On XLE index constituents' social media based sentiment informing the index trend and volatility prediction

Frédéric Maréchal[1,2], Daniel Stamate[1], Rapheal Olaniyan[1] and Jiri Marek[1]

[1] Data Science & Soft Computing Lab, and
Department of Computing, Goldsmith College, University of London, UK
[2] Santander Bank, UK
freddy.marechal@gmail.com

Abstract. Collective intelligence represented as sentiment extracted from social media mining found applications in various areas. Numerous studies involving machine learning modelling have demonstrated that such sentiment information may or may not have predictive power on the stock market trend. This research investigates the predictive information of sentiment regarding the Energy Select Sector related XLE index and of its constituents, on the index and its volatility, based on a novel robust machine learning approach. While we demonstrate that sentiment does not have any impact on any of the trend prediction scenarios investigated here related to XLE and its constituents, the sentiment's impact on volatility predictions is significant. The proposed volatility prediction modelling approach, based on Jordan and Elman recurrent neural networks, demonstrates that the addition of sentiment or sentiment moment reduces the prediction root mean square error (RMSE) to about one third. The experiments we conducted also demonstrate that the addition of sentiment reduces the RMSE for 24 out of the 36 stocks/constituents, representing 87.9% of the index weight. This is the first study in the literature relating to the prediction of the market trend or the volatility based on an index and its constituents' sentiment.

Keywords: Sentiment analysis, Machine learning, Stock market prediction, Volatility, Imputation, Feature selection, Random forest, SVM, Elman and Jordan recurrent neural networks

1 Introduction

Stock market trend prediction is at the centre of investment strategies. The fluctuation of assets and their predictability has been studied by many researchers over several decades. Supporters of the Efficient Market Hypothesis (EMH) and the random walk theory consider that it is impossible to predict future trends. However, some studies have proven the existence of volatility clustering [3] and regime change under certain conditions that made the prediction of market return possible, or at least partially.

Vaiz and Ramaswami [12] investigated the prediction power of technical analysis indicators, on their own, to estimate the future price of a stock traded on an exchange. For this, the daily OHLC price and volume data for the six highest market capitalisation companies of the NSE were gathered for a period spanning from January 2012 to December 2015. Twenty–two technical indicators (e.g. RSI, EMA, MACD, etc.), and three supervised classification tree models, including C5.0, were selected to perform test predictions. The close price and the technical indicators were used respectively as the response and explanatory variables. The author concluded that, in the best–case scenario, these models achieved 85% of the accuracy in predicting the market trend.

The most noticeable feature that has lately attracted researchers' attention is the study of the impact of sentiment-induced variables. Meesad and Li [7] presented a methodology involving the generation of a sentiment score for each of the 4,622 tweets under analysis, based on of bags-of-words, the *SentiWordNet* corpus as well as a feature weighting based on a Term Frequency–Inverse Document Frequency (TF–IDF) methodology. The paper claimed that fitting SVM models, assessed in a Leave–One–Out (LOO) cross–validation, yielded 93.4% prediction test performance.

A more general approach than both above methodologies was developed by Halgamuge [5]. The research presented a model that used both news releases and technical indicators as predicators to enhance the predictability of the daily stock price trends. The experiment consisted of i) building seven technical indicators and ii) tokenising news articles (both company and market specific) to serve as attributes in a SVM model. The predicted response variable was the daily stock price of BHP Billiton Ltd (BHP.AX). The sentiment scoring construction was very similar to the previous approach. The model fit was based on a training and validation set. The prediction test performance showed a 70.1% test accuracy rate for a model using the price, the company and market news.

The above-mentioned works focused on employing machine learning methods to establish the predictive power of sentiment on the stock market trend. However, they failed short of analysing the statistical significance and the stability of the prediction accuracy. The research carried out by Gilbert and Karahalios [4] offered a more robust statistical approach. The study was carried out on a dataset of 20 million posts from the LiveJournal website. The authors proved, using a Granger–causal framework and a Monte Carlo simulation, that negative sentiment tended to influence negatively the S&P500 index. For this, the researchers used a specialised Live–Journal corpus and classified articles' sentences to distinguish between anxious, worried, nervous and fearful versus not anxious sentiment. Their study revealed that high anxiety levels impact negatively on the market. While this methodology demonstrated a more robust statistical approach, there were some limitations in the tools employed to establish the impact of negative sentiment on the stock trend. The linear Granger causality test assumed that the models under analysis were linear. Moreover, the authors recognised that it was sensitive to non–stationary time series, where the mean, variance and autocorrelation varied with time. Besides, the residuals were not normally–distributed.

