistic features, but the distinctive features of Twitter are effective as well.
3. Although classifiers trained by corpus1 and corpus2 both performed well, classifier1 seems to perform better than classifier2 by 0.05 at average value of

accuracy. When we look at the case of best performance for corpus1 and corpus2, there is no any difference between them. Therefore we conclude that classifiers trained by corpus1 and corpus2 have nearly equal performance.

### 5.3.3 Experiments exploiting over sampling method

The performance of SVM drops significantly while facing imbalanced datasets. Some studies have pointed out that it is difficult to avoid such decrease when trying to improve the efficient of SVM on imbalanced datasets by modifying the algorithm itself only. Therefore, as the pretreatment of data, sampling is a popular strategy to handle imbalance dataset problem since it re-balances the dataset directly. In our work, we tried to perform the over-sampling method to improve the efficient of SVM. The method proceeds as follow:

- I. Set the small dataset  $D_0$  containing 250 tweets as a development dataset to tune the parameter of penalty  $c$  in SVM.
- II. Perform 5-fold cross validation on our dataset of test tweets adding corpus1 and corpus2. In each classifier training process, we use the penalty  $c$  which is tuned on the small dataset  $D_0$ .
  - II.1 Perform 5-fold cross validation on test tweets only.
  - II.2 Perform 5-fold cross validation on test tweets adding corpus1 and corpus2 respectively. In this process, we adjust the size of corpus1 and corpus2 in order to keep the proportion of the three categories in test tweets.
  - II.3 Increase the size of test tweets to 5 times size by the oversampling method (copy the original tweets). Then perform 5-fold cross validation on test tweets adding corpus1 and corpus2 respectively.
  - II.4 Increase the size of test tweets to 10, 15, 20 and 25 times size by the oversampling method. Then perform 5-fold cross validation on test tweets adding corpus1 and corpus2 respectively.

We will show the results as follow:

- BOW (Figure 14)
- BOW + POS + polarity (Figure 15)
- BOW + parts of speech + polarity +  $\ln(\text{friends})$  +  $\ln(\text{followers})$  + domain (Figure 16)

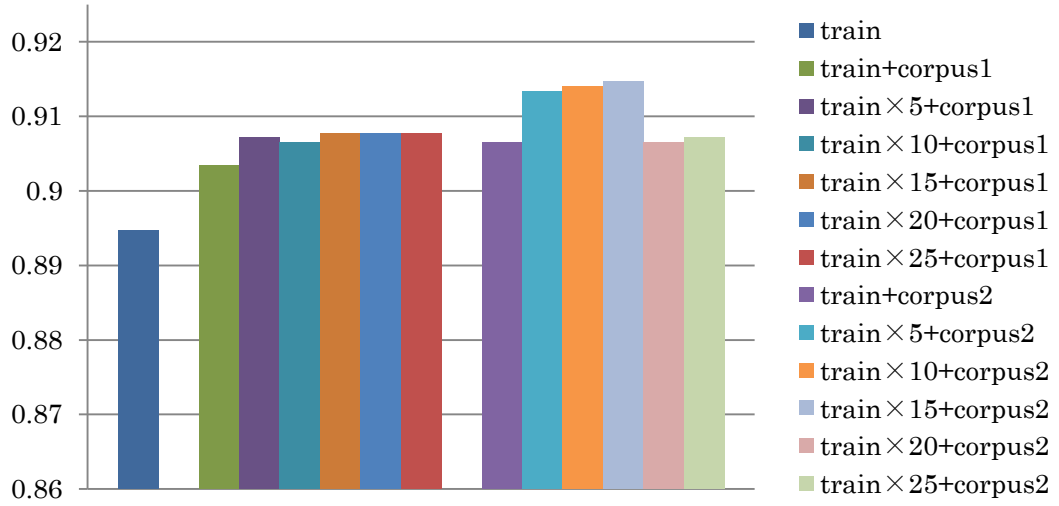


Figure 14. Results of oversampling method exploiting BOW.

