Topic-Dependent Sentiment Analysis of Financial Blogs

Neil O'Hare[1], Michael Davy[2], Adam Bermingham[1], Paul Ferguson[1],
Páraic Sheridan[2], Cathal Gurrin[1], Alan F. Smeaton[1]
[1]CLARITY: Centre for Sensor Web Technologies, Dublin City University, Ireland
[2]National Centre for Language Technology, Dublin City University, Ireland
nohare@computing.dcu.ie

ABSTRACT

While most work in sentiment analysis in the financial domain has focused on the use of content from traditional finance news, in this work we concentrate on more subjective sources of information, blogs. We aim to automatically determine the sentiment of financial bloggers towards companies and their stocks. To do this we develop a corpus of financial blogs, annotated with polarity of sentiment with respect to a number of companies. We conduct an analysis of the annotated corpus, from which we show there is a significant level of topic shift within this collection, and also illustrate the difficulty that human annotators have when annotating certain sentiment categories. To deal with the problem of topic shift within blog articles, we propose text extraction techniques to create topic-specific sub-documents, which we use to train a sentiment classifier. We show that such approaches provide a substantial improvement over full documentclassification and that word-based approaches perform better than sentence-based or paragraph-based approaches.

Categories and Subject Descriptors

H3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing

General Terms

Algorithms

Keywords

Sentiment Analysis, Opinion Mining, Financial Blogs

1. INTRODUCTION

The blogosphere is acknowledged as a source of subjective opinions on a wide variety of topics, as has been recognised in the TREC Blog Track [22]. This track has run since 2006, focussing on the retrieval of subjective text from Blog articles. In the domain of finance, many bloggers publish opinions about specific companies and on markets in general[1]. However, as far as we are aware, no existing work on sentiment analysis in the financial domain has used blogs as sources, instead using traditional news and finance media (e.g. [1],[14]). Blogs have the advantage that their authors are more likely to express opinions and to make predictions about the performance of stocks than traditional media – which are more likely to report news relating to a stock's past performance but may contain few explicit statements of opinion regarding the future.

Our work has developed from a collaboration between Dublin City University (DCU) and an industrial partner working in online stock trading[2]. The aim is to automatically extract the subjective opinions uniquely found on blogs and track the changing sentiment from the blogosphere towards individual stocks and the market in general. This involves crawling financial weblogs, retrieving articles relevant to certain companies and their stocks, running sentiment polarity classification (positive, neutral, negative) on those articles. The extracted sentiment will then be aggregated to obtain a snapshot of the general sentiment of the blogosphere towards that company. We believe that such information will prove useful to users of online stock trading services.

There is a tendency in financial blogs to discuss multiple companies (or their stocks) in a single article, meaning that document level sentiment classification will not always be suited. In this work we explore simple approaches to coping with such topic shift by extracting topic specific subdocuments (i.e. subsets of the documents considered relevant to the topic) and training sentiment polarity classifiers based on these sub-documents. Since there are no existing corpora for sentiment in the financial domain for blogs, we also constructed a new corpus for developing and evaluating sentiment analysis approaches.

The main contributions of this paper are twofold. Firstly, we develop a new corpus of financial blogs, annotated with polarity of sentiment, and analyse this corpus with respect to annotator's ability to create consistent sentiment polarity annotations. Furthermore, we explore the extent to which topic shift within financial blog articles occurs. Having determined that it is a genuine problem, we propose approaches to topic-based text extraction for sentiment polarity classification and evaluate these approaches on this corpus.

[1]blogged.com, for example, lists over 2,000 blogs in the category 'investing'
[2]Zignals: http://www.zignals.com

TSA'09, November 6, 2009, Hong Kong, China.

The remainder of the paper is organised as follows. In the next section we introduce related work in sentiment annotation, topic-based sentiment analysis and sentiment analysis in the financial domain. Then, in Section 3 we discuss the creation of our corpus and present some analysis of it, including examining the ability of humans to annotate certain sentiment labels consistently and the extent to which topic shift is an issue in this corpus. In Section 4 we propose our approach to topic-based sub-document extraction and give details of the sentiment classifier system. Section 5 describes our experiments in sentiment polarity classification and results. Finally, in Section 6 we report conclusions and directions for future work.

2. RELATED WORK

We first present existing work in corpus annotation for sentiment, and then place our work in context of existing topic-based sentiment analysis approaches and existing sentiment analysis work in the financial domain.

