Active Learning for Turkish Sentiment Analysis

Mahmut Çetin, M. Fatih Amasyalı
Department of Computer Engineering
Yıldız Technical University
Istanbul, Turkey
cetinmahmut@msn.com, mfatih@ce.yildiz.edu.tr

***Abstract*—Sentiment analysis/classification is a widely studied problem of natural language processing and data mining. With the availability of social media, there are a lot of data but it's hard to find a labeled training set because of its high cost. The goal of active learning is to get a better or same performance with fewer training data. In this work, the feasibility of active learning scheme for Turkish sentiment analysis is investigated. As a result, the same performance with full training set could be obtained with only half of the training set selected by active learning. Moreover, the affects of different clustering algorithms used at the initial set selection are investigated.**

Keywords-component; Sentiment Classification, Sentiment Analysis, Active Learning, Clustering, Clustering Algorithms, K-means, Self Organizing Maps, Hierarchical Clustering

I. INTRODUCTION

Sentiment classification is one of the hottest research areas among the natural language processing topics. While it aims to detect sentiment polarity of the given opinion, requires a large number of labeled data. However, labeling data takes human effort and long time. To reduce these costs active learning methods have come out recently. There have been several approaches in the literature on this topic. First studies in active learning focused on selecting only one unlabeled instance for each iteration. But recently, there have been batch mode approaches too [1]. Batch mode active learning selects a set of instances at one time in order not to waste resources like time etc. In this study, we will compare classifiers and methods for selecting initial set of instances to start learning a classifier. Selecting initial set of instances affects the whole process [2].

We used a supervised term weighting method called Delta1(D1) method which takes into account the distribution of classes to represent the texts and the reason why it is chosen will be explained in Section 3. In addition these methods have been applied to root of words for Turkish language and Zemberek, Turkish language morphological analyzer, has been used to get roots [3].

The most important thing in our approach is how to decide which instances are more informative. In order to determine instances, we used WEKA data mining tool's classifiers [4]. The most informative instances are difficult to classify, so the prediction rate for these instances are closer than the others' rates. When we annotate these informative instances, they can provide assistance to determine the classes of remaining unlabeled data.

The main steps of Active Learning are listed below:

1. Choosing initial set of instances as training set and labeling them,
2. Evaluating training set on the remaining unlabeled data and decide which to select,
3. Annotate the selected instances and add to training set
4. Turn back to the second step.

These steps are processed until the targeted success ratio on the test instances is obtained.

II. USED DATASET

We have used a dataset which belong to a telecommunication company in our study. The dataset includes 6000 Twitter posts about the company. The dataset is divided into equal sized train and test set. The distributions of train and test sets are the same and they both include 1520 positive tweets, 924 negative tweets, 576 neutral tweets.

III. SUPERVISED TERM WEIGHTING METHODS

Among the text representation methods, the bag of word approach is the most used. Its simplest form is term frequencies (TF). Then, the inverse document frequency weighted text frequency (TF-IDF) was proposed to eliminate the affects of the commonly used terms. In recently, the supervised term weighting methods were started to use [5] because of their superior performances. In conventional TF and TF-IDF methods, the class distributions of the terms are not considered into the weights. In literature, there are several proposed supervised term weighting methods. In this study, we used Delta1 method [5]. It is originally formulized for two-class problems. To implement this method for our three-class problem, we have considered producing a feature for each class. There is the formulation of our method below. While (+) is representing the class of produced feature, (-) represents the other classes.

$$tf\ x\ \Delta_1 = tf\ x\ \log_2 \frac{N^+ x\ n^-(x)}{N^- x\ n^+(x)} \quad (1)$$

978-1-4673-1448-0/12/$31.00 ©2012 IEEE

In the formulation, $n^{+}(x)$ represents the total number of documents containing term x which belongs to related class and N^{+} gives the total number of documents belong to the same class. N^{-} and $n^{-}(x)$ have the same meanings for other two classes.

