Semi-supervised dictionary induction methods for oil price prediction

Candidate Number: 17127 | Github: d151841514526c

Abstract

I study the use of two different semi-supervised induction methods to generate daily sentiment series for oil-related news articles. The documents are scrapped from a major online news provider and cover the period between 2019 and 2020. Comparisons between the induced valences are made and their ability to improve out-of-sample forecasts for forward oil contracts is evaluated. Results show that while induced valences are coherent no improvement on out-of-sample forecasting performance is achieved once the target series is appropriately pre-whitened. Increased market volatility due to Covid-19 may be shadowing the forecasting gains that previous studies in the field found.

Keywords: Sentiment analysis, Dictionary Methods, Oil Prices

1 Introduction

One of the key variables to keep track of the global economy pulse are oil prices. Not only are these related to the production of thousands of different market goods but also highly correlated with global economic activity and foreign exchange rates (Habib, Bützer, and Stracca 2016). In a recent paper Loughran, McDonald, and Pragidis (2019) found strong evidence that traders overreact to the content of oil-related news articles and proposed a novel keyword list that can be used to measure the information content in these articles.

The measures of sentiment used in Loughran, McDonald, and Pragidis (2019) go beyond the traditional dictionary methods by using information related to the context of each token. However, their approach is based on a careful inspection of their particular corpus and it is not easy to assess whether their results are widely applicable. For this reason, in this document I explore whether similar conclusions can be obtained by the use of semi-supervised dictionary induction methods. I focus on the use of the SentiProp algorithm as proposed in Hamilton *et al.* 2016 and the Semantic Axis method of An, Kwak, and Ahn 2018.

One of the advantages in using semi-supervised dictionary induction methods is that the end user only has to provide a set of seed words from which the valence of tokens in the corpus can be inferred. This allows the procedure to be easily used even in slightly different corpora and makes its application straightforward. A further advantage is that the process of induction is done through the use of word embeddings and hence the obtained "dictionary" implicitly takes into account the context of the tokens.

To evaluate the methods above and the original proposal (Loughran, McDonald, and Pragidis 2019) I used a corpus of oil-related news constructed through the use of web scrapping applied to the portal *https://oilprice.com/*. The information scrapped covers the period 2019-2020 and is composed of more than 3400 articles. The information is relevant as this page is currently one of the main information providers for oil market investors and traders.

Results show that the token valence inferred from the semi-supervised induction methods is appropriate. All sentiments series derived from the induced dictionaries share a similar behaviour.

MY 459 (2022)

Candidate Number
17127

Lecturer
Blake Miller

Furthermore, token valence between methods appears to be linearly related as expected. However, derived sentiment series doesn't appear to have additional predictive power on oil prices as measured by out of sample performance once the autocorrelation of the series is taken into account.

The structure of the document is as follows. Section 2 gives a brief literature review. Section 3 introduces the data used and presents a brief exploratory analysis. Section 4 presents the results obtained and section 5 concludes.

2 Motivation

As shown by Kilian (2009) several macroeconomic variables respond differently to oil price movements depending on the origin of the shock. Hence, it is in the interest of economic policy makers to be able to predict oil price trajectories in order to taylor their policy responses appropiately. In fact, oil prices have been show to have a strong relation to economic activity Lardic and Mignon (2008) which strengthens the need to investigate its dynamics.

Interest in sentiment analysis on the finance field as a way to improve forecasting accuracy has been steadily growing since the seminal work of Tetlock, Saar-Tsechansky, and Macskassy (2008). Recent articles in the field have shown that news articles and social media posts indeed have useful information to improve predictions of asset prices, stock performance and market volatility on a broad scale (see Kearney and Liu (2014)).

Starting with the work of Loughran, McDonald, and Pragidis (2019) there has been a huge interest in applying text analysis tools to improve prediction of oil prices. In their seminal paper the authors developed a novel keyword list that enables investors to measure the information content of news articles related to the oil sector. Recently, Datta and Dias (2019) created both oil supply and demand indexes following a similar dictionary approach. The indexes created were shown to contain useful information to predict future oil price movements and used to derive a historical decomposition of large price movements in crisis periods. However, not all studies have shown such positive results, Calomiris, Çakır Melek, and Mamaysky (2020) found that only through data mining it is possible to identify successful out-of-sample forecasting models for oil prices based on text features. The authors remark that is difficult to identify transparent strategies for finding variables that improve out-of-sample performance in this scenario.

