CESifo WORKING PAPERS

7760
2019

July 2019

British Stock Market, BREXIT and Media Sentiments – A Big Data Analysis

Gopal K. Basak, Pranab Kumar Das, Sugata Marjit, Debashis Mukherjee, Lei Yang

CESifo

Impressum:

CESifo Working Papers
ISSN 2364-1428 (electronic version)
Publisher and distributor: Munich Society for the Promotion of Economic Research - CESifo GmbH
The international platform of Ludwigs-Maximilians University's Center for Economic Studies and the ifo Institute
Poschingerstr. 5, 81679 Munich, Germany
Telephone +49 (0)89 2180-2740, Telefax +49 (0)89 2180-17845, email office@cesifo.de
Editor: Clemens Fuest
www.cesifo-group.org/wp

An electronic version of the paper may be downloaded
· from the SSRN website: www.SSRN.com
· from the RePEc website: www.RePEc.org
· from the CESifo website: www.CESifo-group.org/wp

CESifo *Working Paper No. 7760*
Category 14: Economics of Digitization

British Stock Market, BREXIT and Media Sentiments – A Big Data Analysis

Abstract

In this paper we show, using a Machine Learning Framework and utilising a substantial corpus of media articles on Brexit, confirmed evidence of co-integration and causality between the ensuing media sentiments and British currency. The novel contribution of this paper is that along with sentiment analysis using commonly used lexicons, we devised a method using Bayesian learning to create a more context aware and more informative lexicon for Brexit. Moreover, leveraging and extending this we can unearth hidden relationship between originating media sentiments and related economic and financial variables. Our method is a distinct improvement over the existing ones and can predict out of sample outcomes better than conventional ones.

Keywords: digitization, machine learning.

Gopal K. Basak
Stat-Math Unit, Statistical Institute
Kolkata / India
gkb@isical.ac.in

Pranab Kumar Das
Centre for Studies in Social Sciences
Calcutta / India
pkdas@cssscal.org

Sugata Marjit
Centre for Studies in Social Sciences
Calcutta / India
marjit@gmail.com

Debashis Mukherjee
Centre for Training and Research in Public Finance and Policy, Centre for Studies in Social Sciences, Calcutta / India
debum_2k1@yahoo.com

Lei Yang
School of Accounting and Finance
Hong Kong Polytechnic University
Hong Kong
aflei.yang@polyu.edu.hk

1. Introduction and Literature Review

This paper attempts to find the impact of media sentiment on British Stock Market and Currency i.e. on FTSE index, British Pound Exchange Rate with reference to BREXIT focussing on the time period between pre-announcement and pre-referendum to after-referendum in terms of a wide array of news papers. In the process we analyse the long run response of capital and foreign exchange markets to media sentiments when such events take place.

We show, in terms of a framework that leverage an extensive Machine Learning Procedure and utilising substantial print-media data on media sentiments, confirmed evidence of cointegration and causality between Media Sentiments and Currency Exchange Rate. The novel contribution of this paper is that along with conventional tool of extracting media sentiments, using commonly used lexicon, we devise a method which relies on Bayesian Learning. This value addition enables us to create a more context aware and informative lexicon. Moreover, it can be leveraged and extended to unearth relationship between media sentiments and economic and financial variables.

While media sentiments properly quantified through several natural language processing programs may affect stock returns or relative price of foregin exchange, one purpose has been to tag such sentiments with the historic phenomenon of BREXIT. Events such as these have a natural reason to affect stock market as well as forex markets. But anticipated impact of such costs on stock returns (and forex market return) and its intensity must depend on how the print media can affect the perception of investors, small or large. By assessing a large variety of newspapers both from UK and USA (primarily to capture world reactions) and relevant articles there, we have built up a substantial database and applied it to our analysis.

Though it is interesting as such to find out the relationship between stock prices, returns and media sentiments, the novel contribution of this paper is to study such impact with reference to a major event of history. We not only characterise the short term relationship but also show that long term impact of media sentiments cannot be ruled out either.

The literature of media sentiment and financial market behaviour is quite old, dates back to Klein and Prestbo (1974). It developed into a systematic body of research with the advent of computational power of big data and emergence of the method of combining numbers and textual material. Marcus et al (1993) linked the scope of research methodology with statistical tools. Significant research in more recent times in the field includes among others, Das and Chen (2007), Tetlock (2007, Tetlock et al (2008) and Li (2008) etc. A good survey of the literature is covered in Li (2010) and Loughran and McDonald (2016). In a very recent study Fraiberger et al. (2018) extends similar analysis in a cross-country framework with both advanced and emerging market economies in the set. The study also extends the scope of the analysis from asset market to international capital flows.

The foundation of the literature of media sentiment and financial market behavior is grounded in the standard theory of asset valuation. The traditional theory of asset valuation is determined by the fundamentals of the asset and availability of new and relevant information on the part of the fundamentals is reflected in the change in the asset price. A comparison of the business investment in UK for the period 2015 to 2017 with preceding years and with other developed countries show that business investment has substantially decreased in UK. There are many potential factors, of which BREXIT led uncertainty resulting into lower investment, has been given a prominent role by Górnicka (2018). The spirit of the argument in economic theory is rooted in the idea of using the notion of real options because investment in physical capital is irreversible (Bloom et al. 2007; Dixit and Pyndick 1994). Based on the findings of other studies, such as Berden et al (2009) and Crowley et al (2018), it has been argued by Górnicka (2018) that in the 'no deal' scenario if UK exits BREXIT then there will be increased cost of trading. This implies a lowering of return for firm investment, and hence until the resolution of uncertainty firms prefer to refrain from making investment. It is expected that these issues will receive attention in the media reflecting people's sentiments about business prospects and hence on the stock market. The sentiments will, however be of two types - sentiments of the sophisticated investors who form their expectations using some refined method, like Bayesian beliefs and the unsophisticated investors, often called noise traders, who depend on hearsay kind of information and their behaviour can create increased volatility (Campbell et al., 1993; De Long et al., 1990).

The present study is related to Johnman et al (2018) in the sense that it aims at an aggregative analysis of the behaviour of the UK stock market, captured by FTSE 100 index

and its relation with over all market sentiment. Tetlock (2007) and Ferguson (2015) also engaged in this type of analysis. However, the present paper differs from the existing literature at least in four important respects. First, it involves the analysis of the aggregative behaviour of stock and currency market in the context of a particular event, viz. Brexit which is expected to change the structure of the UK economy in the global perspective, thus extends the scope of the analysis of a national economy to a global perspective. Secondly, while the majority of the studies, viz. Tetlock (2009), Ferguson et al (2015) or Johnman et al (2018), focussed on the behaviour of unsophisticated investors - so called noise traders, and how their actions generate fluctuations in the market and temporary deviations from fundamental values while we capture the impact of the aggregate actions of both the sophisticated and unsophisticated investors. Thirdly, the present study addresses the research question in the framework of 'long run' in the sense of time series statistics. As a corollary of this approach we can very easily test for of causality between media sentiment and stock/currency market behaviour which is not possible to obtain when one conducts the statistical analysis, as is the standard practice in the existing literature including Johnman et al (2018), in a simple OLS framework. Finally, we have also built up our own dictionary based on a modified learning mechanism closely resembling Bayesian learning in statistics that is then employed to inquire the nature of relationship between media sentiments and behaviour of British Pound Exchange Rate / FTSE-100. The results of our additional exercise show that long run relationship still exists as in the earlier case. In addition, inclusion of sentiment scores tends to improve forecasting of the British Pound exchange rate. With this introduction the paper proceeds as follows: Section 2 provides a detailed discussion on the data sources used for the statistical analysis of this paper, Section 3 provides the computation of sentiment scores using the conventional methodology of Loughran and McDonald (2011) dictionary and the statistical analysis, Section 4 considers a new dictionary - our own - built using a modified Bayesian learning and additional statistical results using this dictionary. Section 5 describes the Forecasting exercise; Section 6 describes few additional results and insights including distinctive change with different types of news media. Finally, Section 7 concludes the paper.

2. Data Sources

The most important variable for the present study, viz. the sentiment comprises the set of Media Sentiments on BREXIT during the last couple of years starting from the Referendum month of June 2016 and compared Stock market behavior during the same period and analyzed their relationship and investigated existence of any long run relationship. In addition to stock market we also considered foreign exchange market for this study as the latter is generally found to be more sensitive to information flow. This gives us the opportunity to compare sensitivity of stock market vis-à-vis foreign exchange market to media sentiment. The basic theoretical argument as well as methodology remains same for both the cases.

The existence of the long run relationship and the associated short run dynamics, representing the adjustment mechanism when there is movement away from the long run path due to some shock is analysed in the time series framework. Our variables of interest are GBP Exchange Rates with respect to USD and EUR in Currencies and FTSE-100 Index value, representing the stock price behavior and a measure of overall media sentiment. The measure of media sentiments is created from a pool of daily news articles from a group of news papers.

Corpus for this exercise is created using 9108 news articles involving Brexit, sourced from leading news media, e.g. The Guardian, The Financial Times, The Wall Street Journal, The Daily Mail, The Daily Mirror, The New York Times, The Washington Post, The Bloomberg and The Economist. In all the cases, articles involving topic “Brexit” were only selected, covering the time period June 16, 2016 till November 20, 2018. For The Guardian, we used Guardian Media Groups online API (Application Programming Interface). News articles were extracted from this API using a Developer Key. For all the other News Media, their respective online portal was used for article extraction. The articles were crawled (extraction of relevant article links) and then scraped (extraction of the individual Article Contents, in full). As shown below in Fig. 1, more than 80% of this Corpus are sourced from UK based News Media, to give due emphasis to the place of the event. Remaining articles are procured from other parts, including around 15% from US based Media Houses. Top five contributors to this Corpus are The Guardian (35%), The Daily Mail(16%), The Financial Times(14%), The Daily Mirror(14%) and The Wall Street Journal(10%). For the outside UK

set we have The New York Times (6%), The Washington Post (3%) The Economist and The Bloomberg (combined contribution ~ 1%).

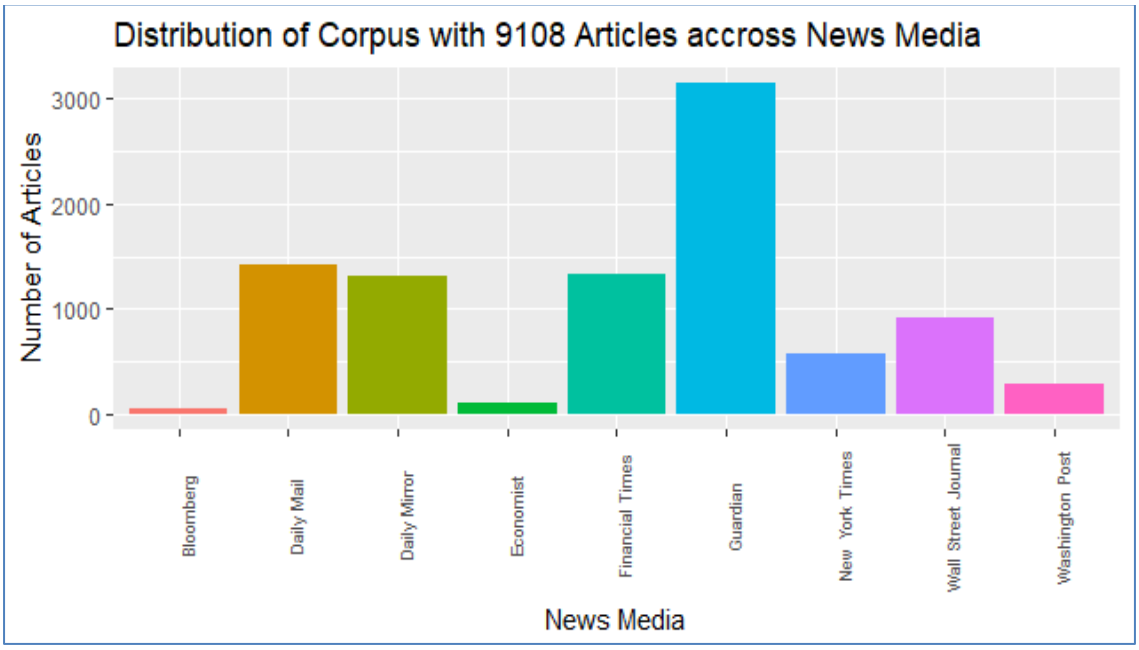


Fig. 1: Corpus Distribution of "Brexit" Articles

We selected the UK Media houses based on their circulations, selected ones are in the top 5 list in terms of their circulations. Corpus also had a mix of News Media in terms of "Pro Leave" like The Daily Mail and "Pro Remain" like The Daily Mirror and The Guardian. There is also coverage of different readership leaning, from "Centre Left" like The Guardian and The Daily Mirror to "pro Right" like The Daily Mail. The set also features news media that covers uninformed investors (so called noise traders), such as The Daily Mail, The Daily Mirror and informed investors, such as The Guardian, The Financial Times, The Wall Street Journal. Fig. 2, below depicts distribution of articles across different sources for the period of analysis. All quarters and almost all days within the set, are found to have coverage from multiple news media sources. As a result, Daily Sentiments will originate from a combination of news media sources, and will not be dominated overwhelmingly by any single news media.

