



HAL
open science

Sentiment analysis using automatically labelled financial news

Michel Génèreux, Thierry Poibeau, Moshe Koppel

► **To cite this version:**

Michel Génèreux, Thierry Poibeau, Moshe Koppel. Sentiment analysis using automatically labelled financial news. LREC 2008 Workshop on Sentiment Analysis: Emotion, Metaphor, Ontology and Terminology, Jun 2008, Marakech, Morocco. hal-00346996

HAL Id: hal-00346996

<https://hal.science/hal-00346996>

Submitted on 13 Dec 2008

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Sentiment analysis using automatically labelled financial news

Michel Génèreux¹, Thierry Poibeau², Moshe Koppel

Laboratoire d'informatique de Paris-Nord – Université Paris 13^{1,2}, Department of Computer Science – Bar-Ilan University
99 avenue Jean-Baptiste Clément – 93430 Villetaneuse – France^{1,2}, 52900 Ramat-Gan Israel
{genereux,poibeau}@lipn.univ-paris13.fr, koppel@cs.biu.ac.il

Abstract

Given a corpus of financial news 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 is a good indicator for labelling financial news, and that a machine learning algorithm can be trained on those news to build models detecting price movement effectively.

1. Introduction

The aim of this research is to build on work by (Koppel and Shtrimberg, 2004) and (V. Lavrenko and Allan, 2000) to investigate the subjective use of language in financial news about companies traded publicly and validate an automated labelling method. More precisely, we are interested in the short-term impact of financial news 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 or balance sheet analysis, 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 the new useful and insightful area of research in NLP (Devitt and Ahmad, 2007; Mishne, 2007). According to (Wilson and Wiebe, 2003), 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 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 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, 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*, neg-

ative with *I met him by accident*, not negative) and context dependent (*There was a decline*, negative for *finance* but positive for *crimes*).

2. Experiments

Based on previous work in sentiment analysis for domains such as movie reviews and blog posts, this first series of experiments aim at selecting 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 (Koppel and Shtrimberg, 2004): 6277 news averaging 71 words covering 464 stocks listed in the Standard & Poor 500 for the years 2000-2002. The automated labelling process is described in section 2.4. We have opted for a linear *Support Vector Machine (SVM)* (Joachims, 2001) approach as our classification algorithm with the software Weka¹. All experiments have been cross-validated ten times.

2.1. Feature Types

We consider five types of features: *unigrams*, *stems*, *financial terms*, *health-metaphors* and *agent-metaphors*. The news are tokenized with the help of a POS tagger (Schmid, 1994). *Unigrams* consist of all nouns, verbs, adjectives and adverbs² that appears 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 (Das et al., 2005). The list comprises 420 words and their variants created by graduate students who read through messages³ and selected words they felt were relevant for finance (not necessarily most frequent)⁴. *Health metaphors* are a list of words identified by (Knowles, 1996) in a six million word

¹<http://www.cs.waikato.ac.nz/ml/weka/>

²This list is augmented by the words *up*, *down*, *above* and *below* to follow (Koppel and Shtrimberg, 2004).

³The corpus was a random selection of texts from on Yahoo, Motley fool and other financial sites.

⁴Sanjiv Das, personal communication.

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 (Morris et al., 2007) shows 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 WordNet⁵ to filter out elements which were not hyponym of synset number 100005598⁶, 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.

2.2. Feature Selection

We consider three feature selection methods that (Yang and Pedersen, 1997) 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 χ^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 χ^2 statistic, except for a few features ranking order in the top ten.

2.3. Counting Methods

There are two methods worth considering for valuing each feature appearance in each news: 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 in movie reviews (Pang et al., 2002). 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).

2.4. Classifying news using cotemporeous prices: [-1 day,+1 day]

To construct our 500 positive examples we used similar criteria as (Koppel and Shtrimberg, 2004), based on contemporaneous price changes (stock price at opening the first market trading day after the news was published - stock price at closing the first trading day before the news was published):

- price change superior to the overall S&P index price change,
- price change in the interval [-4%,+4%] and
- price superior to \$10.

For instance, the following news 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 opening on the 24-May-2002, price reached \$48.43, whereas at closing on the 22-May-2002 it was \$38.71. Therefore, there is a positive price change of

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

or almost 20%, the news 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 results of this experience is presented in table 1. The

| | Features | F-Selection | F-Count |
|--------|----------|----------------|-----------|
| Unig. | 67.5% | IG 67.5% | Bin 67.5% |
| Stems | 66.9% | DF 59.4% | TF 67.6% |
| Fin.-T | 59.2% | χ^2 66.1% | |
| Hea.-M | 52.4% | | |
| Age.-M | 66.4% | | |

