Domain Adaptation using Stock Market Prices to Refine Sentiment Dictionaries

Andrew Moore, Paul Rayson, Steven Young

May 23, 2016

School of Computing and Communications, Department of Accounting and Finance,
Lancaster University, UK.

Table of contents

Introduction

The approach

To adapt automatic sentiment dictionaries using news articles based on stock market prices as indicators of sentiment.

Powered by
theguardian

Motivation

What already exists?

General word lists

MPQA [3].

Financial word lists

Loughran and McDonald [1].

How far to adapt?

Loughran and McDonald

'The question we address in this paper is whether a word list developed for psychology and sociology translates well into the realm of business' [1]

How far to adapt?

Where as Loughran and McDonald looked at the **financial** domain as a whole, we examine whether it is necessary to further refine dictionaries in that domain.

We create sentiment dictionaries for the **industry sector** and the **company**.

Pre-Processing

Companies

BP

Royal Dutch Shell

Volkswagen

Data

Time frame: 30th September 2013 to the 1st of October 2015.

The Guardian News API[1].

Quandl[2] for stock market prices.

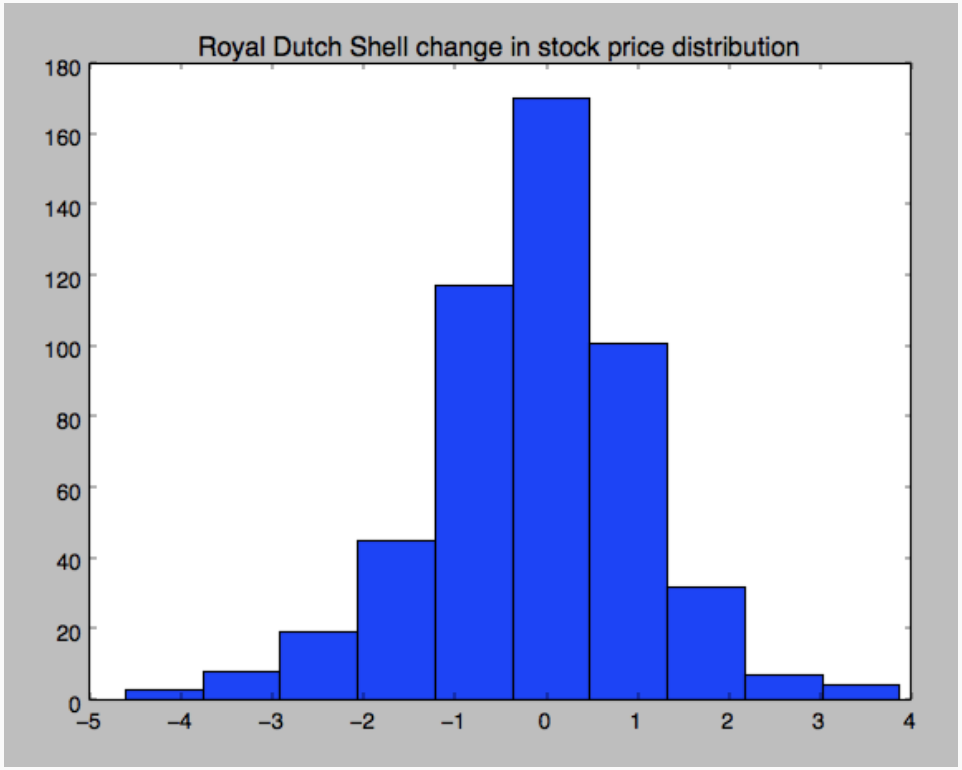
[1] http://open-platform.theguardian.com/
[2] https://www.quandl.com/

Pre-Processing

Stock price change

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

Distribution of Shell's stock price changes



Method

Evaluation

We use the MPQA as a general word list, Loughran and McDonald as a general financial word list, we create an oil sector adaptable word list by combining words from BP and Shell news articles and finally we create company adaptable word lists for all companies.

Evaluation

These word lists are then compared by how well they can find the overall sentiment of news articles based on these companies. The evaluation is ten fold cross validation over a specified time period where we trained on all the data apart from the period we are testing out of the periods that are to be tested in the ten fold cross validation.

