Text Classification for Intelligent Agent Portfolio Management

Young-Woo Seo
Carnegie Mellon University
5000 Forbes Ave
Pittsburgh, PA 15213
ywseo@cs.cmu.edu

Joseph A. Giampapa
Carnegie Mellon University
5000 Forbes Ave
Pittsburgh, PA 15213
garof+@cs.cmu.edu

Katia P. Sycara
Carnegie Mellon University
5000 Forbes Ave
Pittsburgh, PA 15213
katia@cs.cmu.edu

ABSTRACT
In the application domain of stock portfolio management, software agents that evaluate the risks associated with the individual companies of a portfolio should be able to read electronic news articles that are written to give investors an indication of the financial outlook of a company. There is a positive correlation between news reports on a company's financial outlook and the company's attractiveness as an investment. However, because of the volume of such reports, it is impossible for financial analysts or investors to track and read each one. Therefore, it would be very helpful to have a system that automatically categorizes news reports that reflect positively or negatively on a company's financial outlook. To accomplish this task, we treat the unsupervised reading and understanding of news articles as an automatic text classification problem. In this paper, we propose a text classification method that we call, "domain experts" and "self-confident" sampling technique, and compare it with naive Bayes with expectation maximization (EM). We evaluate these learning techniques in terms of how well they improve with unlabeled data after being initially trained on a small number of human-labeled articles and how well they classify the latest financial news articles. The significance of this work lies in the new classification method that we propose and in the sampling technique we used for improving classification accuracy.

1. INTRODUCTION
In the application domain of stock portfolio management, there is a large volume of information about a company and its financial performance for humans to effectively attend to and manage while making decisions. To address this problem, we proposed a multi-agent system, called *Warren*[1] [16], [5] that helps the user track information on a portfolio of companies of interest. Warren is composed of different agents that help the user track the stock price, performance history, earnings summaries, and Beta value (risk) associated with the individual holdings in their stock portfolio, and to proactively advise the user whenever the portfolio may be too risky for the user's preferred tolerance to risk. To supplement the data on a company, the user has access to a *Breaking News agent*, which performs an information retrieval for news from on-line sources such as Reuters, CNN Financial Network, Business Wire, Forbes.com and others. The work that we present in this paper describes our endeavor to create an agent that analyzes news articles that were retrieved by the Breaking News agent for their content about a company's financial well-being, for presentation to the user in a meaningful way.

To accomplish this task, we proposed a new text classification algorithm that classifies financial news into the predefined five categories: "good", "good, uncertain", "neutral", "bad, uncertain", and "bad" and a sampling algorithm that predicts the label of unlabeled data by information theoretically. Our hypotheses for how this goal can be achieved are as follows: (1) frequently co-located phrases can permit the classifier to estimate the category of a news article with high probability; (2) the performance of proposed method can be improved by training with unlabeled data; (3) the *domain experts* algorithm can perform well as a classifier for this problem; and (4) using *vote entropy* (VE) as the sampling measure can boost the process of classifier training.

Our task is similar to the established approaches on the text categorization domain in that text categorization is a task to assign each given document to the appropriate categories, based on the semantic content of the document. In this regard, a number of statistical and machine learning methods have been applied to this domain in recent years including nearest neighbor classification [17], naive Bayes with EM (Expectation Maximization) [12] [15], Winnow with active learning [10], Support Vector Machines (SVMs) [7], [8], Maximum-Entropy model [14]. It is, however, slightly different with others in that our task deal with more objective and confined categories, such as "good" or "bad" for a company's financial outlook than categorization of news articles into "politics" or "economics."

The proposed method, domain experts, predicts the label of a financial news articles through voting process among experts that are frequently co-located phrases. A frequently

[1]The system is named after Warren Buffet, a famous American investor and author about investment strategies.

co-located phrase (FCP) is very often appeared in a certain category (e.g. good) and allows the proposed method to discriminate categories. The method for selecting co-located phrases is described at Section 3 in detail.