3 Data and Methodology

News articles are web scrapped using the Beautiful Soup library in Python parsing the html from *https://oilprice.com/Latest-Energy-News/World-News*. Documents were written in the period between January 1st of 2019 to 31 of December of 2020. The corpus has 3460 articles by 41 different authors.

The source is relevant as this website is visited approximately by 100.000 users each day and publishes more oil-related news than any other online page. Furthermore, the team at OilPrice provides news and analysis to mainstream providers such as CNN Money, Business Insider, NASDAQ among others. The site is mostly used by investors, traders and hedge fund managers that keep track of the oil market in USA, UK and Canada.

To keep only the relevant news content I filtered out the information on the editor and the site by using regular expressions. All words are converted to lowercase. Punctuation, numbers, symbols and trailing white spaces are removed.

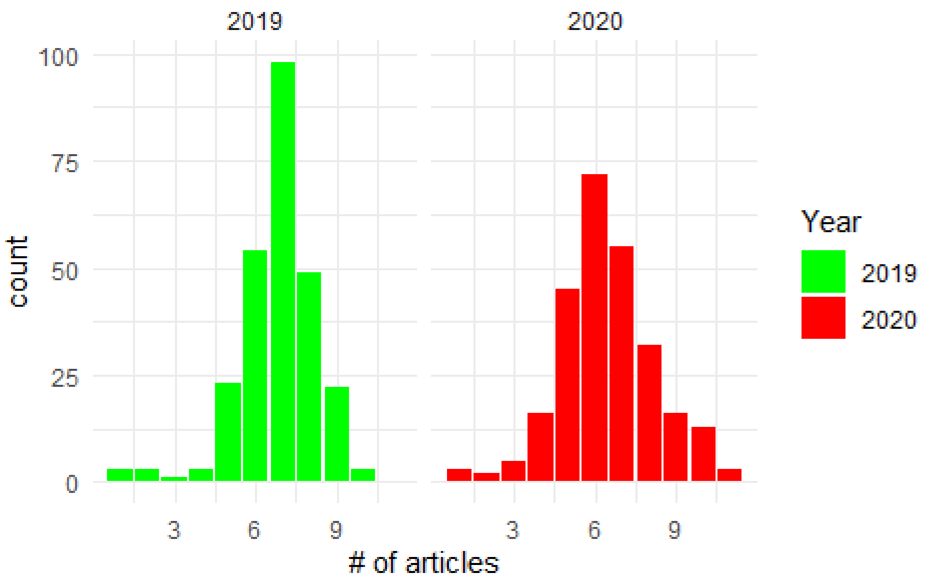

Figure 1. Distribution of articles written each day

It is important to note that the distribution of the number of daily published articles in the two years is very similar. However, we can see a slight tendency to publish more articles per day in 2020 than in 2019. This is expected as the Covid-19 caused a major disruption in financial and commodities markets.

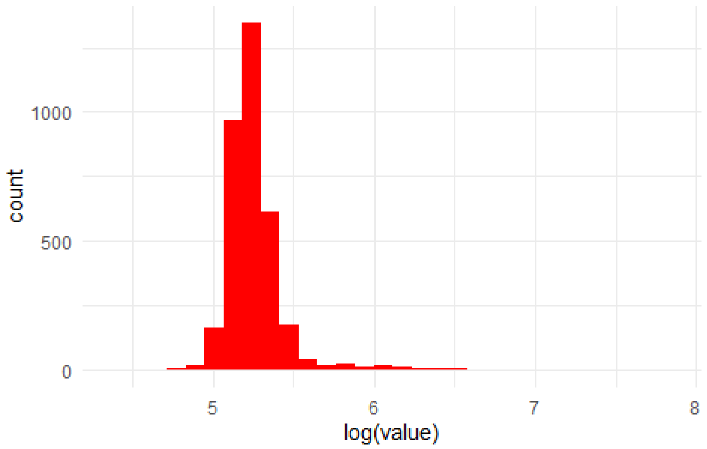

Figure 2. Distribution of tokens by article

Figure 2 shows the number of tokens per article in the log scale. The distribution at the raw count levels is highly skewed and even with the logarithm some skewness remains. The article with the highest count of tokens has 2373 while the lowest one has 81.

