# Chapter 9
# Sentiment analysis using automatically labelled financial news items

Michel Généreux, Thierry Poibeau and Moshe Koppel

**Abstract** Given a corpus of financial news items labelled according to the market reaction following their publication, we investigate 'cotemporeneous' and forward-looking price stock movements. Our approach is to provide a pool of relevant textual features to a machine learning algorithm to detect substantial stock price variations. Our two working hypotheses are that the market reaction to a news item is a good indicator for labelling financial news items, and that a machine learning algorithm can be trained on those news items to build models detecting price movement effectively.

## 9.1 Introduction

The aim of this research is to build on work by [7] and [15] to investigate the subjective use of language in financial news items about companies traded publicly and validate an automated labelling method. More precisely, we are interested in the short-term impact of financial news items on the stock price of companies. This is a challenging task because although investors, to a certain extent, make their decision on the basis of factual information such as income statement, cash-flow statements

Michel Généreux
Laboratoire d'informatique de Paris-Nord – Université Paris 13, 99 avenue Jean-Baptiste Clément – 93430 Villetaneuse – France. Now at Complexo Interdisciplinar da Universidade de Lisboa Av. Prof. Gama Pinto, 2, 1649-003 Lisboa – Portugal e-mail: genereux@clul.ul.pt

Thierry Poibeau
Laboratoire d'informatique de Paris-Nord – Université Paris 13, 99 avenue Jean-Baptiste Clément – 93430 Villetaneuse – France. Now at LaTTiCe-CNES âĂŞ Ecole Normale Supérieure et Université Paris 3 âĂŞ 1 rue Maurice Arnoux, 92210 Montrouge âĂŞ France, e-mail: thierry.poibeau@ens.fr

Moshe Koppel
Department of Computer Science – Bar-Ilan University, 52900 Ramat-Gan Israel, e-mail: koppel@cs.biu.ac.il

or balance sheet analysis. However, there is an important part of their decision which is based on a subjective evaluation of events surrounding the activities of a company. Traditional Natural Language Processing (NLP) has so far been concerned with the objective use of language. However, the subjective aspect of human language, i.e. sentiment that cannot be directly inferred from a document's propositional content, has recently emerged as a new useful and insightful area of research in NLP [3, 8, 16]. According to [17], affective states include opinions, beliefs, thoughts, feelings, goal, sentiments, speculations, praise, criticism and judgements, to which we may add attitude (emotion, warning, stance, uncertainty, condition, cognition, intention and evaluation); they are at the core of subjectivity in human language. We treat short financial news items about companies as if they were carrying implicit sentiment about future market direction made explicit by the vocabulary employed and investigate how this *sentimental* vocabulary can be automatically extracted from texts and used for classification. There are several reasons why we would want to do this, the most important being the potential of financial gain based on the exploitation of covert sentiment in the news items for short-term investment. On a less pragmatic level, going beyond literal meaning in NLP would be of great theoretical interest for language practitioners in general, but most importantly perhaps, it would be of even greater interest for anyone who wishes to get a sense of what are people feelings towards a particular news item, topic or concept. To achieve this we must overcome problems of ambiguity and context-dependency. Sentiment classification is often ambiguous (compare *I had an accident* (negative) with *I met him by accident* (not negative)) and context dependent (*There was a decline*, negative for *finance* but positive for *crimes*).

## 9.2 Data and Method

The automated labelling process is described in section 9.2.4. We have opted for a linear *Support Vector Machine (SVM)* [4] approach as our classification algorithm and we have used the Weka[1] software package.

### *9.2.1 Training and Testing Corpus*

Based on previous work in sentiment analysis for domains such as movie reviews and blog posts, we have selected an appropriate set of three key parameters in text classification: feature *type*, *threshold* and *count*. Our goal is to see whether the most suitable combinations usually employed for other domains can be successfully transferred to the financial domain. Our corpus is a subset of the one used in

[1] `http://www.cs.waikato.ac.nz/ml/weka/`

[7]: 6277 news items averaging 71 words covering 464 stocks listed in the Standard & Poor 500 for the years 2000-2002.