The text categorization task has several characteristics that make it a difficult domain for the use of machine learning, including a very large number of input features, high levels of attribute and class noise, and a large percentage of features that are irrelevant. In this respect, the exploitation of supervised learning requires a relatively large number of labeled examples. When it is given a small set of labeled training data, classification accuracy will suffer because the variance in the probability distribution of data will be high. However it is expensive to obtain labeled training data, while unlabeled data are cheaply available. Several methods have been used for coping with the problem which comes from insufficient labeled data, such as Expectation Maximization (EM) [12], [15], selective-sampling [3], sub-sampling and uncertainty sampling [9]. The proposed sampling method, self-confident sampling, picks out more promising data from unlabeled data sets to improve a classifier's performance. It is similar to uncertainty-sampling in that it predicts the label of unlabeled data on the basis of the learner's confidence which is acquired during training phase. The examples that are predicted with the least uncertainty will be added to the training set in the next iteration. The overall procedure of self-confident sampling is described in Section 4.2.

The paper is organized as follows. Section 2 will give the overview of our task in terms of the text categorization context. Section 3 details the method of selecting co-located phrases. In Section 4, we describes the overall procedure of categorization with consideration of the company's financial well-being. Section 5 provides the experimental results and compares them with those of existing methods. Section 6 and 7 discuss the results and the future work respectively.

2. CATEGORIZATION OF FINANCIAL NEWS ARTICLES

Concisely, our task is to build an algorithm that categorizes each given news article into the predefined categories in terms of the referred company's financial well-being. When we read a financial news article, we are able to classify it as good or bad news to the current financial status of the company. For example, we have a news article at our hands which has one of sentences as "Shares of ABC Company rose 1/2 or 2 percent on the Nasdaq to $24-15/16..." From this clues, we can infer that ABC company's financial status at that spot is good.

Most financial news articles that we gathered for experiments were categorized into several categories by considering how explicitly they mentioned the company's financial status which we want to know. Considering the nature of financial news article and our task, the following five categories are considered to be pertinent:

GOOD News scripts which show good evidences of the company's financial status explicitly.
e.g.) ... Shares of ABC Company rose 1/2 or 2 percent on the Nasdaq to $24-15/16. ...

GOOD, UNCERTAIN News scripts which refer to predictions of future profitability, and forecasts.
e.g.) ... ABC Company predicts fourth-quarter earnings will be high. ...

NEUTRAL News scripts which did not mention anything about the financial well-being of the company explicitly.
e.g.) ... ABC and XYZ Inc. announced plans to develop an industry initiative. ...

BAD, UNCERTAIN News scripts which refer to predictions of future losses, or no profitability.
e.g.) ... ABC (Nasdaq: ABC) warned on Tuesday that Fourth-quarter results could fall short of expectations. ...

BAD News scripts which show bad evidences of the company's financial status explicitly.
e.g.) ... Shares of ABC (ABC: down $0.54 to $49.37) fell in early New York trading. ...

The financial news articles that do not mention anything about a company's clues explicitly were classified into "neutral" category because we could not determine its current financial status. In case of "uncertain" categories, one may be allowed to decide it as a good (or bad) news, but we could not be sure of it (i.e. uncertain.) The prediction of future earning by other reliable source is the very example of these categories: "good, uncertain" and "bad, uncertain."

Most Information Retrieval task including text categorization made use of well-indexed data set, i.e. some human experts on the domain read carefully and categorize (or classify) the collection of data into the proper category in advance. In our case, since there is no data set available for this task, we have had to prepare our own text data. There is a well-known problem in constructing text data set. When two different humans must take a decision on whether to classify a news article under the given classes, they may disagree. It is called "inter-indexer inconsistency" [1]. We, however, could fortunately be allowed to ignore this problem due to the facts that most of the financial news articles has very objective contents and it is easier to discriminate a news between good and bad than to classify a news into the predefined subjects such as economics or politics. Irrespective of objective semantics, we carefully did put a label on each of news articles data set before using it. Figure 1 shows an example of news article used for our experiments.

3. CO-LOCATED PHRASE AS A DOMAIN EXPERT

The proposed text categorization algorithm predicts the label of a financial news articles through voting process among experts that are frequently co-located phrases. A co-located phrase is sequence of nearby, but not necessarily consecutive words. Thus, a frequently co-located phrase is able to use for discriminating the category of financial news articles because it is very often appeared in the category. For example, *Shares* and *rose* can be selected as a frequently co-located phrase for "good" category from a sentence in news article, which is "Shares of Company A rose 1/2 or 2 percent on

