Domain Adaptation using Stock Market Prices to Refine Sentiment Dictionaries

Andrew Moore*, **Paul Rayson***, **Steven Young**†
*School of Computing and Communications, †Department of Accounting and Finance
Lancaster University, UK
{a.moore, p.rayson, s.young}@lancaster.ac.uk

Abstract

As part of a larger project where we are examining the relationship and influence of news and social media on stock price, here we investigate the potential links between the sentiment of news articles about companies and stock price change of those companies. We describe a method to adapt sentiment word lists based on news articles about specific companies, in our case downloaded from the Guardian. Our novel approach here is to adapt word lists in sentiment classifiers for news articles based on the relevant stock price change of a company at the time of web publication of the articles. This adaptable word list approach is compared against the financial lexicon from Loughran and McDonald (2011) as well as the more general MPQA word list (Wilson et al., 2005). Our experiments investigate the need for domain specific word lists and demonstrate how general word lists miss indicators of sentiment by not creating or adapting lists that come directly from news about the company. The companies in our experiments are BP, Royal Dutch Shell and Volkswagen.

Keywords: Sentiment analysis, Sentiment dictionaries, Domain adaptation

1. Introduction

Sentiment dictionaries such as the MPQA lexicon (Wilson et al., 2005) have been used in the past to capture general sentiment, and manually generated lexicons have been adapted to the financial domain (Loughran and McDonald, 2011), however we argue that this process of adaptation to the financial domain does not go far enough.

Each sector[1] within the financial domain has its own specific vocabulary where the meaning of words can change greatly, for instance the word "crude" might be interpreted negatively depending on the context or the domain that a company is operating in, but for oil companies (e.g. BP) it will mean something entirely different. Clearly, it is important to find the correct domain to understand word meanings so that the sentiment dictionary can be tailored appropriately e.g. oil sector or company level. The method presented in this article uses the stock exchange prices to label all news articles with one of three sentiments: positive, neutral or negative. We automatically create sentiment word lists from the training on news articles for the specific companies to compare against MPQA (Wilson et al., 2005) and Loughran and McDonald (2011).

It could be argued that combining word sense disambiguation approaches with sentiment analysis would help address such challenges, but in our scenario this would not directly address the domain expertise and knowledge of performance of a given company that may be external to the text. Instead, we adopt an approach to model the changing meaning at different levels: general (i.e. not adapted), the entire financial domain, specific market sector and finally company specific. In order to investigate the improvement of sentiment labelling of articles, we carry out our experiments at these multiple levels. Our experimental results show that domain adaptation is required to have higher accuracy than existing word lists when trying to predict the sentiment of a news article.

2. Related Work

There is a vast body of work on sentiment analysis methods and techniques. For example, Pang et al. (2002) found that corpus techniques using machine learning greatly improved sentiment classification of movie reviews in comparison to human generated sentiment word lists. Turney (2002) used PMI-IR (Pointwise Mutual Information and Information Retrieval) to detect sentiment within reviews from four different domains on a phrase level basis.

Recent work has applied sentiment methods to financial text analysis. Chen et al. (2014) correlated negative words in articles from Seeking Alpha[2] and comments of the articles with lower performance using the word list from Loughran and McDonald (2011). Using 8K reports[3] Lee et al. (2014) was able to predict the next day's stock price with 55.5% accuracy using an ensemble of three non-negative matrix factorisation models that used both linguistic and numeric features, with majority voting. Also Lee et al. (2014) found that using linguistic features not just numeric features significantly improved their results. Using the Harvard 4 psychological list of negative words, Tetlock et al. (2008) found and correlated negative words within the Wall Street Journal[4] and Dow Jones News Service with the stock price return. Also, Loughran and McDonald (2011) found that with the bag of words (BOW) method that employing a financial sentiment lexicon instead of a general lexicon, there is a correlation between the number of negative words in a 10K report[5] and negative excess returns.

[1] A sector is an industry or market sharing common characteristics. Characteristics could be the type of resources used and what is produced, in our example the sector is oil. Our third company, Volkswagen, was chosen outside of this domain, but as we knew it would have plenty of recent press coverage.