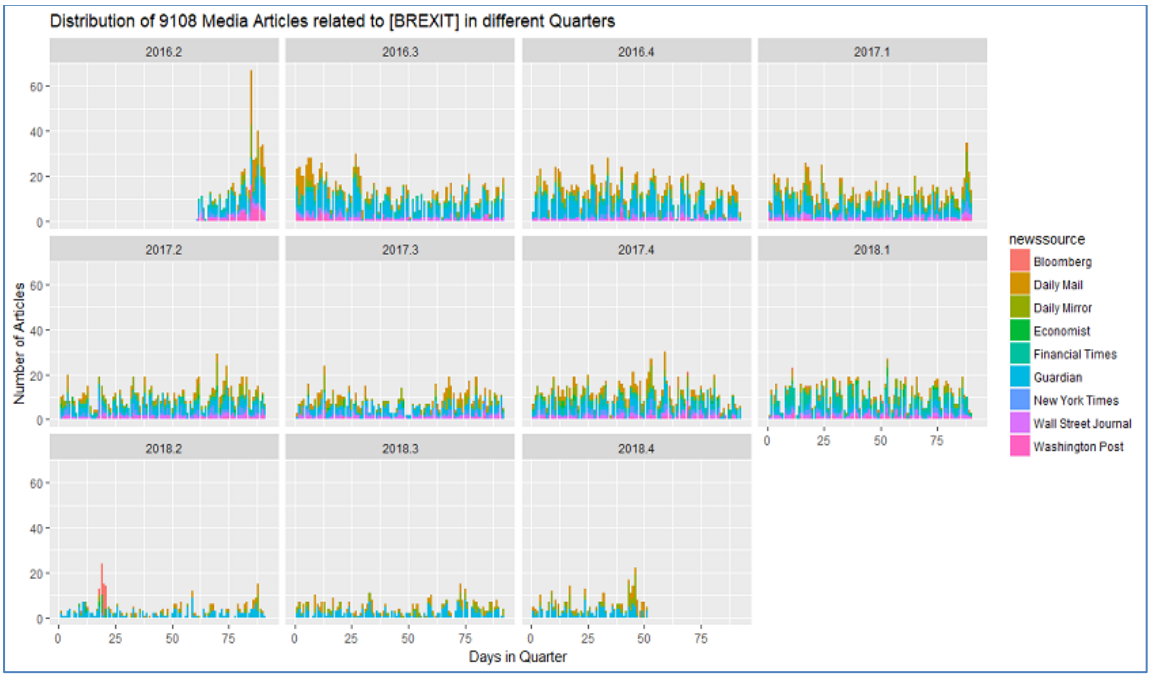


Fig. 2. News Articles by sources across quarters: June 2016 to November, 2018

3. Sentiment score using Loughran and McDonald dictionary and statistical analysis

In the first part we performed Sentiment Analysis using word based lexicon, LOUGHRAN, created by Loughran and McDonald (2011). The first part of this section is devoted to the computation of sentiment score and the second part statistical analysis of the role of sentiment scores in explaining stock price index and currency price.

3.1 Computation of sentiment score using conventional method

To align with Loughran and McDonald lexicon, we have taken words as our tokens. From this set of initial tokens, the information is cleansed by removal of tokens containing English Stopwords (I, we, is, was, the, etc.) and also some other irrelevant tokens coming while extraction from online news media portals. This helps to bring down “noise” in the data. The resulting cleansed set of words is taken as features for the Corpus. In the next step – feature represntation - each feature is assigned a numerical weight, which indicates the number of occurrences of that Word in that article. The Corpus of 9108 articles is accordingly represented, in terms of set of words and their occurrences. Then the corpus of articles is aggregated according to published date of the articles. As our target of analysis are Currency Exchange Rates, Stock Market Index we mapped the article dates against the respective dates when Currency or Stock Exchange is open, mainly we aligned the weekends appropriately. After this aggregation and alignment step, 9108 articles got distributed into 630 days.The next step is the computation of the daily sentiment score whis is achieved by transforming the Articles into a “bag of words” model and depicted as a Document Feature Matrix. Available Days become the Rows and “Cleansed” Word are the Columns and frequency of occurrence of that Word in that Day becomes its cell value. As already mentioned we used, Loughran, a word based lexicon, created by Loughran and McDonald (2011) as the reference dictionary to perform Sentiment Classification. It has been employed in a number of studies in the existing literature for it is better aligned to finance and economics domain and is observed to perform better. The LOUGHRAN lexicon has 4000 plus words, distributed across six Sentiments, *Constraining, Litigious, Superfluous, Uncertainty* along with *Negativeand Positive*. We

focussed our attention on the 2700+ words classified in this lexicon as either, Negative or Positive.

Fig. 3 below depicts the Top 50 words, of both Sentiment (Negative/Positive) using LOUGHRAN lexicon on our Brexit Corpus. We see three the words, “warned”, “cut”, “crisis” appearing at higher frequencies than the others, and all of them are “Negative” words as per LOUGHRAN. This will also indicate Daily Sentiments to be manifested as Negative.

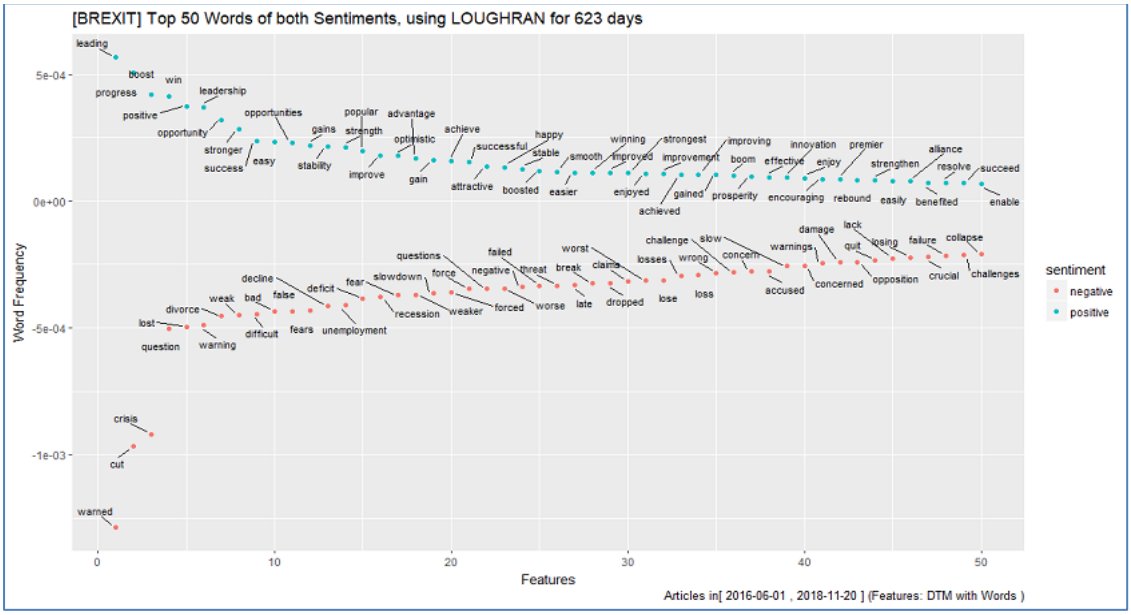


Fig. 3. Distribution of top 50 positive and negative words according to LOUGHRAN

Using Loughran, we calculated for every day, frequency of total positive words and total negative words. Sentiment Score, for any day, is then the difference of positive and negative frequencies.

The other variables for this study are price of foreign exchange and stock prie index. For the former we consider Pound Sterling (GBP) relative to US currency (USD), GBPUSD and relative to EURO, GBPEUR. For stock price index, we have considered FTSE 100 Index, as it is considered a representative index for the London Stock Exchange. We have extracted the Currency Exchange rate and FTSE 100 Price at Close for each day of trading for our period of analysis – June 2016 to November 2018 and also the daily returns for the same period. We used Yahoo Finance to source these data.

Figures below show the Currency Exchange Rates and Return during the analysis period.

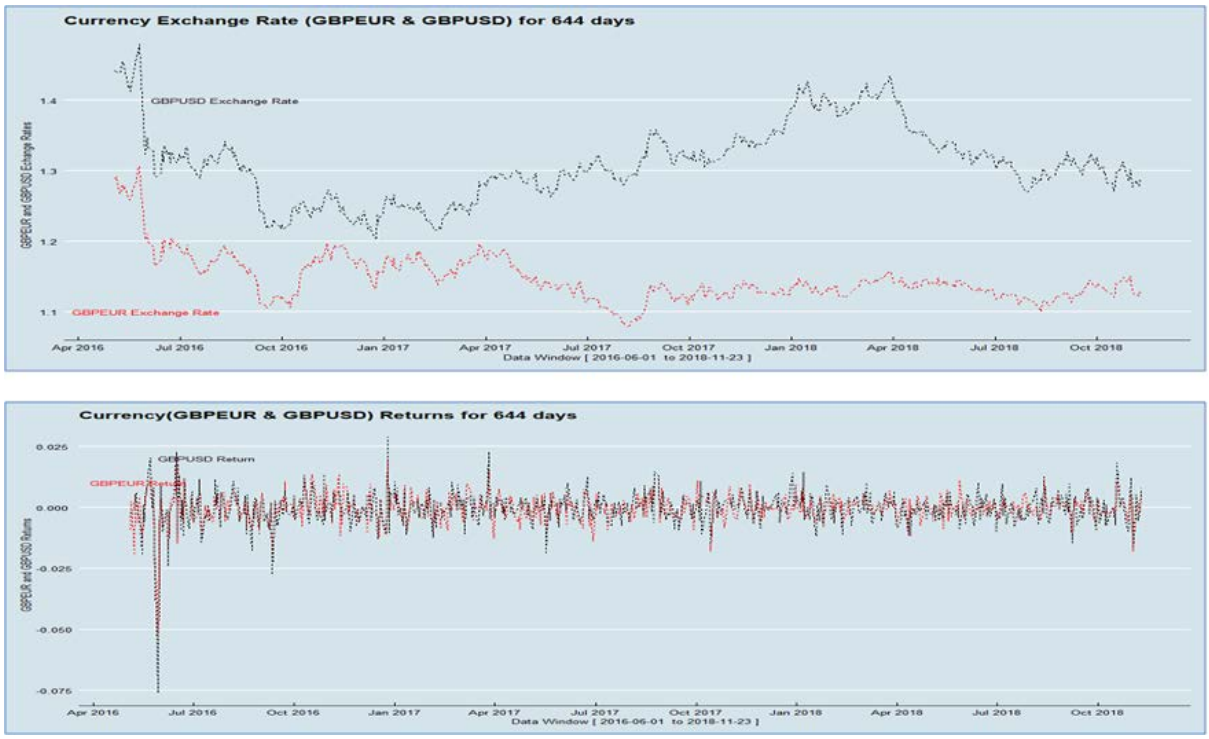


Fig. 4. Exchnage rate of GBP-USD and GBP-EUR and the return

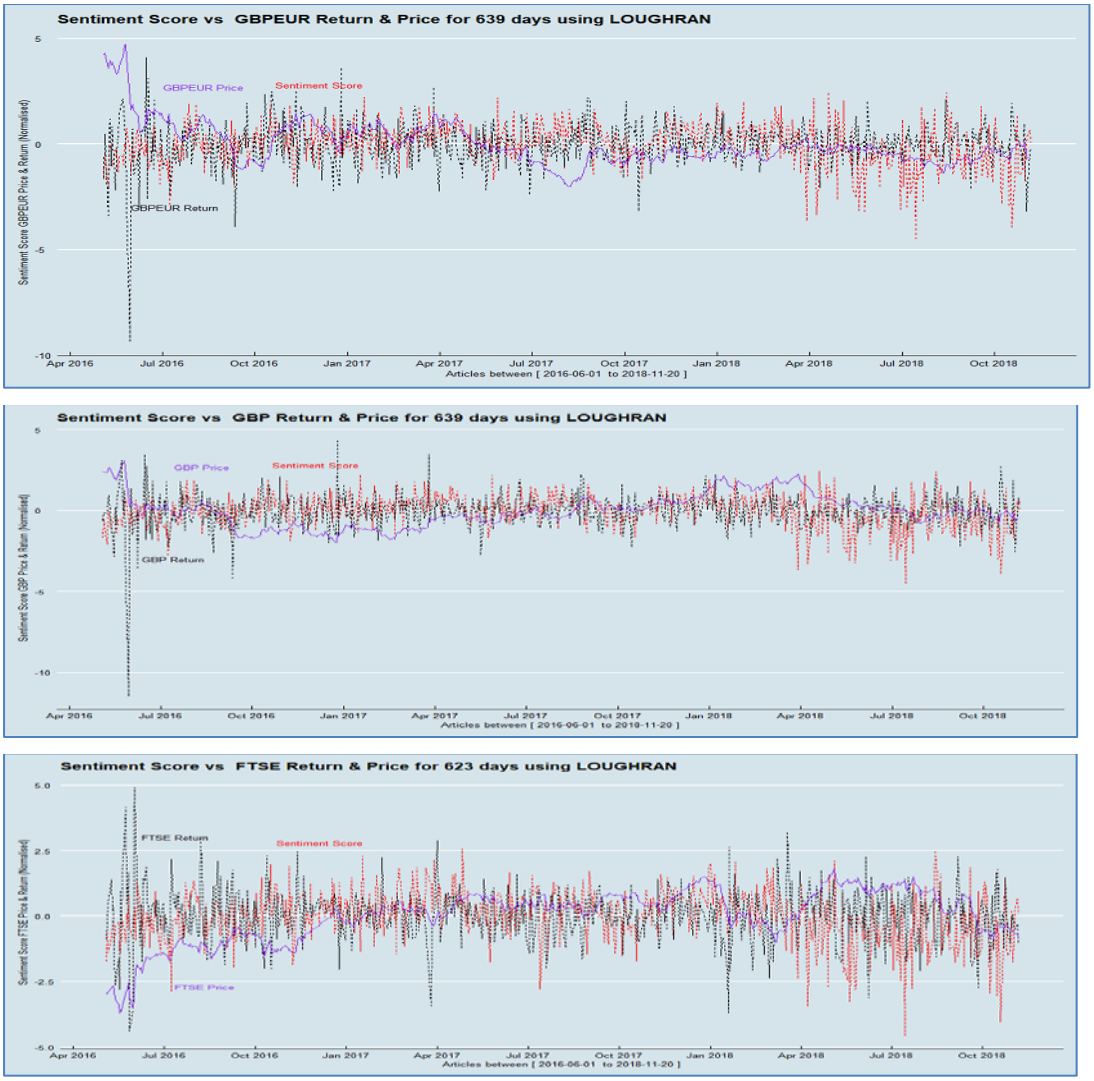


Fig. 5. Sentiment score, FTSE-Index and Currency price and return over time

3.2. Long run between sentiment score based on LOUGHRAN and price series

As we stated in the begining of this paper that the objective of the present study is to inquire the nature of long run relationship between price series and media sentiment where long run is defined in the sense of time series econometrics We consider two price series, viz. currency price of pound sterling relative to some foreign currency (specifically US Dollar and Euro) for the foreign exchange market and FTSE-100 Index for the stock market. Table 1 provides the descriptive statistics of the variables considered for the analysis in this section. It may be noted that long run in the sense of time series econometrics/ statistics hold for very

long period of data while our period of analysis is only slightly more than two years. However, we are using high frequency daily data series and the financial market price series as well as media sentiments adjust very fast, hence the objection regarding long period of data series is not really relevant here. The relevance of long period is important in the context of adjustment when the variables deviate from long run equilibrium which is very long for real variables, such as GDP, investment etc.

