Cross Training and Under Sampling in Categorization of Company Announcements

Cheng G. Weng
School of Information Technologies
The University of Sydney
Sydney NSW 2006, Australia
cheng@it.usyd.edu.au

Josiah Poon
School of Information Technologies
The University of Sydney
Sydney NSW 2006, Australia
josiah@it.usyd.edu.au

Abstract *To process the documents in a share market is crucial. It is because financial activities are socio-economic driven and text documents contain a lot of valuable information. In this paper, we focus on one of these documents, the Company Announcement. Each of these documents requires to be labelled as price sensitive or not before presenting to the general public. In our experiments, we study two specific issues in this text categorization, namely the effectiveness of a feature vector obtained from the corpus belonging to another market sector and the imbalanced nature of the dataset. Our results indicate that the classification can benefit from a different (but related) set of corpus because of a more diversified and generalised nature of the feature set. Regarding the skewness of the dataset, the under-sampling of the majority class in the training process does not have a strong effect on the performance in the test set, while keeping the computational cost minimised.*

Keywords Document Management, Text Categorization

1. Introduction

Information overload is a major problem in this new age of World Wide Web, and the majority of these information are in textual form. The need to manage and utilize textual information has led to the birth of text mining research. In the financial domain, many important decisions making are based on the assessment of text documents, and the timing of these decisions is also crucial. It has, therefore, attracted researchers to work on this area [4, 6, 8].

In this paper, we used company announcements obtained from the Australia Stock Exchange (ASX) website and tried to explore more effective ways of training on text documents. The next section describes some related work, follow by a section that introduces our dataset, then section 4 elaborates the motivations for the experiments describes in section 5. We present the experimental results in section 6 and discuss the observations in section 7. Finally, we will conclude with some future work.

Proceedings of the 10th Australasian Document Computing Symposium, Sydney, Australia, December 12, 2005.

Dataset	Sen	Nonsen	Total	Sen%
Website	1472	9604	11076	13%
Signal G	46530	90100	136630	34%

Table 1. Dataset comparison with [1]. *Website* is the dataset we used for our experiment, and *Signal G* is the name of the dataset used in [1].

2. Related Work

Our approach is based on the traditional statistical text categorization process, which transform text documents into word vectors. Essentially, treating the documents as 'bag-of-words' and ignore other linguistic information. It is a rather superficial approach but it has been shown to be effective in practice [7].

To the best of our knowledge, there was one closely related work done by Calvo and Williams [1]. They also used the announcements, but they had access to a larger dataset, because of their affiliation with the Capital Market CRC[1], which we did not have. They compared the performance of different machine learning algorithms, and concluded "the good performance shows the possibility for commercialization". Table 1 shows the differences of our dataset to theirs.

Since our dataset has a skewed class distribution, we used the same experimental setting as in [1] to check the representativeness of our company announcement sample. We used random under-sampling to accommodate the class imbalance of our dataset, and this sampling technique has been shown useful for non-textual datasets with class imbalance [2].

3. Dataset

Before describing our motivations for the experiments, we first provide some background knowledge about our dataset, the ASX company announcements dataset. These company announcements are manually categorized by their market sensitivity, which is either

[1] It is a joint research centre with stakeholders coming from the industry and the academic institutions, of which ASX is one of the industrial partners.

Code	Sen	Nonsen	Sen%	Unique	Sector
ANN	21	273	7%	11703	H
CSL	25	172	13%	12118	H
SIG	13	86	13%	7229	H
VCR	30	132	19%	6232	H
ANZ	54	637	8%	27312	F
CBA	38	1816	2%	25035	F
MBL	97	2523	4%	68943	F
NAB	80	674	11%	24219	F
AWE	137	428	24%	11374	E
ROC	164	460	26%	12496	E
STO	265	455	37%	11924	E
WPL	113	310	27%	12281	E
BHP	148	325	31%	22168	M
BSL	39	329	11%	16395	M
ORI	65	350	16%	10495	M
RIO	70	158	31%	17692	M
ERG	46	95	33%	14998	I
IFM	16	118	12%	9966	I
MYO	23	189	11%	882	I
VSL	28	74	27%	6460	I

Table 2. Company Statistics.