2.1 Sentiment Annotation

Supervised learning relies on labelled training data to induce and evaluate classification strategies. In some domains documents labelled for sentiment by the document author are available, notable examples being Pang and Lee's work on movie reviews [25] and Dave et al.'s work on product reviews [7]. In domains where author labels are not available, we must rely on human annotators to provide sentiment judgements. Notable manual sentiment annotation efforts include the Blogs06 corpus [20], Wilson's MPQA [31] corpus and the NTCIR corpus for their Multilingual Opinion Analysis Task (MOAT) [27]. Of these, the Blogs06 corpus was annotated at the document level, and the MOAT and MPQA corpora were annotated at the sentence and phrase level respectively. Each corpus required a significant amount of annotator training, in particular the MPQA corpus which featured very detailed annotation.

Annotation of sentiment can be a relatively difficult challenge, as interpretation of sentiment is subject to a number of human factors such as domain expertise, the annotator's private state and inferences the annotator has made into the text to be annotated. The MOAT corpus has, for example, moderately high rates of agreement for Japanese and Chinese but low agreement for English texts, while the MPQA corpus achieved a high level of agreement.

Our own previous work evaluated inter-annotator agreement on a representative subset of the Blogs06 corpus [3]. We evaluated sentiment annotation at the sentence and document level, observing a similar moderate level of agreement for both sentence and document-level annotations. We also found that annotating for the label *mixed* sentiment was troublesome for annotators and agreement for this class was significantly lower than that for the other sentiment classes, a finding which is further explored in the current work.

2.2 Topic-based Sentiment Analysis

Much of the work to date in sentiment analysis has focused on domains where topic relevance is assumed. Examples of this are to be found in the product review [29] [7] [19] and film review [25] domains. This simplifying assumption allows systems to focus specifically on the identification of sentiment, without regarding topic relevance.

With more ad-hoc information sources, such as blogs, topic relevance may not be assumed and relevance determination must be incorporated into the sentiment analysis process. One approach is to first estimate the likelihood of topic relevance using techniques from the field of information retrieval. The relevance probability may then inform the sentiment analysis algorithm, which in turn produces a final topic-sentiment score. This allows us to rank the documents for likelihood of containing topic-directed sentiment. This two-stage approach is the most common approach used in the opinion finding tasks at TREC [22], which evaluated two separate sentiment related tasks: opinion finding (find any opinionated documents) and sentiment polarity ranking (find only positive, or only negative, documents). In some applications, such as the current work, a sentiment ranking is not appropriate and a summary of sentiment in known or assumed relevant documents is required. In this scenario, documents are first labeled for binary topical relevance, and the relevant documents are analysed for sentiment.

Another problem with more freeform domains is topic shift, where several topics are discussed in a single document. A number of proximity-based models have been tried, with some success. The underlying assumption in proximity-based models is that portions of text which co-occur with topic-related terms are likely to be indicative of the sentiment towards that topic. There are many examples of this in the literature which show some degree of success, for example [13] [21]. Zhang et al. [33] propose a method for jointly modelling proximity and sentiment using a generative model and a degradation is observed as the proximity window is made smaller. Rather than incorporate the proximity model into the sentiment analysis directly, an alternative approach is to extract relevant information from the source document before conducting sentiment analysis, with passage retrieval approaches showing some success [16], [2]. Our work differs from such work in that we are interested in sentiment polarity rather than subjectivity detection, and we are interested in hard classification of a set of relevant documents, rather than an information retrieval style ranking.

We also take inspiration from the passage retrieval and proximity models and from the area of topic-based text extraction used in both the citation analysis and Web retrieval fields [15] [12] [5], [4]. In these domains the goal is to identify additional text from the referencing document or web page, and associate this text with the document being referenced. The most common approach to identify words that are to be associated with the document that is being referenced is to take a 'window' of text either side of the reference. We have taken a similar approach in that we identify the areas within the articles that are associated with specific stocks, and then use a variety of windowing techniques in order to identify the text that is associated with that stock. This process of extracting text relating to a specific topic is important as topic shift is a potential problem in our corpus that we are using (see Section 3.2.3) and so we extract topic-based sub-documents for sentiment analysis (Section 4).

2.3 Sentiment Analysis in the Financial Domain

In the financial domain, Ahmed et al. have studied methods for identifying positive and negative news in news streams [1] and for identifying affect in news text [8]. They identify a controversial news event likely to elicit emotive content

and use the subsequent news articles as a corpus. Koppel et al. used market price movements as a ground truth for their financial sentiment analysis [14]. They too used news data as their corpus and achieved an accuracy of 70%. However, their work is unique in the literature in that it is not evaluated against human judgements. The goal of both of the above approaches is to mine sentiment from broadcast news data. We believe that news data is generally objective and not an ideal source for mining and aggregating sentiment. Instead we use blogs which have been shown to be highly subjective [22]. This is particularly true in the financial analysis domain as authors frequently make evaluations and projections targetted at markets, stocks, companies and prominent figures. To the best of our knowledge, our work is the first to mine sentiment in blogs specifically targetted at the financial domain.