Table 1: Feature tuning

reference trio of parameters appears in table 1 between horizontal lines (unigrams, IG and binary). That is, in each successive measurement of accuracy, at least two values of the trio remained unchanged. For example, the classification accuracy when using stems, information gain and binary count is 66.9%. Strictly speaking, the best combination (unigrams, IG and TF) reached 67.6%, a tenth of one percent better than the basic trio (unigrams, IG and Bin). Given this non significant difference in accuracy and a favourable inclination for the binary method in the literature, we keep the basic trio as our parameter values for all other experiments. These results also show that features based on a list of agent metaphors describing market trend

⁵<http://wordnet.princeton.edu/>

⁶*causal agency##1, cause##4 and causal agent##1*

movements appear more useful for the classification of financial news 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 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 (Koppel and Shtrimerberg, 2004) by taking into account negation, i.e. negated words (e.g. not rich) are featured as a single term (not_rich). We also remove all proper nouns as potential feature, our assumption being that a list of features without proper nouns is less tailored to a particular time-period, where some companies happen to be more in the spotlight than others.

Including Osgood’s feature A study by (Mullen and Collier, 2004) have 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 (Osgood et al., 1957; Kamps et al., 2004). 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. This method can be expressed in the following formal manner:

$$\sum_w^{Words} \left(\sum_p^P Rel(w, p) - \sum_n^N Rel(w, n) \right)$$

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 (Banerjee and Pedersen, 2003) and WordNet⁷. The list of polarised terms we used follows:

Pos adjectives: good, rich
 Neg adjectives: bad, poor
 Pos nouns: goodness, richness
 Neg nouns: badness, poverty
 Pos verbs: increase, enrich
 Neg verbs: decrease, impoverish
 Pos adverbs: well, more
 Neg adverbs: badly, less

| Class | SO |
|-------|------|
| 0.00 | -106 |
| 0.25 | -89 |
| 0.50 | -104 |
| 0.75 | -114 |
| 1.00 | -128 |

Table 2: Semantic Orientation

Although the relatedness measure is biased towards negative, as illustrated by all negative semantic orientations, even for positive classes, the trend observed and expected is that positive classes are less negative than positive classes in general (coefficient of correlation is +0.76). This is a result conforing the validity of the automatic labelling technique. However, our result shows no significant improvement on accuracy (69%) if we include semantic orientation as one of our features.

2.5. Horizon Effect

The next experiment looks at the lasting effect of a news 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 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. Table 3 presents the results. Given that classification accuracies are slowly worsening as

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

Table 3: Horizon Effect

we move further away from the day the news 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 interfered with the stock price, this result also reinforced the validity of the automatic labelling technique.

2.6. 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.

| Class 1 | Class 2 | Dist. | Acc. | Aver. |
|---------|---------|-------|-------|-------|
| 0.00 | 0.25 | 0.25 | 62.8% | 62.3% |
| 0.25 | 0.50 | 0.25 | 64.6% | |
| 0.50 | 0.75 | 0.25 | 57.6% | |
| 0.75 | 1.00 | 0.25 | 64.1% | |
| 0.00 | 0.50 | 0.50 | 68.0% | 66.4% |
| 0.25 | 0.75 | 0.50 | 61.8% | |
| 0.50 | 1.00 | 0.50 | 69.3% | |
| 0.00 | 0.75 | 0.75 | 70.3% | |
| 0.25 | 1.00 | 0.75 | 73.1% | 71.7% |
| 0.00 | 1.00 | 1.00 | 69.8% | 69.8% |

Table 4: Polarity effect

Table 4 presents classification accuracy using five classes:

⁷Using the PERL package (Pedersen, 2004).

- 0.00: 400 negative news, price change $< -2\%$
- 0.25: 200 negative plus 200 neutral news
- 0.50: 400 neutral news
- 0.75: 200 neutral plus 200 positive news
- 1.00: 400 positive news, price change $> +2\%$

The five classes above generate four possible combinations of labelling distance: 0.25, 0.50, 0.75 and 1.00. As expected, there is a positive correlation between labelling distance and accuracy (coefficient of correlation +0.89). This reinforces the validity of the automatic labelling technique.

2.7. Range Effect

The range effect experiment explores how the size of the minimum price change for a news to be labelled either as positive or negative influences classification accuracy. The intuition is that the more positive and negative news are labelled according to a larger price change, the more accurate classification should be. Table 5 shows results using contemporaneous price changes. The labelling method yields once

| 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 5: Range Effect

again 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 6%, where classification accuracy improvements beyond 75% seems out of reach. The last column of table 5 reports accuracies for the case where news 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.

2.8. Effect of adding a neutral class on non-cotemporaneous prices: [+1 day,+2 days]

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 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 can be limited. Moreover, the information about overall accuracy of classification is not the most sought after information for investors. Let's examine briefly more useful information for investors:

Positive Precision A news 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 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 should be recognised, but given the potential substantial losses that misrecognition (implying low positive/negative precision) would imply for investors, only a decent level of recall is needed for both.