[title] Goldman Profits Fall 13 Percent [\title]
[date] Mar 20 6:35 PM ET [\date]
[source] Reuters [\source]
[company] Goldman Sachs (GS) [\company]
[body]
Goldman Sachs Group Inc.(NYSE:GS - news), one of Wall Street's top firms, on Tuesday said first-quarter profits fell 13 percent but were above reduced estimates as fees for advising companies on stock sales declined in a slumping market. Investment banking revenue for Goldman, the top seller of new stock offerings in 2000 and 1999, fell 7 percent as the market for initial public offerings dried up.
...
The value of Goldman's principal investments fell $140 million, compared with a gain of $214 million in last year's first quarter. Principal investments were down across the board, Viniar said.
[\body]
[label] bad [\label]

Figure 1: A example of financial news article.

the Nasdaq to $24-15/16..." It is often desirable to consider such a contextual information (i.e. word-collocation) rather than frequency statistics with respect to the characteristics of English text, because "word-collocation" has characteristics of a syntactic and semantic unit, and whose exact and unambiguous meaning of connotation cannot be derived directly from the meaning or connotation of its components [11].

It, however, is not easy to select such a co-locate phrases which express a category well due to the inherent properties of text categorization, such as large features, much noise. To reduce original feature space, we made use of a heuristic that consider a characteristic of financial news report. Since most of financial news articles report several company's news in a news article, they mentioned a company's name (or a company's ticker[2]) explicitly. From this observation, we could build an abridged version of a news article. That is, if a sentence contains a company's name or ticker, it is added to the abridged version from the original news. Indeed, an abridged version of article still has noise, but it also has sufficient information that make a classifier learn the correlation between phrases and a category. Most phrases in abridge version are the candidate of co-located phrases to a category after removing stop-words.

Not all co-located phrases in a sentence are selected as feature due to the existence of the most informative phrases. The most informative co-located phrases are those that would reduce classification error and variance over the distribution of examples. To select the most informative feature, we taken advantage of information gain as the phrase goodness criterion [18]. Let $\{c_j\}_{i=1}^{m}$ denote the set of categories in the target space. The information gain of kth co-located phrase

[2]A ticker is a symbol that usually is used for representing a company's name briefly in stock trade market.

class	selected FCP
$+$	"revenue&rose", "exceeds&expectations", "share&gains|rose", "rose&profit"
$+/?$	"expect|forecasts|anticipate&earnings"
$+/-$	"alliance&company|corp", "fiscal|quarter", "introduce|*", "announces&deal|product"
$-/?$	"warning&profits", "short&expectation"
$-$	"share&off|down|lost", "profit&decrease", "fall&percent", "sales&slump|decrease"

Table 1: Examples of selected FCPs for each category. Each symbol represented at leftmost column stands for five categories: "good", "good, uncertain", "neutral", "bad, uncertain", and "bad" as +, +?, +/-, -?, and - respectively.

in jth category, $fcp_{k,j}$ is defined to be:

$$Gain(\mathrm{fcp}_{k,j}) = -\sum_{i=1}^{m} P(c_i)\log P(c_i)$$
$$+P(\mathrm{fcp}_{k,j})\sum_{i=1}^{m} P(c_i|\mathrm{fcp}_{k,j})\log P(c_i|\mathrm{fcp}_{k,j})$$
$$+P(\bar{\mathrm{fcp}}_{k,j})\sum_{i=1}^{m} P(c_i|\bar{\mathrm{fcp}}_{k,j})\log P(c_i|\bar{\mathrm{fcp}}_{k,j}) \qquad (1)$$

We computed the information gain for each of FCP candidates in the training data and removed from the feature space those phrases which value was less than some predefined threshold. Table 1 shows selected FCPs for each category. A group of domain experts consists of $|K|$, which is determined in experiment empirically, selected FCP and represents a category.

4. CATEGORIZING NEWS BY CONSIDERING FINANCIAL STATUS

To begin with, let us describe text representation. As mentioned earlier, a news article is divided into two group of sentences. One is the condensed version of a news article that consists of those sentences includes company name or ticker. The other is that do not have company name or ticker in the each sentence. Since the condensed news article only has sentences that mentioned the company's name or ticker explicitly, it could not contain co-referred sentence. For instance, "Verizon plans to spend more than US $300 million this year in the New York metro region to enhance its digital wireless network. The company said it will expand coverage and increase capacity in ...." In this case, "Verizon" and "the company" were co-referred, as enhancing its facility. Thus, the condensed news article may have a doubt that it could eliminate such a useful information. But, the experience constructing the data set tells us that the condensed version contains a sufficient amount of information to our task. And then, the condensed version of each news article is represented as a words vector after removing stop-lists:

