Financial Forecasting using Character N-Gram Analysis and Readability Scores of Annual Reports

Matthew Butler and Vlado Kešelj

{mbutler,vlado}@cs.dal.ca, Faculty of Computer Science, Dalhousie University

Abstract. Two novel Natural Language Processing (NLP) classification techniques are applied to the analysis of corporate annual reports in the task of financial forecasting. The hypothesis is that textual content of annual reports contain vital information for assessing the performance of the stock over the next year. The first method is based on character n-gram profiles, which are generated for each annual report, and then labeled based on the CNG classification. The second method draws on a more traditional approach, where readability scores are combined with performance inputs and then supplied to a support vector machine (SVM) for classification. Both methods consistently outperformed a benchmark portfolio, and their combination proved to be even more effective and efficient as the combined models yielded the highest returns with the fewest trades.

Key words: automatic financial forecasting, n-grams, CNG, readability scores, support vector machines

1 Introduction

The Securities and Exchange Commission (SEC) requires that each year all publicly-traded companies supply a third-party audited financial report, which states the company's financial position and performance over the previous year. Contained in these annual reports, inter alia, are financial statements, a letter to the share-holders, and management discussion and analysis. Over the years several research endeavours have been focused on the numbers contained in the financial statements, computing a variety of ratios and price projections without considering textual components of the reports. Peter Lynch, a famous investment "guru," once said that "charts are great for predicting the past," pointing out that there is more to making good investments than just processing the numbers. The textual components give insight into the opinions of the senior management team and provide a direction of where they feel the company is going. This information should not be trivialized or overlooked; it should be processed in a similar way to processing quantitative information, to extract meaningful information to aid in the forecasting process. Up until recently an analyst would have to read an annual report and use their expertise to determine

if the company is going to continue to do well or if there is trouble ahead. They would apply their skill and judgment to interpret what the Chief Executive Officer (CEO) is saying about the company and its direction for the future. This process can be very time consuming and it is a somewhat heuristic approach, considering that two experienced analysts could read the same report and have a different feeling about what it is saying. If an analyst has several companies to consider and even more annual reports to read it could be difficult to take in all the relevant information when it is most likely surrounded by noise and other erroneous information that has no effect on the stock price. Most numeric calculations can be automated to remove human error, and complex data mining and machine learning algorithms can be applied to extract meaningful relationships from them. It would be extremely valuable if the same could be done for the textual components, having a quick, efficient and accurate tool to analyze an annual report and make recommendations on its implications for the stock price over some given time period. This could erase some of the subjective judgments that arise from an individual's interpretation of the report, which could change from person to person. Also, given the sheer amount of annual reports that are produced each year, one would be able to analyze a larger number of companies and have a greater opportunity to find good investments.

In this paper an attempt is made at achieving this goal: two novel approaches to analyzing the text are put forward and then a combined model is also analyzed to see if a union of these approaches is more robust. The first novel technique is to convert the textual components to n-gram profiles and use the CNG distance measure [1] as proposed by Kešelj *et al.* to classify reports. The second is to generate three readability scores (Flesch, Flesch-Kincaid and Fog Index) for each report, and after combining with the previous year's performance, make class predictions using a support vector machine (SVM) method. The combined model will only make a recommendation on a particular annual report when the two models are in agreement; otherwise, the model outputs no decision. The models make predictions whether a company will over- or under-perform S&P 500 index over the coming year. This is an appropriate benchmark as all the companies being analyzed are components of this index. We believe that this is a very meaningful comparison. In some published results, performance of an algorithm was evaluated by measuring how accurately one can predict increase or decrease of a stock price. This evaluation approach may lead us to believe that an algorithm has a good performance, while it may be worse than the index performance. Hence it would be useless to an investor, who could simply invest in the index, achieve higher return, and be exposed to lower risk.