3. FINANCIAL BLOG CORPUS

In this section we outline the creation of our blog articles, followed by an analysis of the corpus.

3.1 Development of Corpus

3.1.1 Crawl and Noise Removal

The corpus we use is made up of financial blog articles collected automatically from a predefined set of sources. We identified 232 financial blog sources, and crawled these sources on two separate occasions: for 3 weeks in February 2009 (Crawl 1), and for 5 weeks from May to June 2009 (Crawl 2). Since there was a significant change in the overall mood between these snapshots (relating to the 2009 global financial crisis), splitting the dataset in this way should capture overall shifts in sentiment in relation to the markets.

After crawling these blog sources and extracting the HTML source for all articles it is necessary to remove irrelevant information contained in those pages, such as links to other pages, advertisements, etc. We use the DiffPost algorithm, proposed by Lee et al [16], to remove noise from the documents in the collection. This approach exploits the fact that, within a given blog feed, the noise, or unwanted content, will tend to be repeated across multiple articles, while the relevant text from the article will be unique to that individual article. Accordingly, each article is first broken up into HTML segments, and each of these segments is compared to segments extracted from articles from the same source. Only unique segments are kept, with non-unique segments being considered as noise and so are removed.

3.1.2 Annotation Granularity

In general the input to the polarity classifier will be documents and sentiment analysis will be first applied at the *document level*. However, it may be the case that finer granularity is required since documents can contain a mixture of sentiments for a variety of topics (e.g. one blog post about a number of different stocks). In addition to document level annotation, we annotated at the *paragraph level*. In the literature sentence and phrase level [31] granularity have been explored. While this mitigates against the problem of mixed topics found at the document level and also paragraph level granularity, a number of new challenges arise. First is the extra demands incurred when annotation is performed at the sentence level. Second, it can be more difficult to accurately label sentences (or even phrases) since the contextual information is not available. The manual annotation effort for paragraph granularity is less than that of sentence or phrase level granularity while contextual information is maintained.

3.1.3 Labels

The labels used for annotation include a five-point scale from *Very Negative* to *Very Positive*: *Very Negative, Negative, Neutral, Positive, Very Positive.* Annotators could also annotate paragraphs or documents as *mixed*, which indicates a mixture of positive and negative sentiment, and *not relevant*. For paragraph-level annotations only, the *noise* label indicates that the paragraph should not be considered to be part of the article body but it an unwanted part of the HTML page containing the article. Finally, we also gave the annotators the option of annotating as *I don't know (IDK)*, which means that the annotator is not confident in making an annotation. We included this class in acknowledgement of the fact that sentiment annotation is an inherently difficult task, and even human annotators sometimes have difficulty annotating documents with confidence.

3.1.4 Annotation Tool

To facilitate the annotation of our corpus, we developed a web-based annotation tool to present annotators with a queue of documents to be annotated with the labels described in Section 3.1.3, and allowed annotators to annotate at the document level and the paragraph level.

3.1.5 Annotators and Training

As sentiment annotation is a difficult task, and since domain knowledge of financial markets is necessary for annotating this corpus, it was important that our annotators were trained before undertaking this annotation task. The corpus was annotated by 7 people, 5 of these being computer science researchers from DCU, and 2 employees of our industrial partner. The training phase involved two rounds of pilot annotations consisting of 5 training documents each, followed by extensive discussions of these annotations, until a consensus annotation was reached. Following this, a set of guidelines for annotations was produced for the annotators.

3.1.6 Topics and Retrieval

We identified the 500 companies that make up the Standard & Poor's **S&P 500** Index as topics of interest for sentiment analysis. In order to retrieve candidate documents for annotation with respect to a certain stock, we ran a case-sensitive phrase search on the name of the stock i.e. relevant articles must contain the whole phrase of the company name, and the case must also match (typically the name is capitalised). Since each document can be annotated with respect to more than one company (or stock), unique annotations are identified by the combinations of document and topic, which we will refer to as a doc-topic or doc-topic pair. In addition to annotating documents with respect to stocks, we are also interested in the sentiment of documents with respect to the market in general. For this reason we annotate a number of documents with respect to their sentiment towards stocks or equities in general: these documents were randomly selected. In total, we annotated 1526 unique doc-topic pairs, 167 of which were annotated for stocks in general, and 164 of which were annotated by two annotators to facilitate inter-annotator agreement analysis.

3.2 Analysis of Corpus

In this section we give details of the corpus, which contains financial blog articles annotated by 7 users.