"market sensitive" or "market non-sensitive". The market sensitivity of an announcement depends on its predicted effect after release to the general public, and this will be judged by the experts in the ASX. If the information could potentially have significant impact on the market, they will be labeled as market sensitive, otherwise market non-sensitive. Because of the subjective nature and the difficulty of knowing the exact impact of an announcement, the experts are more conservative when labeling a document as market sensitive. Therefore, while some announcements are labelled as market sensitive, it is not a guarantee that it did have a significant impact on the market. A sensitive document is considered as a rare and important event in this task, so the cost associated with misclassified sensitive document is high, e.g. ill-informed investors may miss out their best buy/sell timing for their stocks.

The announcements are publicly available in PDF format from the ASX website[2]. We have collected announcements for 20 listed companies on ASX200, each with more than 2-year worth of data, from early 2003 to early 2005. We kept the corpus separated by individual company, and the size of the whole corpus is about 175 MB. Each company belongs to a market sector, which is defined by the Global Industry Classification Standard (GICS).

Table 2 shows the basic statistics for each company. The 'Code' is the trading code of the company on ASX, 'Sen' is the number of sensitive documents, 'Nonsen' is the number of non-sensitive documents, 'Sen%' is the percentage of sensitive documents in the company, and 'Unique' is the number of unique tokens after pre-processing (described in section 5). 'Sector' is the market sector, where *H* stands for Health Care sector, *F* stands for Financial sector, *E* stands for Energy sector, *M* stands for Material sector, and *I* stands for Information Technology sector.

[2] ASX `http://www.asx.com.au`

4. Motivations

The general assumption is that one needs to use a set of training data that resemble the future test data, in order to obtain the best generative model that minimize the test error. We have, therefore, proposed to test this assumption in a text categorization task.

The text categorization task can be viewed as two steps: step one is the selection of features, which we call *TCFeature* (TCF), and step two is the modeling of the data after they are being represented in vector form using the features selected from step one, and we call this step *TCModel* (TCM). In the traditional approach, both TCF and TCM are performed on the same company. So our hypothesis is that a similar performance can be acheived when the TCF is operated on a different company but belonging to the same market sector as the testing company. From here on, the notation TCF and TCM will be used in the rest of the paper.

Next, we tested the effect of under-sampling by randomly remove the non-sensitive documents, the majority class, from the data. We assumed this would give us a better performance on the minority class, which is what we prefer.

5. Experiments

Although both our dataset and [1] came from the Australian Stock Exchange, they are different. Therefore, the aim of the first experiment (**Exp1**) is to find out if the difference will constitute any significant performance variance. For **Exp1**, we used the same setup as in [1], which was a typical process in text categorization. Starting with stopwords removal, the apply the Porter word stemming algorithm [5], perform the feature selection, index the documents with TFIDF [7], and finally build the model with a machine learner. They used document frequency for feature selection to select 1000 features, and Support Vector Machines (SVM) was one of the machine learner they have used.

For our other experiments, we used similar setting, except, we chose information gain as the feature selection method to select 3000 features and SVM as the machine learning method. These choices were made because our focus was not on the feature selection nor the machine learner, and also the previous studies have shown good results when applying the two methods in text categorization [7]. All experiments are evaluated with 10 fold cross-validation.

The second experiment was set up to compare the performance, when different companies are used for TCF. There are two parts to this experiment: in the first

	Micro			Macro		
	p	r	F_1	p	r	F_1
L	0.82	0.82	0.82	0.80	0.79	0.80
S	0.90	0.90	0.90	0.79	0.72	0.75

Table 3. *Exp1:* Comparison with [1]. *L* is the results from [1], and *S* is our results.

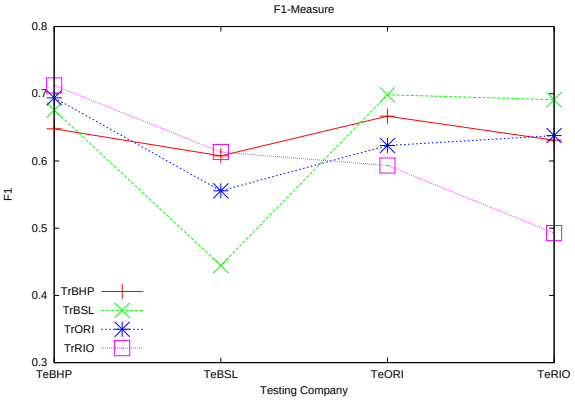

Figure 1. *Exp2a*: results for the Material sector. Each line represents TCF on a company, and x-axis is which company is used for testing.

part the different companies belongs to the same market sector as the test company (**Exp2a**). In the second part the different companies belongs to different market sectors as the test company (**Exp2b**). The control for this experiment is the one when TCF and TCM are both performed on the same company.