Methods

General Word Lists

The general word lists (MPQA and L&M) will use the Bag Of Words method and the majority voting system.

Methods

ABOW

The Adaptable Bag Of Words finds the top 5% most frequent words that are only in that bag. The bags were Positive, Neutral and Negative.

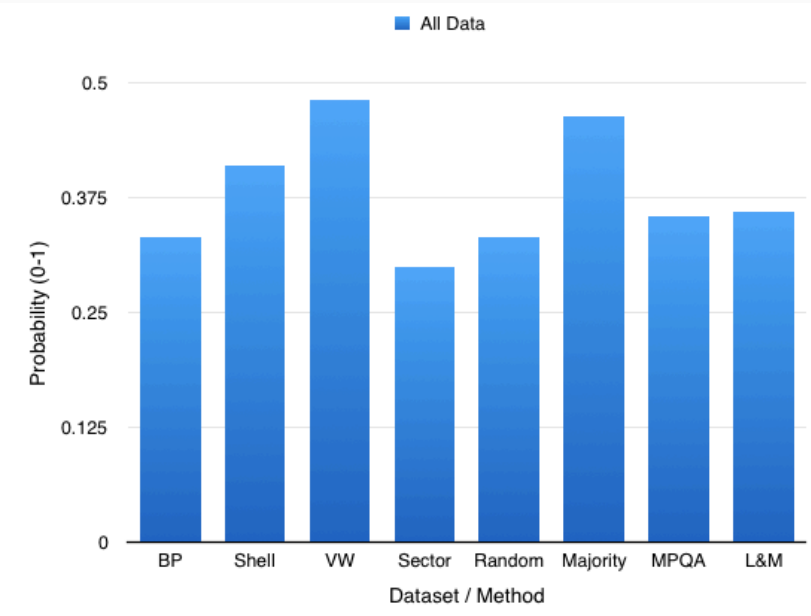
Methods

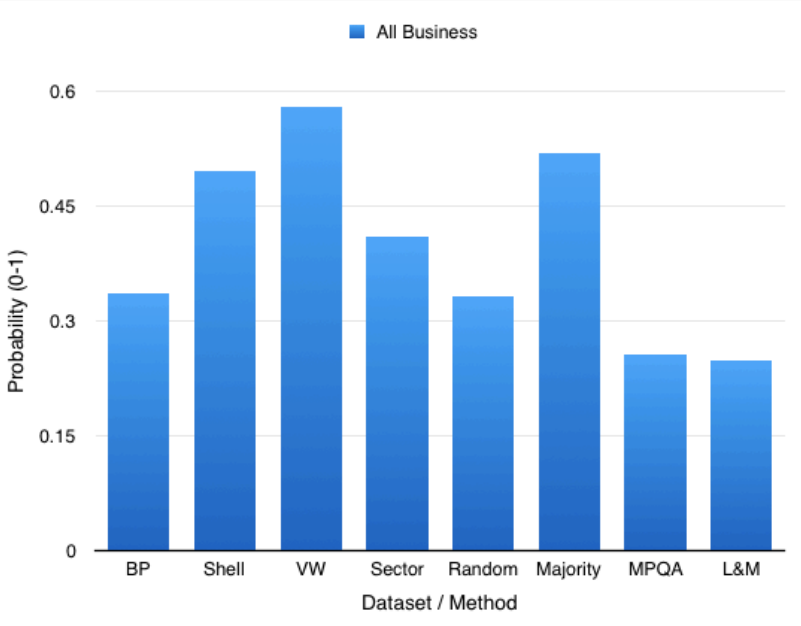
Martineau and Finn [2]