### *9.2.2 Feature Types*

We consider five types of features: *unigrams*, *stems*, *financial terms*, *health-metaphors* and *agent-metaphors*. The news items are tokenized with the help of a POS tagger [14]. *Unigrams* consist of all nouns, verbs, adjectives and adverbs[2] that appear at least three times in the corpus. *Stems* are the unigrams which have been stripped of their morphological variants. The *financial terms* stem from a 'clinical study of investors discussion and sentiment' [2]. The list comprises 420 words and their variants created by graduate students who read through messages[3] and selected words they felt were relevant for finance (not necessarily most frequent)[4]. *Health metaphors* are a list of words identified by [6] in a six million word corpus from the *Financial Times* suggesting that the financial domain is pervaded by terms from the medical domain to describe market phenomena: examples include *addiction*, *chronic* and *recovery*. The full list comprises 123 such terms.

Finally, recent work by [9] suggests that in the case of market trends, investors tend to process *agent metaphors*, when the language treats the market as though it were an entity that produces an effect deliberately (e.g. *the NASDAQ climbed higher*), differently from object metaphors, where the language describe price movements as object trajectories, as events in which inanimate objects are buffeted by external physical forces (e.g. *the Dow fell through a resistance level*) or non-metaphorical expressions that describe price change as increase/decrease or as closing up/down (e.g. *the Dow today ended down almost 165 points*). The same study gives the verbs *jump*, *climb*, *recover* and *rally* as the most frequent indicators of uptrend movement, and *fall*, *tumbled*, *slip* and *struggle* as the most frequent indicators of downtrend movements. The point made in the study is that in the case of agent metaphors, investors tend to believe that the market will continue moving in the same direction, which is not the case for object metaphors or non metaphors. These results are potentially useful for sentiment analysis, as we are trying to find positively correlated textual features with market trends. To construct a list of potential agents, we extracted all nouns from our corpus and used WordNet[5] to filter out elements which were not hyponym of the synset comprising causality and agency words[6], defined as *an entity that produces an effect or is responsible for events or results*: in this way we collected 553 potential agents. To allow those agents to carry out their actions, we completed this list with all 1538 verbs from the corpus.

---

[2] This list is augmented by the words *up*, *down*, *above* and *below* to follow [7].

[3] The corpus was a random selection of texts from Yahoo, Motley Fool and other financial sites.

[4] Sanjiv Das, personal communication.

[5] `http://wordnet.princeton.edu/`

[6] Wordnet synset number 100005598: *causal agency#n#1*, *cause#n#4* and *causal agent#n#1*

### 9.2.3 Feature Selection and Counting Methods

We consider three feature selection methods that [18] reported as providing excellent performance. *Document Frequency* (DF) is the number of documents in which a term occurs. We computed DF for each feature and eliminated features for which DF fell below a threshold (100). In *Information Gain* (IG), features are ranked according to a preferred sequence allowing the classifier to rapidly narrow down the set of classes to one single class. We computed the 100 features with the highest information gain. Finally, the $\chi^2$ statistic measures the lack of independence between a feature and a set of classes. We computed the top 100 least independent features. It is worth mentioning that the same 100 features were selected using either IG or $\chi^2$ statistic, except for a few features ranking order in the top ten.

There are different methods for *counting* values of the features mentioned above. There are two methods worth considering for computing the value of a given feature when a token with the feature is found: the first is the binary method where a value of zero indicates the absence of the feature whereas a value of one indicates the presence of the feature. This method appears to yield good results for movie reviews [27]. The second simply gives a count of the feature in the document and normalise the count for a fixed-length document of 1000 words (TF).

### 9.2.4 News Items and Stock Price Correlation

To construct our 500 positive examples we used similar criteria as [7], based on contemporaneous price changes: the price of a stock was noted at the opening of a market after a news item was published; and, the price of the stock was noted at the closing of the market on the day before a news item was published. Essentially, for a news item to be labelled as a positive example, its positive price change must be greater than a given threshold (we used 4%) and be in excess of the overall S&P index change.