The third experiment tests the effect of random under-sampling of non-sensitive documents from a company. The setup is identical to the second experiment, except the data of each company will be reduced at the training stage. We tried 11 different reduction rates ranging from 0%, the original corpus, to 100%, where only sensitive documents are kept. Again, there are two parts to the experiment: the first part is when TCF and TCM are done on the same company (**Exp3a**), and the second part is when TCF and TCM are done on different companies (**Exp3b**). In the second part of the experiment, because TCF is independent of TCM, the features will remain consistent, while the training data for TCM gets the reduction.

6. Results and Evaluation

The micro and macro averages [7] was reported in [1], hence, we did the same calculation for comparison in **Exp1**. Table 3 shows we have a similar result, which suggests that our dataset is a compatible sample.

Due to space limitation, we will only report subset of the results for our experiments **Exp2** and **Exp3**. But this subset of the results is representative and the reported observations hold for all other results.

The chart shown in Figure 1 is one of the results for experiment **Exp2a**, while Figure 2 shows the results for **Exp2b**. We observed three phenomena:

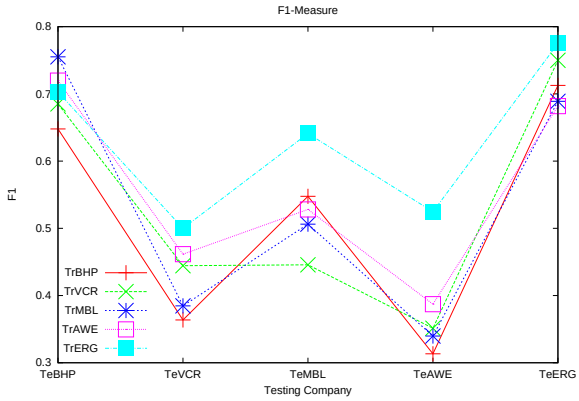

Figure 2. *Exp2b*: Results for TCF on companies in different market sectors. All 5 companies shown here belongs to different market sectors.

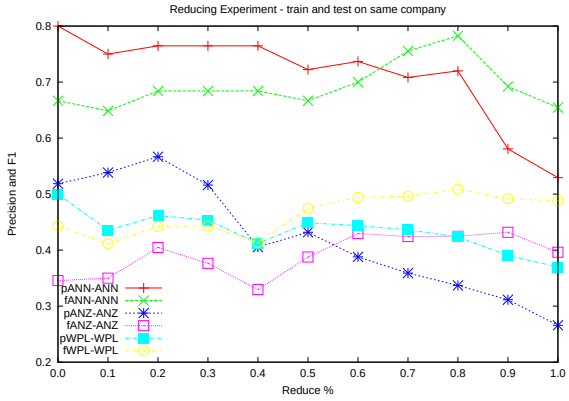

Figure 3. *Exp3a*: Reducing non-sensitive documents for ANN, ANZ, and WPL. *pANN-ANN* stands for the precision of TCF on ANN and testing on ANN, and *fANN-ANN* is the F1 measure. The same notation applies to others. The "Reduce %" is the percentage of non-sensitive documents removed.

Observation 1: Perform TCF and TCM on the same company does not always give the best results, which seems to be counter-intuitive. For all 20 companies, we found only 7 cases, where performing TCF and TCM on itself give above average performance. However, it was never the best, only 2 out of the 7 it was in top 5.

Observation 2: From Figure 2, the classification performance does not have significant difference whether the TCF was done on companies belonging to the same or different market sector.

Observation 3: The performance rises for certain companies but dips for another. The performance of a company being tested seems to be bound by some company dependent variable.