2 Related Work

As text processing techniques become more sophisticated its ability to work in the financial domain becomes more attractive. There has been a few publications in which textual information was analyzed in relation to financial performance. In comparison, the novelty of our approach is in applying character n-gram analysis and readability scores with the SVM method to the annual reports in

making long-term predictions. Pushing the time-horizon for making predictions creates a more practical model, and thus it has a wider appeal in the investment industry. In [2], the effects of news articles on intra-day stock prices are analyzed. The analysis was conducted using vector space modeling and tfidf term weighting scheme, then the relationship between news stories and stock prices was defined with a support vector machine [2]. The experiments produced results with accuracy as high as 83% which translated to 1.6 times the prediction ability when compared to random sampling. Similarly, Chen and Schumaker (2006) [3] compared three text processing representations combined with support vector machines to test which was the most reliable in predicting stock prices. They analyzed the representations based on bag-of-words, noun phrases and named entities, and all of the models produced better results than linear regression; however named entities proved to be the most robust[3]. Other intra-day predictions facilitated through text mining were done by Mittermayer (2004) [4], where he created NewsCATS—an automated system that could day-trade the major American stock indexes. The model was created to automate the trading decisions based on news articles immediately after they are released. Kloptchenko *et al.* (2002) [5] focused on clustering quarterly financial reports in the telecom industry. They were not making predictions on future performance but attempting to use prototype-matching text clustering and collocational networks to visualize the reports. The collocational networks cut down the time required by an analyst to read the report and identify important developments [5]. This work was improved upon for making predictions and the new results (Kloptchenko *et al.* 2004) [6] were released new results, in which prototype-matching text clustering for textual information was combined with self-organizing maps for quantitative analysis. Their analysis was performed on quarterly and annual financial reports from three companies in the telecom industry. The results implied that some indication about the financial performance of the company can be gained from the textual component of the reports; however, it was also noted that the clusters from quantitative and qualitative analysis did not coincide. They explained this phenomenon by stating that the quantitative analysis reflects past performance and the text holds information about future performance and managerial expectations. Before complex text mining methods were developed, the work done by Subramanian, Insley, and Blackwell [7] in 1992 showed that there was a clear distinction between the readability scores of profitable and unprofitable companies. In more recent work by Li [8], he examined the relationship between annual report readability combined with current earnings and earnings persistence, with a firm's earnings. His conclusion was that firms with lower earnings had reports which were more difficult to read and longer.

3 Data Pre-processing

3.1 Data Collection

There are no known publicly available data sets that would contain a pre-processed sample of annual reports to analyze, so the data set was created from scratch. To facilitate this, the website of each company considered was visited

and the relevant annual reports were downloaded from the investor relations section. Prior to downloading, every report's security features were checked to ensure the PDF was not protected; if it was, then it was discarded as the file could not be converted to text (text format is required to apply n-gram and readability programs). Once a sufficient sample size of annual reports was collected, they were converted to text using a Perl script with program `pdftotext`.

3.2 Data Labeling

The most sensitive and time consuming process of the experiment was class labeling of the training and testing data. It is not mandated by the SEC that companies file their annual reports at the same time, so as a result, each performance measure has to be individually calculated for each company, based on different months. To expedite this process, a matrix of relative returns was created based on monthly closing prices for each stock from data obtained from Yahoo! Finance [9]. The returns for each month were calculated as a numeric figure, and introduced as a class attribute as either over or under performing the S&P 500 over the trailing 12 month period. Next, the filling date for the reports was captured from the SEC website and the appropriate text file is labeled. This was done manually for each report.

3.3 Generating N-gram Profiles

The n-gram profiles were created as defined by the CNG method [1] using the Perl n-gram module Text::Ngrams developed by Keselj [10]. The character six-grams and word tri-grams were used, and various profile lengths up to 5000 unique, normalized, most-frequent n-grams from an annual report were used.

3.4 Generating Readability Scores

A Perl script was created that generated the three readability scores from source code developed by Kim Ryan [11] and made publicly at CPAN [12]. The scores for each annual report are combined with the underlying securities' 1-year past performance to form the input attribute set for the SVM. The previous year's performance was represented in two ways: first by its relative performance to the S&P 500, and by an indicator whether or not it decreased or increased in value over the last year. To make the data appropriate for the SVM it was scaled between 0 and 1 to cut down on computation size and transformed into the required format. The three readability scores considered where the Gunning Fog Index, Flesch Reading Ease, and Flesch-Kincaid Grade Level. The Gunning Fog Index developed by Robert Gunning in 1952 is a measure of readability of an English sample of writing, the output is a reading level that indicates the number of years of formal education required to understand the text, and the equation is as follows:

$$\textit{Gunning Fog Index} = 0.4 \cdot \left(\frac{\#\textit{words}}{\#\textit{sentences}} + 100 \cdot \frac{\#\textit{complex words}}{\#\textit{words}} \right)$$

where *#words* is the number of words in text, *#sentences* number of sentences, and *#complex words* number of words that are not proper nouns and have three

or more syllables. The Flesch Reading Ease (FRE) and Flesch-Kincaid Grade Level (FKL) were both created by Rudolph Flesch. The higher the FRE score the simpler the text and the output for the FKL is similar to the Gunning Fog Index, where it generates a Grade Level that reflects the number of years of formal education required to understand it. The two scores are imperfectly correlated and therefore it is meaningful to consider them both. Their respective equations are given below:

$$\textit{Flesch Reading Ease} = 206.835 - 1.015 \cdot \frac{\#words}{\#sentences} - 84.6 \cdot \frac{\#syllables}{\#words}$$

$$\textit{Flesch-Kinkaid Grade Level} = 0.39 \cdot \frac{\#words}{\#sentences} + 11.8 \cdot \frac{\#syllables}{\#words} - 15.59$$

The algorithm for syllable count was implemented as the Perl module Lingua::EN::Syllable [13], with estimated accuracy of 85–90%.

4 CNG Classification of N-gram profiles

The n-gram classification technique was inspired by work done by Kešelj, Peng, Cercone and Thomas [1], where n-gram profiles were used, with a high degree of accuracy, to predict author attribution for a given unlabeled sample of writing. A generalized profile for a given author was generated and then used to gauge a distance calculation from new testing documents. For financial forecasting a general n-gram profile was created from all of the company annual reports for a given class. The classifier would concatenate all the files from one class or another and then generate one overall n-gram profile with the same settings as discussed in the data pre-processing subsection. For each testing year x the training profiles would be generated from years $x-1$ and $x-2$. Once the two generalized profiles are created, one for over-preforming and one for under-performing stocks, the profiles of documents from the testing year are compared with the training profiles using the CNG distance measure:

$$\Sigma_{s \in profiles} \left(\frac{f_1(s) - f_2(s)}{\frac{f_1(s)+f_2(s)}{2}} \right)^2$$

where s is any n-gram from one of the two profiles, $f_1(s)$ is the frequency of the n-gram in one profile, or 0 if the n-gram does not exist in the profile, and $f_2(s)$ is the frequency of the n-gram in the other profile.

5 SVM Classification with Readability Scores

The input attributes to the SVM method where vector representations of the annual reports that contained the three readability scores and the stock's performance over the previous year. An SVM is a very robust classifier that has proven effective when dealing with highly complex and non-linear data, which is indicative of data found in the financial domain. SVM's had been widely experimented with financial forecasting in both classification [14–16] and level

estimation or regression [17] domains. Because the scores are not time sensitive and the SVM does not take into account any time dependencies when evaluating the data, all of the vector representations were used to train the system, except for the particular year it was tested on at any given time. The Support Vector Machine environment utilized was LIBSVM [18]—a very powerful integrated software for support vector classification and regression. It utilizes an SMO-type algorithm[14] for solving the Karush-Kuhn-Tucker (KKT) conditions. A polynomial kernel of degree 3 was used, with the c-SVM approach; i.e., the use of slack variables to allow for "soft" margin optimization.

Five input attributes are used in SVM classification: three readability scores from annual reports, and two performance measures in the previous year: one whether the stock over or under performed, and the second whether the stock price increased or decreased in the previous year.

6 Experimental Results

In general all three individual models and the two combinations preformed well and overall, they each outperformed the benchmark return in the testing period. To display the results, a special attention is given to the three criteria: overall accuracy, overperform precision, rate and investment return. Over-performing precision is a point of interest on its own as positive predictions classify a stock as a future over-performer, and therefore would initiate an investment in the market. This opens the portfolio up to potential losses since an actual position has been taken. However, when the model predicts an under-performing stock, it passes it over for investing and when the prediction is wrong it is only penalized by missing out on a return—an opportunity cost and not an actual dollar loss. Next, we look at each model's performance individually, and then on some comparisons between them and the benchmark. The benchmark portfolio consists of an equal investment in all available stocks in each of the testing periods. The S&P 500 was not used as the experiment sample did not include all underlying assets in the S&P 500 index.