[2] http://seekingalpha.com/

[3] 8K reports are the companies "current report" according to SEC (Securities and Exchange Commission) https://www.sec.gov/answers/form8k.htm

[4] http://www.wsj.com/europe

[5] 10K reports are the companies annual report that "provides a

3. Datasets

Our news article dataset was downloaded from the Guardian newspaper through their API[6]. We gathered 2486, 955 and 306 articles about Shell, BP and Volkswagen respectively. Stock price data for each company was collected through Quandl using their API[7]. The stock prices for BP and Shell were cross checked against stock price on Thomson Reuters using their EIKON application[8] and Volkswagen prices were checked against those shown on the Frankfurt Stock Exchange[9]. The news articles that we used were published online between 30th September 2013 and the 1st October 2015 and the stock prices relate to prices declared between the 1st October 2013 and the 1st October 2015.

3.1. Stock price pre-processing

The stock prices collected were for each company[10] and then processed to calculate the stock price change for each day using equation (1). The stock price changes for each company over the collection time period were distributed normally. We designated the lowest third of stock price changes as decrease, the highest third as increase and the middle third as nominal change.

$$x = \frac{(\text{Closing price} - \text{Opening price})}{(\frac{\text{Closing price} + \text{Opening price}}{2})} \quad (1)$$

3.2. News article pre-processing

The news articles were collected by searching for the company name[11] in the Guardian API. The only restriction was the removal of articles in the media and film sections because a manual inspection revealed that these articles were not relevant to the companies. From each news article only the title and the body of the text were collected after which it was passed through a HTML parser to remove the majority of the HTML tags. The processed text was then Part Of Speech (POS) tagged using the CLAWS POS tagger (Garside and Smith, 1997), in order to tokenise the text, insert sentence boundaries and help remove punctuation.

Finally, each news article was marked with the stock price change (increase, nominal, decrease) via the web publication date and our stock price data collected above. Our assumption is that a news article is most closely related to the stock price change in the next trading day after the article was published. We do not assume that there is a causal link but in general an increase in price is assumed to happen around the same time as good news and vice versa. Therefore articles relating to an increase, nominal change or decrease in stock price are tagged with a sentiment value of positive, neutral or negative respectively. We chose the next working day because Lee et al. (2014) found that linguistic features have the best performance one day after the event, although it should be noted that this was with 8K reports and not news articles.

3.3. Word list pre-processing

The MPQA word list was divided into three lists, one for each sentiment category (positive, neutral and negative). Each sentiment category contained a word as long as its polarity matched the sentiment category and was not stemmed. MPQA ranks words as strong or weak with respect to sentiment, however both ranks were put into the same category and not split producing only three word lists rather than six. The Loughran and McDonald (L&M) word list only contains positive and negative words because the word lists that they produced did not contain a clear neutral category.

4. Method

To determine the sentiment of an article we defined an adaptable bag of words (ABOW) method which finds the top five percent of the most frequently used words in each of the three sentiment categories (positive, neutral and negative) and selects words that appear only in that category, as this will most likely remove common words such as 'the'. The adaptability of the bag of words stems from the fact that the words originate from the text. As more news articles are added to the training set the top five percent of most frequently used words change, thus the model changes with more data. We keep three bags in this ABOW model representing positive, neutral and negative sentiments. The sector list was derived from combining the Shell and BP word lists, Volkswagen did not have a sector list as this was the only company in the car manufacturing industry that we used. We also followed the method by Martineau and Finin (2009) however we used unigrams rather than bigrams as features and we used an SVC (Support Vector Classifier)[12] (Pedregosa et al., 2011) instead of SVM (Support Vector Machine) however both have linear kernels and were used to classify for two-way sentiment (positive and negative).

In the testing phase, each article is subjected to a plurality voting system (Clarkson et al., 2007). Our system determines the sentiment of the article depending on which bag in the ABOW model has the highest count. The total count derives from the frequency of words in each bag occurring in the article. An extra rule was added to the voting system to handle ties.

comprehensive overview of the company's business and financial condition" according to SEC (Securities and Exchange Commission) https://www.sec.gov/answers/form10k.htm

[6] http://open-platform.theguardian.com/

[7] https://www.quandl.com/

[8] http://financial.thomsonreuters.com/en/products/tools-applications/trading-investment-tools/eikon-trading-software.html

[9] http://www.boerse-frankfurt.de/

[10] Both BP and Royal Dutch Shell prices were collected from the Google finance database with the following codes respectively GOOG/LON_BP_, GOOG/LON_RDSB and the Volkswagen prices were collected from the Y finance database with the following code YAHOO/F_VOW.

[11] We searched for bp, shell and volkswagen.