Table 1. Descriptive Statistics – LOUGHRAN Mehodology

Variable	**Obsn.**	**Mean**	**SD**	**Min**	**Max**	Skwewness	Kurtosis
Sentiment score	639	-8.68e-10	1	-4.475158	2.453371	-.7094136	4.389034
GBPEUR	639	1.147045	0.0336065	1.079767	1.305534	1.630569	7.652422
GBPUSD	639	1.310612	0.054039	1.203935	1.47894	.4966115	3.012286
FTSE100	623	7241.442	355.4925	5923.5	7877.5	-1.038875	4.233282

As the first step for time series analysis we next test for non-stationarity of the series and the results are provided in Table 2 below. In order to be sure about the non-stationary nature of the variables we performed the usual tests, viz. Augmented Dickey Fuller(ADF), Phillips and Perron(PP), Kwiatkowski–Phillips–Schmidt–Shin (KPSS) and Dicky-Fuller generalized least square (DF-GLS). It is observed from Table 2 that there are differences in respect of test results for foreign currency price series (GBPEUR,GBPUSD) for different test statistics. As DF-GLS is considered to be more robust and the result implied by the test statistic is consistent with KPSS test statistic we conclude for the presence of unit root for the currency prices. FTSE-100 Index shows presence of unit root for all test results. GBPEUR Returns as well as FTSE100 Returns are found to be stationary in ADF and PP and were confirmed by KPSS and DF-GLS. For GBPUSD Return, KPSS test result indicates presence of Unit Root at 10% significance, with DF-GLS test confirms this time series to be level stationary. Sentiment Score (LOUGHRAN) indicates the underline time series to be I(0) as per ADF and PP but in KPSS it indicates presence of Unit Root at 1% significance. However, as DF-GLS test statistics shows it to be I(0) at least 5% significance.

Table 2. Test of stationarity for LOUGHRAN method

Variable	LEVEL			
	ADF	**PP**	**KPSS**	**DF-GLS**
Sentiment(LOUGHRAN)	-13.0273(***)	-21.3481(***)	0.851(***)	-2.9263(**)
GBPEUR Return	-17.0387(***)	-24.4343(***)	0.0625	-4.0324(***)
GBPUSD Return	-17.8329(***)	-24.5875(***)	0.1377(*)	-7.926(***)
FTSE100 Return	-17.6025(***)	-22.6931(***)	0.0323	-5.273(***)
GBPEUR Price	-4.6581(***)	-4.4451(***)	0.55(***)	-1.233
GBPUSD Price	-3.2604(*)	-3.1915(*)	0.8423(***)	-1.1364
FTSE100 Price	-2.4881	-2.2693	1.1137(***)	-0.7558
	First Difference			
GBPEUR Price	-17.748(***)	-24.1244(***)	0.0588	-7.4363(***)
GBPUSD Price	-17.6216(***)	-24.1139(***)	0.1383(*)	-9.9161(***)
FTSE100 Price	-17.6513(***)	23.2866(***)	0.0301	-8.19(***)
Note : *, **, * indicates significance at 10%, 5% and 1% respectively**				

Now we consider the main research interest of this paper, viz. presence of long run relationship between Media Sentiment and Currency/Stock market performance. Given the fact that either of the price series (GBPEUR/GBPUSD/FTSE-100) is I(1) and Media Sentiment computed with LOUGHRAN lexicon, is I(0) we cannot adopt Johansen test for Cointegration. Instead we follow Bounds test procedure in Auto Regressive Distributive Lag (ARDL) structure proposed by Pesaran, Shin and Smith (2001) using partial sums decomposition in order to ascertain positive and negative long run effects. The procedure is described in brief below. The ARDL equation in two variables is given by the following equation:

$$y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \ldots\ldots + \beta_k y_{t-k} + \alpha_0 x_t + \alpha_1 x_{t-1} + \alpha_2 x_{t-2} + \cdots + \alpha_q x_{t-q} + \epsilon_t \quad (1)$$

where y_t is Price index(GBPEUR/GBPUSD/FTSE-100) and x_t is a measure of Sentiment Score and ϵ_t is the disturbance term. As y_t is I(1) and x_t is I(0) Johansen or Engle-Granger testing procedure for Cointegration cannot be applied while the assumptions for Pesaran,

Shin and Smith(PSS) procedure for testing Cointegration are satisfied. First, the following unrestricted ECM (or conditional ECM in Pesaran, Shin, Smith terminology) is estimated

$$\Delta y_t = \beta_0 + \sum \beta_i \, \Delta y_{t-i} + \sum \gamma_j \, \Delta x_{t-j} + \theta_1 y_{t-1} + \theta_2 x_{t-1} + e_t (2)$$

The highest lag in Δy_{t-i} and Δx_{t-j} are determined by any of the information criterion. We adopted SBC. Since bound test result is dependent on the assumption that e_t is serially independent, we chose optimal lag using the LM test to determine serial independence of e_t . For the dynamic stability of the ARDL model it is checked whether the eigenvalues lie in the unit circle or not.

The bounds test for Cointegration is executed by testing for the null $H_0: \theta_1 = \theta_2 = 0$ against the alternative that H_0 is not true. The relevant test statistic follows F-distribution. A rejection of H_0implies that there is a long run relation between *y* and *x*. Pesaran et al (2001) provides critical values of upper and lower bounds for different number of variables of interest. In eachcase the lower bound is based on the assumption that all the variables are I(0) and the upper bound is based on the assumption that all the variables are I(1). If the computed F-value is less than the lower bound, then there is no Cointegration, if it exceeds the upper bound then there is Cointegration, and hence a long run relation. If the computed F-value lies in between the lower and upper bounds then the test is inconclusive. Additionally we also conduct bounds t-test : $H_0: \theta_1 = 0$ against the alternative that $H_1: \theta_1 \neq 0$.

Table 3. Tabulated values for ARDL Bounds F-test
(Model: Unrestricted intercept, no trend)

Significance level	Lower bound- I(0)	Upper bound - I(1)
10%	4.04	4.78
5%	4.94	5.73
1%	6.84	7.84

Source: Pesaran *et al* (2001)

Table 3.1 Tabulated values for ARDL Bounds T-test
(Model: Unrestricted intercept, no trend)

Significance level	Lower bound- I(0)	Upper bound - I(1)
10%	-2.57	-2.91
5%	-2.86	-3.22
1%	-3.43	-3.82

Source: Pesaran *et al* (2001)

If computed t-value exceeds I(1) bounds provided by Pesaran *et al* (2001) then it establishes the presence of long run relationship. If the computed t-value is less than the I(0) bounds then all the variables are stationary. In the event of non-rejection of Cointegration, we finally estimate the long run relationship:

$$y_t = \alpha_0 + \alpha_1 x_t + u_t (3)$$

and the usual ECM:

$$\Delta y_t = \beta_0 + \sum \beta_i \, \Delta y_{t-i} + \sum \gamma_j \, \Delta x_{t-j} + \phi z_{t-1} + e_t (4)$$

where $z_t = y_t - \hat{\alpha}_0 - \hat{\alpha}_1 x_t$. The long run coefficient is $\left(\frac{-\theta_2}{\theta_1}\right)$.

Table 4. Test of Cointegration – LOUGHRAN Sentiment

Variable set	**F-statistic**	**Remark**
GBPEUR& Sentiment Score	8.417***	Cointegration at 1%
GBPUSD& Sentiment Score	4.816*	Cointegration at 10%
FTSE-100& Sentiment Score	3.937	No cointegration

Note: ARDL test for cointrgartion is conducted *a là*Pesaran *et al* (2001)
*, **, ***: significance levels at 10%, 5%, 1% respectively

Table 4.1 Test of Cointegration – LOUGHRAN Sentiment

Variable set	**t-statistic**	**Remark**
GBPEUR& Sentiment Score	-4.004***	Cointegration at 1%
GBPUSD& Sentiment Score	-2.777	No cointegration
FTSE-100& Sentiment Score	-2.805	No cointegration

The test results are provided in Table 4 above.[6] Starting with an unrestricted ECM (Error Correction Model) the maximum lag structure is determined using BIC (Bayesian Information Criteria). Further it is found that the errors in the model are serially independent and the model is dynamically stable. As the test results show the null hypothesis of cointegration between GBPEUR and Sentiment Score at 1% as per Pesaran, Shin and Smith (2001) bounds test cannot be rejected. For GBPUSD and sentiment score there is conitegration at 10% level while there is no cointegration between FTSE100 and sentiment score. Presence or absence of cointegration asserts the presence or absence of long run relation between any of the price series on the one hand and sentiment score on the other.

[6]All the statistical estimation and tests in this paper were conducted in STATA, Vesrion 15 and R.

Table 5. Long run relation by ARDL – GBPEUR and Sentiment Score

Variable	Co-efficient	P-value	95% Conf. Interval
Score.LOUGHRAN.norm	0.0067118	0.419	[-.0096007, .0230243]

Finally we provide the results of Granger Causality between GBPEUR and Sentiment Score using Toda-Yamamoto procedure. The results are provided in Table 7 below.

Table 7. Granger Causality by Toda-Yamamoto test

Wald test for significance of constant and other variables			
Estimated Equation : SentimentScore = $c_1 + \sum \alpha_i GBPEUR_i$, for i=1..8		$H_0 : c_1 = 0, \alpha_i = 0,$ for i=1..8	
Test Statistic	**Value**	**Degree of Freedom**	**Probability**
χ^2Statistic	1.7	9	1.0
Estimated Equation : GBPEUR = $c_2 + \alpha_j SentimentScore_{j,}$ for j = 1..8		$H_0 : c_2 = 0$, j = 0, for j=1..8	
χ^2- Statistic	18.5	9	0.03

As the Wald test above shows, the null hypothesis of no Granger causality with Sentiment Score as target and GBPEUR as predictor, as probability is 1 can not be rejected. However we can clearly reject the Null Hypothesis, when Target is GBPEUR and Predictor is SentimentScore, at 5% significance. So there is existence of one way Granger Causality, with Sentiment Score Granger Causing GBPEUR Price at 5% significance.

We have shown that there is a long run relation between GBPEUR and Sentiment Score with the latter computed according to the conventional method of Loughran and McDonald (2011). In the next section, as stated earlier, we develop a new methodology for computing Sentiment Score and show its superiority.

4. Sentiment Analysis and Long run with Machine Learning

The previous analysis is based on the standard methodology of LOUGHRAN where same set of words are used to compute Sentiment Score. A serious drawback of this procedure is that role of specific words may not remain same in generating effect on either stock price index or price of foreign exchange. Hence we developed a new method based on self learning drawn from data science. This is then used to implement similar analysis undertaken in the previous section. We also perform comparative study of Forecasting performance of the model using the Sentiment Score computed using our methodology, in subsequent section.

The methodology proposed here differs mainly in two aspects. In the tokenisation of corpus phase, we have taken phrases instead of words. Phrases are created using bigrams. Similar cleansing is undertaken, by removal of phrases containing English stopwords(I, we, is, was, the etc.) and also some other irrelevant tokens coming while extraction from online news media portals, to remove "noise" in the data. As an additional step, we also used Stemming, process of removing suffixes from word to get the originating word (eg. *stopped*, *stopping* both taken as equivalent to the word, *stop*). This helps to reduce sparsity in the subsequent Document Feature Matrix. The resulting cleansed and stemmed set of "phrases" is taken as features for the corpus. After this we, in a similar manner as in LOUGRAN, assign numerical weights to each feature (in this case phrases). Aligned weekends, with respect to target variable (Currency Exchange Rates, Stock Market Index) to finally transform the corpus of 9108 articles into a Document Feature Matrix aligned to 640 days, covering the period, June, 2016 till November, 2018. This section is divided into two parts – first part delas with the computation of sentiment score using our proposed method and the second part with an analysis to show superior performance of the newly developed sentiment score.

4.1 Sentiment Score Machine Learning Pipeline

This is where the process changes as compared to LOUGHRAN. Here we used a Machine Learning Pipeline consisting of Topic Modelling, Random Forest and Support Vector Machine to arrive at the Daily Average Sentiment Scores. This process is executed following the below 4 steps at a high level.

1. ***Phrases, learned*** from the Corpus, using the ***Training*** *Samples*, are tagged with appropriate ***Polarity***(+/-) with a Phrase Score using LDA Topic Model.

2. Features (***Phrases***) are ***ranked***, wrt ***importance*** using a Random Forest Model
3. ***Top*** *N feature* subsets are taken from above, for various N, and ***Weights*** are computed using a SVM model, on the ***Training*** *Samples*.
4. ***Sentiment Scores*** are ***computed*** at ***day*** level, on the ***Out of Sample*** *data*, using the Trained Phrases, Polarity and Weights, using these Top N feature subsets, for various N and for different type of Training splits

Before we elaborate on these steps, we provide short description of the modelling techniques used.

4.2 Short Description of the Machine Learning Individual Components

Topic Modelling using Latent Dirichlet Allocation

LDA(Latent Dirichlet Allocation) is a probabilistic generative model in which documents are represented as mixtures over latent topics. From a Corpus of N documents, the generative process of LDA, treats any document being generated by set of words and phrases, belonging to the documents. First one has to determine How Many Topics are required from the document collection. Assuming there are k topics, through LDA

- Each Document can then be considered as being generated from the weighted combination of these k topics
- Each of these k Topics in turns are generated by set of Words, which originate from the document collection

As an example, if there are 100 documents which we want to segregate into 4 topics, say. Underlying idea could be that we want to classify these documents wrt to these Topics, say "Politics", "Business", "Economy and Finance" and "Social" in terms of their closeness to these Topics. Out of these documents, assume we have 50,000 different words then these set of words and their distribution will create the individual Topics. The word, "party" may found to be prominent for Topic "Politics" whereas "mergers" could be more prominent to "Business" topic.

With this alignment of Words and Topics, any document can be expressed as linear combination of Topics and if one Topic is found to be very dominant than the others, the document can be treated as more related to "Politics" than "Business".

One can also thought of it as somewhat similar to Clustering, an unsupervised Machine Learning method. It does have the favour of clustering however a major difference here is one word can be linked with ALL the Topics and NOT a single Topic, with varying probabilities. Similarly, every document is linked with ALL the Topics, with varying probabilities.

Random Forest

Let us assume we have 15 member council of ministers who are deliberating on a decision, whether to go ahead with it or not. In such cases, every minister will come up with their individual decisions after considering the various options they have at their disposal. The final decision will be arrived with the help of collective outcomes, if we have more “Yes” then the decision will be taken, on the other hand if there are more “No” then the decision will not be taken.

Random Forest, implements this mechanism.

Taking a more related case, assume we have 10,000 observations with 50 features.