In their paper, Olaniyan et al. [9] critically investigated the suitability of using a Monte Carlo simulation as a validation tool for offsetting the shortcomings of the linear Granger Causality test. Using a Monte Carlo inverse transform and a bootstrap sampling method, the authors proved that the empirical and expected F–Statistic were still significantly apart. The researchers also conducted a non–parametric statistical

test, developed by Baeck and Brock [1], on the residuals of the VAR models. This test concluded that the original findings in Gilbert and Karahalios [4], relative to the predictive power of the Anxiety Index on the stock market trend, were biased by the presence of residuals' heteroscedasticity.

Olaniyan et al. [9] inferred that, contrary to the results obtained by Gilbert and Karahalios [4], the Anxiety Index did not possess any significant predictive information on the stock market. In the light of the above conclusions, the researchers re–oriented the previous experiment. First, they introduced a new set of attributes: lagged volatilities and positive/negative sentiment variables. The volatilities were generated via an exponential GARCH (1,1) process (a.k.a. EGARCH). Second, the authors abandoned the Anxiety Index proposed by Gilbert and Karahalios [4]. Instead, they replaced it by the Downside Hedge Twitter Sentiment indicator1[1]. Third, the researchers used i) a non–parametric and nonlinear approach and ii) a hybrid GARCH coupled with artificial neural networks (NN) to test the prediction power of the positive and negative sentiments. As a final experiment, Olaniyan et al. [8] explored the predictive power of sentiment on volatility Q_t. They used an EGARCH lagged volatilities Q_{t-1} and Q_{t-3}, coupled with the positive and negative sentiment P_{t-1}, P_{t-2} and N_{t-1}, N_{t-2} as attribute variables into a feed–forward NN, a Jordan and an Elman recursive NN. The authors concluded that: (i) past volatility was the main contributor to predicting future volatility, (ii) positive sentiment was negatively correlated with future volatility, and (iii) negative sentiment seemed to have the least influence on volatility.

In addition to their clear significant achievements, the above-mentioned methodologies present also a series of limitations. First, the bag-of-words approach does not take the context into consideration. Second, general corpus relies on non–domain specific lexicon. Third, they only consider the impact of the sentiment on the trend prediction, for a small set of technical indicators. Indeed, there is a risk that selecting a small set of technical indicators could generate over optimistic results when it comes to the sentiment true predictive power. Moreover, the use of cross-validation, as done in Meesad and Li [7] is incompatible with time–series prediction. Indeed, the path dependency nature of time–series forbid the leaking of future prices from the validation/test sets into the training set [10]. Finally, none of these above frameworks consider constituents of the analysed index, and sentiment on these constituents. To our knowledge, there is currently no study in the literature relating to the prediction of the market trend or the volatility based on an index and its constituents. The work we propose here is the first to do so, and we investigate the XLE index (US energy index) with its constituents and sentiments on constituents.

First, the paper considers the effect of the sentiment prediction power on the index trend. Second, it examines the effect of sentiment on the predictability of each of the XLE index's constituents trend. Third, it analyses whether the reconstruction of the index prediction when the constituents' sentiment is added, improves the overall index trend predictability. Finally, the research examines the impact of sentiment and sentiment momentum on the volatility predictability both at the index and constituents level.