Table 1. Keyness Chi2 Statistic by Year

2020		2019	
feature	chi2	feature	chi2
pandem	671.86	est	-79.13
coronavirus	564.43	mt	-79.60
demand	419.33	opposit	-82.22
price	354.85	gibraltar	-88.84
covid-19	351.67	al-falih	-92.42
lockdown	309.92	anadarko	-93.20
crash	273.26	lng	-93.55
cut	216.04	cuba	-97.08
gold	203.12	attack	-97.89
outbreak	186.72	guaido	-112.38
recoveri	145.65	pdvsa	-117.91
blockad	131.67	platt	-122.89
loss	113.69	venezuela	-129.31
hydrogen	100.28	us	-135.35
low	99.19	sanction	-156.04
collaps	97.55	ipo	-159.74
opec	97.10	aramco	-232.95
dividend	92.83	waiver	-273.23
job	92.30	iran	-303.48
oilfield	89.31	iranian	-355.50

Table 1 shows the top 20 features ordered according to the chi squared keyness statistic when comparing both years. As expected most of the words that are specific to 2020 are related to Covid-19. A notable exception is the token 'gold' probably alluding to the flight to quality phenomenon observed in financial markets. Some negative adjectives such as crash, cut, loss and collapse are also characteristic of articles written during 2020. Specific vocabulary to 2019 alludes to special situations that occurred related with Iran and Venezuela.

Finally, I present some wordclouds to compare the token usage between 2019 and 2020. Overall the most used tokens remain relatively stable across both years.

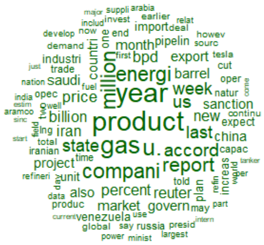

Figure 3. Keywords 2019

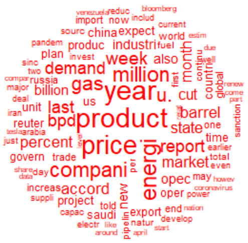

Figure 4. Keywords 2020

4 Results

In this section I explore the results obtained by the use of the semi-supervised dictionary induction methods proposed by An, Kwak, and Ahn (2018) and Hamilton *et al.* (2016). In order to have a strong comparison baseline I first replicated the analysis made in Loughran, McDonald, and Pragidis (2019). There is no prior reference of applying these methods in this field but is probably due to the novelty of both the sentiment analysis in oil-markets and the methods used.

The Loughran, McDonald, and Pragidis (2019) baseline is obtained by using the dictionary of ’positive’ and ’negative’ terms defined in the article along with searching for what they call ’positive’ and ’negative’ modifiers in the context of some keyword terms as defined in their document. I implemented two different versions. The first one uses the original proposal in their paper while the second one simply looks directly for positive and negative modifiers and words regardless of their context. This idea of using the context to determine the valence of a certain token is precisely what is behind semi-supervised induction methods hence this could be a perfect scenario to test their usefulness.

The first step to use both semi-supervised induction methods is to generate word embeddings for the tokens. For this step I used the Word2Vec algorithm to generate static contextual embeddings based on the corpus.

The second step after generating the embeddings is to define a set of ’positive’ and ’negative’ seed words from where the induction procedure begins. I created two set of seed words for each emotion and appended a suffix ’Mc’ to differentiate between them. The first set of ’positive’ and ’negative’ seeds is based on the financial set proposed in Jurafsky and Martin (2020) (chapter 20) while the second one mirrors the set of words proposed by Loughran, McDonald, and Pragidis (2019). In both cases regular expressions are used to match all tokens related within the corpus.

The Semantic Axis method is based on finding the mean vector of each seed set and defining the difference between this two vectors as the axis. Afterwards, the cosine similarity between each token vector and the axis is used to measure the valence of each word. The SentiProp algorithm also uses the cosine similarity to score the valence of tokens but works in a slightly different way. Starting from the seed set, a fixed number of neighbors per token is defined based on this similarity and a weighted graph is built using as edge weights the calculated scores. After this step a random walk algorithm is implemented starting with equal probability from each of the seed words and transitioning to different tokens according to the weight of the edges. The number of times that the random walk visits a certain token is used to define the valence.