Random Forest, uses subsets from these 10k observations using random sampling, assuming 10% subset, it will draw set of 1000 observations and in those observations it will also select a random subset of features, a 20% feature subset will mean selecting 10 features randomly from it. In this process number of Decision trees are built using Random Subsets of Samples and Features. Presence of so many trees, give rise o the word “Forest”. Each of these trees will generate a prediction for the data. Using averaging or “voting” different predictions are combined and is treated as the final prediction.
Combination of random selection of data samples as well as random splits of predictor combined with the “bagging” method make Random Forest an useful and powerful Machine Learning tool. It also includes an innovative method to determine the variable importance for the prediction exercise.

Support Vector Machine

The basic idea of a Support Vector Machine(SVM) is to find a hyper plane which separates the N-dimensional data perfectly into its two classes. Support vectors are the data points nearest to the hyperplane, the points of a data set that, if removed, would alter the position of the dividing hyperplane. Because of this, they can be considered the critical elements of a data set.
However, since example data is often not linearly separable, SVM’s introduce the notion of a “kernel induced feature space” which casts the data into a higher dimensional space where the data is linearly separable.

4.3 Machine Learning Framework and Computation

A – Phrase Polarity Learning

The Corpus days are distributed first into two subsets, one where the Target Currency/FTSE Return(GBPEUR/GBPUSD/FTSE100) was Positive and the other where the Return was Negative. Next using LDA Topic Model, phrases are assigned to K topics for these two Subsets.Assuming we have N Phrases, W ∈ $\{W_i , i = 1..N\}$ mapped to k Topics, T ∈ $\{T_j , j = 1..k\}$, after Topic Modelling using LDA, then we have the following :

For each of the k Topic,

$\sum_{i=1}^{N} p(W_i) = 1$ (7.1)

where $p(W_i)$ is the probability of Phrase W_i, as per Latent Dirichlet Allocations, β Distribution.

$\sum_{j=1}^{k} \sum_{i=1}^{N} p(W_{ij}) = k$ (7.2)

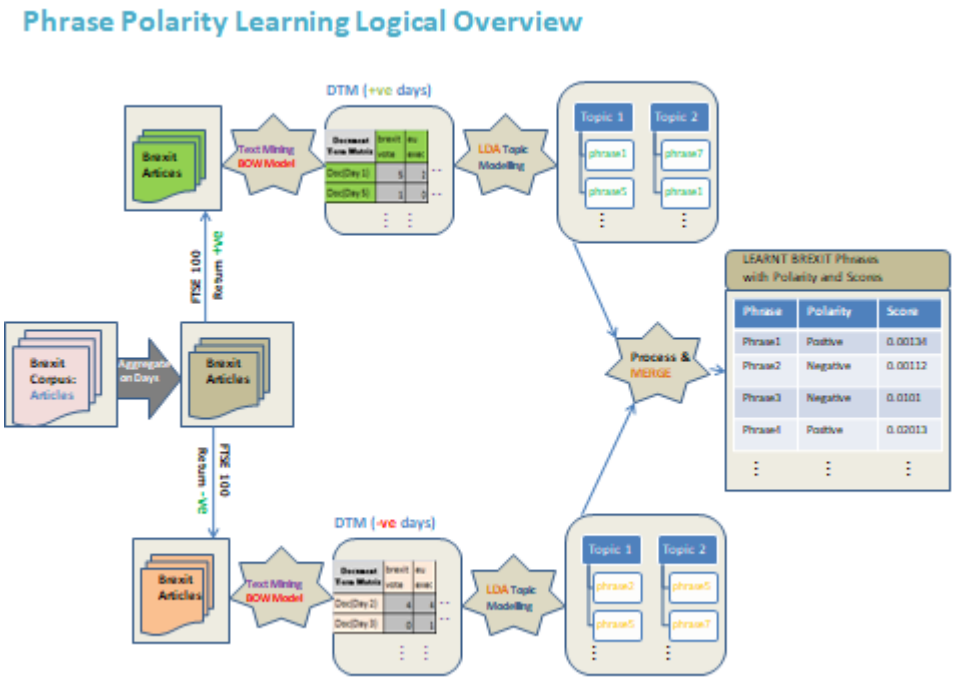
where $p(W_{ij})$ is the probability of Phrase W_i in Topic T_j

With these two constraints, we compute the Total Probability P(W_i) for a phrase W_i as

P(W_i) =($\sum_{j=1}^{k} p(W_{ij})$) / k (7.3)

Due to time gap of the news articles publication and Currency Exchange/Stock Market availability window, some of the articles may likely to have their effect felt the next day. To balance this anomaly, we considered a combination of one day Lag as well as same day effect. So, for Positive Return days, we considered two LDA Topic models, one with LAG1 effect and other with Same Day effect. The final probability of a Phrase for Positive Return days is the average of the two probabilities. Similar combined Lag effect is considered for

Fig. 6. Logical overview of phrase polarity learning



Negative Return days. Thus for any Phrase W_i, the total probability of Positive days will be

$P_{pos}(W_i)$ = Average ($P_{pos_{(Lag1)}}(W_i)$, $P_{pos_{(Lag0)}}(W_i)$) (7.5)

$P_{\neg}(W_i)$ = Average ($P_{\neg Lag1}\ (W_i)$, $P_{\neg Lag0_{(\ )}}(W_i)$) (7.6)

where individual lagged probabilities are computed using equation (7.3). Now every Phrase, $W_i$has a total probability for Positive and Negative Days, depicted as

$P_{pos}(W_i)$ and $P_{\neg}(W_i)$.

Then Polarity of any Phrase W_i is defined by,

$$\text{Polarity}(W_i) = \begin{cases} 1, & if\ P_{pos}(W_i) > P_{neg}(W_i) \\ -1, & if\ P_{pos}(W_i) < P_{neg}(W_i) \end{cases} \quad (7.7)$$

Phrase Score for W_i is obtained by the absolute difference of the respective probabilities for positive and negativedays

$Score$ = ABS $|P_{pos}(W_i) - P_{\neg}(W_i)|$ (7.8)

At the end of this step, for every Phrase W_i, we have learnt it's associated Phrase Score and Polarity given by { $Score$) , Polarity(W_i) } as per equations (7.5) and (7.6)

B – Phrase Importance using Random Forest

In this step we train a Random Forest using the phrases as predictors and Currency/FTSE Return(GBPEUR/GBPUSD/FTSE100) movement (up/ down) as Target. Using 10-fold cross validation on training samples we find the best Random Forest model, in terms of Predicting Accuracy and determine the Importance of the Features(Phrases) in Predicting the Return Movement.Like in Topic Modelling we used Random Forest for the other for same day return movement. We also repeated the exercise with the return movement with one day lag. There is hardly any difference in terms of results for the performance evaluation. Hence we do not report it here. For any Phrase, W_i, it's Importance is computed as the average of two Lagged Importance,

Imp(W_i) = Average ($Imp_{Lag1}(W_i)$, $Imp_{Lag0}(W_i)$) (8)

After this step, the Feature Set is now been ordered according to Predictor Importance as per the trained Random Forest model. Next we use this ordered feature as Predictors to a Support Vector Machine to learn weights of these features i.e Phrases.

C – Computing Daily Sentiment Score

Using the ordered feature set as Predictors and Return Movement as Target a Support Vector Machine is trained using Training Samples. A Linear Kernel is used and the best value of the related hyper parameter is first determined. Best Model is selected using 10-folds Cross Validation and feature weights are determined. Like in earlier two cases of learning, two SVMs are used, one for 1 day Lagged effect and the other for same day effect of Return Movement. At the end of this step we have for every Phrase, W_i, it's associated weights, Weight(W_i).

Combining steps A and C, we now have, using the Training Samples, for every Phrase, W_i, it's combined score indicated as CScore

CScore(W_i) = Weight(W_i) * $Score$) (9)

After Steps A,B & equation (8), we have a triplet as below,

{W_i, Polarity(W_i), CScore(W_i) }, for i = 1..N (10)

Using all the information, so far, we went on to calculate Sentiment Score on Out of Sample data. We start from the Document Feature Matrix based on the Out of Sample data. Assuming M Phrases, W_1', W_2',..,W_M' spread across D days, and it's DFM is represented as

$\{ n_{dj}$; d=1,2,..D, j=1,2,..M $\}$,

where n_{dj} represents the frequency of occurrence of Phrase W_j' in day d.

Next we restrict the DFM to include only those Phrases that are available as Phrases in the Training Sample, as in step A. The DFM for Out of Sample then becomes,

$DFM_{test} = \{ n_{dp}$; d=1,2,..D, p=1,2,..L $\}$, (11)

where n_{dp} represents the frequency of occurrence of Phrase W_p' in day d such that,

$\forall p = 1..L, W_p' = W_r$, for some r in 1..N where 1 < L <= N

Using equation (10) we have the Day Wise Sentiment Score,

$SentScore_{test} = \{ Z_d$; d=1,2,..D $\}$,

where $Z_d = \sum_{p=1}^{L}(n_{dp} * CScore(W_p) * Polarity(W_p))$ (12)

The figures below depict 50 Phrases having maximum weights with Target as GBPEUR, GBPUSD and FTSE respectively. *Custom Union* and *Northern Ireland* are appearing as most "weighty" phrase in all three cases however their Polarity differs. *Custom Union* is observed to be more associated with GBPEUR and FTSE "down" movement but for GBPUSD it is more prone to "Up" movement of its return.

Fig. 13. Feature weights of Top 50 Phrases for prediction of GBPEUR return movement

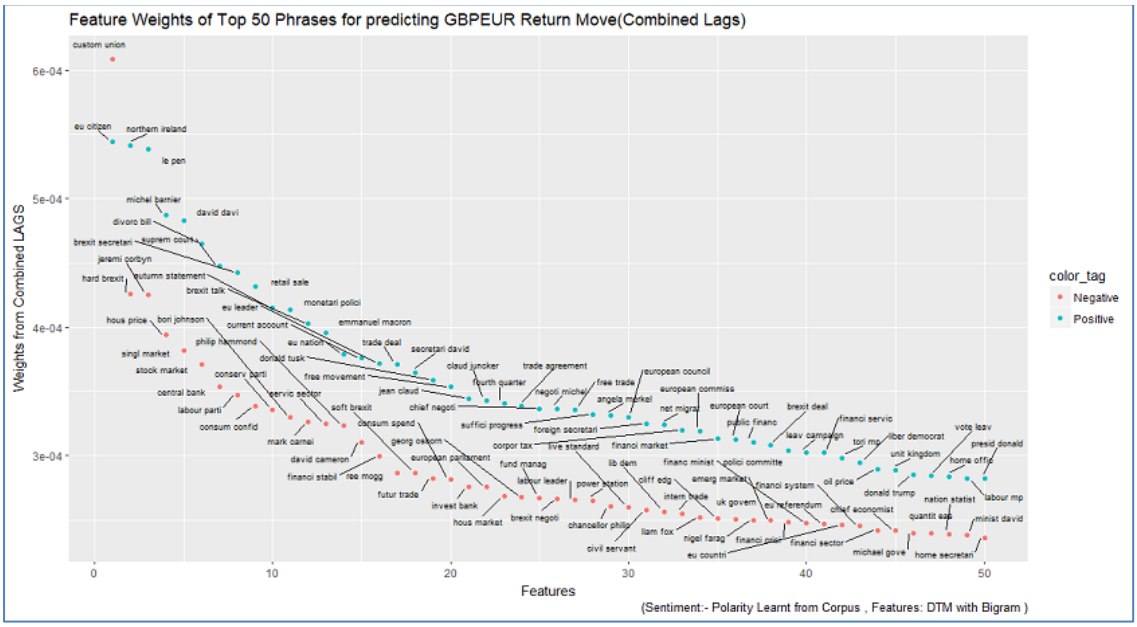


Fig. 14. Feature weights of Top 50 Phrases for prediction of GBPUSD return movement

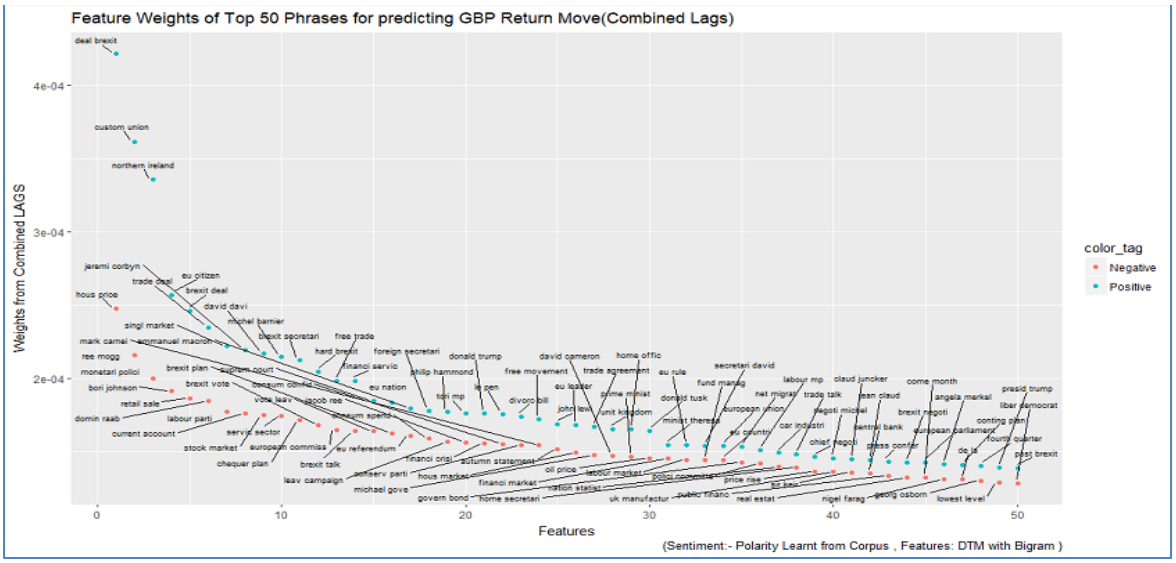
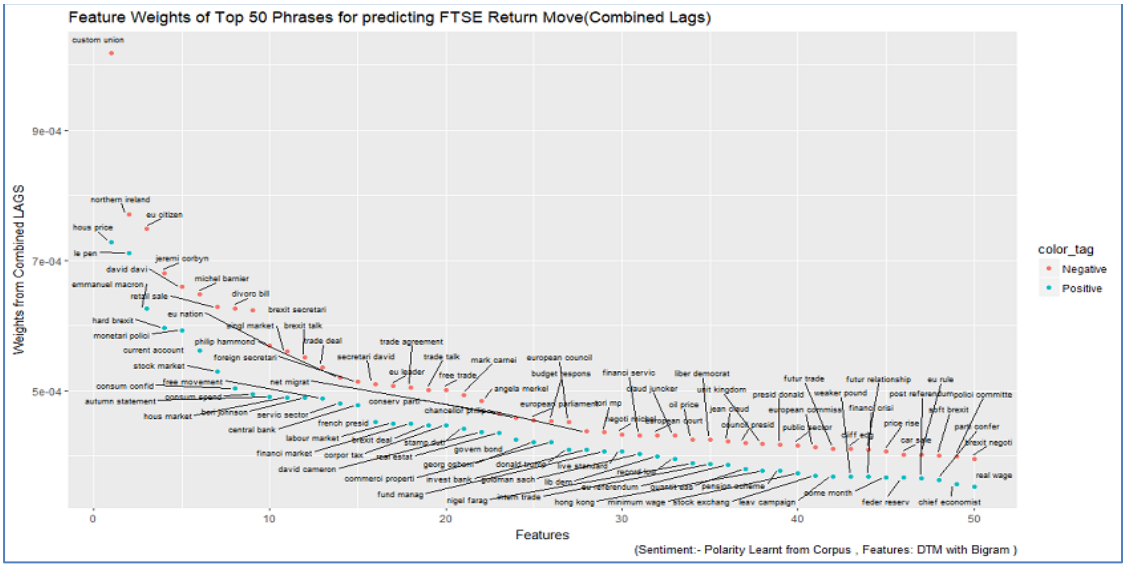


Fig. 15. Feature weights of Top 50 Phrases for prediction of FTSE return movement



Interestingly, we also looked at the top phrases, with two different subsets, one consists of Articles from Tabloids (Daily Mail, Daily Mirror) and the other from "Serious" News Media(Guardian, Financial Times, Wall Street Journal). We observed these two sets indicates some distinctive differences. Related Charts are included in the Section, Additional Insights.