For instance, the following news item about the company *Biogen, Inc.* (symbol BGEN), appeared on May 23rd 2002:

> Biogen, Inc. announced that the FDA's Dermatologic & Ophthalmic Drug Advisory Committee voted to recommend approval of AMEVIVE (alefacept) for the treatment of moderate-to-severe chronic plaque psoriasis.

At the opening on the 24-May-2002, the price reached $48.43, whereas at the closing on the 22-May-2002 the price was $38.71. Therefore, there is a positive price change of

$$\frac{\$48.43 - \$38.71}{\$38.71} = 0.19995$$

or almost 20%, so the news item is classified as being positive. The same reasoning is applied to find 500 negative examples, corresponding to a negative price change of at least 4%.

The corpus of positive and negative news items, contemporaneous with price changes, was used in the training. We have five feature types, three feature selection methods, and two counting methods - 30 different ways to represent news items. To narrow down the training possibilities, we first focused on the features themselves. The *unigram* feature, in combination with the *information gain* feature selection criterion and *binary count* of the feature values, appear to give the highest accuracy (67.5%) when compared to the use of other features including *stems* and the *agent-metaphors*, *financial terms* and *health-metaphors* respectively (see Table9.1). The method of evaluation was 10-fold cross-validation on the news items dataset..

| **Feature** | **Accuracy (%)** |
|---|---|
| Unigram | 67.5 |
| Stems | 66.9 |
| Agent Metaphor | 66.4 |
| Financial Metaphor | 59.2 |
| Health Metaphor | 52.4 |

**Table 9.1** Feature tuning: Performance of various features with Information Gain and Binary Count

The unigram feature, with its inherent simplicity, appears to perform well in combination with certain feature selection and feature counting parameters, We found that the highest classification accuracy, at 67.6%, was obtained by using *unigrams*, *information gain (IG)* and *term frequency (TF)* (Table9.2). The second highest accuracy was obtained by using *unigrams*, *information gain (IG)* and *binary count (Bin)* and reached 67.5%. Given this small, perhaps insignificant difference in accuracy together with the favourable reviews of the binary method in the literature, we believe that this basic trio (*unigrams, IG and binary*) will suffice.

| **Feature** | **Feature Selection** | **Feature Count** | **Accuracy (%)** |
|---|---|---|---|
| Unigram | Information Gain | Term Frequency | 67.6 |
| Unigram | Information Gain | Binary Count | 67.5 |
| Unigram | $\chi^2$ | Binary Count | 66.1 |
| Unigram | Degrees of Freedom | Binary Count | 59.4 |

**Table 9.2** Feature tuning: Performance of unigram features with different feature selection and counting methods

Our results suggest that the features based on a list of agent metaphors describing market trend movements appear more useful for the classification of financial news items than a list of health metaphors or a human-constructed list of financial terms. At closer examination, it appears that most of the contribution is made by the notion of *agent*: only five of the eight most frequent indicators (*recover, climb, fall, slip* and *struggle*) actually appear in our corpus, and only one (*fall*) made the cut

through the top 100 features that bring most information gain. We conjecture that the description of financial news items retains the same agent-based feature as in market trend description, however it is expressed by commentators using a different set of (predicative) terms. In the remaining experiments we depart slightly from [7] by taking into account negation, i.e. negated words (e.g. not rich) are featured as a single term (not_rich). We also excluded proper nouns as a potential feature: In financial news items, proper nouns usually refer to companny names and employees at a specific instance of time and these nouns change over time as new companies and people enter the domain and many leave. Proper nouns do not appear to us as an appropriate feature.