[12] http://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC.html#sklearn.svm.LinearSVC

Company	MPQA	L&M	BP	Shell	Volkswagen	Sector	Random	Majority class
BP	0.409	**0.654**	0.342	0.348	0.195	0.351	0.333	0.450
	0.475	**0.528**	0.351	0.236	0.214	0.202	0.333	0.377
	-	-	0.481	0.495	0.520	0.476	0.5	**0.521**
Shell	0.309	**0.703**	0.308	0.420	0.125	0.414	0.333	0.522
	0.253	0.480	0.238	0.119	0.096	0.192	0.333	**0.545**
	-	-	0.443	0.515	0.444	0.480	0.5	**0.597**
Volkswagen	0.331	0.370	0.311	**0.438**	0.321	0.318	0.333	0.396
	0.100	0.050	0.270	0.170	0.220	0.120	0.333	**0.500**
	-	-	0.444	0.508	0.362	0.416	0.5	**0.633**

Table 1: Results table for positive stock price trend data.

Company	MPQA	L&M	BP	Shell	Volkswagen	Sector	Random	Majority class
BP	0.355	0.360	0.332	0.410	**0.482**	0.300	0.333	0.464
	0.256	0.249	0.336	0.496	**0.580**	0.410	0.333	0.518
	-	-	0.584	0.526	0.647	0.411	0.5	**0.652**
Shell	0.249	0.305	0.303	0.328	0.510	0.290	0.333	**0.526**
	0.254	0.229	0.430	**0.562**	0.535	0.507	0.333	0.442
	-	-	0.536	0.522	0.550	0.527	0.5	**0.697**
Volkswagen	0.261	0.221	0.247	0.429	0.568	0.363	0.333	**0.661**
	0.054	0.027	0.281	0.267	0.371	0.198	0.333	**0.717**
	-	-	0.496	0.409	0.596	0.456	0.5	**0.712**

Table 2: Results table for negative stock price trend data.

Company	MPQA	L&M	BP	Shell	Volkswagen	Sector	Random	Majority class
BP	0.322	**0.440**	0.310	0.338	0.362	0.301	0.333	0.341
	0.253	**0.408**	0.262	0.355	0.333	0.308	0.333	0.368
	-	-	0.532	0.464	0.588	0.442	0.5	**0.609**
Shell	0.297	0.339	0.343	0.300	**0.460**	0.308	0.333	**0.460**
	0.209	0.281	0.389	0.507	0.490	**0.515**	0.333	0.441
	-	-	0.513	0.495	0.517	0.488	0.5	**0.579**
Volkswagen	0.339	**0.419**	0.331	0.312	0.400	0.336	0.333	0.418
	0.100	0.200	0.300	0.300	0.300	0.300	0.333	**0.500**
	-	-	0.508	0.569	**0.583**	0.522	0.5	0.545

Table 3: Results table for generally neutral stock price trend data.

5. Results

The results are shown in tables 1[13], 2[14], and 3[15]. We have divided results into three tables in order to evaluate our system over three time periods representing three differing stock trends (positive: table 1, negative: table 2, and neutral: table 3). After manually sampling ten news articles from the news dataset we found low precision[16] with respect to relevancy of the news articles to the companies financial performance. Therefore, we created a sub-corpus using news articles occurring in the business sections thus reducing the dataset[17] and the number of test data points greatly but with an increase in relevance to financial performance[18]. As seen in the results table each company has three rows. The first row for each company shows the results when using the whole dataset, the second row shows the results when testing on business section data only, finally the third row is the results of the SVC on the whole dataset. Each column represents a different word list that was used on the company data represented in the row; all company names in the columns are word lists that were created from our ABOW. SVC was trained on data from the companies mentioned in the column header, and tested on the company data that is mentioned in the row header.

For each company, we compared our method for finding the sentiment of a news article against the MPQA and L&M dictionaries using ten-fold cross validation. It should be noted that as L&M only have positive and negative word lists, any neutral news articles were ignored for those figures, to ensure they were not penalised for the lack of a

[13]The majority class for BP and shell for the SVC analysis is positive, the other two companies is neutral but Volkswagen SVC is negative. These results are from tests on data between 2013-12-17 and 2014-5-6.

[14]The majority class for all companies is guessing negative. These results are from tests on data between 2015-5-14 and 2015-10-1.

[15]The majority class for all companies is guessing negative apart from business section BP which is guessing positive. These results are from tests on data between 2015-2-6 and 2015-8-5.