Table 1 displays comparative models' performance year over year for percentage return, cumulative dollar returns and accuracy, and over- and under-performance precision of the model.

Character N-grams with CNG (C-grams) method outperformed the benchmark portfolio return overall and in five of the six years.

Word N-grams with CNG Classification (W-grams) model had superior accuracy and over-performance precision to that of the character n-gram model, and it also outperformed the benchmark return.

Readability Scores with SVM (Read) performed well, and in all but one year outperformed the benchmark and the n-gram model.

Combined Readability-scores with Character N-grams (Combo-char) makes a recommendation only when there is an agreement between the two combined methods. In addition to previously mentioned measures, for the combined models we also consider the percentage of cases with no decision due to the disagreement of the models.

Table 1. Detailed Experimental Results

Year	Return (% and $)		Accuracy	Over-perf.	Under-perf.	No Decision
Character N-gram Model						
2003	-6.59%	$9341.18	61.91%	70.59%	25.00%	
2004	47.80%	$13806.26	60.87%	65.00%	33.33%	
2005	20.32%	$16611.11	53.12%	52.63%	53.85%	
2006	31.48%	$21839.65	51.28%	52.38%	50.00%	
2007	34.67%	$29410.73	63.41%	75.00%	58.62%	
2008	-10.33%	$26371.62	41.02%	26.67%	50.00%	
Overall	163.72%	$26371.62	55.27%	57.04%	45.13%	
Word N-gram Model						
2003	-3.00%	$9700.00	71.43%	80.00%	50.00%	
2004	50.53%	$14601.35	56.52%	64.71%	33.33%	
2005	15.82%	$16911.02	50.00%	50.00%	50.00%	
2006	27.94%	$21636.71	53.85%	55.56%	47.62%	
2007	36.60%	$29555.75	70.73%	80.00%	65.38%	
2008	-9.29%	$26808.80	51.28%	41.18%	59.09%	
Overall	168.09%	$26808.80	58.97%	61.91%	50.90%	
Readability Model with SVM						
2003	-2.42%	$9758.33	66.67%	81.82%	44.44%	
2004	30.07%	$12692.34	56.52%	66.67%	37.50%	
2005	25.23%	$15894.71	59.38%	61.54%	57.89%	
2006	48.06%	$23534.11	69.23%	75.00%	65.22%	
2007	19.33%	$28084.04	60.98%	59.26%	64.29%	
2008	-3.13%	$27206.41	64.10%	62.50%	64.52%	
Overall	172.06%	$27206.41	62.81%	67.80%	55.64%	
Combined Readability and Character N-grams						
2003	-2.42%	$9,758.33	68.75%	83.33%	5.88%	5.60%
2004	27.69%	$12,460.64	64.29%	61.54%	25.49%	9.97%
2005	35.22%	$16,849.56	61.11%	66.67%	9.80%	7.72%
2006	73.50%	$29,233.98	78.57%	83.33%	8.82%	7.54%
2007	41.50%	$41,366.08	72.73%	90.00%	11.44%	9.09%
2008	39.00%	$57,498.85	55.56%	100.00%	1.04%	1.06%
Overall	474.99%	$57,498.85	66.83%	76.47%	62.48%	6.83%
Combined Readability and Word N-grams						
2003	-3.55%	9,645.4545	72.22%	83.33%	50.00%	14.29%
2004	26.30%	12,182.2091	63.64%	60.00%	100.00%	52.17%
2005	32.50%	16,141.4270	58.82%	70.00%	42.86%	46.88%
2006	40.50%	22,678.7050	76.47%	66.67%	81.82%	75.86%
2007	43.08%	32,449.4471	78.26%	91.67%	63.64%	43.90%
2008	4.00%	33,747.4250	68.75%	100.00%	66.67%	58.97%
Overall	237.47%	33,747.4250	69.69%	76.47%	65.68%	48.68%

Combined Readability-scores with Word N-grams (Combo-word) performed better than the benchmark, but significantly worse than the Combo-char model.