### *9.2.5 Feature Selection and Semantic Relatedness of Documents*

A study by [10] has suggested that information from different sources can be used advantageously to support more traditional features. Typically, these features characterise the semantic orientation (SO) of a document as a whole [11, 5]. One such feature is the result of summing up the semantic relatedness (Rel) between all individual words (adjectives, verbs, nouns and adverbs) with a set of polarised positive (P) and negative (N) terms, for the domain of interest, here finance. The semantic relatedness of a document can be defined as:

$$\sum_{w}^{Words} (\sum_{p}^{P} Rel(w,p) - \sum_{n}^{N} Rel(w,n))$$

Note that the quantity of positive terms P must be equal to the quantity of negative terms N. To compute relatedness, we used the method described in [1] and WordNet[7]. The list of polarised terms we used follows:

| Adjectives | | Nouns | | Verbs | | Adverbs | |
|---|---|---|---|---|---|---|---|
| Positive | Negative | Positive | Negative | Positive | Negative | Positive | Negative |
| good, rich | bad, poor | goodness, richness | badness, poverty | increase, enrich | decrease, impoverish | well, more | badly, less |

**Table 9.3** Polarised terms used in our study

Our experiments suggest that the relatedness measure produces a negative semantic orientation even for documents that are labelled as positive. Even more worryingly, the negative score of positively labelled documents is greater than the negative score of a negatively labelled documents. This suggests that the polarity orientation of financial news items cannot be captured by techniques based on semantic

[7] Using the PERL package [13].

relatedness and our result shows no significant improvement on accuracy (69%) if we include semantic relatedness as one of our features.

## 9.3 Results

### *9.3.1 Horizon Effect*

The next experiment looks at the lasting effect of a news item on the stock price of a company. Using 300 positive examples and 300 negative examples with a $\pm 2\%$ price variation, we computed classification accuracies for non-cotemporeneous, more precisely subsequent, price changes. Therefore, news items were classified according to price changes from the opening the first open market day after the news to X number of days after the news. We consider the following values for X: 2, 3, 7, 14 and 28. (See Table 9.4). Given that classification accuracies are slowly worsening

| **Horizon** | **Accuracy** |
|---|---|
| [+1,+2] | 69.5% |
| [+1,+3] | 68.8% |
| [+1,+7] | 67.5% |
| [+1,+14] | 68.0% |
| [+1,+28] | 66.3% |

**Table 9.4** Horizon Effect

as we move further away from the day the news item first broke out (coefficient of correlation is -0.89), we conclude that some prices are getting back to, or even at the opposite of, their initial level (i.e. before the news broke out). Assuming that in the interval no other news items interfered with the stock price, this result also reinforced the validity of the automatic labelling technique.

### *9.3.2 Polarity effect*

This experiment looks at the effect on accuracy a change in the labelling distance between two classes produces. The intuition is that the more distant two classes are from each other, the easiest it is for the classifier to distinguish among them, which translates as a higher accuracy.

To test the accuracy of our classifier, we created five sub-corpora each comprising 400 labelled news items: We have three sub-corpora which exclusively contain positive (price change greater than +2%), negative (price change smaller than -2%) or neutral text respectively, the other two corpora are a mixture of negative and neutral news items and positive and neutral news items respectively (Table9.5).

| **Sub-corpus** | **Neg***ative* | **Neu***tral* | **Pos***itive* | **Total** |
|---|---|---|---|---|
| I Negative only | 400 | | | 400 |
| II Neg+Neu | 200 | 200 | | 400 |
| III Neutral only | | 400 | | 400 |
| IV Neu+Pos | | 200 | 200 | 400 |
| V Positive only | | | 400 | 400 |

**Table 9.5** Subcorpora for testing polarity effects

One can imagine a line of unit length along which we can classify the subcorpora in Table9.5: The sub-corpus I, with purely negative texts, lies at the origin and sub-corpus V, with purely positive news items, lies at unit distance from the subcorpus I. The *distance* between corpora I and II (and II and III, III and IV, IV and V) is assumed to be 0.25; the distance between I and III (neutral news items) is 0.5 and I and IV (positive plus neutral news items) is 0.75 and so on. When our classifier is presented with a mixture of subcorpora with greatest distance, then its accuracy should be the best and contrariwise for a mixture of subcorpora that lie close to each other will lead to reduced accuracy of classification. The *distance* between two subcorpora changes the classification accuracy by 8% on average: when the mixture is of proximate subcorpra (*distance* of 0.25) the accuracy is over 62% and rises to 70% when the distance approaches unity (Table9.6). This is further evidence of the utility and accuracy of the automatic labelling technique.

