Classifying Business Types from Twitter Posts Using Active Learning

Chanattha Thongsuk, Choochart Haruechaiyasak, Phayung Meesad

Department of Information Technology
Faculty of Information Technology
King Mongkut’s University of Technology North Bangkok (KMUTNB)
1518 Pibulsongkarm Rd., Bangsue, Bangkok 10800

Human Language Technology Laboratory (HLT)
National Electronics and Computer Technology Center (NECTEC)
Thailand Science Park, Pathumthani 12120, Thailand

Department of Teacher Training in Electrical Engineering
Faculty of Technical Education
King Mongkut’s University of Technology North Bangkok (KMUTNB)
1518 Pibulsongkarm Rd., Bangsue, Bangkok 10800

s4970290054@kmutnb.ac.th
choochart.haruechaiyasak@nectec.or.th
pym@kmutnb.ac.th

Abstract: Today, many companies have adopted Twitter as an additional marketing medium to advertise and promote their business activities. One possible solution for organizing a large number of posts is to classify them into a predefined category of business types. Applying normal text categorization technique on Twitter is ineffective due to the short-length (140-character limit) characteristic of each post and a large number of unlabeled data. In this paper, we propose a text categorization approach based on the active learning technique for classifying Twitter posts into three business types, i.e., airline, food and computer & technology. By applying the active learning, we started by constructing an initial text categorization model from a small set of labelled data. Using this text categorization model, we obtain more positive data instances for constructing a new model by selecting the test data which are predicted as positive. As shown from the experimental results, our proposed approach based on active learning helped increase the classification accuracy over the normal text categorization approach.
1 Introduction

Twitter, a well-known micro-blogging website, has recently gained a lot of popularity among the Web 2.0 community. Increasingly, many businesses use Twitter as a new channel to promote their products and services including other related activities. For example, many airlines use Twitter to post special flight discount or promotions for their followers. As with many social networking websites, Twitter is considered an important part of Web 2.0 community. Web 2.0 is a departure from traditional websites, and represents a large Internet social networking group which is constantly collecting a lot of online information. In a social networking website, people are allowed to follow other users based on their personal interests. Advertising on social networking websites is growing and interesting because it can reach a lot of customers with low overhead costs.

Twitter provides an attractive platform for advertisers to promote their brands. The customer will get information and promotions from the companies. Moreover, the customers can post their opinions or complaints to the companies. Therefore, Twitter acts as a third-party provider where partners may place advertisements on their products and services. Today many brands and companies are using Twitter to advertise, get feedback from the customers and gain more revenue.

With a large number of posts, one approach for organizing them is to apply a text categorization model. Previous research works on text categorization considered textual documents such as news articles, publications and web pages. These documents typically contain a large number of words in the range of hundreds or thousands of words. Applying a text categorization model for Twitter is very challenging due to the following reasons.

1. Twitter is a micro-blogging website which allows only short posts of no more than 140 characters. The average number of terms in each post from our corpus is approximately equal to 12.
2. Most posts are often colloquialism and consist of acronym.
3. There are a lot of the junk posts.

To construct a text categorization model, training data set is required. However, to prepare a large labeled data set (assigning each document with a class label) is time consuming and expensive. One approach to improve the performance of traditional supervised learning is by applying the active learning technique. In this paper, we apply the active learning to automatically increase the number of labelled training instances for classification. We use Twitter posts from three business types, i.e., airline, food and computer & technology. Starting with a small size of training data set, an initial classification model is constructed. To increase the training data size, we iteratively accumulate more posts which are classified with positive label. Once more posts are automatically obtained, we construct a new improved model. The final text categorization model can be used to organize the posts into different business types.
The rest of this paper is organized as follows. Section 2 gives a review of previous research works related to different machine learning approaches especially the active learning technique. Section 3 describes our solution based on the active learning approach. Section 4 presents the evaluation results of our experiments. Finally, Section 5 depicts our conclusions.

2 Related Works

Text categorization is a well-known and widely applied machine learning technique for classifying textual documents into a predefined set of categories. Previous text categorization approaches were applied on document corpora such as news articles and web pages. The typical document corpus contains a large labelled data set in which each document instance contains hundreds or thousands of terms. One of the problem issues in text categorization is the preparation of labelled data set. To construct an effective model, a large number of documents is needed. However, manually labelling each document instance into a predefined category set is very time consuming.