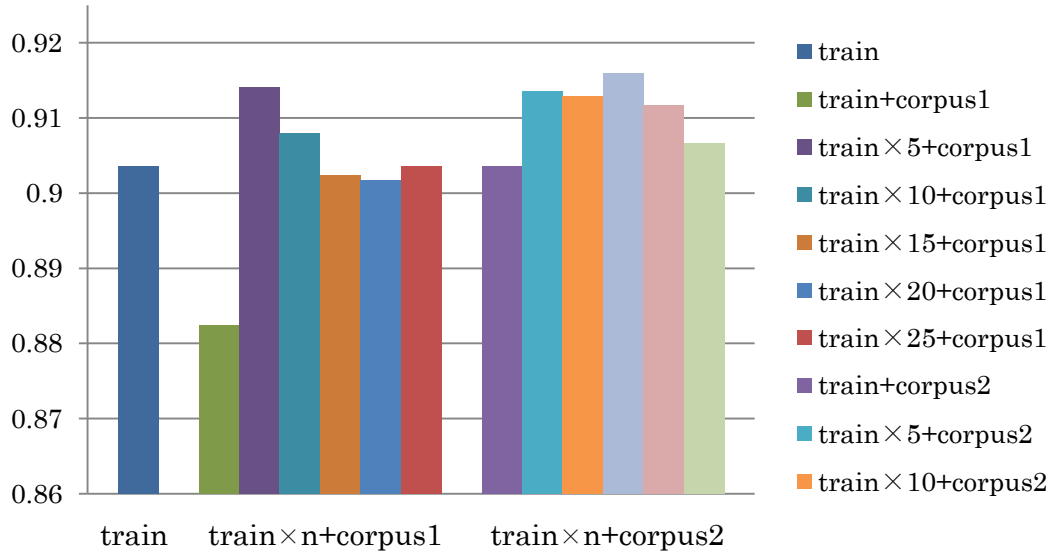
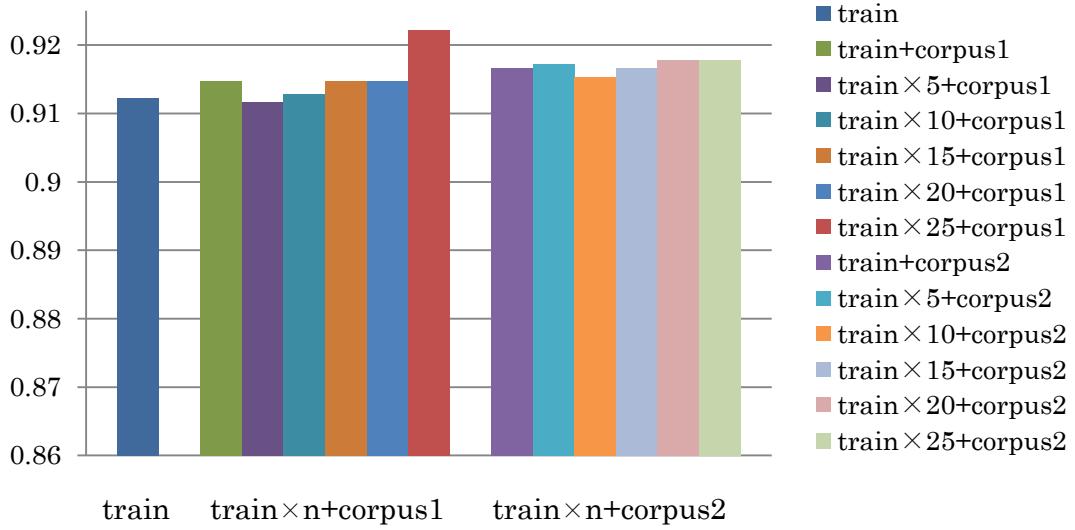


Figure 15. Results of oversampling method exploiting BOW + POS +polarity.



**Figure 16. Results of oversampling method exploiting BOW + POS + polarity  $\ln(\text{friends}) + \ln(\text{followers}) + \text{domain}$ .**

As we can see from the three figure (from Figure 14 to Figure 16) 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 Figure 16,  $\text{train} \times 25 + \text{corpus1}$ ) to show the details of the classified tweets.

**Table 14. The confusion matrix of the best case.**

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

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. But the Commercial tweets are still hard to classify correctly.