Table 2 shows the top 10 positive words ordered by valence obtained by each of the semi-supervised methods. We note that starting from the seed words proposed in Loughran, McDonald, and Pragidis (2019) leads to slightly different results that starting from the financial seed set. Differences also arise due to the selection of the inductive method. Analogously, table 3 shows the top 10 negative words obtained by each method. In this case results from the two seed sets are more similar. We note that most of the word coined as ’negative’ are related to violence. This includes explosions, fires, attacks, killed, airstrikes among many others. This appears to be related to the political context in middle east where a great part of the oil market trades.

Table 2. Top 10 Positive Words by Valence

Sentiprop	SentipropMc	SemanticAxis	SemanticAxisMc
improvements	discovery	profitably	reserves
gained	discoveries	accelerate	extra
profitably	surplus	portfolio	grade
profits	repair	enhancing	minimum
successfully	repairing	society	sized
excellent	overproduction	executing	discovery
beneficial	discovered	reshape	medium
profiting	overproducers	financially	resource
positively	oversupplied	fortescue	discoveries
successful	repairs	progressing	feedstock

Table 3. Top 10 Negative Words by Valence

Sentiprop	SentipropMc	SemanticAxis	SemanticAxisMc
blocked	strike	airstrikes	health
los	explosions	endangered	diego
failed	saboteurs	floods	rebellion
damaged	fire	killing	alameda
damage	explosive	attacks	condemned
failing	explosion	isis	protest
blockaded	sabotage	gibraltar	halting
blockades	strikes	killed	occupation
shutdown	attacks	militia	arguments
blockade	attack	seized	blocking

An interesting comparison of the algorithms can be made by examining the pairplot as shown in 5. We note that the distribution of token valence obtained through the use of the Sentiprop algorithm is much more skewed than the one obtained through the SemanticAxis method. This means that the former makes a sharper separation between the valence of the tokens in the corpus. However, it is not clear whether this is beneficial as less words get scored and hence the sentiment index derived may be relatively flat. Scatter plots between the scores show a relatively linear relationship between the methods with some presence of high-valence clusters.

After obtaining the induced dictionaries I built a sentiment index at the daily level by weighting each document according to the valence of the tokens included. For that purpose a document feature matrix is created and weighted according to the relative frequency of tokens. Then, for each document the valence is multiplied by the relative frequency of the tokens and results are aggregated. Figure 6 shows the sentiment series obtained using a Henderson smoother (Dagum and Bianconcini 2016), series are smoothed because data at the daily level is highly volatile and no apparent trend is discernible.