We build the time series for currency prices and Stock Index for the duration of our analysis, June 2016 till November 2018, as already described in the LOUGHRAN process.Next we merge Sentiment Scores and Currency / FTSE Price & Returns for the respective days. Articles published on Weekends and other days when Currency Exchange or Stock Market is not open, are linked to next available day of Stock Market or Currency Exchange.Post this alignment and merge we have the below triplets,

$\{ Z_d, R_d , P_d ,\ \forall$d=1,2,..D $\}$, (13)

whereZ_d is Sentiment Score of day d, as in equation (12).

R_d denotes Daily Returns of respective Financial Parameter (GBPUSD, GBPEUR or FTSE 100) on day d,

P_d denotes Price of respective Financial Parameter (GBPUSD/ GBPEUR Exchange Rate or FTSE 100 Price) on day d.

Figures below plot Sentiment Score computed by proposed methodology with the price and the corresponding return for the 30% sample in 70:30 split of training and testing. Both Sentiment Scores and Price/Returns are normalized for uniformity.

Fig. 16. Sentiment score vs. GBPEUR Price and Return (30% sample in 70:30 split)

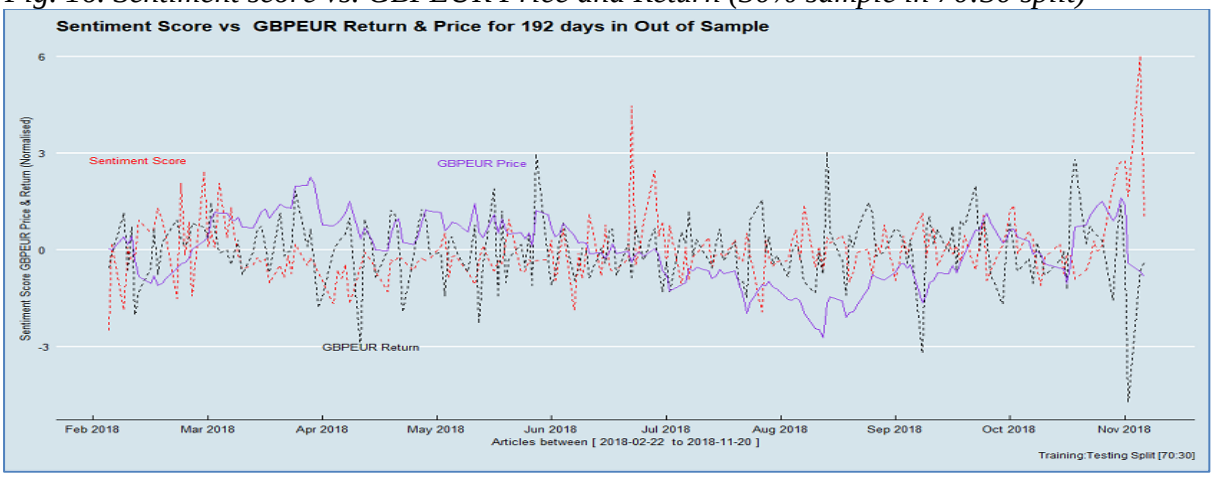


Fig. 17. Sentiment score vs. GBPUSD Price and Return (30% sample in 70:30 split)

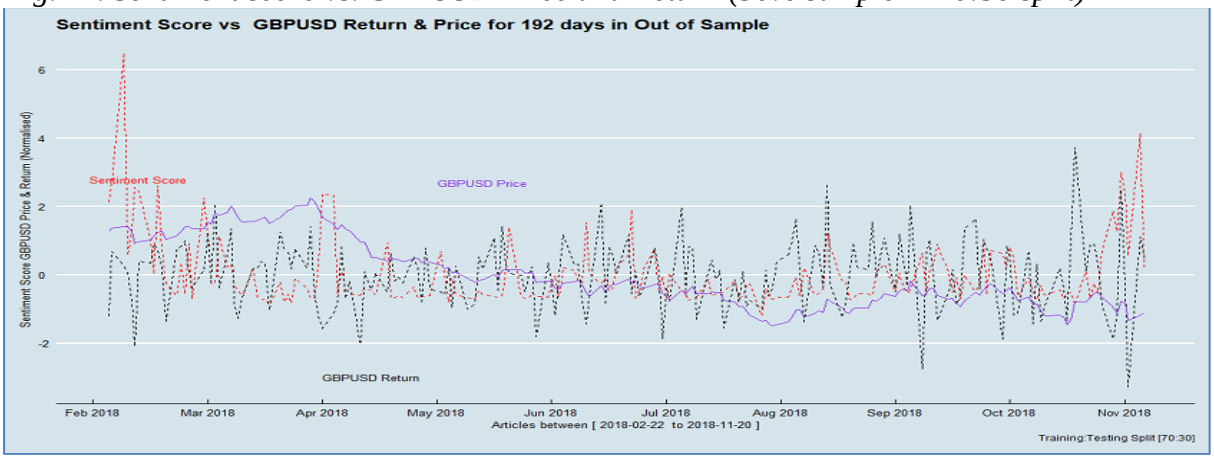
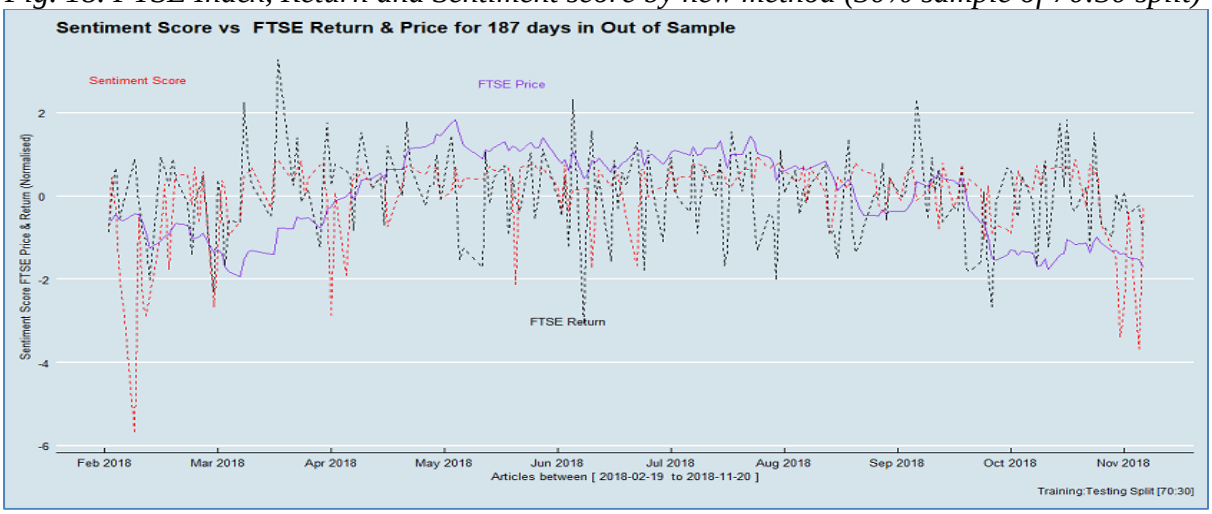


Fig. 18. FTSE Index, Return and Sentiment score by new method (30% sample of 70:30 split)



4.4 Statistical analysis of Long run

The descriptive statistics for the Sentiment scores with the newly proposed methodology, henceforth called Score_NewEUR, Score_NewUSD, Score_FTSE and the price series are given below for the period of estimation (i.e. 30% sample in 70:30 split) are given in Table 8 and the results of the tests for stationarity in Table 9 below. It may be noted that the newly computed sentiment score will be different corresponding to the three price series. It may be noted that we also computed sentiment score using new methodology with one day lag for the sentiment calculation. All the results, graphical representation as well as statistical tests are similar. Hence we reported our analysis with sentiment score of the same day only. The two scores cannot be included together in the analysis as they are highly correlated, hence subject to multicollinearity problem.

Table 8. Descriptive statistics – Self learning sample (30% sample)

Variable	Mean	SD	Min	Max	Skewness	Kurtosis
Score_NewEUR	-2.24e-09	1	-3.070092	4.557474	.5500636	5.600033
Score_NewUSD	-1.25e-09	1	-3.410589	4.606915	.7324663	6.340054
Score_NewFTSE$^{\$}$	-1.43e-10	1	-6.543618	1.903303	-2.963825	18.4022
GBPEUR	1.13207	.011512	1.100625	1.158006	-.2770568	2.479347
GBPUSD	1.335094	.0438756	1.26939	1.434206	.6377441	2.17748
FTSE$^{\$}$	7398.962	261.8878	6888.7	7877.5	-.2034508	1.674757

Note: No of observations are 187 for these, for the rest 192.

Table 9. Tests of stationarity for self learning sample (30% sample)

Variable	LEVEL			
	ADF	**PP**	**KPSS**	**DF-GLS**
Sentiment(Ref:GBPEUR Return)	-8.395(***)	-13.0743(***)	0.065	-4.8672(***)
Sentiment(Ref:GBPUSD Return)	-8.1646(***)	-11.9537(***)	0.0595	-4.2505(***)
Sentiment(Ref:FTSE Return)	-9.1916(***)	-12.6668(***)	0.0316	-4.553(***)
GBPEUR Return	-10.07(***)	-13.3323(***)	0.0451	-3.6386(***)
GBPUSD Return	-10.5744(***)	-14.1827(***)	0.0455	-3.1889(**)
FTSE100 Return	-10.6236(***)	-15.2413(***)	0.0641	-4.4812(***)
GBPEUR Price	-2.9062	-2.8296	0.3878(***)	-2.1475
GBPUSD Price	-2.0232	-2.1372	0.5409(***)	-1.3386
FTSE100 Price	-1.0225	-1.0845	0.7954(***)	-1.0373
	First Difference			
GBPEUR Price	-9.9991(***)	-13.6679(***)	0.043	-4.7535(***)
GBPUSD Price	-10.0774(***)	-14.0089(***)	0.0601	-4.4028(***)
FTSE100 Price	-9.9447(***)	-14.2125(***)	0.0574	-4.3753(***)
Note : *, **, * indicates significance at 10%, 5% and 1% respectively**				

It is evident from Table 9 that all the sentiment score series are I(0) while three price series are I(1). Hence as in the previous case we conducted test for the long run relation by ARDL method of Pesaran *et al* (2001) explained in the previous section. The results are provided in Table 10 below. We are interested in the pairwise long run relation between a particular price series and the corresponding sentiment score. Table 10 shows that there exists only long run relation between GBPEUR and Sentiment_NEWEUR. The corresponding estimated long run coefficients and the ECM are given in Tables 10 and 11. Rest of the analyses are undertaken for these two series only.

Note : The F-test clearly indicates cointegration. However, with additional t-test, cointegration is narrowly missed. We explored also with an 80:20 split, and we observed, in that case, both F-test and t-test indicates cointegration. Results for 80:20 split are included in Section 6.

Table 10. Test of Cointegration –Sentiment Score by New Method

Variable set	**F-statistic**	**Remark**
GBPEUR & Score_NEWEUR	6.093**	Cointegration at 5%
GBPUSD & Score_NEWUSD	1.234	No cointegration
FTSE-100 & Score_NEWFTSE	0.535	No cointegration

Note: ARDL test for cointrgartion is conducted *a là*Pesaran *et al* (2001)
*, **, ***: significance levels at 10%, 5%, 1% respectively

Variable set	**T-statistic**	**Remark**
GBPEUR& Sentiment Score	-2.761	Just short of Cointegration at 10%
GBPUSD& Sentiment Score	-1.266	No cointegration
FTSE-100& Sentiment Score	-0.914	No cointegration

Table 11. Long run relation for the ARDL Model – GBPEUR and Score_NewEUR

Variable	**Co-efficient**	**P-value**	**95% Conf. Interval**
Score.GBPEUR.norm	-.0097425	0.066	[-.0201271, .0006421]

Table 13 below provides the Granger causality test by Toda-Yamamota procedure. It is evident from Table 13 that Score_NewEUR Granger causes GBPEUR and not vice versa. Hence Score_NewEUR is exogenous variable of the system.