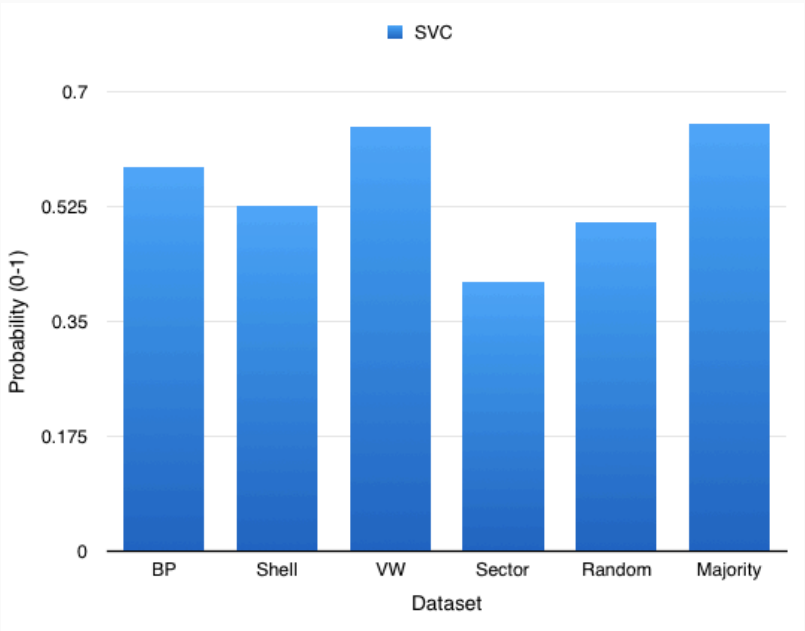
This used a sentiment TF-IDF with a vector for a Support Vector Classifier (SVC) and a linear kernel.

