Can we predict political poll results by using blog entries?

Manabu Okumura, Tetsuya Motegi, Tokyo Institute of Technology
{koku, motegi}@lr.pi.titech.ac.jp
Tetsuro Kobayashi, Keizo Oyama, National Institute of Informatics
{k-tetsu, oyama}@nii.ac.jp
Takahisa Suzuki, The Graduate University for Advanced Studies
s-takahisa@nii.ac.jp

Abstract
Blogs have become an important medium for people to publish their opinions and ideas on the Web. However, it is still not clear whether we can analyze political public opinions from blogs. There have been some recent work on political viewpoint classification, but most only classified political blog entries or sites into opposing viewpoints such as conservative/liberal or Israeli/Palestinian. However, to predict a broader range of political opinions, we need to analyze a wide variety of blogs. Therefore, we constructed a dataset of general blogs that are connected to political poll results. With the dataset, we conducted experiments to predict political poll results by using the blog entries. Our prediction methods are based on a supervised learning algorithm, Support Vector Machines (SVM), with features in blog sites. We also attempted manual prediction with three human subjects as the upper bound of the system performance, and found that such a task is rather difficult even for humans and that the system performance can outperform that of humans.

1. Introduction
Blogs have become an important medium for people to publish their opinions and ideas on the Web. Individuals and organizations have been interested in information on the blogosphere, and have analyzed reputation of products and services on the Web. However, it is still not clear whether we can analyze political public opinions from blogs.

There have been some recent work on political viewpoint classification. One of the earliest work was done by Lin [6], who used the bitterlemon corpus consisting of Israeli and Palestinian author editorials. Malouf and Mullen [7] did a similar work by classifying users of an online community according to their political viewpoints. Kale [5] presented political-viewpoint classification based on link analysis. For classifying a political blog as either liberal or conservative, Jiang and Argamon [4] presented a subjectivity analysis-based method. They first identified subjective sentences and extracted features to build political-leaning classifiers from subjective sentences. They showed that using features from subjective sentences can significantly improve the classification performance. While Jiang and Argamon [4] presented the work that classifies blog sites, Oh et al. [8] presented the work on blog entries classification. Durant and Smith [1] also explored techniques for predicting the political orientation (liberal or conservative) of blog entries. Efron [2] utilized co-citation information to estimate the political orientation of blog sites.

Most of the previous work only classified blog entries or sites into opposing viewpoints: conservative/liberal, Democratic/Republican, Israeli/Palestinian, or left/right. In this work, we analyze finer-grained differences in political public opinions. The target of classification in the previous work was restricted to only political blogs. However, to analyze more general public opinions, we need to analyze a wider variety of blogs. Therefore, as we explain in the next section, we constructed a dataset of general blogs that are connected to political poll results. We asked bloggers on Yahoo!blog (http://blogs.yahoo.co.jp/) to be poll respondents and provide all their blog entries.

With this dataset, we conducted experiments to predict bloggers’ political poll results by using their blog entries. Our prediction methods are based on a supervised learning algorithm, Support Vector Machines (SVM), with the features in their blog sites. We take into account the following ideas to improve prediction accuracy:

1. Since general blogs contain many entries that are not related to political topics, we automatically removed such blog entries that are considered to be not helpful for the prediction and used only political ones.
2. Since political poll results can be considered to be related to the description of whether a political party or politician is favorable or not, we used the lexicon of polarity-bearing words and calculate the number of
positive and negative words as features.

3. As we mention in Section 2, our data collection is severely imbalanced regarding the number of examples in different classes. Therefore, we thought we should tackle this problem by using an under-sampling method. Under-sampling is an imbalanced data learning method which uses only a subset of major class examples.

The reason why we devised a method to automatically predict political poll results by using blog entries is that it is now desirable to have a method to predict political poll results from a collection of public opinions on the Web, since it has recently become more difficult to thoroughly collect responses in a political survey while we now have a plenty of public opinions described on the Web.

We also attempted manual prediction with three human subjects as the upper bound of the system performance, and found that this task is rather difficult even for humans and that the system performance can outperform that of humans.

In Section 2, we explain our data collection in which blog sites are connected to their bloggers’ political poll results. In Section 3, we explain our method of predicting political poll results by using blog entries. Lastly, in Section 4, we show the experimental results where we apply our methods to our data collection.

2. Political Blog Data Collection

We constructed a dataset of general blogs that are connected to political poll results in Japan. We asked bloggers on Yahoo!blog (http://blogs.yahoo.co.jp/) to be poll respondents and provide all their blog entries. 2,619 bloggers kindly agreed with our request. Twelve questionnaire items were asked from 20th to 31st of March, 2008. The questionnaire items asked their political attitudes and their attitudes to some points of an issue, other than their demographic attributes, such as their sex, age, and educational qualification. The following are samples of the questionnaire items asked:

q1: Do you like a politician or a political party? (4 classes; dislike/slightly dislike/slightly like/like)

Liberal Democratic Party(q1-1), Democratic Party(q1-2), Yasuo Fukuda(q1-3), Ichiro Ozawa(q1-4)

q2: Did you vote in the last House of Councilors election in 2007? (2 classes; yes/no)

q3: For which party did you cast a ballot? (2 classes; yes/no)

For the LDP(q3-1), the DP(q3-2), and other parties(q3-3)

q4: Which political party do you support? (2 classes; yes/no)

the LDP(q4-1), the DP(q4-2), and other parties(q4-3)

q5: Do you support the Fukuda cabinet? (3 classes; yes/no/no opinion)

Therefore, for those 2,619 bloggers, their poll results can be connected to their blog sites.