[1] Available at www.downsidehedge.com/twitter-indicators/

2 The data and modelling framework

2.1 The data modelling

For the purpose of this study we obtain historical price/volume information from *Yahoo!Finance,* as well as the sentiment data from *Quandl*[2]. From the price (P_t) information, the daily return is produced, i.e. $\log(P_{t+1}/P_t)$. This serves as the base for the generation of the categorical response variable, which takes two labels *Up* and *Down*. The price/volume information is used to generate the explanatory variables as proxies for technical indicators such as, but not limited to, the momentum indicators (e.g. Rate of Change), trend indicators (e.g. Simple Moving Average), volatility indicators (e.g. Average True Range). The sentiment (S_t) and sentiment momentum ($SM_t = S_t – S_{t-1}$) are also added to the model. In total more than 50 variables and their lags are retained in the original model.

For the volatility prediction case, the data modeling is different. Indeed, according to Brownlees et al. [2], standard statistical tools for forecasting volatility are the GARCH models. Since the volatility (the response variable) is unobserved, a reasonable proxy for the volatility is calculated from the square of the daily return [10]. The explanatory variables are represented by the lag versions of the exponential GARCH (1,1) process volatility and the volume, for the period t-1, t-2 and t-3. These variables were chosen following the positive results produced by Olaniyan et al. [9].

2.2 The trend prediction methodology

The trend prediction methodology is summarised in the following steps:

- The preprocessing step includes the response variable class rebalancing, zero variance explanatory variable elimination, and highly correlated explanatory variables exclusion. When the supervised model requires it, a Box-Cox transformation, data imputation or normalisation are performed [6, 11].
- One of the feature selection methodologies applied uses a wrapper method and a filter method to reduce the space of explanatory variables for each machine learning model, over a 20-year long training data set. Fig. 3 illustrates the outcome of applying the wrapper method based on random forest with 500 trees with Gini index attribute selection criterion, and the outcome of applying the filter method based on Relief and permutation test [6].
- The machine learning models that we produced were based on 8 algorithms including linear and quadratic discriminant analysis, bagging, random forest, support vector machine, multi-layer perceptron network with weighted decay, and the Elman and Jordan recurrent neural networks.

[2] Quandl collect content of over 20 million news and blog sources real–time. They retain the relevant articles and extrapolate the sentiment. The sentiment score is generated via a proprietary algorithm that uses deep learning, coupled with a bag-of-words and n-grams approach.

- For each index/constituent, the optimised model (and its feature selection list), that produces the highest test accuracy rate, is retained as part of the base scenario, which contains no sentiment variable. The test accuracy rate is obtained from the average of test accuracy rates of 100 sliding time windows. Fig. 1 illustrates this process. Each contiguous and non-overlapping time window contains 220 training records, 66 validation records and 5 test records. The model is trained and optimised (on a pre-defined hyper parameter grid) using the training and validation sets, and evaluated on the test set, for each time window. The latter slides 5 days, the process is repeated 100 times, and performances on the test set are averaged.
- As the sentiment is not available for the XLE index itself, a proxy for the index's sentiment is constructed from the sentiments of the index's constituent, denoted SIS_t. The proxy index sentiment is simply defined as the sum of the products of the sentiment scores (SS_i) and the weights (W_i) of the index constituents, for each day t, as shown below.

$$SIS_t = \sum_{i=1}^{n} (SS_i * W_i)_t, \text{ where n is the number of index's constituent} \quad (1)$$

- Numerous scenarios are created involving the sentiment (S_t), the sentiment momentum ($SM_t = S_{t-} - S_{t-1}$), and their respective lags at the top of the *Base* scenario. Each scenario produces a scenario accuracy rate for each constituent, named *SAi*.
- A proxy index weighted accuracy rate (PIWA) is generated for the constituents, both for the base and the sentiment scenarios. The PIWA, defined in (2), is the sum of the products of the constituent weights (W_i) and their accuracy rate (SA_i).

$$PIWA = \sum_{i=1}^{n} (SA_i * W_i), \text{ where n is the number index' constituents} \quad (2)$$

- At the end of the process, the index test accuracy rate, named IAR_{sent}, generated from technical indicators and the proxy sentiment (SIS_t) variables, is compared to the proxy index weighted test accuracy rate in (2). When the proxy index test accuracy rate (PIWA) is greater than index test accuracy rate accuracy rate (IAR_{sent}), we then conclude that the sentiment applied to each XLE constituents has more predictive power compared to the sentiment applied directly to the index.