[16]BP, Shell and Volkswagen had precision of 20%, 10% and 40% respectively.

[17]BP, Shell and Volkswagen have 327, 347 and 80 news articles respectively.

[18]BP, Shell and Volkswagen had precision of 40%, 80% and 90% respectively.

neutral word list.

Sentiment	BP	Shell	Volkswagen
Positive	**0.357**	0.337	0.294
Neutral	0.308	0.322	0.232
Negative	0.335	**0.342**	**0.474**

Table 4: Distribution of all company articles

As shown in the results tables, all companies apart from BP performed well against the existing and sector-level word lists thus demonstrating the need for adapting sentiment word lists to company level. Interestingly, the general word lists (MPQA and L&M) perform best when the data is less skewed, as shown by the majority class having a lower probability. The most likely reason why the Volkswagen list performs better on negative trend data is because of the unbalanced nature of the Volkswagen articles towards negative sentiment during our sampling period as shown by the distribution table 4[19]. We observed in some of the experiments that the word lists performed better on the smaller business section data indicating that more relevant data is required to enhance performance and quality of word lists. Although the general majority classifier beats all other classifiers we have shown improvement of sentiment word lists by domain adaptation using stock market prices relative to existing static lists. A better machine learning algorithm with a non-linear kernel may further improve these results.

6. Conclusion and Future Work

Our results show promising improvement over existing sentiment dictionary methods but could be further improved using more advanced machine learning methods such as Lee et al. (2014). We also intend to investigate word embedding and vector space techniques for improving sentiment analysis as shown by Maas et al. (2011) and Loughran and McDonald (2011) since these should help the system to take account of local and document level context. Instead of using the entire article, we may improve results by only using subjective sentences (Pang and Lee, 2004) or simple negation (Pang et al., 2002). Rather than assuming that all words in an article and all articles mentioning the company by name have equal importance in terms of stock price change, we will investigate relevance metrics to better model influence and trust relationships for readers of the texts. Finally, as the precision sampling was on a small subset of the whole dataset more work is needed to see how large a problem relevancy is in the Guardian dataset and other news sources. All word lists created for this research are made freely available[20].

7. Acknowledgements

This research is funded at Lancaster University by an EPSRC Doctoral Training Grant.

[19]The distribution of just the business section articles is similar apart from BP which has marginally more negative rather than positive articles.

[20]http://ucrel.github.io/ABOW/

8. References

Chen, H., De, P., Hu, Y. J., and Hwang, B.-H. (2014). Wisdom of crowds: The value of stock opinions transmitted through social media. *Review of Financial Studies*, 27(5):1367–1403.

Clarkson, M. R., Chong, S., and Myers, A. C. (2007). Civitas: A secure voting system. Technical report, Cornell University.

Garside, R. and Smith, N. (1997). A hybrid grammatical tagger: CLAWS4. *Corpus annotation: Linguistic information from computer text corpora*, pages 102–121.

Lee, H., Surdeanu, M., MacCartney, B., and Jurafsky, D. (2014). On the Importance of Text Analysis for Stock Price Prediction. *Proceedings of LREC-2014*.

Loughran, T. and McDonald, B. (2011). When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. *The Journal of Finance*, 66(1):35–65, February.

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., and Potts, C. (2011). Learning Word Vectors for Sentiment Analysis. *Proceedings of ACL'11*, pages 142–150.

Martineau, J. and Finin, T. (2009). Delta TFIDF: An Improved Feature Space for Sentiment Analysis. In *ICWSM*.

Pang, B. and Lee, L. (2004). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of ACL'04*, page 271. Association for Computational Linguistics.

Pang, B., Lee, L., and Vaithyanathan, S. (2002). Thumbs up? Sentiment Classification using Machine Learning Techniques. *arXiv:cs/0205070*, May. arXiv: cs/0205070.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Tetlock, P. C., Saar-Tsechansky, M., and Macskassy, S. (2008). More Than Words: Quantifying Language to Measure Firms' Fundamentals. *The Journal of Finance*, 63(3):1437–1467, June.

Turney, P. D. (2002). Thumbs Up or Thumbs Down?: Semantic Orientation Applied to Unsupervised Classification of Reviews. In *Proceedings of ACL'02*, ACL '02, pages 417–424, Stroudsburg, PA, USA. Association for Computational Linguistics.

Wilson, T., Wiebe, J., and Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of ACL'05*, pages 347–354. Association for Computational Linguistics.