## Discussion

In our experiments, we found that many Commercial tweets are classified into other two categories, especially the News category, which results in low recall value of Commercial category and low precision value of News and Private categories. We have two interpretations to explain this problem:

- Although we restricted that one tweet only can be assigned one label, some tweets simultaneously have characteristics of news and commercial tweets. For instance: “【予約開始】ワンピース D.P.C.F シリーズ第6弾 / トニートニー チョッパー ウエスタン Ver. <http://t.co/rAO57pr>” is a Commercial tweet in our dataset, meanwhile it can be regarded as a news of a product release. Actually, it's more suitable to actual circumstances if we can assign a tweet by two labels.
- Tweets are so small in text size that it's hard to classify them by contents of the tweets. In our work, we exploited features of number of user's friends and followers, but in the datasets number of accounts in each category is imbalanced. Accounts of Private category are more than that in News and Commercial categories, which may cause a bias in the classification.

What's more, some News and Private tweets are misclassified to each other for some of them share overlapping characteristics. For instance, experiences of people sometimes can be regarded as news like “重大！ フランス放射線防護委員会が福島避難圏の外側に激しい汚染地域があり、さらに7万人以上の避難が必要と通告した！”. In some cases, it's hard to make a distinction between News tweets and Private tweets.

At last, we build a tool to show the result of labels which each tweet is assigned by. From , I show classified tweets collected by keywords “地震”, “AKB” and “CM”. From classified tweets showed in the News and Commercial categories, we succeed to classify some public information into the two categories.

News	Commercial	Private
 <p>【気象庁速報】03日 19時 34 分頃 最大震度 4 の地震が発生。 ( <a href="http://t.co/akoLdVEB">http://t.co/akoLdVEB</a> ) 震度 4: 茨城県南部他 / 震度 3: 茨城県北部他 #saigai #eqjp #earthquake #jishin</p>	 <p>気象庁発表の震度でなく、この建物のこの場所の震度が知りたい。そんな時に「フリースタイル 家庭用地震計 グラグラフ」を Amazon でチェック! <a href="http://amzn.to/gnKzQH">http://amzn.to/gnKzQH</a></p>	 <p>[渡り廊下走り隊] 心配です。なっちゃん: 日本ですごい地震があったことをTV のニュースで知りましたが 火災なども起きてるみたいですごく心配です</p>
 <p>拡散希望 東松島市宮戸室浜避難所 炊き出し募集 100 名分 支援物資 カットパン 缶詰 カップラーメン 直接地震避難所に宅急便で支援可能ですその際窪田和美からと送り状に書いて頂ければ受け取り可能です 八月中旬まで受付しております 室浜避難所一階石田様宛 お願いします</p>	 <p>■東北地方太平洋沖地震被災者支援チャリティーライブ ■3/30(水)『ここにあかりが灯るまで』knave ■入場無料 <a href="http://p.twipple.jp/AZpch">http://p.twipple.jp/AZpch</a></p>	 <p>今の地震発車前の電車の中だったけど、電車はだいぶ揺れますね。怖かったです</p>
 <p>【速報】福島第一原発と福島第二原発から、半径 20 キロの範囲に新たな避難指示の要請</p>	 <p>真夏の節電対策に通気性の優れた生地の商品が人気です。タイで生まれた通気性抜群のオーガニックコットンを使った「ブリークurai」のワンピースやタイパンツがオススメ! 今年の夏はリラックスファッションで夏を乗り切りましょう♪ <a href="http://ow.ly/5bNsP">http://ow.ly/5bNsP</a> #mitozakka #節電 #地震</p>	 <p>地震大丈夫ですか、3 時津波来るっていつてるから気をつけて!!</p>

Figure 17. Classified tweets collected by “地震”.