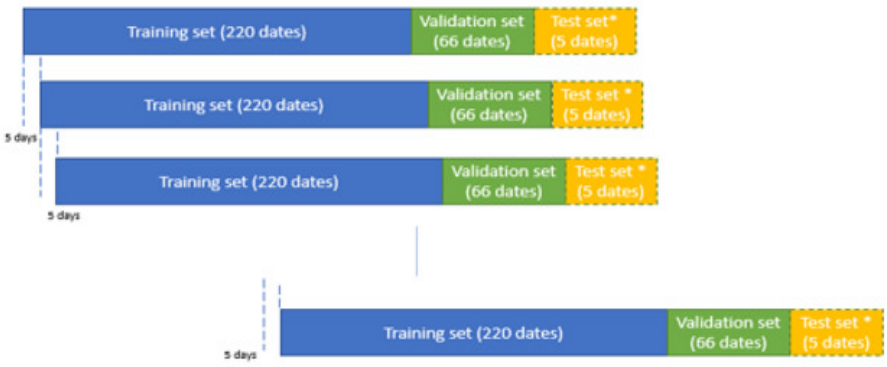
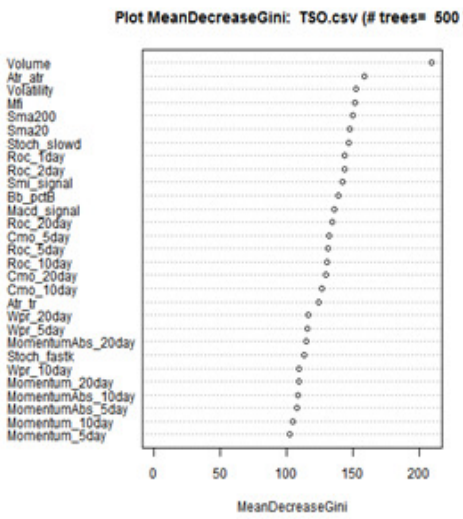


Fig. 1. Model training, optimization and evaluation using a time sliding window approach, with average model performance over the 100 repetitions



	perm.standardized
Roc_20day	1.604437113
Wpr_20day	1.554924154
Cmo_10day	1.270080985
Roc_5day	1.119483781
Cmo_5day	1.087991962
Bb_pctB	1.032308008
Momentum_10day	1.008765804
Wpr_10day	0.904513715
MomentumAbs_10day	0.529141177
Volatility	0.440550265
Macd_signal	0.427807479
Smi_signal	0.307154621
Momentum_20day	0.219783662
Atr_tr	0.173885894
Mfi	0.035131085
Stoch_fastk	0.019966381
Wpr_5day	-0.0384412
Sma200	-0.069096232
Momentum_5day	-0.125518065
Roc_1day	-0.137689472
Cmo_20day	-0.195720512
Roc_10day	-0.228013808
Roc_2day	-0.303251854
Stoch_slowd	-0.343510548
Sma20	-0.539738255
MomentumAbs_5day	-0.628299126
Atr_atr	-0.631340653
Volume	-0.859099049
MomentumAbs_20day	-1.078179694

Fig. 2. Feature selection. Left: Wrapper method based on random forest with 500 trees with Gini index attribute selection. Right: Filter method based on Relief and permutation test

2.3 The volatility prediction methodology

The volatility prediction methodology is summarised as follows:

- The feature selection was greatly simplified, as the number of explanatory variable is relatively small. The aim is to eliminate correlated predictors for those showing a correlation threshold of 95% or above.
- The best average test result is obtained by selecting the smallest RMSE generated by a Jordan and Elman Recurrent Networks on a sliding time window, as per (3) below. This constitute the *Base* scenario.

$$A_i = \text{Min}\,(\text{RMSE}_{i\,\text{Jordan}}, \text{RMSE}_{i\,\text{Elman}}),\ i \text{ represents an index or a constituent} \qquad (3)$$

- For each index/constituent feature selection and the selected neural network algorithm (3), the same test results are generated with the added sentiment and its lags. A parallel experiment is carried out with the sentiment momentum and its lags.

The base test results (no sentiment) and scenario test results (containing sentiment or sentiment momentum) are compared to establish whether the sentiment adds prediction power on the volatility prediction.