$$\vec{d_i} = < t_1, t_2, ..., t_k, ..., t_{|T|} >, \qquad (2)$$

where t_k is k th component (i.e. word) of i th document, d_i, which is made up of $|T|$ number of words. Equation 1 was applied for estimation of the importance of each phrase after making candidates, which are made by combining each word in condensed version and five consecutive words toward the end of a sentence. One of the five FCP candidates which has the highest value of information gain is selected for a domain experts for a category.

4.1 naive Bayes Classifier with EM

One of the most effective classifiers in terms of predictive performance is the naive Bayes classifier. It learns from training data the conditional probability of each attribute w_k given the class c_j. Classification is then done by applying Bayes rule to compute the probability of c_j given the particular instance of $w_1, ..., w_n$, and then predicting the class with the highest posterior probability. In equation 3, $freq_{k,j}$ is the number of times word t_k is occurred, $|J|$ is the total number of unique words in category j, and $|V|$ is the total number of unique words in data set. This computation is made possible by making a strong independence assumption: all the attributes w_k are conditionally independent given the value of class c_j. Thus, the naive Bayes classification is the classification that maximizes the probability of observing the words that were actually found in the document, subject to the usual naive Bayes independence assumption.

$$\begin{aligned} Pr(c_j|d_i) &= \arg\max_{c_j \in C} Pr(c_j) \prod_k Pr(w_k|c_j) \\ &\approx \frac{Pr(c_j)\prod_{k=1} Pr(w_k|c_j)}{\sum_{j=1}^{|C|} Pr(c_j)\prod_{k=1} Pr(w_k|c_j)} \qquad (3) \\ & where, \\ w_k &= \frac{1 + freq_{k,j}}{|J| + |V|} \end{aligned}$$

The naive Bayes classifier estimates the probability that a new article is a member of a certain category using the probabilities of terms occurring in the category.

There, however, are some concerns for applying naive Bayes classifier to the real world problem. It is that naive Bayes classifier relies heavily on a strong independence assumption among the features. Fortunately, there is a good evidence of which allows us not to consider this concern seriously. In [6], they showed why the violation of the feature independence assumption does little damage to classification accuracy.

Another is that the its performance could be decreased by variance from training data. When it is given just a small set of labeled training data, classification accuracy will suffer because variance in the probability distribution of data will be high [12]. It is expensive to acquire a sufficient number of labeled data for training. In [12], they tried to decrease a variance in classifying unseen data by a combination of a variant of Active Learning and Expectation Maximization (EM). Active learning is used to actively select documents for labeling, then EM with a naive Bayes model further improves classification accuracy. EM is a class of iterative algorithms for maximum likelihood estimation in problems with incomplete data. Given a model of data generation, and data with some missing values, EM will converge to a set of generative parameters that locally maximizes the likelihood of both the labeled and unlabeled data. In this paper, EM is applied for an iterative two-step process: E-step calculates probabilistic class labels, $Pr(c_j|d_i)$, for given a set of unlabeled document using a current estimate of a category $Pr(c_j)$ and M-step calculates a new maximum likelihood estimates for $Pr(c_j)$ using all the labeled data, both original and probabilistically labeled.

Input:
Training data $D \leftarrow D_l$, where, D_l: labeled news articles

Feature Selection
Choose the co-located phrases among the current labeled news articles.

Do while performance of classifier improve
1. Generate unlabeled data D_u
   Pick news articles randomly from unlabeled data set
2. Use the current classifier and feature sets to estimate the class of each unlabeled article D_u.
   - $\begin{cases} D_u \leftarrow D_u + \lambda d_i & \text{if } d_i > \theta_{ve} \\ \text{skip } d_i \text{ to the next iteration} & \text{otherwise} \end{cases}$
   - $D \leftarrow D_l + D_u$
3. Re-select feature sets by new data set D.

Output:
A classifier that takes an unlabeled articles and predicts a class label.

Figure 2: The self-confident sampling for domain experts algorithm.