6.1 Model Results Comparison

To adequately compare the models we present in this subsection performances graphically on a combined plot. Figure 1 plots the year over year percentage accuracy of the five models. We can see that the word-combo model had better accuracy in all six years including 2008 when the market experienced a major trend shift. It is worth noting that the character-gram model slipped below the 50% margin in the last year during the trend change in 2007–2008. This was the only occurrence of any of the models performing below 50% accuracy.

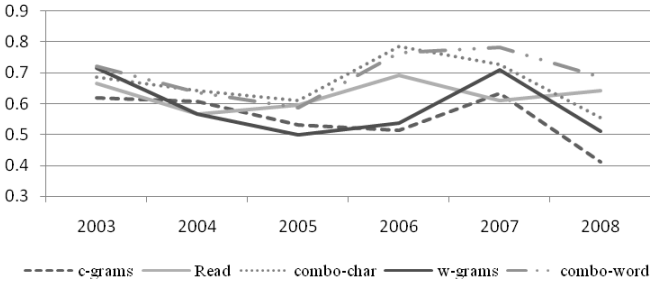

Fig. 1. Year over year accuracy

Figures 2 and 3 chart the percentage return and overall dollar return respectively for the five models and the benchmark portfolio.

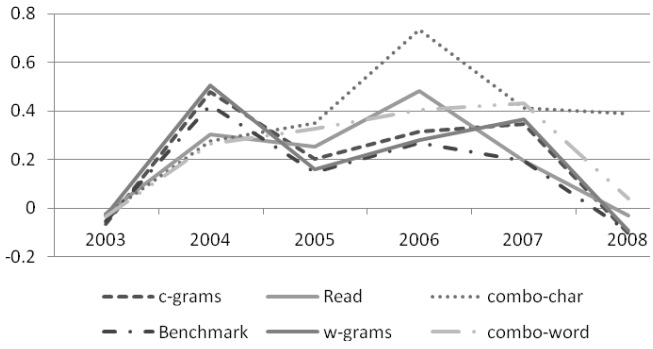

Fig. 2. Year over year % returns

Comparing the plots between the models and the benchmark portfolio it appears that their trends all match a general shape, only that in the majority of the years the benchmark is the poorest performer. In 2008 the only models to produce a positive return were the combined models and this was achieved when the benchmark lost nearly 10%.

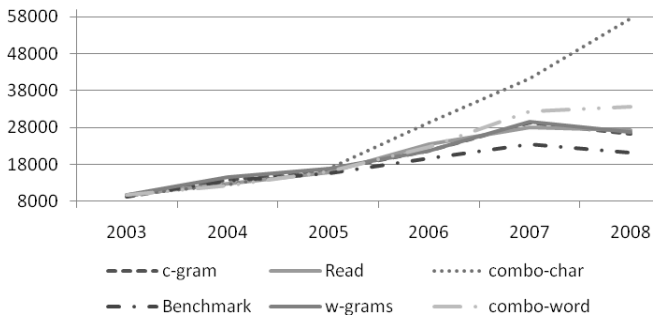

Fig. 3. Cumulative investment returns (in dollars, with initial investment $10,000)

By a large margin the character n-gram combination model had the superior investment strategy. For the first three years all 4 portfolios were quite close but in 2006 the character n-gram combination model pulled away and in 2008 picked up its most significant relative gain. This 2008 return is a direct result from the benefit of having a perfect overperformance precision rate.

7 Discussion and Conclusions

In general, the endeavour put forth here is an attempt to automate the analysis of annual reports. The expected benefit is that one could quickly evaluate the textual component and remove some of the uncertainty that arises from analysts having different opinions. More specifically, two novel NLP techniques are applied to solving the aforementioned problem. This section details the results, and gives some explanations as to what worked and what did not.