News	Commercial	Private
 <p>AKB48「Everyday、カチエーシヤ」撮影時使用衣装 10億円突破  <a href="http://bit.ly/n2QdVa">http://bit.ly/n2QdVa</a></p>	 <p>Video:【PV】大声ダイヤモンド / AKB48 [公式]          (by AKB48)  <a href="http://tumblr.com/xoz31xdvmy">http://tumblr.com/xoz31xdvmy</a></p>	 <p>Hey! Say! JUMP and AKB48 appear on MUSIC STATION 25th sp. tonight. #Hey! Say! JUMP</p>
 <p>30位 大矢 29位 大家          28位 山本彩 27位 みゃ          お 26位 平嶋 25位 多          田 24位 仲川 23位 高          柳 22位 梅田          #AKBvote2011</p>	 <p>8/22 20時30分～ アキバから生放送 ※秋葉原のレンタルスペース AKB162からアキバで働く生主とアキバの仲間がアキバからニコ生放送! ※観覧無料!          #col148629  <a href="http://t.co/7sflnLk">http://t.co/7sflnLk</a> ※観覧 入場 20:15~ 退場 21:45まで。お題は秋葉大喜利</p>	 <p>【拡散希望】          『AKB48?AKBのスカイツリーは誰?〜』選挙を開催します♪          ※参加メンバー前田敦子・小嶋陽菜・宮澤佐江・大島優子・柏木由紀・渡辺麻友          ※投票は1人1票1回だけ※気軽に投票お願いします※投票は@ryuLOVEakbまで?</p>
 <p>「AKB48、NMB48から謹慎続出」  <a href="http://t.co/y9jysbk">http://t.co/y9jysbk</a>          #yjfc_akb48 (AKB48)</p>	 <p>【秋元康氏「やっぱり、きれいだ」AKB こじはるハワイで水着や生着替え】  <a href="http://shr.im/016F">http://shr.im/016F</a> 小嶋陽菜 大胆ソロ写真集  <a href="http://amzn.to/i1qq21">http://amzn.to/i1qq21</a>          #geinou</p>	 <p>@AKBfafaafan なんぞ? AKESHOP?原宿のは当たらないと行けないよ?</p>

Figure 18. Classified tweets collected by “AKB”.

News		Commercial		Private	
	<p>〈カップヌードル〉新CMは井上雄彦「バガボンド」&amp;ザ・クロマニヨンズを起用  <a href="http://tower.jp/article/news/76903">http://tower.jp/article/news/76903</a></p>		<p>【CM】オリジナル・ラブ約5年ぶりのニューアルバム『白熱』好評発売中！珠玉のポップスをあなたの生活に。  <a href="http://amzn.to/n0fg9P">http://amzn.to/n0fg9P</a>  <a href="http://bit.ly/qM5wzn">http://bit.ly/qM5wzn</a>            iTunes Storeから配信も。</p>		<p>これをTV等メディアでCMがわりに使いましょう            @takapon_jp            @tnatsu</p>
	<p>『AKB48 “CM選抜総選挙”を実施！AKB、SKE、NMB初共演CMをファン投票で制作』  <a href="http://t.co/RnmvtIB">http://t.co/RnmvtIB</a></p>		<p>Fairlight CMI アプリがiPhoneとiPadに登場：伝説のFairlight Computer Musical InstrumentがiPhoneとiPadに登場しました。80年代の有名なアーティストと同じ...  <a href="http://bit.ly/e0oeA0">http://bit.ly/e0oeA0</a></p>		<p>テレビで使う電力の約40%は利権と無駄に支えられています。QT            @tosa_guigei 東京電力テレビCM  <a href="http://is.gd/WCM7r0">http://is.gd/WCM7r0</a>  <a href="http://is.gd/N0oCwJ">http://is.gd/N0oCwJ</a> 首都圏で使う電力の約40%は福島と新潟が支えています。</p>
	<p>向井理:レッチリ新作のCMに出演 エアベースを披露 上司に土下座も...  <a href="http://t.co/RFpVzCL">http://t.co/RFpVzCL</a></p>		<p>桑田佳祐 sg 発売決定！docomo CM曲とチーム・アミューズ！！“LET'S TRY AGAIN”の別 Ver などを収録したトリプルA面sgです。予約受付開始しました♪  <a href="http://bit.ly/p3MFLR">http://bit.ly/p3MFLR</a> 今なら旧譜作品購入でうちわ差し上げてます【ラゾーナ川崎&gt;伯爵令嬢】</p>		<p>JRの新幹線CMのリレーが何かいい感じだなあ  <a href="http://ow.ly/1sLKwA">http://ow.ly/1sLKwA</a></p>

Figure 19. Classified tweets collected by “CM”.

## 6 Conclusion and future work

Twitter is an ideal environment for the dissemination of public information directly from the news source to the geographical location of events. Some of them contain important and valuable information 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 so messy even on the same one subject that public information is hard to provide directly for people. Therefore, in the research, we propose a new concept: Information Publicness. According to the content, we can divide tweets into two parts: tweets with publicness and tweets without publicness. We name tweets without publicness “Private tweets”. And we subdivided tweets with publicness into two parts: tweets for profit and tweets for non-profit, which we call Commercial and News tweets respectively. In order to support user’s searching on Twitter, we set a task that is how to classify tweets based on information publicness: news, commercial messages and private tweets.