| **Class 1** | **Class 2** | **Accuracy %** | **Nominal distance** |
|---|---|---|---|
| Neg+Neu | Pos | 0.703 | 0.75 |
| Neg | Pos | 0.698 | 1 |
| Neu | Pos | 0.693 | 0.5 |
| Neg | Neu+Pos | 0.693 | 0.75 |
| Neg | Neu | 0.68 | 0.5 |
| Neg+Neu | Neu | 0.646 | 0.25 |
| Neu+Pos | Pos | 0.641 | 0.25 |
| Neg | Neg+Neu | 0.628 | 0.25 |
| Neg+Neu | Neg+Pos | 0.618 | 0.5 |
| Neu | Neu+Pos | 0.576 | 0.25 |

**Table 9.6** Accuracy for mixture classification

### *9.3.3 Range Effect*

The range effect experiment explores how the size of the minimum price change for a news item to be labelled either as positive or negative influences classification accuracy. The intuition is that the more positive and negative news items are labelled according to a larger price change, the more accurate classification should be. Ta-

ble 9.7 shows results using cotemporeneous price changes. The labelling method

| Range | Nb examples | 2-class | 3-class |
|---|---|---|---|
| ±0.02 | 1000 | 67.8% | 46.3% |
| ±0.03 | 1000 | 67.1% | 47.9% |
| ±0.05 | 800 | 69.5% | 46.8% |
| ±0.06 | 600 | 74.0% | 50.1% |
| ±0.07 | 400 | 76.3% | 50.1% |
| ±0.10 | 200 | 75.0% | 51.3% |

**Table 9.7** Range Effect

once again yields expected results: for two classes (positive and negative), the more comfortable the price change margin gets, the more accurate classification is (coefficient of correlation is +0.86). However, accuracies appear to reach a plateau at around 67%, and classification accuracy improvements beyond 75% seems out of reach. The last column of Table 9.7 reports accuracies for the case where news items whose price change is falling between the range are labelled as *neutral*. Although accuracies are, as expected, lower than for two classes, they are significantly above chance (33%). The same positive correlation is also observed between the price change margin and accuracies (coefficient of correlation is +0.88). In the next experiment we examine more in depth the effect of adding a neutral class on precision.

### *9.3.4 Effect of adding a neutral class on non-cotemporaneous prices: One- and two-days ahead*

In all but one of the experiments so far, we have considered classes with maximum polarity, i.e. with a neutral class separating them. On one hand this has simplified the task of the classifier since news items to be categorised belonged to one of the positive or negative extremes. On the other hand, this state of affairs is somewhat remote from situations occurring in real life, when the impact of news items can be limited. Moreover, the information about overall accuracy of classification is not the most sought after information by investors. Let's examine briefly more useful information for investors:

**Positive Precision** A news item which is correctly recognised as positive is a very important source of information for the investor. The potential winning strategy now available is to buy or hold the stock for the corresponding range. Therefore, it is very important to build a classifier with high precision for the positive class, significantly above 50% to cover comfortably transaction costs.

**Negative Precision** A news item which is correctly recognised as negative is also an important source of information for the investor. The potential saving strategy now available to the investor, given that he or she owns the stock, is to sell the stock before it depreciates. Therefore, it is important to build a classifier with

high precision for the negative class, significantly above 50% to cover safely transaction costs.

**Positive and Negative Recall** Ideally, all positive and negative news items should be recognised, but given the potential substantial losses that misrecognition (implying low positive/negative precision) would imply for investors, high recall should not be a priority.