The third experiment **Exp3a** is shown in Figure 3, and Figure 4 shows the results for **Exp3b**. They have exhibit another 2 phenomena:

Observation 4: Reducing, up to 50% of, the non-sensitive documents does not deteriorate the performance. When the reduction is greater than

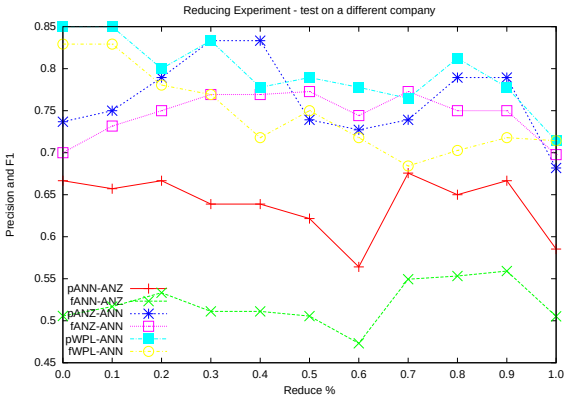

Figure 4. *Exp3b*: Reducing experiment for testing on other companies. The notation used here is the same as in figure 3.

50% the F1 measure does not drop, because the recall boost compliments the precision drop.

Observation 5: When the TCF is done on other companies, even if the non-sensitive documents are absent, the TCM can still be done with reasonable performance.

7. Discussion

In the previous section, we have shown a similar result compare with [1]. The higher micro average is due to the difference in performance of our two classes. The performance for non-sensitive class is much higher than the sensitive class. Next, we will discuss the observations made in our experiment.

Observation 1 suggests the feature vector can be constructed from another company has better generalization, i.e. the features are more diversified and generalised, and observation 2 suggests there is generic attribute that does not change across market sectors.

Observation 3 suggests the possibility of company dependent variables effecting the performance. So we attempted to look for a correlation between the performance and varies company dependent statistics, but none has shown a strong correlation for conclusion. The statistics we have tried are the class distribution, the number of documents, the percentage of sensitive documents, the number of unique tokens, the number of unique tokens in sensitive and non-sensitive document separately, and the overlapping of unique tokens of sensitive and non-sensitive documents.

From observations 4 and 5, we see a discriminative feature vector can maintain the performance even when the documents in the majority class is removed. The word usage can still be modeled without compromise the performance. When we looked at the features selected from different reduction rate in **Exp3a**, we found high overlapping of features among the feature vectors. This suggests similar features can still be selected without most of the non-sensitive documents.

8. Conclusion and future work

Our hypothesis in Section 4 aimed to explore the building a feature vector from a different corpus, and also the effect of random under-sampling on the dataset. We discovered even though the feature vector was constructed from a different corpus, they still give good performance, and frequently outperform the feature vector generated by the training corpus itself. We also found that random under-sampling of majority class does not deteriorate the F1 measurement, even when the majority class is completely removed. These empirical evidences suggests the possibility of a more effective text categorization process, by obtaining feature vectors from a better source, and model the word usage with a smaller subset of documents.

An interesting future work should be testing the same procedure on other standard text categorization copra and see if our findings still occurs. Also, the imbalance nature of the dataset would be another interest direction for further investigation [3].

References

[1] R.A. Calvo and K. Williams. Automatic categorization of announcements on the australian stock exchange. In *7th Australasian Document Computing Symposium*, Sydney, Australia, 2002.

[2] N. Chawla, K. Bowyer, L. Hall, and P. Kegelmeyer. Smote: Synthetic minority over-sampling technique. In *International Conference on Knowledge Based Computer Systems*, 2000.

[3] N. Japkowicz and S. Stephen. The class imbalance problem: A systematic study. *Intelligent Data Analysis*, 6(5):429–450, 2002.

[4] Antonina Kloptchenko, Tomas Eklund, Jonas Karlsson, Barbro Back, Hannu Vanharanta, and Ari Visa. Combining data and text mining techniques for analysing financial reports. *Intelligent Systems in Accounting, Finance and Management*, 12:29–41, 2004.

[5] Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.

[6] G. Pui Cheong Fung, J. Xu Yu, and Wai Lam. Stock prediction: Integrating text mining approach using real-time news. In *Computational Intelligence for Financial Engineering, 2003. Proceedings. 2003 IEEE International Conference on*, pages 395–402, 2003. TY - CONF.

[7] Fabrizio Sebastiani. Machine learning in automated text categorization. *ACM Comput. Surv.*, 34(1):1–47, 2002.

[8] Dongsong Zhang and Lina Zhou. Discovering golden nuggets: data mining in financial application. *Systems, Man and Cybernetics, Part C, IEEE Transactions on*, 34(4):513–522, 2004. TY - JOUR.