3.2.1 Annotation Statistics

Table 1 summarises the document-level annotations; since a number of documents were annotated more than once (i.e. with respect to different topics) the number of unique documents annotated is much less than the total number of annotations. There is a clear bias towards negative sentiment in *Crawl 1*, with approximately twice as many negative labels as positive labels, while *Crawl 2* shows the opposite bias. Overall, though, there is a roughly even balance between positive, negative and neutral annotations. Comparatively few documents are annotated as Very Positive or Very Negative, and 90 annotations (just over 5%) were *I don't know*.

	Crawl 1	**Crawl 2**	**Total**
Total Annotations	541	1150	1691
Unique Documents	311	668	979
Unique Doc/Topic Pairs	476	1050	1526
Very Positive	21	47	68
Positive	54	251	305
Neutral	80	187	267
Negative	124	177	301
Very Negative	43	56	99
Mixed	27	75	102
Not Relevant	154	305	459
I don't know	38	52	90

Table 1: Statistics for document-level annotations.

3.2.2 Inter-Annotator Agreement

Table 2 shows the Kappa score for inter-annotator agreement for various levels of granularity. The 7-point scale, made up of all document level annotations (except *I don't know*, which we interpret as abstaining from annotating) has a Kappa of 0.462, indicating only a moderate level of agreement. This increases to 0.593 for a *5-point* scale, suggesting a low level of agreement for annotation of degree or strength of polarity, as merging positive and very positive, and negative with negative, greatly improves the agreement.

Since our sentiment analysis classifier will not be interested in learning the not relevant class (indeed, it would not be feasible to create separate relevance classifiers for all 500 topics) it is also worth looking at the agreement with the not relevant class excluded. The *4-point* granularity, which removes the not relevant class, has a Kappa of 0.592, equivalent to the *5-Point* Kappa, suggesting that relevance was annotated consistently. Combining the mixed and neutral classes to create the *3-PointMN* granularity gives a kappa of 0.596. Removing the mixed class completely, however, leads to a Kappa of 0.712, a huge improvement which suggests that the mixed category is a difficult one for annotators to agree on (of the doubly annotated doc-topics, of 10 *mixed* annotations, only 1 of these was annotated consistently by both annotators). If we only look at the positive and negative classes, there is perfect agreement in the annotations.

Based on these agreement scores, we believe that it is most appropriate to train a polarity classifier at either the *3-PointN* or the *binary* granularity, and it is at these granularities that we will evaluate our topic-dependent sentiment analysis in Section 5.

Granularity	**Kappa**
7-Point (VP / P / Nu / N / VN / M / NR)	0.466
5-Point (VP&P / Nu / N&VN / M / NR)	0.59
4-Point (VP&P / Nu / N&VN / M)	0.59
3-PointMN (VP&P / Nu&M / N&VN)	0.6
3-Point (V&P / Nu / N&VN)	0.712
Binary (V&P / N&VN)	1

Table 2: Kappa score for document-level inter-annotator agreement at various levels of granularity.

3.2.3 Topic Relevance

To determine the level of topic shift within the documents in our collection, we analysed the relevance statistics of the documents in *Crawl 2* of our collection. For our purposes here a document is considered relevant if it is retrieved by an initial retrieval process, as described in Section 3.1.6. We can see from Table 3 that, although the number of doc-topic pairs is roughly equal to the number of documents in the crawl, only about 30% (2,249) of these are relevant to at least one stock meaning that, when a document is relevant to a stock, it is, on average, relevant to 3 stocks. The table also shows that over half of the relevant documents are relevant to 2 stocks or more, and approximately a quarter of them are relevant to 4 stocks or more. This indicates that, as expected, this dataset does contain a lot of topic shift within relevant documents, and that documents that mention one stock will very often mention other stocks also, supporting our argument that sub-document extraction for sentiment classification are necessary.

Total Documents	6,561
Doc-Topic Pairs	6,614
Docs Relevant to at least:	
1 Stock	2,249
2 Stocks	1292
3 Stocks	820
4 Stocks	560
5 Stocks	403
6 Stocks	284
7 Stocks	173
8 Stocks	137
9 Stocks	110
10 Stocks	86

Table 3: Article Relevence Statistics for Crawl 2.

Of course, looking at the annotation statistics in Table 3, we can see that 459 out of 1,691 annotations (approx. 27%) are non-relevant. However, we do not believe that this biases the observations made above, but rather inflates the number of relevant documents reported. Anecdotal reports from the annotators suggest that the majority of these non-relevant

documents are retrieved due to failure of the noise removal component, and we believe that improving the noise removal would alleviate this problem.

4. TOPIC-BASED SENTIMENT ANALYSIS

In this Section we introduce our approach to topic-based sentiment classification, first introducing our topic-based text extraction approaches, and then outlining the sentiment analysis classifier used.