There have been some solutions proposed in previous works. Cabrera Rafael Guzman et al. [CG08; CRG09] proposed the automatic extraction of unlabeled examples from the web using self-training approach to classify documents without requiring a predefined set of unlabeled data. In addition, they also proposed a method using the semi-supervised learning to solve an ambiguous word to the correct sense using unlabeled examples extracted from the web. Cheng Yong et al. [CZ09] proposed self-training classifier based on local consistency eliminating the noise and discovering the unlabeled data to join the training set by classifier to be consistent with the local neighborhood. Watson Rebecca et al. [WBC07] applied the statistical parser with sentences partial-bracketing self-training. Mao Ching-Hao et al. [MLP09] proposed a co-training method for multi-view intrusion detection and identify unlabeled data in ambiguous parts. Mojdeh Mona et al. [MC08] applied the semi-supervised learning to spam filtering. Wu Xianchao et al. [WOT09] to mining Chinese-English lexicons from large amounts of Chinese Web pages but the algorithm still had the problem of appending new mined entries into the existing seed lexicon. Yang Bishan et al. [YSW09] predicted the possible labels of the unlabeled data and the expected loss multi-label data according to the confident result of label prediction. Zheng Yabin et al. [ZTL08] discovered the constant common knowledge and built a model to fit the distribution test set by adding confident instance of unlabeled test set to training set until convergence.

We apply the active learning concept to improve the performance of classification algorithm using a small initial training set and built a better classifier.
3 The Proposed Active Learning Solution

To construct an effective classification model, a large sample size of labelled training data set is usually required. To perform data selection, normal sampling methods are not very effective, since some instances with incorrect class labels could be added into the corpus. In this paper, we focus on finding a better solution for effectively selecting high-quality training data instances. We apply the active learning concept which can help increase the performance of classification algorithm using a small initial training set to iteratively increase more training data instances.

The details of the proposed active learning technique are as follows.

1. Build the elementary classifier of each corpus using a small initial training set.
2. Classify the unlabeled set using a classification algorithm.
3. Select \( N \) instances (posts) per class from the unlabelled data set pool.
4. Append the selected instances into the initial training set.
5. Build the classifier using the improved training set and evaluate the classification accuracy at that point. (The points refer to amount of training instances after combined training set.)
6. Repeat Step 2 until the combined training set is complete.

Figure 1: The Proposed Active Learning Process.
In this paper, we observed six iterations for evaluating the trend of the classification accuracy of each business type, e.g., 500, 600, 700, 800, 900 and 1,000 instances per class.

4 Experiments and Discussion

This section presents the experimental evaluation of the proposed method. This evaluation was carried out in three different business types in Twitter such as “airline”, “food”, and “computer & technology”.

Data Collection. We performed experiments using a collection of direct posts in microblogging of followers obtained from a selection of ten companies, which are screen names in Twitter, of the three business types: airline, food, and computer & technology. Direct Post is composed of a screen name of the receiver, which appears between the symbols “@” and wrote post together, for example, if the screen name is Google then the screen name of the receiver is @Google@. We collected the follower list of each company using Twitter API and collected direct posts from follower’s blogs using java application. In this paper, we collected the direct posts from only ten companies. And we prepared the initial training set for each corpus consisting of 1,000 posts (500 direct posts for the class label to itself, e.g., airline, food, and computer & technology and 500 normal posts labeled with “other” class). The initial training set will be plotted in the first iteration.

Data Pre-Processing. We removed all punctuation marks, numerical symbols and screen name of the receiver (consists with “@” symbol and receiver’s screen name) i.e., “@Google@”. After that, we converted all words to lowercase.

Learning Algorithm. We selected Support Vector Machines (SVM) to classify business type from the original corpus and the manual selected corpus, because, based on state-of-the-art literature, the SVM algorithm is an appropriate algorithm for many applications. We used the WEKA machine-learning environment as the experiment tool.

Active Learning Experiment. We separated the experiment into two sections. The first is the experiment for the manual method and the second is the experiment for the active learning technique.

For the initial iteration, we selected 500 instances for the individual class and 500 instances for the “other” class in the initial training set. We inputted the initial training set to the manual method of each corpus and separate our experiment into two experiment sections. Both experiments received the same data (the initial training set) then built the elementary classifier of each corpus using a small initial training set. Therefore, the classification accuracy values of the first iteration from the manual method and active learning technique all have the same values.
For the next iteration, additional instances are increased in each iteration. The manual method, we selected 100 instances per class from the corpus pool that have the relevant terms related to the business type from each corpus, and then these are inserted into the manual method experiment corpus. But in the active learning, we classify the unlabelled instances from the corpus pool using the elementary classifier and selected 100 instances per class which are predicted as positive to insert into the active learning experiment corpus. After that, we rebuild classifier and used the SVM algorithm to classify the instances in both active learning and manual experiments to evaluate the classification accuracy of each iteration. We repeat the same process until the number of positive training instances reaches 1,000 instances per class. (Each completed corpus has 1,000 instances of the “positive” class and 1,000 instances of the “other” class)

The results of the both experiment sections are as follows.