In Table 1, there is a comparison of weighting methods and four different classifiers. The mentioned weighting methods are term frequency (TF), term frequency weighted by inverse document frequency (TF-IDF) and Delta1 (D1). The used classifiers are Naïve Bayes (NB), one nearest neighbor (IB1), decision tree (J48), Random Forest (RF). All methods use all train set and being tested with 3000 tweets.

TABLE I. TERM WEIGHTING METHODS' SUCCESS RATIOS(%)

Methods	NB	IB1	J48	RF	Avg.
TF	53,7	49,1	53,7	57,5	53,5
TFIDF	53,7	49,1	53,7	57,5	53,5
D1	62,6	57,6	62	60	60,6

It can be seen that Delta1 method has the best performances for all classifiers. So, we can say that the usage of supervised term weighting method is a better way for Turkish sentiment analysis. At the rest of the paper, Delta1 method is used at all of the experiments.

IV. ACTIVE LEARNING

In Active Learning literature, two main problems are described. The first one is how to select the instances. The most used method is firstly classifying unlabeled instances and then selecting the instances having similar class prediction probabilities. In this way, the labeling cost of uninformative or redundant instances is reduced. We used the same procedure.

Another problem is the first selecting process because of its effect to the followings. We have implemented randomly selection and selection by clustering. A disadvantage for randomly selection is getting different results in all executions and the process of selection can't be controlled. We have decided to use clustering algorithms instead of random selection to get the control of process and tested them.

V. EXPERIMENTAL RESULTS

After representing the texts mentioned Section 2 by the Delta1 (D1) method in Section 3, we compared classifiers and selecting initial set in different ways.

Firstly, we investigated the active learning scheme is suitable for Turkish sentiment analysis or not. We compared active learning and random selection. We used two classifiers (Simple Logistic–SL, and Naïve Bayes–NB) giving class prediction probabilities. The required numbers of labeled instances were found to reach a success ratio on test set. At the each iteration, we selected 100 training instances randomly or according to their class prediction probabilities. In Table 2, full random selection method, random initial set selection then active learning, and initial set selection by clustering then active learning were compared. Each experiment is repeated by 10 times and their averages were shown.

In Table 2, each cell in the table represents the average number of required labeled texts (instances) and standard deviations to reach the given test success ratio. For K-Means algorithm, number of iteration (epoch) is determined as 1000. K number is determined as 100.

TABLE II. COMPARISON OF CLASSIFIERS ACCORDING TO THE NUMBER OF REQUIRED LABELLED INSTANCES

Target Test success Ratio	Initial Set Selection + instance set Selection Method	# of Required Labeled Instances for Simple Logistic	# of Required Labeled Instances for Naïve Bayes
60%	Initial set Randomly + Active Learning	730 ± 206	740 ± 143
	Initial set K-Means + Active Learning	783 ± 228	665 ± 177
	Full Randomly	850 ± 212	940 ± 232
61%	Initial set Randomly + Active Learning	850 ± 227	950 ± 158
	Initial set K-Means + Active Learning	913 ± 244	805 ± 160
	Full Randomly	1110 ± 303	1190 ± 242
62%	Initial set Randomly + Active Learning	1120 ± 193	1090 ± 223
	Initial set K-Means + Active Learning	1123 ± 184	1055 ± 142
	Full Randomly	1560 ± 267	1780 ± 571
63%	Initial set Randomly + Active Learning	1420 ± 225	1350 ± 259
	Initial set K-Means + Active Learning	1473 ± 261	1215 ± 224
	Full Randomly	2140 ± 347	-
64%	Initial set Randomly + Active Learning	1640 ± 236	1590 ± 197
	Initial set K-Means + Active Learning	1763 ± 320	1525 ± 388
	Full Randomly	-	-

When we look at the Table 2, we have following results:

- Active learning can effectively reduce the number of required labeled instances. To reach 63% test success, randomly selection needs 2140 instances while active learning needs only 1215 instances.
- Randomly selection never reaches 64% test success, while active learning does by learning 1525 instances.
- NB is more compatible (requires less labeled instance) than SL for active learning.
- The usage K-Means clustering algorithm instead of randomly selecting initial set of instances requires less training data especially for NB.