4.1 Topic-Based Text Extraction

Since blog articles often contain discussion of multiple topics, it is useful to extract those segments from the documents that are most relevant to the topic of interest. This topic-based text extraction will enable sentiment analysis to be carried out at a sub-document level, ensuring that we restrict our analysis to the portions of a document relevant to a specified topic, and this should alleviate the topic shift problems discussed in Section 3.2.3. As discussed in Section 2.2, there has been similar prior work carried out which attempts to calculate an optimal window of text around a topic word in order to retain the most relevant words associated with that topic. We investigate the use of three different sub-document extraction approaches, while also using the output for the task of polarity detection, which is quite distinct from work carried out by [33] for the task of opinion finding – our task is essentially a classification task, as opposed to a retrieval oriented task. In addition, we thoroughly investigate a spreading window-size approach that uses a number of different extraction methods to find the most effective input for our sentiment classifier.

Each of our three segmentation algorithms take a topic (a text string containing one or more terms) and extracts sub-segments of the document that occur adjacent to any of the topic terms. We implemented the following approaches:

- *N-word extraction.* Based on natural sequence of words in article, we extract a given number, n, of words either side of any topic word. Figure 1 shows an example of n-words (N=1 and N=3) extracted from either side of a topic word.
- *N-sentence extraction.* Similar to n-word segmentation, n-sentence segmentation extracts n sentences side of a sentence containing a topic term.
- *N-paragraph extraction.* Extracts n paragraphs adjacent to paragraphs containing any of the topic terms.

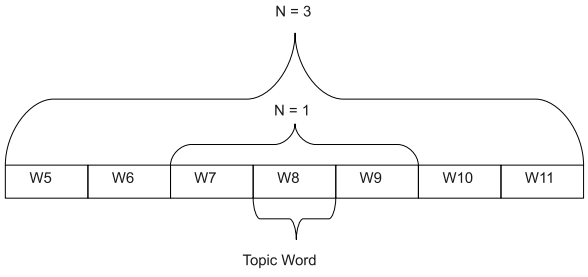

Figure 1: N-word text extraction

4.2 Sentiment Classification

Two distinct approaches to automatic sentiment polarity classification have been proposed in the literature. The first uses domain independent lexical resources to classify text [29, 6, 8], while the other builds domain dependent models using machine learning techniques [23, 17, 11]. In this work we focus on the latter, and use two alternative classifiers. We use a multinomial naïve Bayes (MNB) classifier, since it has been shown to give strong performance [28] without requiring parameter tuning. The second classifier is a Support vector machine (SVM), the current state-of-the-art in topic classification, which has also been shown to perform well in the task of sentiment polarity classification [23, 11].

The classification task attempts to model a function $f : X \mapsto Y$ which maps from doc-topic pairs (X) to a set of predefined categories (Y). We explore two classification tasks:

- *Binary classification*, which predicts whether an article is either positive or negative to a given topic ($Y \in \{positive, negative\}$).
- *3-Point classification*, which is a finer level of classification granularity. In this case we include neutral documents ($Y \in \{positive, negative, neutral\}$).

Of the two classification tasks performed, *3-Point* classification is considered more challenging than *binary*.

As a pre-processing step, the dataset was firstly tokenised on whitespace, digits and punctuation characters. Following this, we removed stopwords (using the list from the RCV1 [18] corpus), stemmed all tokens (using the Porter stemming algorithm [26]), and transforming all tokens to lowercase. From this we used the bag-of-words representation to construct feature vectors for each document and sub-document. A binary weighting scheme was employed, since it has been found to outperform traditional weighting schemes (such as tf-idf) for sentiment classification [10, 24].

	Trivial	SVM	MNB
Binary	50.876	66.0601	69.5447
3-Point	38.143	49.719	54.454

Table 4: Baseline Accuracy

5. EXPERIMENTS

We evaluate our proposed sentiment polarity classification approaches using the corpus described in Section 3. Examples not having the labels Y (see Section 4.2 above) are discarded, while those examples that were labelled inconsistently by more than one annotator are also discarded. This gives a total of 687 labelled documents for binary classification and 917 labelled documents for 3-Point classification.

We consider the different representations of a doc-topic pair given by each of the text extraction techniques outlined in Section 4.1, and compare the accuracy obtained by constructing a classifier trained on each of the approaches. We compare three classifiers: a multinomial naïve Bayes, a Support Vector Machine [30] and a baseline trivial classifier. For the SVM classifier, we used a linear kernel with default parameters (C = 1). The trivial classifier predicts the mode of the classes in the training data, and is included as a