Table 6 gives a first glimpse of the sort of positive (+precision) and negative (-precision) precision we can expect if we built a 3-class classifier. In order to get closer to real classification conditions, we remove the constraint that stock prices must be greater than \$10. Results show that

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

Table 6: 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. Clearly, this demonstrate that if we are to build a financial news classifier satisfying at least high precision for the positive news, we must abandon the approach using three classes.

2.9. Conflating two classes

In section 2.8. we underlined the importance of high precision for the classification of positive and negative news 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 7 displays three classification measures for the case where the classes neutral and negative have been conflated to a single class. Table 8 displays three classification measures for the case where the classes neutral and positive have been conflated to a single class. We used a range of ± 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 manoeuvre to overcome this impediment, especially in the case of the positive classifier (table 7).

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

Table 7: Positive against all others

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

Table 8: Negative against all others

2.10. Positive and Negative features

Closer examination of the features resulting from the selection process paints a different picture from the one presented in (Koppel and Shtrimberg, 2004). Recall that (Koppel and Shtrimberg, 2004) 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, (Koppel and Shtrimberg, 2004) 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 sentimental orientation (positive or negative) of each feature as the class in which the feature appears the most often. Table 9 shows the top ten positive features and table 10 shows the top ten negative features. The *Pos* column indicates the position of the

| Pos | Feature | +b/-b | +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: Positive Features

feature in the top 100 ranking resulting from the information gain screening. The +b/-b column displays the number of documents (examples) in which the feature appears at least once (+b for positive and -b 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.

| Pos | 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 10: Negative Features

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 all negative examples. Below is one highly positive news (+11% price change) and one highly negative news (-49% price change) with positive features inside square brackets and negative features inside braces. The following news 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 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.

3. 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 overtraining. 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 and 10, 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, 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.

4. Conclusion and future work

We have revisited a method for classifying financial news 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 prognostic as to the validity of such an approach for forward-looking investment. We indicate a number of elements where extensive research should be carried on to test the approach within a practical and realistic framework. To this end, our next step is to use our system coupled with a virtual trading site⁸ to monitor financial news to invest 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.

5. References

S. Banerjee and T. Pedersen. 2003. Extended gloss overlaps as a measure of semantic relatedness. In *Proceed-*

ings of the Eighteenth International Joint Conference on Artificial Intelligence, Acapulco, 2003, pp. 805–810.

Sanjiv Das, Asis Martinez-Jerez, and Peter Tufano. 2005. e-information: A clinical study of investor discussion and sentiment. *Financial Management*, 34(5):103–137.

Ann Devitt and Khurshid Ahmad. 2007. 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. ACL.

Thorsten Joachims. 2001. *Learning to Classify Text Using Support Vector Machines*. Kluwer Academic Publishers.

J. Kamps, M. Marx, R. Mokken, and M. de Rijke. 2004. Using wordnet to measure semantic orientation of adjectives. In *In LREC 2004, volume IV, pages 1115–1118*.

Francis Knowles. 1996. Lexicographical aspects of health metaphors in financial texts. In *Proceedings Part II of Euralex 1996*, pages 789–796, Department of Swedish, Göteborg University.

Moshe Koppel and Itai Shtrimerberg. 2004. 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.

Gilad Mishne. 2007. *Applied text analytics for blogs*. Ph.D. thesis, University of Amsterdam.

Michael W. Morris, Oliver J. Sheldon, Daniel R. Ames, and Maia J. Young. 2007. 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.

Tony Mullen and Nigel Collier. 2004. Sentiment analysis using support vector machines with diverse information sources. In *Empirical Methods in NLP*.

Charles E. Osgood, George J. Suci, and Percy H. Tannenbaum. 1957. *The Measurement of Meaning*. University of Illinois.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? sentiment classification using machine learning techniques. In *Proc. of the 2002 Conf. on Empirical Methods in Natural Language Processing*.

Ted Pedersen. 2004. Wordnet::similarity - measuring the relatedness of concepts. In *In Appears in the Proceedings of the Nineteenth National Conference on Artificial Intelligence (AAAI-04)*, 2004.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Int. Conference on New Methods in Language Processing*, Manchester, UK.

D. Lawrie P. Ogilvie D. Jensen V. Lavrenko, M. Schmill and J. Allan. 2000. Mining of concurrent text and time series. In *6th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, August 2000, August.

T. Wilson and J. Wiebe. 2003. Annotating opinions in the world press. In *In SIGdial-03*.

Yiming Yang and Jan O. Pedersen. 1997. 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. Morgan Kaufmann Publishers, San Francisco, US.

⁸<http://vse.marketwatch.com/>