In Figure 1 and 2, we have visualized performs of methods. Methods have been renamed as their first letters in the graphics.

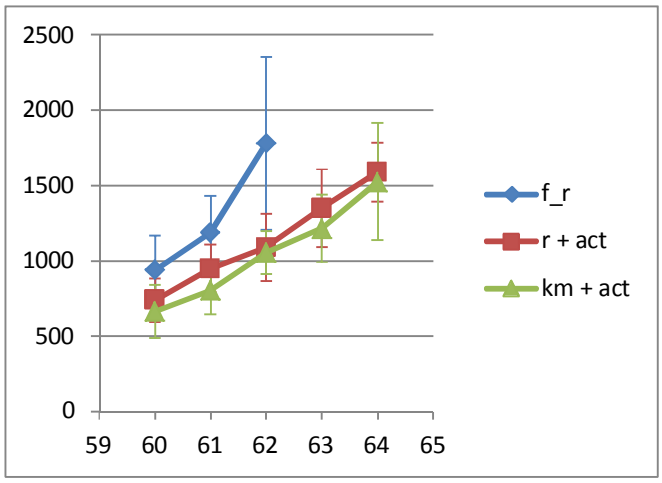

Figure 1. Naive Bayes

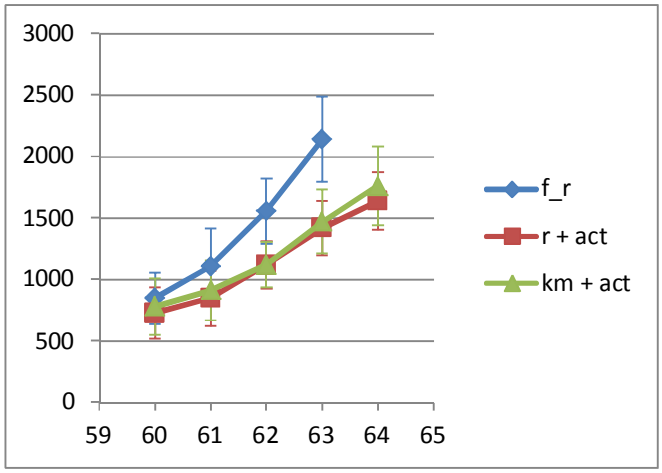

Figure 2. SimpleLogistic

The benefit of using a clustering algorithm for the initial set selection process brings to our minds a question: Which clustering algorithm is most suitable for this selection? To answer this question, we have designed an experiment to compared four clustering algorithms (K-means, Self Organizing Maps – SOM, and two Hierarchical Clustering algorithms - Single Linkage, Complete Linkage). The results are shown in Table 3. Each cell of the table shows the average number of instances to perform the given success ratio and standard deviation. Single Linkage and Complete Linkage algorithms are static algorithms, so these algorithms perform the same for each execution. K-Means and SOM clustering algorithms are repeated 10 times. The results are averaged as in the table.

TABLE III. COMPARISON OF CLUSTERING ALGORITHMS ACCORDING TO THE NUMBER OF REQUIRED LABELLED INSTANCES

Target Test success Ratio	Initial Set Selection Method	# of Required Labeled Instances for Naïve Bayes
60%	Single Linkage	900
	Complete Linkage	591
	SOM	792 ± 268
	K-Means	665 ± 177
61%	Single Linkage	900
	Complete Linkage	691
	SOM	901 ± 132
	K-Means	805 ± 160
62%	Single Linkage	1300
	Complete Linkage	791
	SOM	1052 ± 205
	K-Means	1055 ± 142
63%	Single Linkage	1600
	Complete Linkage	1391
	SOM	1231 ± 132
	K-Means	1215 ± 224
64%	Single Linkage	1900
	Complete Linkage	2591
	SOM	1551 ± 276
	K-Means	1525 ± 388

Figure 3 shows the number of required labeled instances for clustering algorithms by a graphic. Algorithms have been renamed as their first letters in the graphics.