Table 13. Granger causality by Toda-Yamamoto test

Wald Coefficient test for significance of constant and other variables			
Estimated Equation : SentimentScore = $c_1 + \alpha_1\ GBPEUR_t + \alpha_2\ GBPEUR_{t-1}$		$H_0 : c_1 = 0, \alpha_1 = 0, \alpha_2 = 0$	
Test Statistic	**Value**	**Degree of Freedom**	**Probability**
x^2 - Statistic	3.4	3	0.34
Estimated Equation :GBPEUR = $c_2 + \alpha_3\ SentimentScore_t + \alpha_4\ SentimentScore_{t-1}$		$H_0 : c_2 = 0, \alpha_3 = 0, \alpha_4 = 0$	
x^2 - Statistic	10.6	3	0.014

As in the previous section test results show that Score_NewEUR is exogenous to the system. Next we perform a comparative analysis of forecasting in traditional ARIMA framework and in a dynamic regression model with Score_NewEUR and its lagged values as an additional regressors. This can be executed in two alternative ways – ARIMAX and ARDL model. Since Score_NewEUR is exogenous to the system we can implement by ARDL framework (please see de Brouwer and Ericsson, 1998; Hendry and Ericsson, 1991; Davidson *et al*, 1987). So we estimate ARIMA model for GBPEUR series by Box-Jenkins Methodology and compare its forecasting performance with the dynamic regression model estimated by ARDL. The dynamic regression model in ARDL framework to be estimated is given by the following regression equation:

$$GBPEUR_t = \alpha + \sum_1^m \beta_j\, GBPEUR_{t-j} + \gamma_0 Score\_NewEUR_t + \sum_1^n \gamma_k Score\_NewEUR_{t-k} + \epsilon_t.$$

One can include current period value of *Score_NewEUR*in the above specification as it is exogenous, hence its determination does not depend on current and lagged values of *GBPEUR*. In case of the ARIMA model the best fit is obtained using standard technique of Box-Jenkins methodology. For the dynamic regression model a very important aspect is the lag structure. It is decided using two criteria – significance of regressors as obtained from t-value and model's overall F-Statistic. The related plot is given in the following section. We performed the comparison of the Forecast for two cases, using Sentiment Score computed using Machine Learning and Sentiment Score using LOUGHRAN Word Lexicon.

5 Forecasting Comparison with different Sentiment Score flavors

5.1 Forecasting with Machine Learnt Sentiment Score

After we evidenced cointegration for GBPEUR Price and Sentiment Score, with 70:30 split, and also observed a one way Granger Causality with Sentiment Score => GBPEUR Price, we went ahead and perform Forecasting on the Test Samples, which is the last 30% among the 639 days between, June-2016 till November-2018.

We used two Forecasting Models, ARIMA and Dynamic Regression using ARDL.

For ARIMA, we determined the best fit ARIMA model on the Test Set after confirming that the errors are indeed iid.

TABLE 5.1.1 ARIMA Model Selection on Test Samples

ARIMA Model	AIC	BIC	Log Likelihood	Test on ARIMA Residuals `Ljung-Box test`
Arima(0,1,0)	-1531.45	-1528.19	766.72	data: Residuals from ARIMA(0,1,0) Q* = 16.839, df = 10, p-value = 0.078 Model df: 0. Total lags used: 10
Arima(1,1,0)	-1529.45	-1522.94	766.72	data: Residuals from ARIMA(1,1,0) Q* = 16.697, df = 9, p-value = 0.05368 Model df: 1. Total lags used: 10
Arima(1,1,1)	-1527.6	-1517.84	766.8	data: Residuals from ARIMA(1,1,1) Q* = 16.795, df = 8, p-value = 0.03232 Model df: 2. Total lags used: 10
Arima(2,1,1)	-1530.39	-1517.38	769.2	data: Residuals from ARIMA(2,1,1) Q* = 10.646, df = 7, p-value = 0.1548 Model df: 3. Total lags used: 10
Arima(2,1,2)	-1528.19	-1511.93	769.1	data: Residuals from ARIMA(2,1,2) Q* = 14.765, df = 6, p-value = 0.02216 Model df: 4. Total lags used: 10
Arima(3,1,2)	-1527.03	-1507.52	769.52	data: Residuals from ARIMA(3,1,2) Q* = 9.4672, df = 5, p-value = 0.09182 Model df: 5. Total lags used: 10

Based on the above result, we took Arima(0,1,0), Arima(2,1,1) and Arima(3,1,2) and compute the MSPE in all the three cases, the Forecast Errors are given in the below table

TABLE 5.1.2 Forecast Error(MSPE) on ARIMA Models

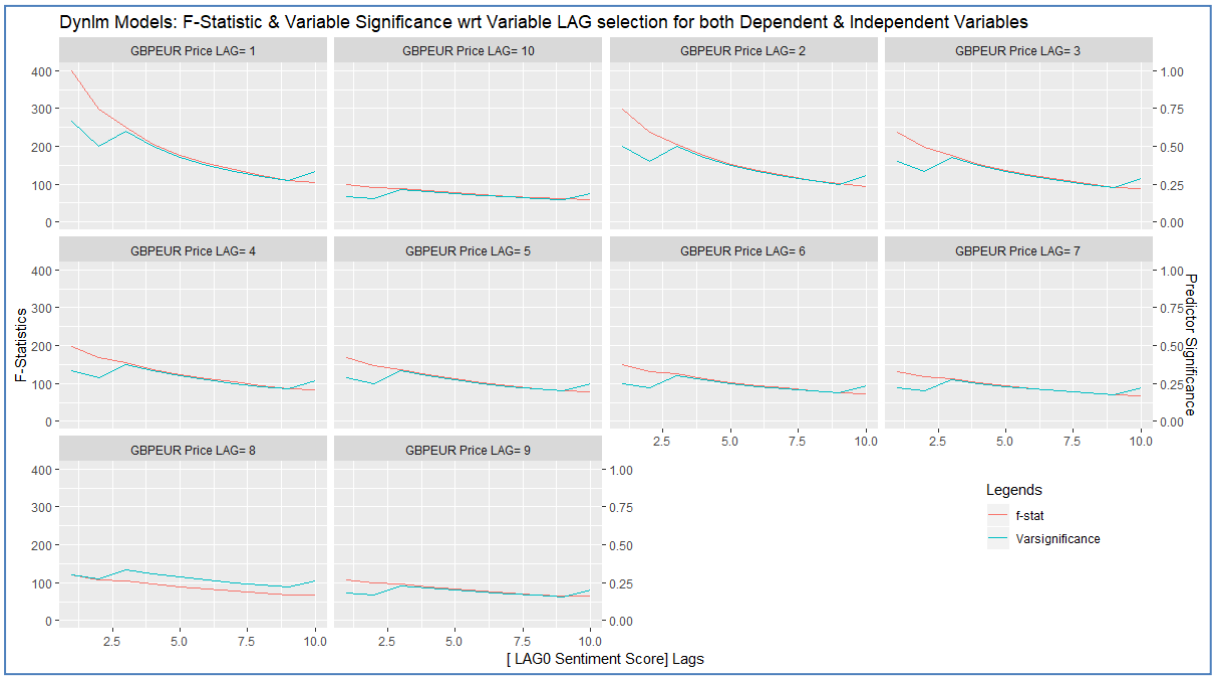
ARIMA Model	ar1	ar2	ar3	ma1	ma2	**Forecast Error (MSPE)**
Arima(0,1,0)	NA					1.899632e-05
Arima(2,1,1)	0.8831	-0.0781	NA	-0.9041	NA	1.849424e-05
Arima(3,1,2)	0.4510	0.3081	-0.0921	-0.4750	-0.3551	1.84313e-05

Next we perform Forecasting using Dynamic Linear Model. To do this we first determine the LAG Structure for both Dependent Variable GBPEUR Price and Exogenous variable Sentiment Score. We started with sufficient LAGs(upto 10) for both the variable GBPEUR Price and Sentiment Score and continue reducing their LAGs to find the best model.

The best model is determined using two criteria; Most Significant regressors (in terms of p-value) and Model's F-Statistic. Using these two criteria, we select the best model for both, Previous Day and Same Day, Sentiment Score.

From the plot below, we determine LAG structure for both the dependent variable, GBPEUR Price and Exogenous variable, Sentiment Score to be 1. Sentiment Score, itself, also appears as a regressor in this model.

Figure: 5.1.1 Determining optimum LAG structure for GBPEUR Price and Sentiment Score in dynlm model



To confirm the optimum LAG structure, we also looked at the AIC value along with it's F-statistics, and as the below table reveals, we see LAG = 1 gives the best model.

Table 5.1.3 : F-statistics Value of different dynlm Models at various LAGS

F-Statistic at Various LAGS					
GBPEUR Price LAGS → Sentiment Score Lags ↓	1	2	3	4	5
1	**400.3161**	297.2435	236.9932	197.1154	167.9106
2	298.9621	238.0524	197.5803	169.1761	147.0219
3	249.2084	206.661	176.1673	155.1517	137.2711
4	206.2219	175.8597	153.0231	137.4815	123.2396
5	175.5003	152.7955	135.0613	122.8241	111.4624

Table 5.1.4 : AIC Value of different dynlm Models at various LAGS

AIC at Various LAGS					
GBPEUR Price LAGS → Sentiment Score Lags ↓	1	2	3	4	5
1	**-1536.21**	-1525.24	-1515.54	-1505.96	-1495.85
2	-1526.18	-1524.31	-1514.51	-1505.08	-1494.87
3	-1523.8	-1521.89	-1519.9	-1511.68	-1501.8
4	-1513.35	-1511.41	-1509.42	-1510.09	-1500.32
5	-1503.04	-1501.14	-1499.15	-1499.77	-1498.41

Next we explore with three of these models based on F-stat and AIC values and as per the table below and check for Serial Correlation.

TABLE 5.1.5 Dynlm Models with Serial Correlation Test Outcomes

	GBPEUR Price LAGS	Sentiment Score LAGS	Serial Correlation Test Breusch-Godfrey test for serial correlation
Dynlm Models	1	1	data: Residuals LM test = 11.633, df = 10, p-value = 0.3104
	1	2	data: Residuals LM test = 12.286, df = 10, p-value = 0.2664
	2	1	data: Residuals LM test = 12.036, df = 10, p-value = 0.2827
	2	2	data: Residuals LM test = 18.128, df = 10, p-value = 0.05284

Next we chose the first three models, which doesn't have any serial correlation and compute the Forecast Errors using MSPE.

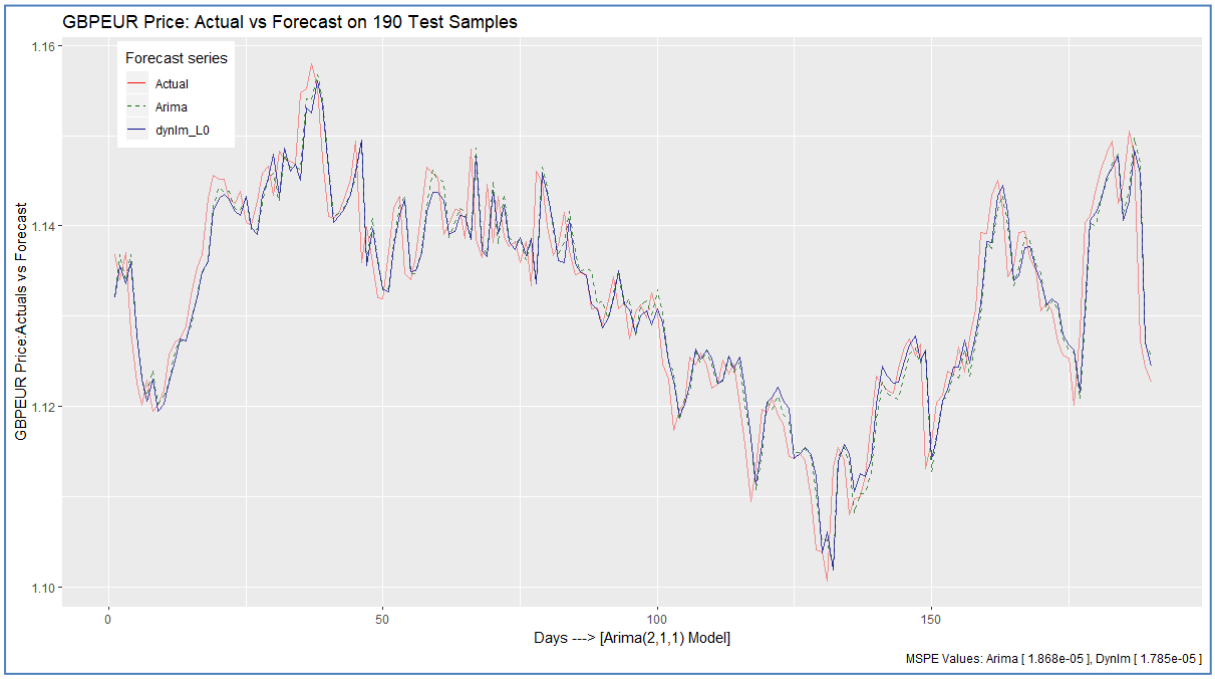
TABLE 5.1.6 DYNLM Models, Coefficients and Forecast Error(MSPE)

DYNLM Model	**intercept**	**GBPEUR_l1**	**GBPEUR_l2**	**Score**	**Score_l1**	**Score_l2**	**Forecast Error (MSPE)**
Y_lag=1, X_lag=1	0.0872375	0.9228986	NA	-0.0007284	-0.0001860	NA	1.785536e-05
Y_lag=1, X_lag=2	0.0911989	0.9194005	NA	-0.0006951	-0.0001755	-0.0003203	1.784865e-05
Y_lag=2, X_lag=1	0.0890291	0.9454559	-0.0241371	-0.0007214	-0.0001735	NA	1.793796e-05

We observe that best performance of dynlm model is seen with GBPEUR Price having 1 lag and Sentiment Scores with 2 lags.

Below plot depicts the best Forecast from both the ARIMA and Dynlm models.

Figure 5.1.2 GBPEUR Price Actual AND Forecast Series on Test Sample with Self Learnt Sentiment



5.2 Forecasting using Loughran Sentiment Score

We finally explore Forecasting using LOUGHRAN Sentiment Score. We evidenced cointegration and one way causality for 50:50 split. We follow similar steps, as above, and compare Forecast in the later 50% of the sample, which consists of later 320 days out of total 639 days.