4.2 Domain Experts with Self-confident Sampling

A group of domain experts is made up of a number of co-located phrases (i.e. FCP) which are consistent with a version space of a category. That is, the k th domain expert of j th domain (i.e. category) expert group, $fcp_{j,k}$, is consistent with a set of training documents D_j if and only if $fcp_{j,k}(d_{i,j}) = c_j(d_{i,j})$ for each example $fcp_{j,k} \in d_{i,j}$ and $< d_{i,j}, c(d_{i,j}) > \in D_j$, where $d_{i,j}$ is an instance of jth category and D_j is data set of the category. Thus, from the definition of version space which is the subset of all hypotheses consistent with the observed training data, a domain experts group maintains the version space of a category [13]. Figure 3 describes the predictions procedure of domain experts in detail. When it make prediction of a news article, it make use of voting among the group of domain experts and then learns by altering the weight associated with each domain expert. One attractive property of the proposed algorithm is that it is able to accommodate inconsistent hypothesis as well as consistent hypothesis. In other words, it does not eliminate a hypothesis that is found to be inconsistent with some training documents, but rather reduces its weight with the degree of β. Since we made use of our own text data set, we can tell that there are little phrase which appears only a category. Domain experts algorithm is similar with Sleeping experts algorithm [2], in that they consider each of selected "phrases" as a consistent expert (or hypothesis) to

$C = \{$"good", "good, uncertain", "neutral",
"bad, uncertain", "bad"$\}$
$fcp_{k,j}$ denotes k th expert in j th domain (or category).
$w_{k,j}$ denotes the weight associated with $fcp_{k,j}$.
E_i denotes the most probable category of i th document.

- For each training example $< d_i, c(d_i) >$
  - Initialize e_j to 0
  - For each domain expert $fcp_{k,j}$
    $e_j \leftarrow e_j + w_{k,j}$ if $fcp_{k,j} \in d_i$
  - Predict
    $E_i(d_i) = \arg\max_j \frac{e_j}{\sum_j \sum_k w_{k,j}}$
  - Update weight
    $w_{k,j} \leftarrow \beta w_{k,j}$ if $e_j \neq c(d_i)$ and $fcp_{k,j} \in e_j$

Figure 3: Domain-Experts algorithm.

the category. But Sleeping expert did not allow a classifier to have the inconsistent hypotheses.

The self-confident sampling method which we have proposed in figure 2 shares a property of the uncertainty-sampling [9], in that it predicts the label of an unlabeled data on the basis of the learner's confidence which is obtained through the training phase. The examples that are categorized with the least uncertain will be added to the training set in the next iteration. Unlike uncertainty-sampling, our method rely only on the vote by each of member of domain experts group, which has knowledge induced from the labeled data. We, however, could not rely on its knowledge completely. In this regards, λ is introduced for regulating the degree of reliance on learner's experience. Empirically, the proposed sampling method shows the best performance at 70 % confidence.

The voting process is determined by the value of vote entropy. Vote entropy is the entropy of the class label distribution resulting from having each committee member deterministically "vote" for its winning class [4]. Let $V(j)$ be the number of domain experts which are involved in 'voting' for d_i for the category j:

$$VE(d_i) = -\sum_{j}^{|C|} \frac{V(j)}{|K|} \log \frac{V(j)}{|K|} \quad \text{if } fcp_k \in d_i, \tag{4}$$

where $|K|$ is the number of domain experts which took parts in voting of ith data, d_i which is ith data from the unlabeled data set.

While the vote entropy is 0 if a number of domain experts participating in the vote belong to the same category, the vote entropy is 1 when the vote committee is consist of an equal number of each category. We found empirically that the average vote entropy for category assigned correctly was 0.25, whereas the average entropy for incorrectly classified data was 0.66. This approach is similar with active learning in terms of voting among the constituent of a classifier. Our sampling method specifies a process that determines the label of a new article, whereas active learning guides the direction that focuses what kind of instance is useful for learning.

Data	+	+?	+/−	−?	−	Total
Labeled	239	70	526	60	344	1239
Unlabeled	–	–	–	–	–	5000
Total	239	70	526	60	344	6239

Table 2: The number of news articles for each category. Relatively smaller data in "not, certain" category could be explained by the objective contents of financial news article in terms of "good" or "bad."

5. EXPERIMENTAL RESULTS

In this section, we describe the results of the proposed methods, as compared with conventional methods.