3 The results

This section details impact of sentiment on the trend and volatility test prediction. The test accuracy rate is generated for the trend prediction, whereas the RMSE is computed for the volatility prediction.

3.1 The trend prediction results

Table 1 and Table 2 present the summary of the results obtained for the trend prediction test accuracy rates. The greyed row identifies the results obtained for the index. The following rows correspond to the results generated for each of the three most important index's constituents. The total row displays the sum of weighted test accuracy rates for all constituents. Each table details respectively the selected 'best' feature selection and model applied to the index and the index's constituents. The feature selection column in these two tables shows 'f' or 'w' to respectively represent a filter or wrapper feature selection strategy. The 'Model' column lists the model selected for the index and each index's constituent. The base column shows the original trend prediction accuracy rates without sentiment predictors.

Table 3, Scn.1/2/3/4 represent the scenarios for the test accuracy rates corresponding to the trend prediction when the sentiment at t-1, t-2, t-3 or t-4 is present in the model. Table 4, Scn.5/6/7/8 represent the scenarios for the test accuracy rates for the trend prediction when the sentiment momentum is accumulated, i.e. SM_t, $SM_t + SM_{t-1}$, $SM_t + SM_{t-1} + SM_{t-2}$ and $SM_t + SM_{t-1} + SM_{t-2} + SM_{t-3}$.
The results indicate that the additions of the sentiment or the sentiment momentum have, in this experiment, not a favorable impact on the index trend prediction.

Table 1. Trend prediction under different sentiment scenarios

Code	Weight	Feat. Select.	Model	Base	Scn.1	Scn.2	Scn.3	Scn.4
XLE		f	elman	**0.54**	**0.53**	**0.50**	**0.49**	**0.52**
XOM	16.80%	w	pda	0.09	0.09	0.09	0.09	0.09
CVX	14.81%	w	svm	0.08	0.08	0.08	0.08	0.08
SLB	8.19%	w	mlp	0.05	0.04	0.04	0.04	0.04
…								
Total				**0.54**	**0.53**	**0.53**	**0.53**	**0.53**

Table 2. Trend prediction under different sentiment momentum scenarios

Code	Weight	Feat. Select.	Model	Base	Scn.1	Scn.2	Scn.3	Scn.4
XLE		f	elman	**0.49**	**0.45**	**0.45**	**0.48**	**0.48**
XOM	16.80%	w	pda	0.09	0.09	0.09	0.09	0.09
CVX	14.81%	w	svm	0.08	0.08	0.08	0.08	0.08
SLB	8.19%	w	mlp	0.04	0.04	0.04	0.04	0.04
…								
Total				**0.53**	**0.52**	**0.52**	**0.52**	**0.52**

3.2 The volatility prediction results

Table 3 shows the test RMSE for each constituent/index and their best selected model, i.e. either Elman or Jordan recurrent neural networks. It also shows the improve-

ment/deterioration of the RMSE under the different scenarios. Column 4 lists the test RMSEs when there is no sentiment variable. Column 5 and 7 show the test RMSEs for scenarios where the sentiment and sentiment momentum are present, respectively. Columns 6 and 8 represent the delta of test performance RMSEs between one of the scenario (column 5 or 7) and the original RMSE (column 4).