TABLE 5.2.1 ARIMA Model Selection on Test Samples on 50:50 Split

ARIMA Model	AIC	BIC	Log Likelihood	Test on ARIMA Residuals `Ljung-Box test`
Arima(0,1,0)	-2503.8	-2500	1252.9	data: Residuals from ARIMA(0,1,0) Q* = 18.436, df = 10, p-value = 0.04804 Model df: 0. Total lags used: 10
Arima(1,1,0)	-2501.87	-2494.34	1252.94	data: Residuals from ARIMA(1,1,0) Q* = 18.831, df = 9, p-value = 0.02667 Model df: 1. Total lags used: 10
Arima(1,1,1)	-2503.83	-2492.53	1254.91	data: Residuals from ARIMA(1,1,1) Q* = 15.547, df = 8, p-value = 0.04935 Model df: 2. Total lags used: 10
Arima(2,1,1)	-2501.83	-2486.76	1254.91	data: Residuals from ARIMA(2,1,1) Q* = 13.077, df = 7, p-value = 0.07025 Model df: 3. Total lags used: 10
Arima(2,1,2)	-2500.92	-2482.12	155.47	data: Residuals from ARIMA(2,1,2) Q* = 13.212, df = 6, p-value = 0.03979 Model df: 4. Total lags used: 10
Arima(3,1,2)	-2498.06	-2475.47	1255.03	data: Residuals from ARIMA(3,1,2) Q* = 12.795, df = 5, p-value = 0.02538 Model df: 5. Total lags used: 10
Arima(3,1,3)	-2505.16	-2478.8	1259.58	data: Residuals from ARIMA(3,1,3) Q* = 8.4597, df = 4, p-value = 0.07612 Model df: 6. Total lags used: 10

Based on the above result, we took Arima(2,1,1) and Arima(3,1,3), as these don't have any serial correlation, and compute the MSPE in all the two cases, the Forecast Errors are given in the below table

TABLE 5.2.2 Forecast Error(MSPE) on ARIMA Models for 50:50 Split, Test Sample

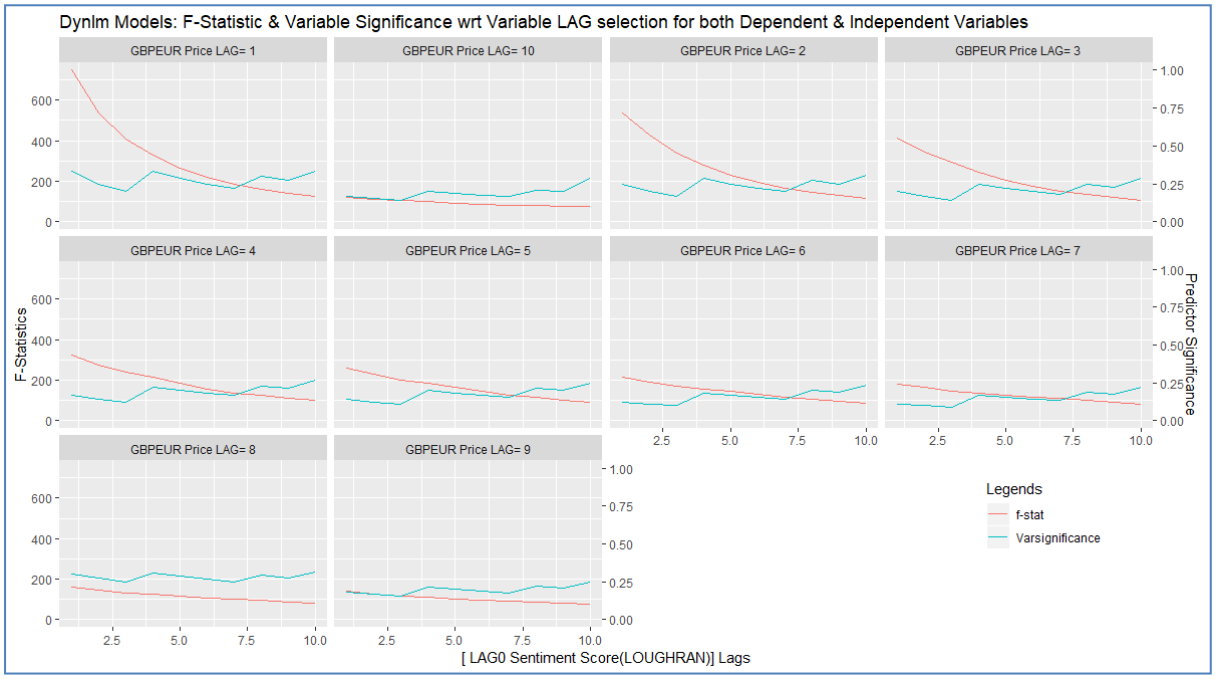
ARIMA Model	ar1	ar2	ar3	ma1	ma2	**ma3**	**Forecast Error (MSPE)**
Arima(2,1,1)	0.4215	-0.0939	NA	-0.4439	NA	NA	2.234911e-05
Arima(3,1,3)	0.399	0.0324	-0.8131	-0.4164	-0.0753	0.8494	2.141798e-05

Next we perform Forecasting using Dynamic Linear Model, following similar steps, as above.

The best model is determined using two criteria; Most Significant regressors (in terms of p-value) and Model's F-Statistic. Using these two criteria, we select the best model for both, Previous Day and Same Day, Sentiment Score.

From the plot below, we determine LAG structure for both the dependent variable, GBPEUR Price and Exogenous variable, Sentiment Score to be 1. Sentiment Score, itself, also appears as a regressor in this model.

Figure: 5.2.1 Determining optimum LAG structure for GBPEUR Price and LOUGHRAN Sentiment Score in dynlm model



To confirm the optimum LAG structure, we also looked at the AIC value along with it's F-statistics, and as the below table reveals, we see LAG = 1 gives the best model.

Table 5.2.3 : F-statistics Value of different dynlm Models at various LAGS with LOUGHRAN Sentiment Score

F-Statistic at Various LAGS					
GBPEUR Price LAGS → Sentiment Score Lags ↓	1	2	3	4	5
1	**747.3756**	534.8753	410.9004	322.69	260.2973
2	536.0298	427.5343	342.2847	276.4234	227.6456
3	409.6768	340.4192	292.4409	241.0897	201.6912
4	326.7436	279.24	244.6492	216.8982	183.5135
5	262.9762	229.4245	204.2801	183.3422	166.314

Table 5.2.4 : AIC Value of different dynlm Models at various LAGS with LOUGHRAN

AIC at Various LAGS					
GBPEUR Price LAGS → Sentiment Score Lags ↓	1	2	3	4	5
1	**-2513.71**	-2503.11	-2497.81	-2487.76	-2478.78
2	-2503.71	-2501.76	-2496.59	-2486.48	-2477.53
3	-2496.99	-2495.08	-2494.59	-2484.49	-2475.53
4	-2491.17	-2489.25	-2488.49	-2486.66	-2477.36
5	-2481.54	-2479.63	-2478.97	-2477.11	-2475.41

Next we explore with three of these models based on F-stat and AIC values and as per the table below and check for Serial Correlation.

TABLE 5.2.5 Dynlm Models with Serial Correlation Test Outcomes with LOUGHRAN

	GBPEUR Price LAGS	Sentiment Score LAGS	Serial Correlation Test Breusch-Godfrey test for serial correlation
Dynlm Models	1	1	data: Residuals LM test = 17.718, df = 10, p-value = 0.0599
	1	2	data: Residuals LM test = 17.398, df = 10, p-value = 0.066
	2	1	data: Residuals LM test = 17.966, df = 10, p-value = 0.05555
	2	2	data: Residuals LM test = 18.462, df = 10, p-value = 0.04765

Next we chose the first three models, p-value Not < 0.05, we cannot reject Null Hypothesis of No Serial Correlation. We compute the Forecast Errors using MSPE for all of them.

TABLE 5.2.6 DYNLM Models, Coefficients and Forecast Error(MSPE) with LOUGHRAN Sentiment Score

DYNLM Model	intercept	GBPEUR_l1	GBPEUR_l2	Score	Score_l1	Score_l2	Forecast Error (MSPE)
Y_lag=1, X_lag=1	0.0891480	0.9211627	NA	0.0001173	0.0001672	NA	2.14624e-05
Y_lag=1, X_lag=2	0.0916244	0.9189758	NA	0.0001625	0.0002068	-0.0002191	2.146894e-05
Y_lag=2, X_lag=1	0.0919426	0.9292869	-0.0105934	0.0001223	0.0001539	NA	2.150936e-05

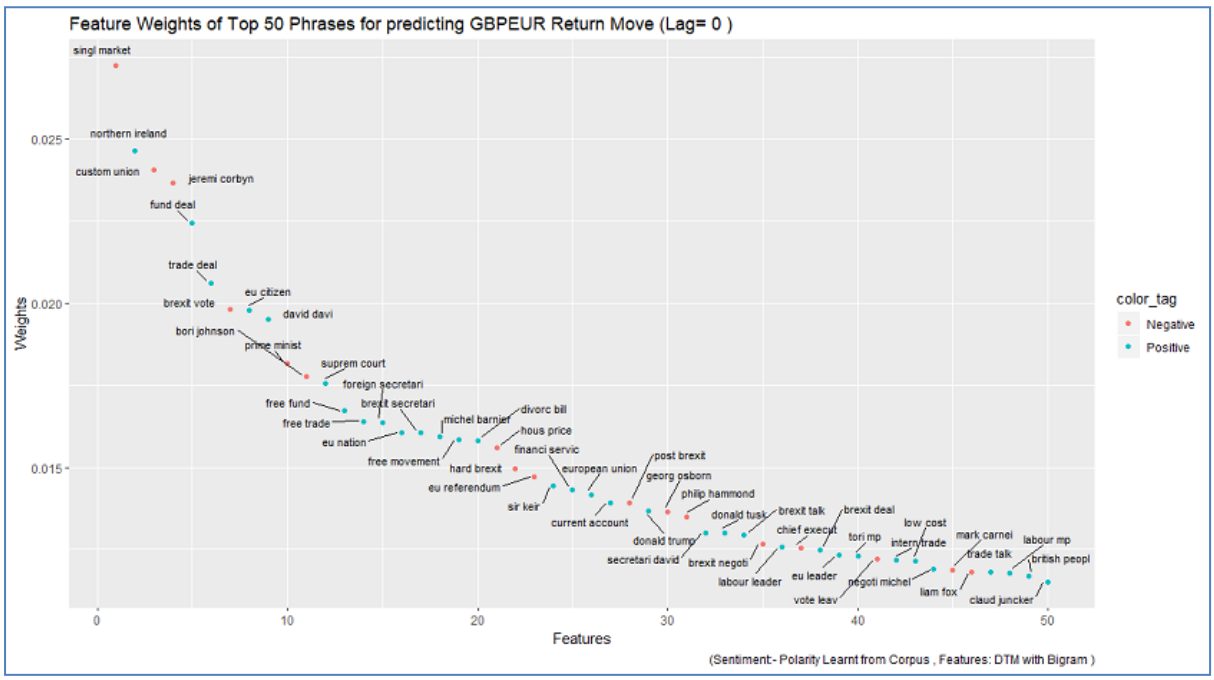
6 Additional Results and Insight

6.1 Comparison of Top Phrases for UK Tabloids and Non Tabloid Media

From our entire Corpus, we created two subdivisions one having articles from UK Tabloids, like Daily Mail and Daily Mirror and the other subset, involving The Guardian, Financial Times and Wall Street Journal. The intention was to explore whether articles read by "Commoners" vis-a-vis "Intellectuals/Educated" are similar or having some difference. Phrases like Customs Union, Northern Ireland etc. are some common more prominent phrases.
Even then, Top Phrases from these two sets also gives a contrasting picture. In UK Tabloids, we see strong appearances of Political figures like Jeremy Corbyn, Boris Johnson and other personalities. Boris Johnson is almost among the Top 10 when Same day Sentiment is concerned, moves a bit lower within the Top 20s with Previous Day Sentiments, showing greater importance and prominence.

On the other hand, for Non Tabloid Newspapers, we see more analytical terms involving economy and related, with terms such as Consumer Confidence, Consumer Spending, House Price, Trade Agreement, Hard Brexit etc. dominating the discussions. Among personalities, we see Mark Carnei(Governer of ECB) and Philip Hammond and interestingly no Boris Johnson or Jeremy Corbyn.

Figure 6.1.1 Top 50 Phrases from UK Tabloids, with Sentiment Score of Same Day and One Day Previous