As mentioned earlier, experiments were performed using the text data which we had made by ourselves. The labeled financial news articles data set amounts to 1,239 financial news articles gathered from various electronic news sources: CNN Financial Network, Forbes, Reuters/Reuters Securities, NewsFactors, Motley Fool, CNet, ZDNet, Morningstar.com, Business Week, AP Financial, Business Wire, PR News Wire, and Associated Press. Table 2 describes the distributions of news articles for each category. A phenomenon that "neutral" category has more data than others can be explained by the majority of report-type news article. That is, larger part of them are classified into the neutral category because they did not mentioned anything about the company's financial well-being explicitly, but deal with general information about the company.

Experiments aimed to verify the proposed methods in terms of two performance criterion: how well it make use of unlabeled data for improving categorization accuracy and how accurately it categorize the latest news articles into predefined categories. Firstly, we evaluated whether the proposed sampling method would improve categorization performance rates than those trained by conventional methods. A experiment was performed to show the performance of domain experts with self-confident sampling, naive Bayes with EM, domain experts with EM and naive Bayes with self-confident sampling. Through the experiments, about 25% of the labeled data was used for testing. The rest of labeled set were used by the learning algorithm to produce classifiers. After training phase, each methods was tested in terms of classification accuracy: the proportion of the number of news articles classified correctly to the number of total news articles used. A domain experts group is empirically made up of 200 domain experts (i.e. FCP).

Figure 4 and 5 show results of testing the accuracy performance of each sampling method with different number of labeled data. Total 50 iterations were carried out for each method. At each iteration, 50 unlabeled news article were given to each methods and used for improving its performance. When 690 out of 1239 labeled data feeds on training, the performance of the proposed method, the combination of Domain Experts and Self-Confident Sampling, is

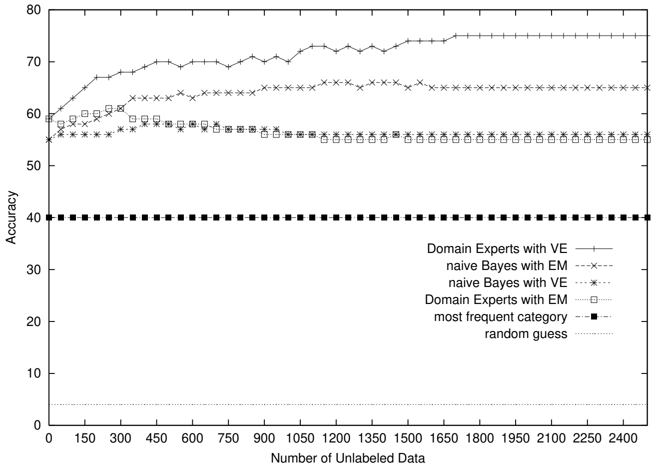

Figure 4: A result of sampling experiment was represented after training each of methods with 690 labeled data. "Most frequent category" and "random guess" are base line of performance. Since about 40% (526/1239) news articles out of labeled data are "neutral" category, we can assume that a method is able to gain 40% accuracy when it answers consistently the label of each news article in test set as "neutral."

	+	+/?	+/-	-/?	-	average
DM&VE	.81	.58	.93	.65	.8	75.42
bayes & EM	.64	.38	.94	.45	.76	63.4
bayes & VE	.59	.46	.88	.29	.6	56.5
DM & EM	.5	.19	.79	.19	.5	44

Table 3: The accuracy of each combination to each category. Each column specifies the proportions of the number of news article classified correctly to the number of news articles for each category when it shows the best accuracy performance.

going up until sampling of 1,750 unlabeled news articles, and shows the best performance on accuracy measure at the point. Table 3 describes the accuracy for each category when each method shows its best performance. Very similar performance was shown when all labeled data was used for training in 5. From this, we assumed that around 2,000 news articles allow us to make a classifier with 75 % accuracy because it seems largely depend on the fact that most of news companies used to deliver financial news with a restricted vocabulary set. With self-confident sampling, 16% accuracy is improved with 56 % of labeled data (690/1239) and 35% of unlabeled data (1,750/5000) from the result in figure 4. As another goal of our task is to categorize the label of the on-line financial news articles, the second experiment was performed to show the accuracy of categorization of the latest financial news data. A online data set is made up of the news articles that can be gathered from the same news sources as the labeled data set and reports the latest financial news at the experimental time. Since some of financial news articles, however, deal with general report-type news, such as the introduction of new technology, the announcement of merger and personnel management, and