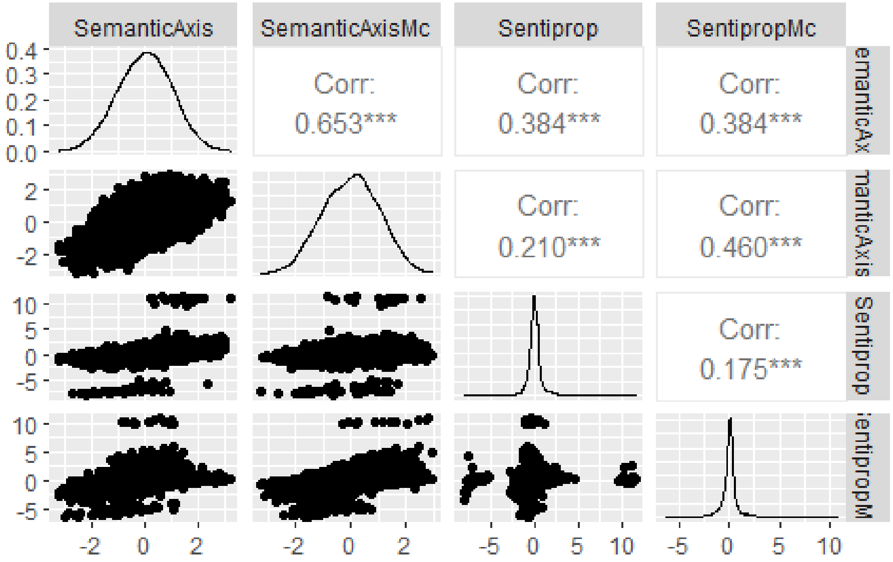

Figure 5. Distribution and comparison of induced dictionaries

Figure 6 shows the daily sentiment series obtained by the use of the Sentiprop and the SemanticAxis algorithms along with the one derived based on the (Loughran, McDonald, and Pragidis 2019) method. We see that all series derived from semi-supervised methods show a sharp fall during the first months of 2020 followed by a weak recovery until april just to abruptly fall during may. This pattern is coherent with the reaction that the oil market had in response to the Covid-19 pandemic. We note a fall in valence during the months that the Covid-19 outbreak originated followed by a slight recovery and a subsequent sharper decrease during the last part of april/may period in which oil forward prices reached negative numbers. From the 5th month onwards we note some swinging of sentiment probably associated with intermittent lockdowns around the world. Overall the series based on the Loughran, McDonald, and Pragidis (2019) measures look quite different with an evident fall during the first semester of 2020 just to start showing some signs of recovery in the second semester.

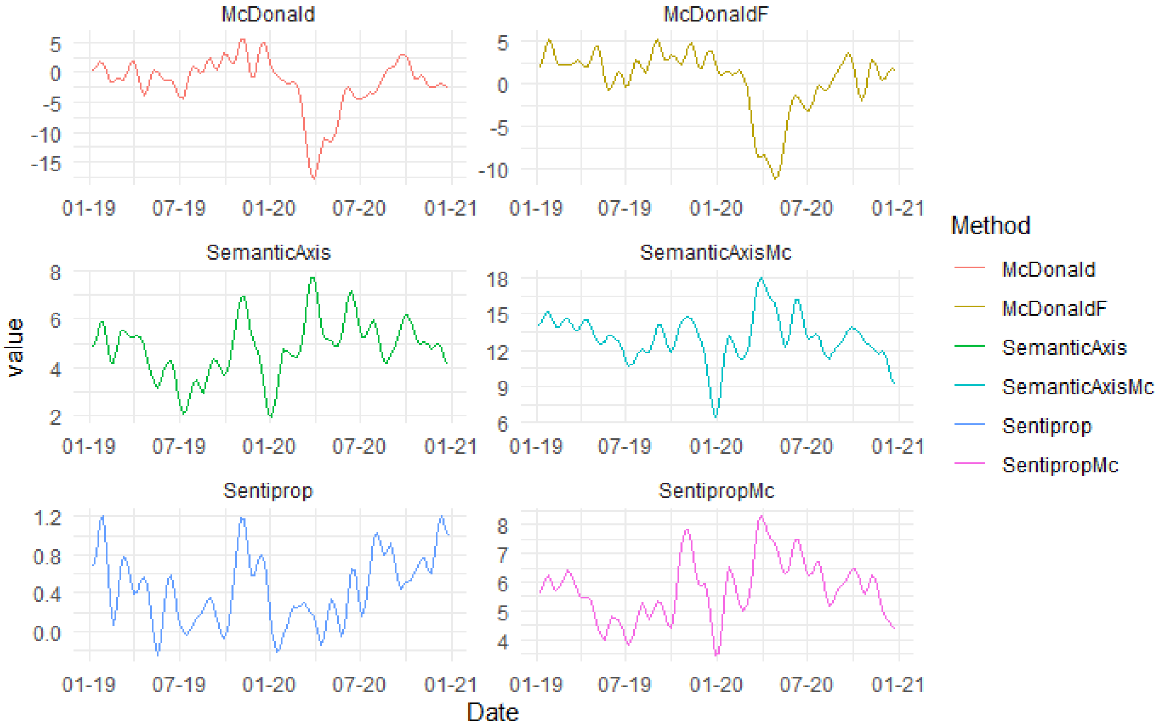

Figure 6. Smoothed Weekly Sentiment Series

Finally, in order to check whether the resulting sentiment series are useful to improve the prediction of oil prices I implemented a model following Loughran, McDonald, and Pragidis 2019 specification. The target variable is the difference in the logarithm of the price of the NYMEX Cushing, OK Crude Oil Future Contract 1, in U.S. dollars per barrel possible controls are the spot price of gold, 10-year treasury constant maturity rate, the Chicago Board Options Exchange Market Volatility Index (VIX) and the number of articles written each day in the corpus. The NYMEX Cushing prices are prewithened[1]. All other variables are differenced to account for unit-root behavior.