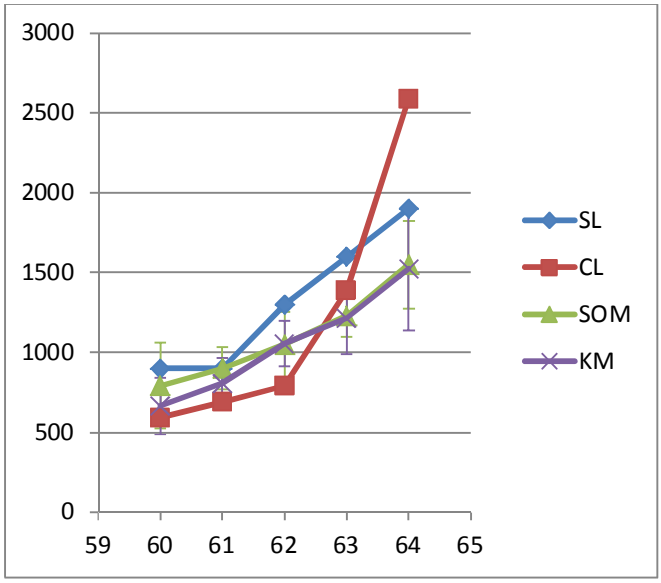

Figure 3. Performs of clustering algorithms for NaïveBayes

According to the Table 3 and Figure 3, Complete Linkage is more compatible (requires less labeled instance) than others for lower success ratios. But, K-means and SOM are more compatible for the highest ratio.

VI. Conclusion

Sentiment analysis has become the most popular application of text categorization recently, because of its commercial use. It aims to find out whether the text is positive, negative or neutral. The increasing use of social media increases the need for this application, but while building a sentiment analysis tools everyone come across same challenges. To handle these challenges, active learning is a promising way.

One of the challenges in a sentiment analysis system is training a good classifier. This process requires a sufficient number of annotated training data. But, annotating training instances takes a long time and human effort. According to our experiment we have reduced the number of required labeled instances by half for Turkish Sentiment Analysis problem. This means the process of training classifier takes half of the human effort and time which has been spent before. In addition, the better test success ratio can be obtained with fewer training data selected by active learning (64%) than using all training set (62.6%) with Naïve Bayes classifier. We also compared several initial set selection methods. According to the comparison SOM and K-Means algorithms perform better than the other clustering algorithms. We also found that a supervised term weighting method (Delta1) is more suitable than conventional unsupervised methods such as TF, TF-IDF for the Turkish sentiment text representation.

Acknowledgment

The research work described in this paper has been supported by Ericsson Turkey.

References

[1] Y. Guo and D. Schuurmans, "Discriminative batch mode active learning", In Advances in Neural Information Processing Systems (NIPS), number 20, pages 593–600. MIT Press, Cambridge, MA, 2008.

[2] Burr Settles, Synthesis Lectures on Artificial Intelligence and Machine Learning, Morgan & Claypool Publishers, 2012.

[3] Mehmet Dündar Akın and Ahmet Afşin Akın, "Türk Dilleri İçin Açık Kaynaklı Doğal Dil İşleme Kütüphanesi" , Electricity Engineering – Elektrik Mühendisliği, vol 431 pp.38-44, 2007

[4] Ian H. Witten, Eibe Frank and Mark A. Hall, Data Mining Practical Machine Learning Tools and Techniques, Morgan KauffMan Publishers, 2011.

[5] Justin Martineau and Tim Finin, "Delta TFIDF: An Improved Feature Space for Sentiment Analysis", Third AAAI International Comverence on Weblogs and Social Media, San Jose CA, 2009.