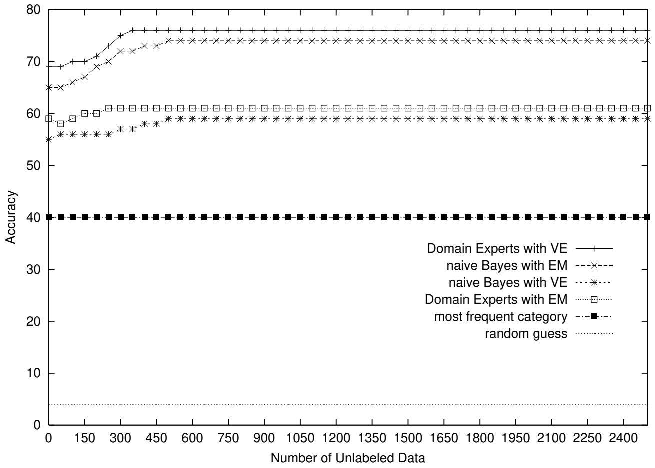

Figure 5: A result of sampling experiment was represented after training each of methods with 1239 labeled data.

etc., it is necessary to filter out non-financial news from the online data set. Figure 6 shows an interface of which is called "GoodNews". In this interface, the left side on top shows a tree-type classification of non-financial news, such as "strategy", "lawsuit", "M&A", "product", and "miscell." the right side of on top is a stack that presents the retrieved news articles and gets evaluations from user. Each clicked news article is displayed on the bottom part.

With filtering out non-financial news, we could get 549 financial news articles out of the latest 1,168 news articles which are made up of 2 trials for each company out of 20 companies. At each trial, a company name is given to 6 from Warren agent and 30 news articles was gathered from various news sources. However news article about a minor company could not meet the number of test documents at a trial (i.e. 30) The second row of Table 4 tells us the distribution of online test set after filtering out non-financial news. As a result, the proposed method has 79% averaged accuracy, which means 433 out of 549 total financial news articles were categorized correctly. Table 4 shows the accu-

Figure 6: The interface of GoodNews.

categories	+	+/?	+/−	−/?	−	total
articles	85	1	243	0	220	549
DE	.76	1	.8	–	.78	.79
naive	.61	0	.68	–	.62	.65

Table 4: Accuracy measure of each category to the online news data. Each column at the third and fourth row represents the accuracy of each category in terms of the proportion of the number of news article classified correctly to the total number of news article for the category. That the column about "good, uncertain (+/?)" category has the value of 1 means that only one news article which is labeled by human to that category was categorized correctly. There is no news articles about "bad, uncertain (−/?)" category.

racy of tested methods per each category.

6. DISCUSSIONS

Comparing with the result of the combination of naive Beyes and EM, the better result of the proposed method could be explained by the fact that the proposed method of which exploits the collocation property and vote entropy are appropriate to such an objective domain. In other words, categorization of news articles into the predefined categories of which their decision boundaries would be relatively objective, such as "good" or "bad" and confined with a specific company's name is not difficult as much as categorization of news articles into categories which its boundary would be relatively unclear depending on a perspective, such as politics, health, agriculture, etc.

But the proposed method has several weak points that prevent it from reaching the performance above 75 % accuracy. One is the difficulty in determining the label of news article of which made up of commensurate number of co-located phrases of each category. To illustrate, "Shares of company B rose 5 % in contrast with company A of which shares fall 7 %." In this example, domain experts may fail to predict "good" for company B. Because both phrases, which are "shares" with "rose" and "shares" with "fall", are very strong indicators of company's financial well-being at the moment, even though they are not assigned with the same phrase weight during the training phase. As mentioned earlier, another weak point is that the proposed method does not consider the co-referred sentence. In other words, that it does not consider sentences, which did not mention company's name or ticker explicitly, as the financial evidence. For example, "Company C expects to boost revenue next quarter, Chief Operating Officer xxx said Wednesday. Despite of these anticipation, the company's shares fall again." In here, the prediction by the proposed method could be "good, uncertain", even though the true label of this example might be "bad" because "Company C" and "the company" are co-referred as the company's current financial well-being is not good.