	Paragraph		Sentence		Words		
N	**SVM**	**MNB**	**SVM**	**MNB**	**N**	**SVM**	**MNB**
0	67.9462	**73.3429**	69.2377	71.8958	5	68.9565	71.2904
1	64.5829†	71.8679301	**70.7022**	**72.5925**	10	72.6119	72.1599
2	66.3369	70.99617	68.0999	72.1534	15	72.1901	73.6156
3	66.9230	70.855509	70.5656	72.5839	20	73.3301	74.6280
4	67.0724	69.83894466	67.9247	71.7143	25	**74.3683**	74.0460
5	67.7949	69.09919†	66.4883	70.5636	30	71.8807	**75.0691**
6	**68.2383**	69.38905†	66.7825	71.4331	40	72.1728	74.4787
7	64.8618†	69.8217	64.8791†	70.4143	50	71.5779	74.0482
8	63.4127†	69.96663	67.7925	69.8367	60	68.3722	74.3511
9	65.1604†	69.966631	66.1920	70.2758	70	68.3301	73.3190
10	64.8749†	70.260748	65.4586†	69.6789	80	69.0925	73.1722
Baseline	66.0601	69.5447	66.0601	69.5447		66.0601	69.5447

Table 5: Binary classification results for paragraph, sentence and word text extraction. Maximum accuracy is represented by bold text, while accuracy below that of the baseline are indicated with †

baseline to show that the more advanced classification techniques offer significant advantage in terms of effectiveness. The WEKA [32] machine learning library implementation of each classifier was used in all experiments.

Ten-fold cross validation was used for each of the segmentation experiments, with the results averaged over the ten folds. We use classification accuracy as the performance metric, with a baseline measurement being calculated using the entire document for each doc-topic pair (with the topic terms removed). Table 4 displays the baseline results, showing that for both *binary* and *3-Point classification*, the document-level SVM and naïve Bayes classifiers achieve a large improvement in performance over the trivial classifier.

5.1 Results

Binary classification results using sentence-, paragraph- and word- based sub-document extraction are shown in Table 5. Each row corresponds to the results at the given level of the extraction (e.g. N=10 indicates that 10 paragraphs / sentences / words either side of a topic term are included). All approaches are shown to give a large improvement over the baseline performance of 69.54% accuracy for full document classification with the MNB classifier (66.06% for SVM), with the paragraph achieving accuracy of over 73% and the sentence approach achieving accuracy of over 72%. The largest improvement was achieved by word-based extraction, with a classification accuracy of 75.07% using the MNB classifier, an improvement of almost 5.52% in absolute terms (or a relative improvement of 8%).

Although performance is lower for the 3-point classification task, as shown in Table 6, the improvement over document-based classification is similar, improving from 54.45% to 59.46% (a relative improvement of over 9%) for word-based text extraction (N=30), with sentence- and paragraph-based approaches also giving large improvements over the baseline.

The performance of the naïve Bayes classifier was consistently better than that of the SVM, which may be due to the fact that a linear kernel used in conjunction with default parameter values were not appropriate for this domain, and which warrants further investigation.

We conducted a detailed analysis of the binary classification results from the optimally performing word-based (n=30) text extraction approach, examining variation in performance as the amount of training data for specific stocks is increased. Figure 2 shows the classification accuracy grouped by the number of training instances for specific stocks. There is some variation in performance, with stocks represented by 1 training instance achieving a classification accuracy of 75.78% (95 stocks), and those represented by 11 or more documents achieving a classification accuracy of 80.9% (6 stocks). Nevertheless, there is no clear trend towards higher accuracy for stocks that are over-represented in the training set, encouragingly suggesting that our classifier is not overtly biased towards those stocks, but rather is general enough to perform similarly for all stocks.

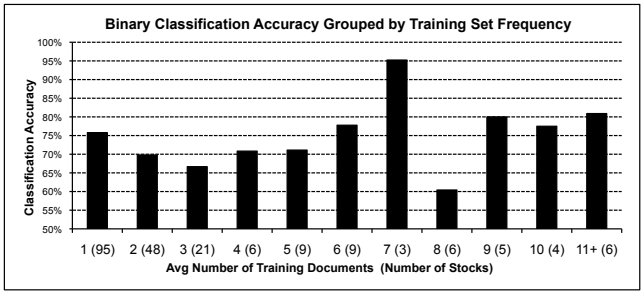

Figure 2: Binary classification accuracy, grouped by the number of times a topic stock was present in the training set. Since 10-fold cross-validation was used, these figures are averaged across the 10 folds.