7.1 N-grams with CNG Classification

It has been shown that this methodology can be effective the problem of authorship attribution. In changing from the authorship attribution task to recognizing language indicative to one type of behaviour to another is a bit of a stretch. The belief is that certain language and phrases are used when the outlook is bleak and is measurably different than that when the outlook is positive. Overall, both the n-gram models were the weakest of the five models constructed, however they were still superior to the benchmark portfolio and that fact alone makes the experiment a success. The two n-gram based models had similar results, with the word-grams performing slightly better in overall accuracy and investment return. Although neither n-gram approach could capture all the information in the report, it was able to model a portion of it, such that, sufficient enough to give above average returns. The n-grams proved to be least effective when the market trend drastically shifted in 2007–2008. This may not necessarily be a short-coming of the n-grams themselves but the classification approach applied to them. It would be interesting to use a SVM for the n-gram profiles as a comparison to the CNG method. The overall accuracy of the models were about 55% and 59% for character-grams and word-grams respectively which is quite

typical of investment models and is good evidence that it is better that random guessing.

7.2 Readability-scores with SVM

As noted earlier, SVM's have proven very effective at producing robust investment models and dealing with the highly complex and non-linear data that is inherent in financial forecasting. Part of the success of this model could be attributed to the SVM choice of the classifier. Based on our preliminary tests, some other algorithms such as Artificial Neural Networks or Naïve Bayes could not achieve the same accuracy. Readability scores and their relation to stock performance have been well documented and the favourable results of this method are not unexpected as this model combined a proven linguistic analysis technique with a powerful classification algorithm. This model outperformed the n-grams technique and the benchmark portfolio on investment return (percentage and dollars) and in overperform precision, which made for more efficient trades. The overall accuracy and over-performance precision was 62.81% and 67.80% respectively, giving evidence that the model was more than just random guessing. This technique also demonstrated an ability to partly understand the text in the annual reports and learn what it indicated for future performance.

7.3 Combined Models

Choosing to only make decisions when the models agreed proved to be a valuable approach. This approach could be characterized as an ad hoc ensemble approach. It is evident that the three individual models were each able to explain part of the relationship between performance and the textual components of the annual reports and that what they learned was not completely overlapping. The combined models consistently outperformed the individual models and the benchmark portfolio. The combined models were also the most efficient as they made only about half the number of trades as the other three. This fact is evident from the "no decision" figures in table 1, where on average 40% (character n-grams combo) and 48% (word n-gram combo) of the time the two models did not agree and therefore no position was taken. Having the two models agree introduced a further confidence factor into the combined model which makes it more robust to noise in the market. In the majority of the years and overall the combined models proved superior in terms of investment return (dollar and percentage), overperformance precision, accuracy, and efficiency of investments. The most significant difference came in 2008 when the other three portfolios posted negative returns and the combined models made a positive gain of 39% (character n-gram combo) and 4% (word n-gram combo). It is also interesting that in this year the character combined model was not as accurate as the Readability model but it did, like the word n-gram combo, have a perfect 100% for overperform precision and therefore made no poor choices when an actual position in the market was taken. This abnormal investment return in 2008 is a bit of an anomaly and is not entirely realistic and will be discussed in the next section.

7.4 The 2008 Investment Anomaly

An overperformance precision of 1 and an investment return of 39% or 4% when the market losses almost 10% seems very good, however the problem is the models are suppose to build a portfolio of investments to spread the risk. Due to the volatile nature of the markets in 2007-2008 the two models were only able to agree once on an over-performer and therefore only made one investment each in the market. In reality an investment manager would most likely not have accepted this response and either moved some of the assets to the money market or conducted further analysis on the companies to find other suitable investments. The annual reports that the 2008 returns are calculated from are the 2006 annual reports produced sometime in 2007. Figure 4 illustrates the massive shift of market momentum in 2007. The arrow labeled '1' represents

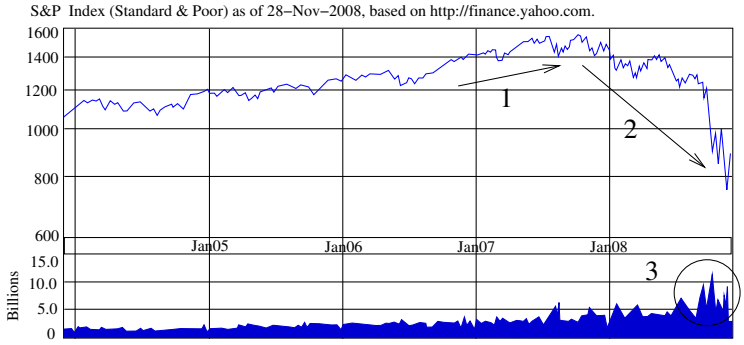

Fig. 4. S&P 500 Index Returns

the time period when the 2006 annual reports were being published and arrow '2' represents the time period when the actual performance was being evaluated. It is quite clear that the market environment drastically changed between those two time periods and the increase volatility is supported by the large increase in market volume highlighted by the circle labeled '3'.