![The result of Active Learning using SVM algorithm from the airline corpus](image)

Figure 2 : The graph presents the classification accuracy comparison between the manual method and active learning method of the airline corpus using SVM Algorithm.

The trend of the airline corpus using the highest accuracy with the SVM algorithm was stable. Although the future, we will increase the new instances to appended to the initial training set continuously and using active learning technique. The classification accuracy was not lower.
<table>
<thead>
<tr>
<th>Airline</th>
<th>Accuracy (%)</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DT</td>
<td>NB</td>
<td>SVM</td>
</tr>
<tr>
<td>Original</td>
<td>65.8</td>
<td>78.5</td>
<td>85.4</td>
</tr>
<tr>
<td>Active Learning</td>
<td>78.4</td>
<td>85.8</td>
<td>98.3</td>
</tr>
<tr>
<td>Manual</td>
<td>83.1</td>
<td>87.8</td>
<td>94.9</td>
</tr>
</tbody>
</table>

Table 1: The classification accuracy (F1-Measures) of the six iteration (1,000 instances per class) of the airline corpus.

The classification accuracy comparisons of three algorithms (DT, NB and SVM) in the airline corpus are shown in Table 1. The classification accuracy of the active learning using SVM Algorithm is higher than the original method up to 12.9% (The original method was the random selection) which is still higher than the manual method, up to 3.4%. The results of Active Learning using Decision Tree and Naïve Bayes of each corpus shows high percentage of classification accuracy based on the last iteration (at 1,000 instances per class) from Active Learning.

![The result of Active Learning using SVM algorithm from the food corpus](image)

Figure 3: The graph presents the classification accuracy comparison between the manual method and active learning technique of the food corpus using SVM Algorithm.

The trend of the active learning of food corpus is very fast growing when the data set has a lot of instances.
Table 2: The classification accuracy (F1-Measures) the six iteration (1,000 instances per class) of the food corpus.

Table 2 shows the classification accuracy comparisons among the original, active learning, and the manual methods in the food corpus. The classification accuracy of the active learning is higher than the original method up to 18.4% and higher than the manual method which is up to 8.9%.

The result of Active Learning using SVM algorithm from the computer & technology corpus

Figure 4: The graph presents the classification accuracy comparison between the manual and active learning technique of the computer & technology corpus using SVM Algorithm.

The trend of the active learning of the computer & technology corpus was steadily growing like the airline corpus but smooth and continues unlike the airline corpus.
<table>
<thead>
<tr>
<th>Computer &amp; Technology</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DT</td>
</tr>
<tr>
<td>Original</td>
<td>64.6</td>
</tr>
<tr>
<td>Active Learning</td>
<td>74.8</td>
</tr>
<tr>
<td>Manual</td>
<td>77.5</td>
</tr>
</tbody>
</table>

Table 3: The classification accuracy (F1-Measures) of the six iteration (1,000 instances per class) of the computer & technology corpus.

Table 3 shows the classification accuracy comparisons of the three methods perform on the computer & technology corpus. The classification accuracy of the active learning using the SVM algorithm is higher than the original method up to 16.3% and higher than the manual method up to 7.6%.

From the results of the experiments, the active learning technique of direct posts in Twitter can improve the classification accuracy higher than the original method (random selecting) and still higher than the manual method by humans selection by observing the relevant words that appear in the posts.

5 Conclusion and Future Works

In this paper, we focus on the problem issue of data selection for building a classification model. The problem is due to, firstly, the extremely small labelled training set and, secondly, acquiring a large high-quality training set with high overhead costs. We applied the active learning technique to solve those problems. The additional training instances were acquired using the proposed active learning method to append the initial training set with 500 instances per class for improving the classification accuracy. From the experimental results, the performance of active learning method, based on the F1 measure, is higher than the original method (random selection). The three data sets: airline, food, and computer & technology, active learning method improves 12.9%, 18.4%, and 16.3%, respectively, the active learning method performs better with higher accuracy than the manual method (manually selection) up to 3.4%, 8.9%, and 7.6%.

For future works, we plan to change the number of initial training set, i.e., 100 instances per class, initial training set is a key factor to grow the classification accuracy rate, and maybe use the Latent Dirichlet Allocation (LDA) algorithm to build the topic model classifier and the training set from the active learning technique to improve the classification accuracy.
References