The answers for some of the questionnaire items were severely imbalanced and some classes had many more examples than others. For example, in the above samples of the questionnaire items, the following are the typical ones, for which we show the number of examples in the classes:

- Did you cast a ballot for the LDP?
  - yes 368/no 2251
- Did you cast a ballot for the DP?
  - yes 583/no 2036
- Do you support the LDP?
  - yes 560/no 2059
- Do you support the DP?
  - yes 577/no 2042.

3. Our Methods of Predicting Poll Results

The methods for predicting political poll results by using blog entries are based on a classification model trained with SVM, and a data set consisting of entries from blog sites that are connected to poll results. We adopted SVM because SVM learning algorithms have been successfully used in text classification in the past.

To learn a model for predicting political poll results from blog sites using supervised learning algorithms, the simplest way is to use ‘bag of words' (BoW) in the blog sites as features. However, to improve prediction accuracy, we take into account the following ideas and devise better features:

1. Since general blogs tend to contain many entries that are not related to political topics and are useless for prediction, we automatically remove those blog entries and used only political ones.

First, we collected words that appear in the title of the articles of 61 political categories in Wikipedia as

1 LDP leader at the time of the poll
2 DP leader at the time of the poll

3 61 categories were manually selected. The categories are ‘politics,’ ‘politician,’ ‘political group,’
keywords. In total, 3,983 words were collected as keywords. If the entries include at least two keywords and satisfy the following condition, they can be considered as 'political blog entries':

\[
\frac{\text{the character length of keywords in the entries}}{\text{the character length of the entries}} > 0.05
\]

or

\[
\text{the character length of keywords in the entries} > 30.
\]

The above parameters were heuristically determined. We assume that those political blog entries more often include political topics. At last, 2,857 entries for 318 bloggers were selected as political entries.

2. Since political poll results can be considered as related to the description of whether a political party or politician is favorable or not, we use the lexicon of polarity-bearing words and calculate the number of positive and negative words as features. When we try to identify opinions for a political party in blog entries, the text fragments of the blog entries around the name of the party are considered more informative. Therefore, we first extract the five words just before and after the topical word, such as the name of the political party, as other features. Then, we use the lexicon of polarity-bearing words and calculate the number of positive and negative words that appear in a ten-word window as features. The sentiment lexicon available on the Web was used as our lexical resource for polarity-bearing words. The lexicon was constructed with the method developed by Takamura et al. [9].

3. Our data set is severely imbalanced regarding some of the questionnaire items, i.e. some classes have many more examples than others. Learning algorithms that do not take into account class imbalance tend to be overwhelmed by the majority class and ignore the minority one, and the overall accuracy of conventional learning algorithms will degrade significantly, since the classifiers from them are greatly biased towards the majority class. Therefore, for the questionnaire items with an imbalanced number of examples, under-sampling [10] is used and training is performed with equal number of positive and negative examples. Sampling is a class of methods that alters the size of training sets. Under-sampling is an imbalanced data learning method that uses only a subset of majority class examples. Under-sampling changes the training sets by randomly sampling examples from the training set of a majority class and making it smaller. The level of imbalance is reduced, with the hope that a more balanced training set can give better results. Among various class-imbalance learning methods, under-sampling has been commonly used. The method of under-sampling that we adopted is to cluster the examples in the majority class first and to randomly select the examples from those clusters.

4. Experiments

In this section, we report on the experimental results using our prediction methods described in the last section on our data collection. We used the standard precision and recall for the positive class, and the F1 measure, the harmonic mean between precision and recall, to evaluate prediction in binary classification. In addition, we calculated accuracy as the number of correctly classified blogs divided by the total number of blogs to be classified. For multi-class classification, we used only the accuracy for the evaluation. We used the libsvm implementation of SVM with the linear kernel in 10-fold cross-validation.

We compared the following different feature sets for prediction:

1. BoW in all the blog entries (baseline feature set)
2. BoW in political entries (M1)
3. BoW in political entries + BoW in the window around the topical words (M2)
4. BoW in political entries + BoW in the window around the topical words + features for sentiment words (M3)

The experiments were performed with and without under-sampling, except for the baseline feature set.

As the upper bound of the system performance, we also attempted manual prediction with three human subjects. For 100 randomly selected blog sites, the three subjects try to predict the bloggers' poll results by reading their blog entries. The average accuracy for the results of the three subjects is also shown for comparison (shown as 'Human' in Table 1).

In Table 1, we show the prediction results

---

*political party,* 'election,' 'Japanese politics,' 'Japanese politician,' 'Japanese political group,' 'Japanese Diet member,' 'Japanese election,' 'Japanese electoral zone,' 'Japanese prime minister,' and so on.

for the questionnaire items described in Section 2 with our methods. Due to space limitation, we only show the accuracy, and the F1 measure for binary classification. The F1 measure is shown in parentheses for binary classification. The best performance is shown in bold for each questionnaire item, where the F1 measure is focused on for binary classification since higher accuracy can be easily obtained when the data is greatly imbalanced in favor of the majority class.

In questionnaire items where the examples are not necessarily imbalanced, under-sampling degrades the performance. However, for the imbalanced questionnaire items (q3 and q4), under-sampling succeeds to detect more examples in the positive class and improves the performance (the F1 measure).

By comparing the feature sets, we found that removing unrelated entries and using only political entries improve the performance (M1 > Baseline). Setting the window around topical words and using polarity-bearing words inside the window can also contribute to improving the performance (M2, M3 ≧ M1).

Interestingly, human prediction accuracy which tends to be considered as the upper bound for system performance was not so good in this task, and the system performance is much better than that of humans. Therefore, we can conclude that this task is rather difficult even for humans and the system performance can greatly outperform that of humans.

For future work, we need to improve prediction performance by devising more intelligent features with the information such as bloggers’ news sources and links from their blog sites.

6. References