Results

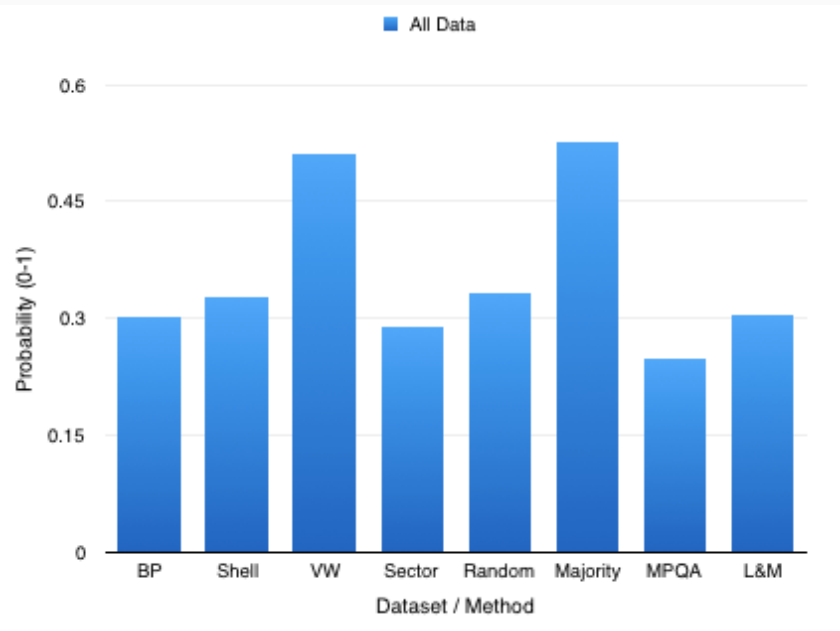
BP

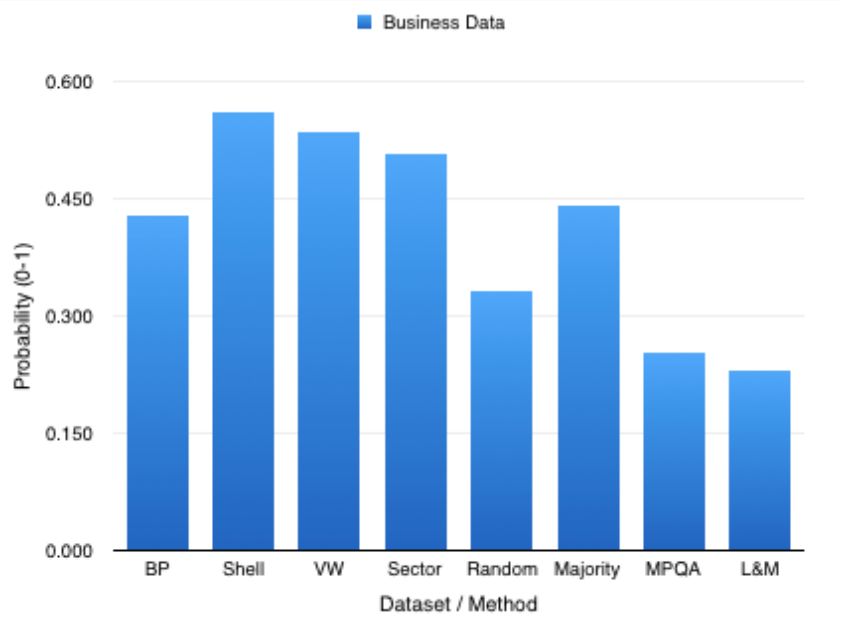

All Data
0.5
0.375
0.25
0.125
0
Probability (0-1)
BP
Shell
VW
Sector
Random
Majority
MPQA
L&M
Dataset / Method

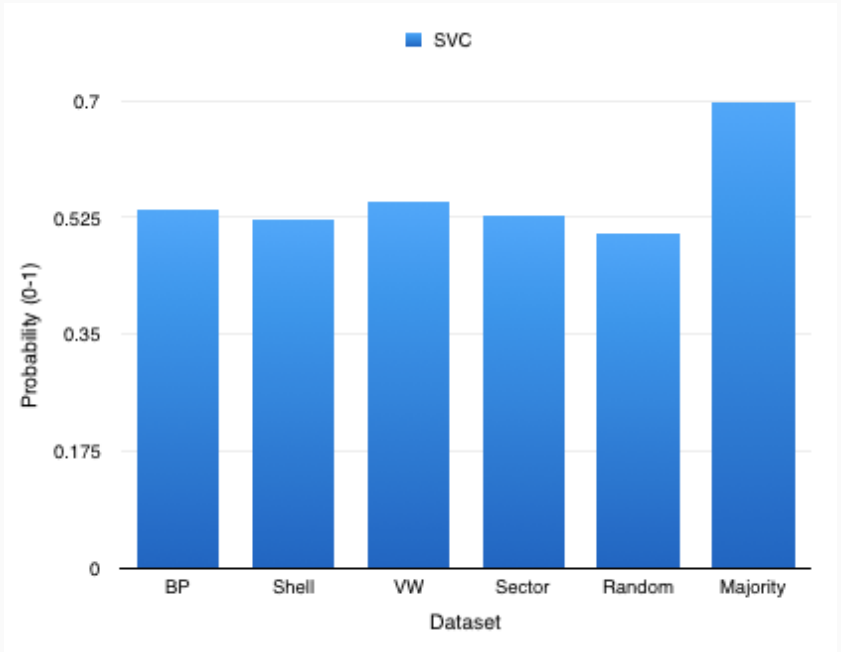

All Business
0.6
0.45
0.3
0.15
0
Probability (0-1)
BP
Shell
VW
Sector
Random
Majority
MPQA
L&M
Dataset / Method


SVC
0.7
0.525
0.35
0.175
0
Probability (0-1)
BP
Shell
VW
Sector
Random
Majority
Dataset

Shell

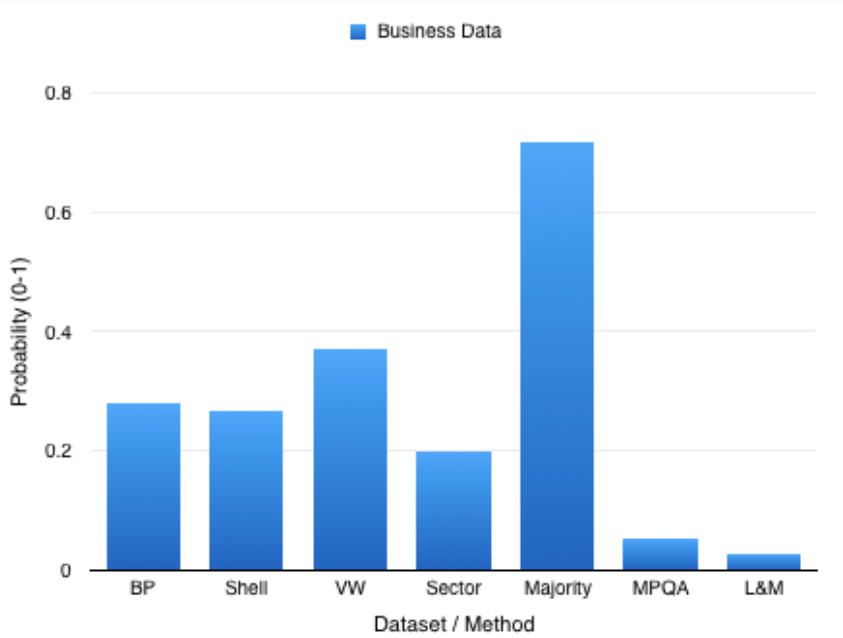

All Data
0.6
0.45
0.3
0.15
0
Probability (0-1)
BP
Shell
VW
Sector
Random
Majority
MPQA
L&M
Dataset / Method