Table 3. Volatility RMSE under different scenarios

1	2	3	4	5	6	7	8
Code	Weight	(E)lman/ (J)ordan	Base Rmse	Scenario1 Rmse	Scneario1-Base	Scneario2 Rmse	Scneario2-Base
EOG	4.64%	J	0.002467	0.001921	-0.000545	0.001970	-0.000496
HAL	3.66%	J	0.000695	0.000552	-0.000143	0.000499	-0.000196
NBL	1.58%	J	0.001819	0.001373	-0.000446	0.001187	-0.000632
OXY	3.14%	J	0.001513	0.001312	-0.000201	0.001319	-0.00019411
APA	1.91%	E	0.003884	0.004511	0.000627	0.003371	-0.000513
APC	2.98%	E	0.003164	0.001604	-0.001560	0.002468	-0.000696
BHI	2.43%	E	0.002403	0.001289	-0.001114	0.003156	0.000753
CHK	0.47%	E	0.024763	0.027091	0.002328	0.023402	-0.001362
COG	1.52%	E	0.001145	0.004158	0.003013	0.005131	0.003985
COP	3.12%	E	0.002378	0.002303	-0.000075	0.001902	-0.000476
CVX	14.81%	E	0.002511	0.000776	-0.001735	0.002209	-0.000302
CXO	1.30%	E	0.003664	0.001202	-0.002462	0.003506	-0.000158
DVN	1.88%	E	0.005442	0.005405	-0.000036	0.006627	0.001185
EQT	0.79%	E	0.003731	0.002250	-0.001481	0.003624	-0.000107
FTI	0.94%	E	0.003178	0.002883	-0.000296	0.002658	-0.000520
HES	1.40%	E	0.003978	0.001374	-0.002603	0.005157	0.001179
HP	0.58%	E	0.003612	0.001188	-0.002424	0.002332	-0.001280
KMI	2.65%	E	0.004272	0.003467	-0.000805	0.004209	-0.000063
MPC	1.70%	E	0.002575	0.007050	0.004475	0.003366	0.000791
MRO	1.20%	E	0.005308	0.004910	-0.000398	0.020095	0.014788
MUR	0.48%	E	0.010084	0.013026	0.002942	0.012608	0.002524
NFX	0.60%	E	0.001124	0.005571	0.004447	0.007022	0.005897
NOV	1.25%	E	0.003555	0.004473	0.000918	0.004815	0.001260
OKE	0.80%	E	0.001328	0.005660	0.004332	0.003316	0.001989
PSX	2.55%	E	0.002021	0.000391	-0.001630	0.002315	0.000294
PXD	4.78%	E	0.002676	0.001096	-0.001580	0.003810	0.001134
RIG	0.37%	E	0.009760	0.012175	0.002415	0.010180	0.000420
RRC	0.68%	E	0.004048	0.004161	0.000113	0.002834	-0.001214
SE	2.53%	E	0.002238	0.000493	-0.001745	0.001583	-0.000655
SLB	8.19%	E	0.002676	0.000811	-0.001865	0.002879	0.000202
SWN	0.46%	E	0.012091	0.013626	0.001535	0.013101	0.001010
TSO	2.22%	E	0.002471	0.001081	-0.001390	0.002812	0.000341
VLO	2.84%	E	0.002300	0.000565	-0.001735	0.001838	-0.000462
WMB	1.87%	E	0.003640	0.061584	0.057944	0.009966	0.006325
XEC	0.86%	E	0.002758	0.001057	-0.001701	0.002920	0.000162
XLE		E	0.002014	0.000646	-0.001367	0.000691	-0.001323
XOM	16.80%	E	0.002274	0.000307	-0.001967	0.002162	-0.000112

The analysis of Table 2 indicates that the presence of sentiment provokes a reduction in test RMSE for 24 stocks out of the 36. These 24 stocks represent 87.87% of total index weight. The sum of the constituents' weighted RMSE is 0.0029. There is also a reduction in test RMSE for XLE index. The RMSE moves from 0.002014 to 0.000646 in the scenario 1, where S_{t-1} is added to the model. When the sentiment momentum, SM_{t-1} and SM_{t-2} (Scenario 2), is added as a predicator, there is a reduction in test RMSE for 19 stocks out of the 36. These 19 stocks represent 65.42% of total index weight. The sum of the constituents' weighted RMSE is 0.003191. There is also a reduction in test RMSE for XLE index. The RMSE decreases from 0.002014 to 0.000691.

These results indicate that sentiment has a significant impact on the index volatility prediction, decreasing the RMSE at about one third of the RMSE value obtained initially when the sentiment was not used. The sentiment variable (S_{t-1}) seems to have a greater impact in increasing prediction than the sentiment momentums (SM_{t-1} + SM_{t-2}), at both at the index and the constituents' level.

Although, the proxy for the index volatility was not re-generated from the constituent, 87% of the index weight show an improvement of volatility prediction after the addition of sentiment (S_{t-1}). This result suggests that the proxy index volatility prediction, generated from the constituents, is improved when sentiment is present.