For each of the sentiment series I used a stepwise procedure to select the best possible model including it and all the controls mentioned above. All variables are included up to the third lag to give the model more flexibility. In none of the cases were the sentiment measures selected in the stepwise procedure.

Results are in sharp contrast with those obtained by Loughran, McDonald, and Pragidis 2019. Even though in most of the regressions implemented the coefficient of the sentiment index is statistically significant the fact that is never chosen by the stepwise procedure suggests that it doesn't have additional predictive power nor does it cause movements on the price in the granger sense.

Several explanations can be given as to why results obtained are different. The first one is that the Covid-19 pandemic induced such volatility in the market that historical links between variables were shadowed by the huge increase in variance. Hence, traditional non-robust statistical methods have difficulties in capturing the hidden patterns. The second one is that the inclusion of a pre-whitening procedure before the modelling alters significantly the results obtained. However,

[1] Using an ARIMA$(2, 0, 2)$ model and a dummy variable to control the negative prices observed in april of 2020

However, carrying this procedure is necessary as what we are truly interested in is knowing whether sentiment analysis has additional explanatory power to predict changes in prices. Hence, analysis should be done conditional on controlling for the series autocorrelation.

5 Conclusion

In this document a daily sentiment series for oil-related news articles is generated based on the use of two different semi-supervised induction methods (An, Kwak, and Ahn 2018, Hamilton *et al.* 2016). The 3460 used articles are scrapped from https://oilprice.com/Latest-Energy-News/World-News, a major online news provider, and cover the period between 2019 and 2020.

Some comparisons between the induced valences are made and their ability to improve out-of-sample forecasts for forward oil contracts is evaluated. Obtained results are compared with a baseline given by the method of Loughran, McDonald, and Pragidis 2019.

The analysis shows that while induced valences are coherent, no improvement on out-of-sample forecasting performance is achieved once the target series is appropriately pre-whitened. A stepwise selection procedure discards all sentiment related predictors. The main reason behind the lack of improvement in forecast accuracy could be increased market volatility due to Covid-19. The turmoil generated by the global pandemic may be shadowing the forecasting gains that previous studies in the field had found.

References

An, J., H. Kwak, and Y.-Y. Ahn. 2018. “SemAxis: A lightweight framework to characterize domain-specific word semantics beyond sentiment.” *arXiv preprint arXiv:1806.05521.*

Calomiris, C. W., N. Çakır Melek, and H. Mamaysky. 2020. “Mining for Oil Forecasts.” *Federal Reserve Bank of Kansas City Working Paper,* nos. 20-20.

Dagum, E. B., and S. Bianconcini. 2016. *Seasonal adjustment methods and real time trend-cycle estimation.* Springer.

Datta, D. D., and D. A. Dias. 2019. “Oil Shocks: A Textual Analysis Approach.” *manuscript, Federal Reserve Board.*

Habib, M. M., S. Bützer, and L. Stracca. 2016. “Global exchange rate configurations: do oil shocks matter?” *IMF Economic Review* 64 (3): 443–470.

Hamilton, W. L., K. Clark, J. Leskovec, and D. Jurafsky. 2016. “Inducing domain-specific sentiment lexicons from unlabeled corpora.” In *Proceedings of the conference on empirical methods in natural language processing. conference on empirical methods in natural language processing,* 2016:595. NIH Public Access.

Jurafsky, D., and J. H. Martin. 2020. *Speech and Language Processing.*

Kearney, C., and S. Liu. 2014. “Textual sentiment in finance: A survey of methods and models.” *International Review of Financial Analysis* 33:171–185.

Kilian, L. 2009. “Not all oil price shocks are alike: Disentangling demand and supply shocks in the crude oil market.” *American Economic Review* 99 (3): 1053–69.

Lardic, S., and V. Mignon. 2008. “Oil prices and economic activity: An asymmetric cointegration approach.” *Energy Economics* 30 (3): 847–855.

Loughran, T., B. McDonald, and I. Pragidis. 2019. “Assimilation of oil news into prices.” *International Review of Financial Analysis* 63:105–118.

Tetlock, P. C., M. Saar-Tsechansky, and S. Macskassy. 2008. “More than words: Quantifying language to measure firms’ fundamentals.” *The Journal of Finance* 63 (3): 1437–1467.