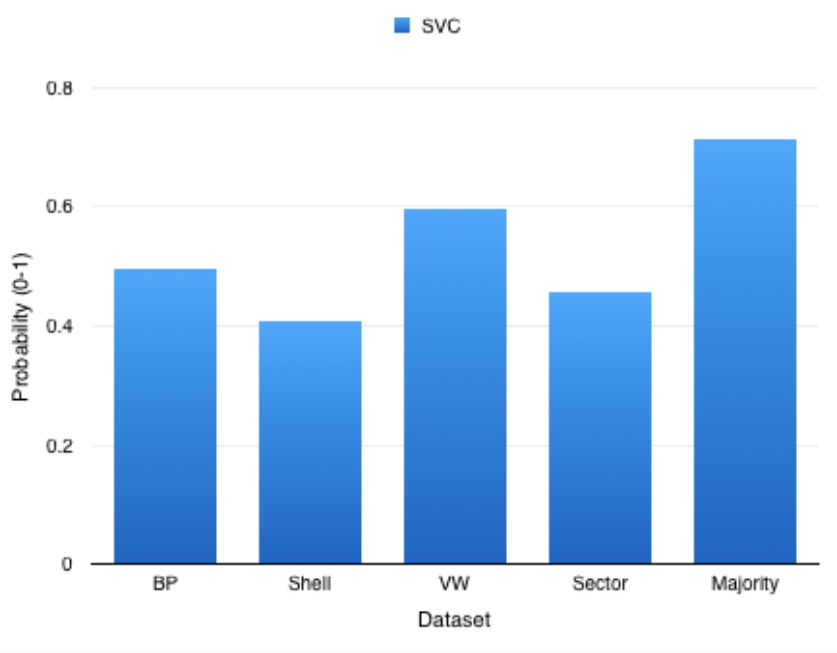

Business Data
0.600
0.450
0.300
0.150
0.000
Probability (0-1)
BP
Shell
VW
Sector
Random
Majority
MPQA
L&M
Dataset / Method


SVC
0.7
0.525
0.35
0.175
0
Probability (0-1)
BP
Shell
VW
Sector
Random
Majority
Dataset

VW


All Data
0.7
0.525
0.35
0.175
0
Probability (0-1)
BP
Shell
VW
Sector
Majority
MPQA
L&M
Dataset / Method


Business Data
0.8
0.6
0.4
0.2
0
Probability (0-1)
BP
Shell
VW
Sector
Majority
MPQA
L&M
Dataset / Method


SVC
0.8
0.6
0.4
0.2
0
Probability (0-1)
BP
Shell
VW
Sector
Majority
Dataset

Company Vs. Sector Vs. General

Sector

Word such as “spill” (BP and Sector) represent possible negative words in the oil sector however in general and financial word lists this could be misrepresented.

Company

Words such as “mexico” (BP and Sector), “deepwater” (BP and Sector), “obama” and “2008” (Shell), would represent negative words with respect to BP and Shell.

Conclusion

Novelty

To automatically generate domain specific sentiment lexicons using stock market data and news articles.

Conclusion and Future Work

1. Better quality news article collection.
2. Better Machine Learning method may improve results.
3. Using only subjective sentences.
4. Relevance metrics for news articles based on trust and influence.
5. Using rules to better model the context of the words collected.
6. Expand the number of companies and the industry sectors.

GitHub storing all of the word lists created in this paper:
http://ucrel.github.io/ABOW/

Questions?

Tweet us:
@apmoore94, @perayson, @UCREL_Lancaster

Lexicon at:
http://ucrel.github.io/ABOW/

References


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


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


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