In general, the results indicate that it is possible to achieve large improvements over document-based sentiment classification using quite simple text-extraction approaches to extract the most relevant segments of those documents. For both *binary* and *3-point* classification, the best results were achieved when word-based text extraction approaches were used, suggesting that for this dataset at least, paragraphs and sentences do not necessarily correspond to the unit of expression. This result differs from the result obtained by Zhang el al [33], who found that the full document gave

	Paragraph		Sentence		Words		
N	**SVM**	**MNB**	**SVM**	**MNB**	**N**	**SVM**	**MNB**
0	**53.3143**	**57.7242**	51.7928	54.9983	5	50.0378	55.5285
1	49.9402	56.9427	53.5476	56.8389	10	53.2130	56.1830
2	51.2468	55.6225	52.5535	56.2869	15	52.9150	56.6178
3	50.9255	55.5175	**53.8737**	**57.4825**	20	54.3244	57.8258
4	50.7082	55.1987	51.5643	56.9378	25	55.4064	58.0346
5	52.2348	54.8665	50.4724	55.6396	30	55.3001	**59.4621**
6	51.46914	55.4124	51.8036	57.1662	40	**56.6019**	58.2458
7	50.1744	55.4075	50.0402	56.0767	50	55.0522	58.1443
8	49.3822†	55.2988	51.0207	55.8642	60	52.7524	58.5877
9	50.48714	55.7373	50.9120	55.9717	70	54.8311	58.3605
10	49.7301	55.6334	50.6812	56.0743	80	53.7307	58.5779
Baseline	49.7190	54.4540	49.7190	54.4540		49.7190	54.4540

Table 6: 3-Point classification results for paragraph, sentence and word text extraction. Maximum accuracy is represented by bold text, while accuracy below that of the baseline are indicated with †

the best performance. They are interested in opinion detection, however, not sentiment polarity classification, and the information retrieval paradigm in which they conduct their evaluation means that relevance and opinion detection are being evaluated simultaneously. Our work, on the other hand, is exclusively concerned with sentiment polarity classification and shows the performance that can be achieved if topic relevance is assumed. It would be of interest to explore whether this result is specific to this dataset, or if similar a approach would prove useful in alternative domains.

6. CONCLUSIONS AND FUTURE WORK

In this paper we have explored the use of blog sources for sentiment analysis in the financial domain, and developed a corpus of over 1,500 document-level annotations. Analysis of this annotation effort suggests that humans have particular difficulty annotating for strength or degree of polarity, and in annotating documents as having mixed sentiment. Topic shift, which is where a single blog article discuss more than one topic, was identified in a significant percentage of the blog articles collected, with articles relative to at least one stock also relevant to an average of 2 other stocks.

In order to tackle the problem of topic shift, we proposed and evaluated simple text-extraction approaches to extract the most relative segments of a document with respect to a given topic, then trained and tested sentiment classifiers on the extracted sub-document representation. Empirical evaluation revealed that word-, sentence- and paragraph-based text extraction all achieved improvements over baseline (full document) effectiveness, with the best performance recorded when word-based text extraction techniques are used. Paragraph-based appoaches performed slightly better than sentence-based approaches, suggesting that, in this dataset at least, the paragraph is a more natural unit for the expression of sentiment than the sentence.

The features (bag of words) that we use for our classifier are quite simple compared to what has been used by other researchers in sentiment analysis. We plan to explore the use of linguistic features and domain independent resources (such as SentiWordNet [9]) in subsequent experiments. We have made exclusive use of document-level annotations in this paper, even though we have annotated our corpus at the paragraph level, and we plan to use the paragraph-level annotations in future work.

As discussed previously, this work is part of a project to monitor the overall sentiment of the blogosphere towards individual companies and the market in general. Accordingly, in addition to improving our sentiment polarity classifier, we will explore methods of aggregrating these sentiment results, and we are currently developing a user interface for displaying and interacting with this data.

Acknowledgements

This work is supported by Science Foundation Ireland under grant 07/CE/I1147, and by Enterprise Ireland under grant IP/2008/0549.

7. REFERENCES

[1] K. Ahmad, D. Cheng, and Y. Almas. Multi-lingual sentiment analysis of financial news streams. In *Proceedings of the 1st International Conference on Grid in Finance*, Palermo, 2006.

[2] G. Attardi and M. Simi. Blog mining through opinionated words. In E. M. Voorhees and L. P. Buckland, editors, *TREC*, volume Special Publication 500-272. National Institute of Standards and Technology (NIST), 2006.

[3] A. Bermingham and A. F. Smeaton. A study of inter-annotator agreement for opinion retrieval. In *SIGIR '09: Proceedings of the 32nd annual international ACM SIGIR conference on Research and development in information retrieval*, 2009.

[4] S. Bradshaw. Reference directed indexing: Redeeming relevance for subject search in citation indexes. In *Proc. of the 7th ECDL*, pages 499–510, 2003.

[5] S. Chakrabarti, B. Dom, P. Raghavan, S. Rajagopalan, D. Gibson, and J. Kleinberg. Automatic resource compilation by analyzing hyperlink structure and associated text. In *WWW7: Proceedings of the seventh international conference on World Wide Web 7*, pages 65–74, Amsterdam, The Netherlands, The Netherlands, 1998. Elsevier Science Publishers B. V.