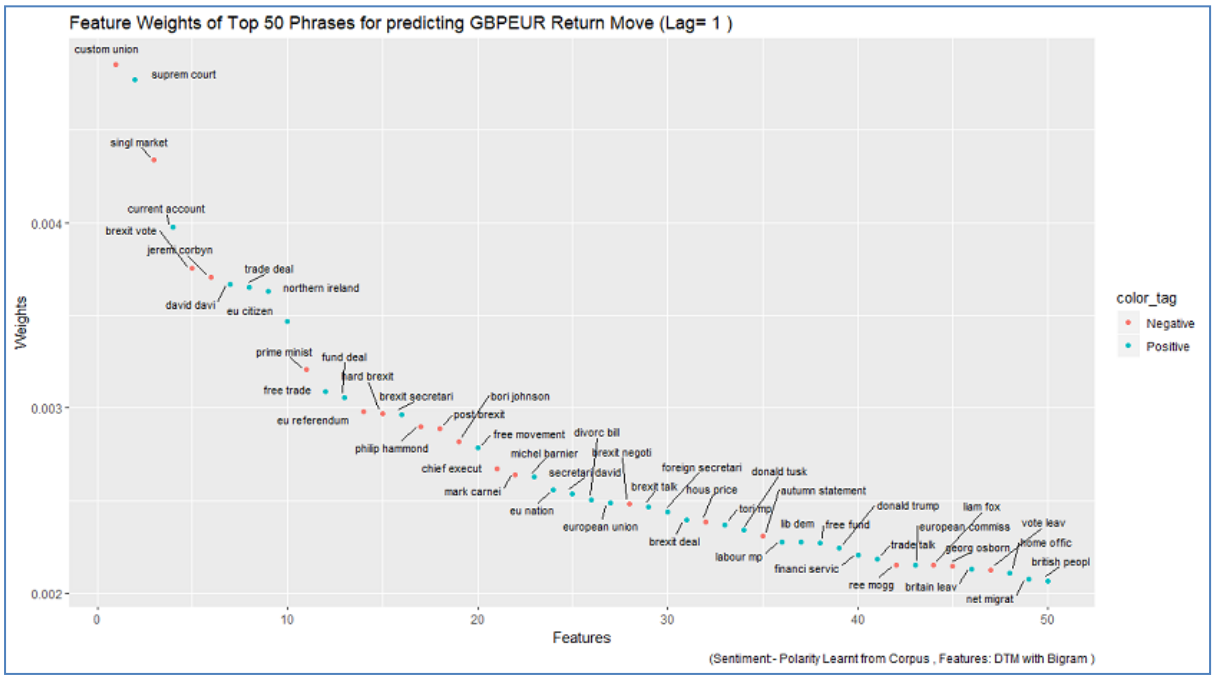
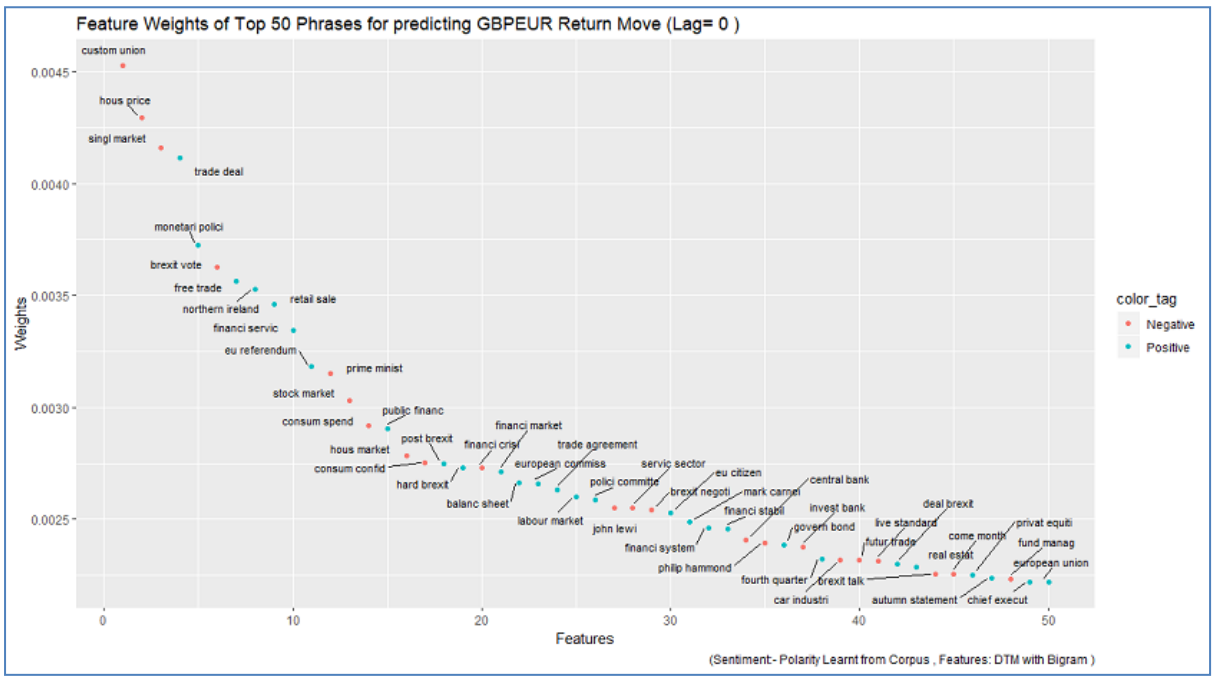
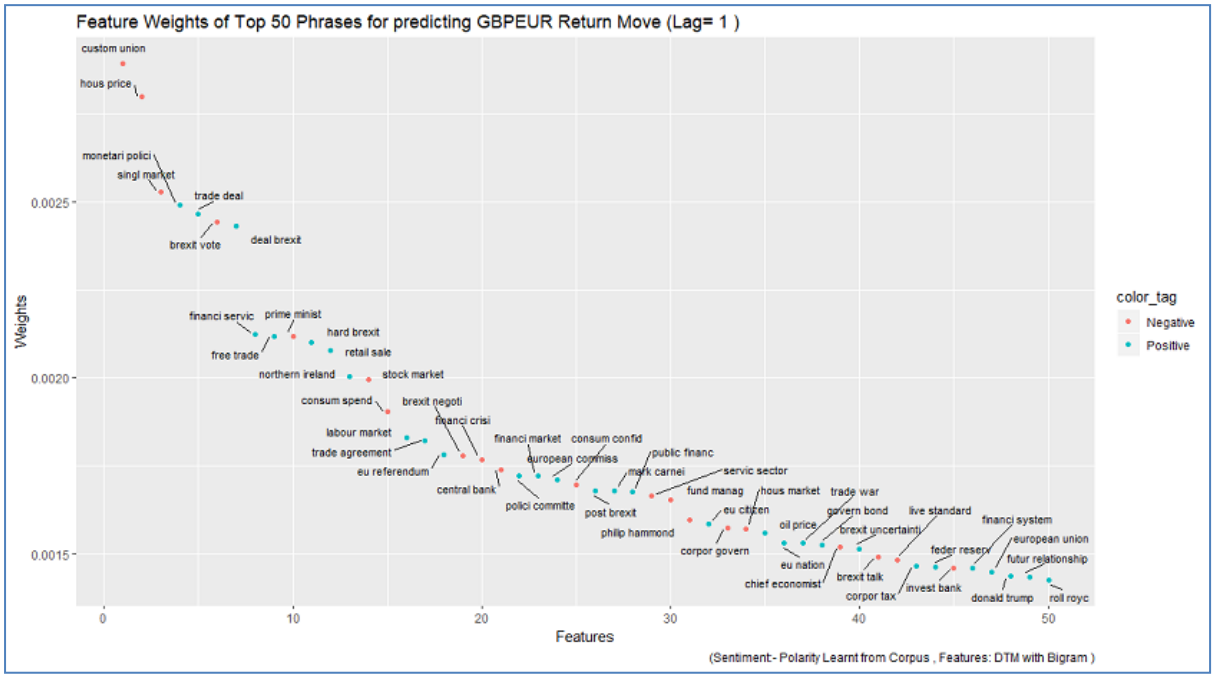


Figure 6.1.2 Top 50 Phrases from UK NON Tabloids, with Sentiment Score of Same Day and One Day Previous





6.2 Cointegration Outcome with GBPEUR 80:20 split

TABLE 6.2.1 Cointegration Results with 80:20 split

Split , 80:20, sequential

Cointegration for Price ~ Sentiment Score(Self Learning) using PREVIOUS Day

Target Variable	F-Statistic	Criticality	I(0)	I(1)
GBPEUR	6.488(**)	10%	4.04	4.78
GBPUSD	2.617	5%	4.94	5.73
FTSE 100	0.34	1%	6.84	7.84

Target Variable	T-Statistic	Criticality	I(0)	I(1)
GBPEUR	-3.021(*)	10%	-2.57	-2.91
GBPUSD	-2.093	5%	-2.86	-3.22
FTSE 100	-0.592	1%	-3.43	-3.82

Note : *, **, * indicates significance at 10%, 5% and 1% respectively**

Cointegration for Price ~ Sentiment Score(Self Learning) using SAME Day

Target Variable	F-Statistic	Criticality	I(0)	I(1)
GBPEUR	6.299(**)	10%	4.04	4.78
GBPUSD	2.588	5%	4.94	5.73
FTSE 100	0.331	1%	6.84	7.84

Target Variable	T-Statistic	Criticality	I(0)	I(1)
GBPEUR	-2.992(*)	10%	-2.57	-2.91
GBPUSD	-2.102	5%	-2.86	-3.22
FTSE 100	-0.585	1%	-3.43	-3.82

Note : *, **, * indicates significance at 10%, 5% and 1% respectively**

7. Conclusion

Using Brexit Corpus of Media Articles, covering 2016-June till 2018-November we looked at Sentiment Analysis, first with "off-the-shelve" LOUGHRAN word based lexicon, then augment Currency and Stock Price/Returns with the obtained daily Sentiment Score and went on to perform Econometric Analysis, to investigate cointegration and Long Run relationship between the various flavours of Prices(GBPEUR/GBPUSD/FTSE) and Sentiments.

We observed cointegration between GBPEUR Price and Sentiment, using Pesaran, Smith, Shin Bounds Test, at 1% significance, while using the entire Corpus . Subsequently, using Toda-Yamamoto procedure, we also observe that Sentiment Score Granger Causes GBPEUR Price at 5% significance.

For other Target, GBPUSD/FTSE we could not evidence any cointgeration.

Next, we repeat the entire exercise, the change being, Sentiments are computed with a Machine Learning Pipeline. In this case our feature becomes, Phrases(bigrams), instead of words, as in LOUGHRAN. We also determined Phrase Polarity and Weights using Machine Learning within the selected Training Corpus subset and using them computed the Daily Sentiment Scores in Out of Sample observations.

We observed that the Phrases generated using our Machine Learning model is more flexible and extensible, we used Bigrams for our exercise which can also extend to trigrams and more, if needed, whereas in Loughran the Lexicons are single Words.

We also notice, the Learnt Phrases are more "context aware", with "Single Market", "Customs Union", "Northern Ireland" is correctly identified by the Machine Learnt model, as phrases of prominence for Brexit. Loughran fixed lexicons doesn't have this flexibility , it identifies words like "warned", "cut", "crisis" as high important words. It is evident that the Brexit Lexicon created by Machine Learning Model is more informative and topical. The Machine Learnt Phrases also distinguishes between "Tabloid" and "Non-Tabloid/Serious" Media in terms of their Top Phrases and identified personalities.

This Self learning is carried out with the Corpus, using various splits for Training and subsequently investigated Cointegration and Long Run relationship in the Out of Sample data.

We perform Self learning, using various splits for Training and Out of Sample and subsequently investigated Cointegration and Long Run relationship in the Out of Sample data.
For Sequential 70:30 Splits, Cointegration and Long Run relationship evidenced for GBPEUR and Sentiment Score in the 30% Test Sample, Cointegration is observed at 5% criticality.
We also evidence one way Granger Causality between Sentiment Score => GBPEUR Price, using Toda-Yamamoto procedure, at 5% criticality.
Finally, we perform Forecasting to observe the effect Sentiment Score have on GBPEUR Price Forecast. Using 70:30 Split, for GBPEUR Price Forecast, we used the 30% Test Sample and build two Forecasting Models. First, with ARIMA to Forecast GBPEUR Price and then use Dynamic Regression including Sentiment Scores as variable. We observe Dynamic Regression with Machine Learnt Sentiment Score, have smaller Forecast Error, compared to ARIMA.
We attempt to simulate above experiment using LOUGHRAN Sentiment Score, 70:30 split doesn't indicate cointegration. We found cointegration and one way causality in 50:50 split with larger Test Samples. In this case, also, Dynamic Regression including Sentiment Score has lesser Forecast Error as compared to ARIMA Forecast of GBPEUR Price.

In conclusion, we saw that with our BREXIT Corpus, using Machine Learning, we could unearth "context aware" phrases and their measure of impact which is not available with LOUGHRAN based lexicon. For Split Samples, our Machine Learning based Sentiment Scores performs a bit better in uncovering cointegration and Long Run relationships, and identifies the same for various splits 70:30, 80:20 etc. where as LOUGHRAN, it is more constrained, at 50:50 and nonexistent when Test Sample becomes smaller in size .
Both flavors of Sentiment Scores improve the GBPEUR Forecast compared to its vanilla ARIMA counterpart.
Finally, this Machine Learning Pipeline, with small customisation, can be applied to other Socio Economic events and situations to start from an abstract space of Text Content (News/Articles etc.) and then build a more "context aware" and "informative" understanding and to further analyse and identify hidden relationships and impacts to more concrete Financial and Economic indicators.

Bibliography

Berden, K., Francois, J., Tamminen, S., Thelle, M. and P. Wymenga (2009): Non-Tariff Measures in EU-US Trade and Investment – An Economic Analysis, **Ecorys report prepared for the European Commission**, Reference OJ 2007/S180219493.

Bernanke, B. (1983): Irreversibility, Uncertainty, and Cyclical Investment, **Quarterly Journal Economics, 98**, 85-106.

Bloom, N., Bond, S. and J. van Reenen (2007): Uncertainty and Investment Dynamics, **Review of Economic Studies, 74**, 391–415.

de Brouwer, G. and N. Ericsson (1998): Modelling Inflation in Australia, **Journal of Business and Economic Statistics**, **16**, 433-449.

Campbell, J., S. Grossman and J. Wang (1993): Trading volume and serial correlation in stock returns, **Quarterly Journal of Economics, 108**, 905–939.

Cortes C. and Vapnik V(1995): "Support vector networks", **Machine Learning**, **20,**1-25.

Crowley, M., Exton, O. and L. Han (2018): Renegotiation of Trade Agreements and Firm Exporting Decisions: Evidence from the Impact of Brexit on UK Exports, **Cambridge-INET Working Paper** No: 2018/10, Cambridge, UK.

Das, S. R., and M. Y. Chen (2007): Yahoo! for Amazon: Sentiment Extraction from Small Talk on the Web, **Management Science**, **53**,1375–88.

D. Blei, A.Ng, and M. Jordan (2003) : Latent Dirichlet Allocation, **Journal of Machine Learning Research, vol. 3,** 993–1022

Davidson, J., D. F. Hendry, F. Srba and S. Yeo (1978): Econometric Modelling of the Aggregate Tise-Series Relationship Between Consumers' Expenditure and Income in the United Kingdom, **The Economic Journal**, **88**, 661-692.

De Long, J. B., A. Shleifer, L. H. Summers, and R. J. Waldmann (1990): Noise trader risk in financial markets, **Journal of Political Economy**, **98**, 703–738.

Dixit, A. and R. S. Pyndick (1994): Investment under Uncertainty, Princeton University Press, Princeton, USA.

Ferguson, N. J., D. Philip, H. Y. T. Lam, and J. M. Guo (2015): Media content and stock returns: the predictive power of press, **Multinational Finance Journal**, **19**, 1–31.

Fraiberger, S. P., D. Lee, D. Puy and R. Ranciere (2018): Media Sentiment and International Asset Prices, **IMF Working Paper**, No. WP/18/274, International Monetary Fund, Washington D. C., USA.

Górnicka, Lucyna: (2018): Brexit Referendum and Business Investment in the UK, **IMF Working Paper,** No. WP/18/247, International Monetary Fund, Wishington D.C.

Johnman, M., B. J. Vanstone and A. Gepp (2018): Predicting FTSE 100 returns and volatility using sentiment analysis, **Accounting and Finance**, **58**, S1, 253-274.

Hendry, D. F. and N. Ericsson (1991): An Analysis of UK Money Demand in 'Monetary Trends in the United States and the United Kingdom by Milton Friedman and Anna J. Schwartz', **American Economic Review**, **81**, 8-39.

Klein, F. C., and J. A. Prestbo (1974): **News and the Market**, Regnery, Chicago, IL.

L. Breiman (2001): Random Forests,**Machine Learning**, **45,**5–32.
Li, F. (2008): Annual Report Readability, Current Earnings, and Earnings Persistence, **Journal of Accounting and Economics**, **45**,221–47.
Li, F. (2010): Textual Analysis of Corporate Disclosures: A Survey of the Literature, **Journal of Accounting Literature**, **29**,143–65.
Tetlock, P. C. (2007): Giving content to investor sentiment: the role of media in the stock market, **The Journal of Finance**, **62**, 1139–1168.
Tetlock, P. C., M. Saar-Tsechansky, and S. Macskassy (2008): More than words: quantifying language to measure firms' fundamentals, **The Journal of Finance**, **63**, 1437–1467.
Toda, H. Y. and T. Yamamoto (1995): Statistical inferences in vector autoregressions with possibly integrated processes, **Journal of Econometrics**, **66**, 225-250.