8 Drawbacks, Limitations, and Future Work

Although the results are persuasive that the techniques presented are effective at analyzing annual reports, there still is a need for more thorough testing with an expanded data set that contains more of the companies in the S&P 500 index. The n-gram profiles were set size 6 and 3 for the character grams and word grams respectively taking up to the top 5000, these settings are most likely a local optimum and require fine tuning to optimize the model. With all the recent turmoil and volatility in the financial markets it will be worth applying the models to the newly released annual reports over the coming year to see how the models hold up under such extreme conditions. There is also a lot of information that is generated and can be learned from the experiment and deeper drilling down through the data could reveal more interesting information. For example, it would be interesting to know if there are some companies that produce more easily read annual reports making them more transparent, and

therefore a safer investment, or if the distance score that the CNG classifier reports is an indication of how sure the model is and could a threshold be introduced to improve overall accuracy and overperform precision. Finally the labeling process should be automated to cut down on pre-processing time and human error.

References

1. Keselj, V., Peng, F., Cercone, N., Thomas, C.: N-gram-based author profiles for authorship attribution. In: Proceedings of the Conference Pacific Association for Computational Linguistics, PACLING'03, Dalhousie University, Halifax, Nova Scotia, Canada (August 2003) 255–264
2. Falinouss, P.: Stock trend prediction using news articles. Master's thesis, Lulea University of Technology (2007) ISSN 1653-0187.
3. Schumaker, R., Chen, H.: Textual analysis of stock market prediction using financial news articles. In: Proc.from the America's Conf.on Inform.Systems. (2006)
4. Mittermayer, M.: Forecasting intraday stock price trends with text mining techniques. In: Proc.of the 37th Hawaii Int'nal Conf.on System Sciences. (2004)
5. Kloptchenko, A., Magnusson, C., Back, B., Vanharanta, H., Visa, A.: Mining textual contents of quarterly reports. Technical Report No. 515, TUCS (May 2002) ISBN 952-12-1138-5.
6. Kloptchenko, A., Eklund, T., Karlsson, J., Back, B., Vanharanta, H., Visa, A.: Combined data and text mining techniques for analysing financial reports. Intelligent Systems in Accounting, Finance and Management **12** (2004) 29–41
7. Subramanian, R., Insley, R., Blackwell, R.: Performance and readability: A comparison of annual reports of profitable and unprofitable corporations. The Journal of Business Communication (1993)
8. Li, F.: Annual report readability, current earnings, and earnings persistence. Journal of Accounting and Economics (2008)
9. Yahoo! Inc.: Yahoo! finance. http://ca.finance.yahoo.com/ Last access 2008.
10. Kešelj, V.: Text::Ngrams Perl module for flexible ngram analysis. http://www.cs.dal.ca/~vlado/srcperl/Ngrams/Ngrams.html (2003–9) Ver,2.002. Avail.at CPAN.
11. Ryan, K.: Lingua::EN::Fathom Perl module for measuring readability of english text. Available at CPAN (2007)
12. CPAN Community: CPAN—Comprehensive Perl Archive Network. http://cpan.org (1995–2009)
13. Fast, G.: Lingua::EN::Syllable Perl module for estimating syllable count in words. Available at CPAN, http://search.cpan.org/perldoc?Lingua::EN::Syllable (1999)
14. Fan, R., Chen, P., Lin, C.: Working set selection using second order information for training SVM. Journal of Machine Learning Research **6** (2005) 1889–1918
15. Fan, A., Palaniswami, M.: Stock selection using support vector machines. In: Proceedings of IJCNN'01. Volume 3. (2001) 1793–1798
16. Huang, W., Nakamori, Y., Want, S.Y.: Forecasting stock market movement direction with support vector machine. Computers and Operations Research **32** (2005) 2513–2522
17. Kim, K.: Financial time series forecasting using support vector machines. Neurocomputing **55** (2003) 307–319
18. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines (2001) Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.