[6] S. Das and M. Chen. Yahoo! for amazon: Sentiment extraction from small talk on the web. *Management Science*, 53(9):1375–1388, 2007.

[7] K. Dave, S. Lawrence, and D. M. Pennock. Mining the peanut gallery: opinion extraction and semantic classification of product reviews. In *WWW '03: Proceedings of the 12th international conference on World Wide Web*, pages 519–528, New York, NY, USA, 2003. ACM.

[8] A. Devitt and K. Ahmad. Sentiment polarity identification in financial news: A cohesion-based approach. In *Annual Meeting of the Association of Computational Linguistics*, volume 45, page 984, 2007.

[9] A. Esuli and F. Sebastiani. Sentiwordnet: A publicly available lexical resource for opinion mining. In *5th Conference on Language Resources and Evaluation*, pages 417–422, 2006.

[10] G. Forman. An extensive empirical study of feature selection metrics for text classification. *Journal of Machine Learning Research*, 3(1):1533–7928, 2003.

[11] A. Funk, Y. Li, H. Saggion, K. Bontcheva, and C. Leibold. Opinion analysis for business intelligence applications. 2008.

[12] T. H. Haveliwala, A. Gionis, D. Klein, and P. Indyk. Evaluating strategies for similarity search on the web. In *WWW '02: Proceedings of the 11th international conference on World Wide Web*, pages 432–442, New York, NY, USA, 2002. ACM.

[13] B. He, C. Macdonald, J. He, and I. Ounis. An effective statistical approach to blog post opinion retrieval. In *CIKM '08: Proceeding of the 17th ACM conference on Information and knowledge management*, pages 1063–1072, New York, NY, USA, 2008. ACM.

[14] M. Koppel and I. Shtrimberg. Good news or bad news?: Let the market decide. In *AAAI Spring Symposium on Exploring Attitude and Affect in Text*, pages 86–88. Springer, 2004.

[15] R. Kraft and J. Zien. Mining anchor text for query refinement. In *WWW '04: Proceedings of the 13th international conference on World Wide Web*, pages 666–674, New York, NY, USA, 2004. ACM.

[16] Y. Lee, S.-H. Na, J. Kim, S.-H. Nam, H.-Y. Jung, and J.-H. Lee. Kle at trec 2008 blog track: Blog post and feed retrieval. In *TREC*, 2008.

[17] K. Lerman, A. Gilder, M. Dredze, and F. Pereira. Reading the markets: Forecasting public opinion of political candidates by news analysis. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 473–480, Manchester, UK, August 2008. Coling 2008 Organizing Committee.

[18] D. Lewis, Y. Yang, T. Rose, and F. Li. Rcv1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research*, 5:361–397, April 2004.

[19] B. Liu, M. Hu, and J. Cheng. Opinion observer: analyzing and comparing opinions on the web. In *WWW '05: Proceedings of the 14th international conference on World Wide Web*, pages 342–351, New York, NY, USA, 2005. ACM.

[20] C. Macdonald and I. Ounis. The trec blogs06 collection : Creating and analysing a blog test collection. Technical report, University of Glasgow, Department of Computing Science, 2006.

[21] N. Nicolov, F. Salvetti, and S. Ivanova. Sentiment analysis: Does coreference matter? In *Symposium on Affective Language in Human and Machine*, 2008.

[22] I. Ounis, C. MacDonald, and I. Soboroff. Overview of the TREC-2008 Blog Track. In *The Seventeenth Text REtrieval Conference (TREC 2008) Proceedings*. NIST, 2008.

[23] B. Pang and L. Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. pages 115–124, 2005.

[24] B. Pang and L. Lee. *Opinion mining and sentiment analysis*. Now Publishers, 2008.

[25] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up?: Sentiment classification using machine learning techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 79–86, 2002.

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

[27] Y. Seki, D. K. Evans, L. Ku, L. Sun, H. Chen, and N. Kando. Overview of multilingual opinion analysis task at NTCIR-7. 2008.

[28] H. Tang, S. Tan, and X. Cheng. A survey on sentiment detection of reviews. *Expert Systems With Applications*, 2009.

[29] P. Turney. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 417–424, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.

[30] V. Vapnik. The nature of statistical learning theory. 1995.

[31] T. Wilson, J. Wiebe, and P. Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 347–354. Association for Computational Linguistics Morristown, NJ, USA, 2005.

[32] I. H. Witten and E. Frank. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2nd edition, 2005.

[33] M. Zhang and X. Ye. A generation model to unify topic relevance and lexicon-based sentiment for opinion retrieval. In *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 411–418, New York, NY, USA, 2008. ACM.