In our research, we focus on approaches of collecting automatic corpora for training classifiers. We proposed two approaches: one is based on a small number of typical Twitter user accounts, and the other one is an expansion of the first one that is 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. As we exploited more and more complicated features, the value of accuracy did go up which means that besides the linguistic features, especially the distinctive features of Twitter are effective for achieving the best accuracy in the experiments. And in the experiments exploiting over sampling, 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.

For the future work, we can develop our study to classify tweets into more specific categories. For instance, the Private category can be subdivided 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.

## References

1. TaylorChris. Social networking 'utopia' isn't coming. : CNN. Retrieved December 14, 2011, 2011.
2. McDonoughNg K., Jeanrenaud P., Gish, H., Rohlicek J.R.J.,. Approaches to topic identification on the switchboard corpus. : Acoustics, Speech, and Signal Processing, 1994. ICASSP-94., 1994 IEEE International Conference, 1994.
3. HsuehP.-Y. , MooreJ.D. AUTOMATIC TOPIC SEGMENTATION AND LABELING IN MULTIPARTY DIALOGUE. : Spoken Language Technology Workshop, 2006. IEEE , 2006.
4. Bo PangLee and Shivakumar VaithyanathanLillian. Thumbs up?: sentiment classification using machine learning techniques.: Proceeding EMNLP '02 Proceedings of the ACL-02 conference on Empirical methods in natural language processing - Volume 10, 2002.
5. D. IraniWebb, 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.
6. Alexander PakParoubekPatrick. Twitter as a Corpus for Sentiment Analysis and Opinion Mining. : Proceedings of the Seventh conference on International Language Resources and Evaluation LREC, 2010.

7. BanerjeeRamanathan, K., and Gupta, AS.,. Clustering short text using Wikipedia. : Proc. SIGIR, 2007.
8. 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.
9. Takeshi SakakiOkazaki , Yutaka MatsuoMakoto. Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors. : In Proceedings of the Nineteenth International WWW Conference, 2010.
10. SankaranarayananSamet, H., Teitler, B. E.,J.,. TwitterStand: news in tweets. : In Proc. ACM GIS, 2009.
11. Alec GoBhayani, and Lei HuangRicha. Twitter Sentiment Classification using Distant Supervision. : 2011 analysis semantic seminar talk twitter winter by coca.alina, 2009.
12. Yegin GencSakamoto, and Jeffrey V. NickersonYasuaki. Discovering Context: Classifying Tweets through a Semantic Transform Based on Wikipedia. : D.D. Schmorrow and C.M. Fidopiastis (Eds.): FAC 2011, HCII 2011, LNAI 6780, pp. 484–492, 2011., 2011.
13. Bharath SriramFuhry, Engin Demir, Hakan Ferhatosmanoglu and Murat DemirbasDavid. Short Text Classification in Twitter to Improve Information Filtering. : SIGIR'10, July 19–23, 2010, Geneva, Switzerland., 2010.
14. WuHofman, J.M., Mason, W.A., Watts, D.J.S.,. Who Says What to Whom on Twitter. : 20th Annual World Wide Web Conference, ACM, 2011.
15. Xiaojin ZhuGhahramaniZoubin. Learning from Labeled and Unlabeled Data



with Label Propagation. : Technical Report CMU-CALD-02-107, 2002.

16. Steven BirdKlein, Edward LoperEwan. Natural Language Processing. 2001.

17. VapnikVladimir. The Nature of Statistical Learning Theory, . : Springer, 1995.

## Publications

1. Hongguang Zheng, Nobuhiro Kaji, Naoki Yoshinaga, and Masashi Toyoda. Classification of Microblog Posts Based on Information Publicness. The 4th Forum on Data Engineering and Information Management (DEIM2012) (to appear).
2. 鄭 洪光, 鍛冶 伸裕, 吉永 直樹, 豊田 正史. 典型的情報発信者に着目した Microblog 分類器学習の低コスト化に関する一考察. 日本データベース学会第二回ソーシャルコンピューティングシンポジウム(SoC2011), ポスター11
3. 鄭 洪光, 鍛冶 伸裕, 吉永 直樹, 豊田 正史. Twitter List を用いた低コストな Tweet 分類器の学習. 言語処理学会 NLP 若手の会第 6 回シンポジウム(YANS2011)