Table 9.8 gives an indication of the kind of effect positive (+precision) and negative (-precision) precision we can expect if we built a 3-class classifier. Results show that

| Range | Nb examples | -Precision | +Precision |
|---|---|---|---|
| $\pm$0.01 | 1000 | 77% | 51% |
| $\pm$0.02 | 800 | 41% | 53% |
| $\pm$0.03 | 400 | 69% | 54% |

**Table 9.8** Effect of adding a neutral class on non-cotemporaneous prices

precision is either worryingly close to 50% (the positive case), or is very volatile and could swing precision level well below 50% on too many occasions. This perhaps demonstrates that if we were to build a financial news items classifier satisfying at least high precision for the positive news items, it appears important to avoid the three-class classification approach.

### *9.3.5 Conflating two classes*

In section 9.3.4 we underlined the importance of high precision for the classification of positive and negative news items and concluded that a 3-class categoriser was unlikely to satisfy this requirement. In this section we conflate two of the three classes into one and examine the effect on precision and recall. Table 9.9 displays three classification measures for the case where the classes neutral and negative have been conflated to a single class. Table 9.10 displays three classification mea-

| Measure/Class | POS | NEG+NEU |
|---|---|---|
| Precision | 0.857 | 0.671 |
| Recall | 0.555 | 0.908 |
| Accuracy | 0.7313 | |

**Table 9.9** Positive versus combined neutral and negative news items

sures for the case where the classes neutral and positive have been conflated to a single class. We used a range of $\pm$0.02, a forward-looking horizon of [+1,+2] days with 800 training examples. It is difficult to evaluate precisely what the cost of trading represents, but there seems to be enough margin of maneuver to overcome this impediment, especially in the case of the positive classifier (Table 9.9).

| Measure/Class | NEG | POS+NEU |
|---|---|---|
| Precision | 0.652 | 0.805 |
| Recall | 0.870 | 0.535 |
| Accuracy | 0.7025 | |

**Table 9.10** Negative versus combined neutral and positive news items

### *9.3.6 Positive and Negative features*

Closer examination of the features resulting from the selection process paints a different picture from the one presented Koppel and Shtrimberg [7]. These authors have used all words that appeared at least sixty times in the corpus, eliminating function words with the exception of some relevant words. We kept only adjectives, common nouns, verbs, adverbs and four relevant words, *above*, *below*, *up* and *down*, that appear at least three times in the training corpus. In a nutshell, [7] found that there were no markers for positive stories, which were characterised by the absence of negative markers. As a result, recall for positive stories were high but precision much lower. Our findings are that negative and positive features are approximately equally distributed (53 negatives and 47 positives) among the top 100 features with the highest information gain and that recall and precision for positive stories were respectively lower and higher. We define a *polarity* orientation (positive or negative) of each feature as the class in which the feature appears the most often. Table 9.11 shows the top ten positive features and table 9.12 the top ten negative features. The *Rank* column indicates the position of the feature in the top 100 ranking resulting from the information gain screening. The *+df/-df* column displays the number of documents (examples) in which the feature appears at least once (*+df* for positive and *-df* for negative). The *+tf/-tf* column displays the number of times the feature appears in the entire set of documents (*+tf* for positive and *-tf* for negative), while the *+n/-n* column displays the same values normalised to a constant document length of 1000 words. For example, the feature *common* appears in 29 positive examples and 8 negative examples. It also appears 33 times in all positive examples and 13 times in

| Rank | Feature | +df/-df | +tf/-tf | +n/-n |
|---|---|---|---|---|
| 1 | common | 29/8 | 33/13 | 1318/390 |
| 2 | shares | 33/11 | 48/17 | 2014/640 |
| 3 | cited | 20/4 | 20/4 | 427/49 |
| 5 | reason | 18/4 | 18/4 | 411/69 |
| 8 | direct | 7/0 | 7/0 | 163/0 |
| 9 | repurchase | 15/3 | 26/3 | 818/115 |
| 10 | authorised | 17/4 | 18/5 | 596/177 |
| 11 | drug | 6/0 | 6/0 | 114/0 |
| 13 | partially | 6/0 | 6/0 | 89/0 |
| 14 | uncertainty | 6/0 | 6/0 | 102/0 |