7. CONCLUSIONS

We introduced an application of text categorization that categorizes financial news articles by considering referred company's financial well-being from their contents. The proposed algorithm which observed co-located phrase of certain category from news contents and predicted the label with Weighted-Majority voteing outperform naive Bayes classifier about 14 %. For further improvement of accuracy, we proposed a sampling technique of which determine the category of an unlabeled news article with its entropy value. With the proposed sampling method, self-confident sampling, 16% accuracy is improved with 56 % of labeled data (690/1239) and 35% of unlabeled data (1,750/5000).

The successful results from sampling test and online test supports that the proposed algorithms effectively works in this task, even though the promising results are partially come from the task characteristics of which its decision boundaries are relatively objective and are confined with a specific company's name.

In future work, we will compare the performance of the proposed methods with the established ones, such as k-NN, SVMs, and Maximum Entropy model. For the purpose of verifying the applicability of proposed method, we also are about to try to apply the proposed method to the domains of which has similar characteristics to our task. To cope with the problems mentioned in 6 section, we consider to employ some natural language processing techniques, such as the consideration of more wide range of a sentence and resolution of co-reference.

8. ACKNOWLEDGEMENTS

We would like to thank Massimo Paolucci and Sean Owens for their maintenance of the Warren financial portfolio agent system. This research has been sponsored in part by DARPA Grant F-30602-98-2-0138 and the Office of Naval Research Grant N-00014-96-16-1-1222.

9. REFERENCES

[1] C. Cleverdon. Optimizing convenient online access at biliographic database. *Information Services and Use*, pages 37–47, 1984.

[2] W. W. Cohen and Y. Singer. Context-sensitive learning methods for text categorization. In *Proc. of Int'l ACM Conf. on Research and Development in Information Retrieval (SIGIR-96)*, pages 307–315, 1996.

[3] D. A. Cohn, L. Atlas, and R. Ladner. Improving generalization with active learning. *Machine Learning*, 15(2):201–221, 1994.

[4] I. Dagan and P. Engelson. Committee-based sampling for training probabilistic classifiers. In *Proc. of Int'l Conf. on Machine Learning (ICML-95)*, 1995.

[5] K. Decker, K. Sycara, A. Pannu, and M. Williamson. Designing behaviors for information agents. In *Proc. of Int'l Conf. on Autonomous Agents (AA-97)*, pages 404–413, 1997.

[6] P. Domingos and M. Pazzani. Beyond independence: Conditions for the optimality of the simple bayesian classifier. *Machine Learning*, 29:103–130, 1997.

[7] S. Dumais. Using svms for text categorization. *IEEE Intelligent Systems*, 13(4), 1998.

[8] T. Joachims. Text categorization with support vector machines: Learning with many relevant features. In *Proc. of European Conf. on Machine Learning*, pages 137–142, 1998.

[9] D. D. Lewis and W. Gale. Training text classifiers by uncertainty sampling. In *Proc. of Int'l ACM Conf. on Research and Development in Information Retrieval*, pages 3–12, 1994.

[10] R. Liere and P. Tadepalli. Active learning with committees for text categorization. In *AAAI-98 Workshop on Learning for Text Categorization*, 1998.

[11] C. D. Manning and H. Schutte. *Foundations of Statistical Natural Language Processing*. MIT Press, 1999.

[12] A. McCallum and K. Nigam. Employing em and pool-based active learning for text classification. In *Proc. of Int'l Conf. on Machine Learning (ICML-98)*, pages 359–367, 1998.

[13] T. Mitchell. *Machine Learning*. MIT Press, 1997.

[14] K. Nigam, J. Lafferty, and A. McCallum. Using maximum entropy for text categorization. In *IJCAI-99 Workshop on Machine Learning for Information Filtering*, pages 61–67, 1999.

[15] K. Nigam, A. McCallum, S. Thrun, and T. Mitchell. Text classification from labeled and unlabeled documents using em. *Machine Learning*, 39(2/3):103–134, 2000.

[16] K. Sycara, K. Decker, and A. Pannu. Distributed intelligent agents. *IEEE Expert*, 1996.

[17] Y. Yang and X. Liu. A re-examination of text categorization methods. In *Proc. of Int'l ACM Conference on Research and Development in Information Retrieval (SIGIR-99)*, pages 42–49, 1999.

[18] Y. Yang and J. O. Pedersen. A comparative study on feature selection in text categorization. In *Proc. of Int'l Conf. on Machine Learning (ICML-97)*, pages 412–420, 1997.