**Table 9.11** Positive Features

| Rank | Feature | +b/-b | +tf/-tf | +n/-n |
|---|---|---|---|---|
| 4 | change | 0/8 | 0/11 | 0/206 |
| 6 | work | 1/11 | 1/12 | 33/183 |
| 7 | needs | 0/7 | 0/7 | 0/139 |
| 12 | material | 0/6 | 0/6 | 0/128 |
| 15 | pending | 0/6 | 0/7 | 0/100 |
| 16 | gas | 8/23 | 13/34 | 229/571 |
| 19 | cut | 1/9 | 1/10 | 15/216 |
| 20 | ongoing | 1/9 | 2/14 | 14/201 |
| 25 | e-mail | 0/5 | 0/5 | 0/68 |
| 26 | week | 0/5 | 0/5 | 0/72 |

**Table 9.12** Negative Features

all negative examples. Below is one highly positive news item (+11% price change) and one highly negative news item (-49% price change) with positive features inside square brackets and negative features inside braces. The following news item about the company Equifax Inc. (symbol EFX) appeared on the 20th of September 2001. Its stock price jumped from $18.60 at opening on the 21st of September 2001 to $20.70 on the 24th of September 2001, for a price change of 11.29%:

> Equifax Inc. announced that it is repurchasing [shares] in the open market, pursuant to a previous [repurchase] authorisation. The [Company]'s board of directors had [authorised] a repurchase of up to $250 million of [common] stock in the open market in January 1999, of which approximately $94 million remains available for purchase.

The following news item about the company Applied Materials, Inc. (symbol AMAT) appeared on the 15th of April 2002; its stock price plummeted from $53.59 at opening on the 16th of April 2002 to $27.47 on the 17th of April 2002, for a price change of -48.74%:

> Applied Materials, Inc. announced two newly granted U.S. Patents No. 6,326,307 and No. 6,362,109, the [Company]'s third and fourth patents covering the use of hexafluorobutadiene (C4F6) {gas} chemistry for critical dielectric etch applications. A high-performance etch process chemistry, C4F6 used in an Applied Materials etch system, enables the industry's move to the 100nm chip generation and beyond.

## 9.4 Discussion

The surprisingly encouraging results we have presented for a forward-looking investment strategy should not be viewed outside its specific experimental setup conditions. In what follows we highlight a number of points worth considering:

### Lack of independent testing corpus

Cross-validation is a method which can provide a solid evaluation of the overall accuracy of a classifying method. However, a more accurate evaluation should involve an independent testing corpus, ideally covering a distant time-period to avoid overfitting or over-training. Nevertheless, we have attempted to avoid these caveats by keeping a small number of features compared to the number of training examples and by avoiding the use of proper nouns as features.

### Pool of features

Our pool of features was selected among the entire training set, which includes the cross-validated sections. Although to a small degree, this may have caused a *data-snooping* bias, where features were selected among the testing examples. On the other hand, as can be observed in tables 9.11 and 9.12, the interpretation of positive and negative features is not straightforward, which suggests that portability among different domains and even time periods could be problematic.

### Size of documents

Clearly, the size of documents is crucial for classification. The corpus we used averaged just over 71 words per document, which in general should be long enough to collect enough statistics. Nevertheless, if we look at our top ten positive stories (those with the highest positive price change), we found that half of them contained no feature at all, whereas three out of our top ten negative examples were similarly deprived of features. Given that this situation is likely to worsen if we train and test on different domains and periods, this is a potential area where a default bias can be difficult to avoid (i.e. a document without features will systematically be classified in the same class). One solution would be to increase the number of features.

### Trading costs

If the minimum transaction level to overcome fixed and relative trading costs is high, this brings upon the investors a burden of risk which he or she may not be able or willing to bear. The classifier should be characterised clearly by its level of precision matched with an estimate of the trading costs that would guide the investor in its decision.

## 9.5 Conclusion and future work

We have revisited a method for classifying financial news items using automatically labelled data. Our findings give a different picture of the set of features best suited for the task and a somewhat less pessimistic prognosis as to the validity of such an approach for forward-looking investment. We have suggested some topics where further research should be carried out for testing the automatic labelling approach within a practical and realistic framework. To this end, our next step will be to use our system coupled with a virtual trading site[8] to monitor financial news items with a view to use the analysis for investing in companies. This should give us a better idea of the effect of the transaction costs as well as the portability of the features and model developed during our experiments.

## References

1. S. Banerjee and T. Pedersen. Extended gloss overlaps as a measure of semantic relatedness. In *Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence, Acapulco, 2003, pp. 805–810.*, 2003.
2. Sanjiv Das, Asis Martinez-Jerez, and Peter Tufano. e-information: A clinical study of investor discussion and sentiment. *Financial Management*, 34(5):103–137, 2005.
3. Ann Devitt and Khurshid Ahmad. Sentiment polarity identification in financial news: A cohesion-based approach. In *Proceedings of ACL-07, the 45th Annual Meeting of the Association of Computational Linguistics*, pages 984–991, Prague, CZ, June 2007. ACL.
4. Thorsten Joachims. *Learning to Classify Text Using Support Vector Machines*. Kluwer Academic Publishers, 2001.
5. J. Kamps, M. Marx, R. Mokken, and M. de Rijke. Using Wordnet to measure semantic orientation of adjectives. In *In LREC 2004, volume IV, pages 1115–1118.*, 2004.
6. Francis Knowles. Lexicographical aspects of health metaphors in financial texts. In *Proceedings Part II of Euralex 1996*, pages 789–796, Department of Swedish, Göteborg University, 1996.
7. Moshe Koppel and Itai Shtrimberg. Good news or bad news? let the market decide. In *AAI Spring Symposium on Exploring Attitude and Affect in Text*, pages 86–88. Stanford University, March 2004.
8. Gilad Mishne. *Applied text analytics for blogs*. PhD thesis, University of Amsterdam, 2007.
9. Michael W. Morris, Oliver J. Sheldon, Daniel R. Ames, and Maia J. Young. Metaphors and the market: Consequences and preconditions of agent and object metaphors in stock market commentary. *Journal of Organizational Behavior and Human Decision Processes*, 102(2):174–192, March 2007.
10. Tony Mullen and Nigel Collier. Sentiment analysis using support vector machines with diverse information sources. In *Empirical Methods in NLP*, 2004.
11. Charles E. Osgood, George J. Suci, and Percy H. Tannenbaum. *The Measurement of Meaning*. University of Illinois, 1957.
12. Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up? sentiment classification using machine learning techniques. In *Proc. of the 2002 Conf. on Empirical Methods in Natural Language Processing*, 2002.

[8] `http://vse.marketwatch.com/`

13. Ted Pedersen. Wordnet::similarity - measuring the relatedness of concepts. In *In Appears in the Proceedings of the Nineteenth National Conference on Artificial Intelligence (AAAI-04), 2004.*, 2004.
14. Helmut Schmid. Probabilistic part-of-speech tagging using decision trees. In *Int. Conference on New Methods in Language Processing*, Manchester, UK, 1994.
15. D. Lawrie P. Ogilvie D. Jensen V. Lavrenko, M. Schmill and J. Allan. Mining of concurrent text and time series. In *6th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, August 2000, August 2000.
16. J.G. Shanahan, Y. Qu, J. Weibe. (Eds.) Computing Attitude and Affect in Text: Theory and Applications. Dordrecht: Springer. 2006.
17. T. Wilson and J. Wiebe. Annotating opinions in the world press. In *In SIGdial-03.*, 2003.
18. Yiming Yang and Jan O. Pedersen. A comparative study on feature selection in text categorization. In Douglas H. Fisher, editor, *Proc. of ICML-97, 14th Int. Conf. on Machine Learning*, pages 412–420, Nashville, US, 1997. Morgan Kaufmann Publishers, San Francisco, US.