4 Conclusion and future work

This work proposes a new approach to verify the prediction power of social media inferred sentiment when predicting the XLE index, and its 36 constituents trend and volatility, over a 5 years period. Although the trend prediction methodology was based on a robust time-series machine learning approach involving i) a "2-way" feature selection, ii) a sliding times windows in replacement for cross-validation, iii) a basket of 8 machine learning algorithms, including recurrent neural network, it did not produce comparable performance results achieved in the case of researches using other data.

However, the second approach involving i) lagged EGARCH volatility, and ii) lagged volume variables feeding into a Jordan and an Elman recurrent neural networks, on the data of the XLE index and each of its 36 constituents, proved to yield far better results. On the index alone, the RMSE drops to 0.000646 (about one third) when past sentiment was added. Furthermore, more than 65% of the index constituents also show a reduction in the RMSE, when the sentiment variable is added.

This second approach that we propose demonstrates that the predictive power of sentiment on the future volatility should be investigated further. First, the constituents' volatility prediction should be generated using the following formula (4), where $w_1 \ldots w_2$ represent the weights and $\sigma_{11} \ldots \sigma_{1N}$ are the stock volatilities.

$$\sigma_w^2 = w^T S w = [w_1, \ldots, w_N] \begin{bmatrix} \sigma_{11} \ldots \ldots \sigma_{1N} \\ \sigma_{21} \ldots \ldots \sigma_{2N} \\ \ldots \ldots \ldots \ldots \ldots \\ \sigma_{1N} \ldots \ldots \sigma_{NN} \end{bmatrix} \begin{bmatrix} w_1 \\ \ldots \\ \ldots \\ w_N \end{bmatrix} \quad (4)$$

The main source of complexity of this study comes from the high dimensionality of explanatory variables, generated by technical analysis indicators, the sentiment scores, and especially the lags of these variables. The novelty emanates from the analysis of the trend and volatility predictions for each constituent of a stock index, using sentiment scores over a 5 year period.

As future potential extensions of this work, the volatility prediction results (see Subsection 3.2) could be compared to the index volatility prediction when augmented with the sentiment variables, investigating whether sentiment has more prediction power on the index volatility than on the weighted sum of the index constituents' volatility. Second, the analysis could be extended to other indices, across different sectors and basket size, such as the XLF (US financial index) or the S&P500. A third line of potential extension of the research is to consider constituent weights as being time-dependent, rather than being static as it is in the framework presented here. We currently look into investigating these directions.

References

1. Baek E., Brock W.: A general test for nonlinear Granger causality: bivariate model. https://www.ssc.wisc.edu/~wbrock/Baek%20Brock%20Granger.pdf, Working paper. Accessed: 15–Mar–2017.
2. Brownlees C., Engle R., Kelly B.: A practical guide to volatility forecasting through calm and storm, The Journal of Risk. Volume 14/Number 2, pp. 3-22 (2012).
3. Cont R.: Empirical properties of asset returns: stylized facts and statistical issues, Quantitative Finance, Volume 1, p223-236 (2001).
4. Gilbert E., Karahalios K.: Widespread worry and the stock market In Proceedings of the 4th International Conference on Weblogs and Social Media, pp. 58–65 (2010).
5. Halgamuge S.K.: Combining News and Technical Indicators in Daily Stock Price Trends Prediction. Conference: Advances in Neural Networks – ISNN 2007. 4th International Symposium on Neural Networks. ISNN 2007 Proceedings. Part III (2007).
6. Kuhn M., Johnson K, Applied Predictive Modeling, Springer, 2013.
7. Meesad P., Li J.: Stock trend prediction relying on text mining and sentiment analysis with tweets. In: 2014 Fourth World Congress on Information and Communication Technologies (WICT), pp. 257–262 (2014).
8. Olaniyan R., Stamate D., Lahcen O., et al.: Sentiment and stock market volatility predictive modelling – A hybrid approach, In: Gaussier E, Cao L, Gallinari P et al., Data Science and Advanced Analytics, 36678 2015, ACM/IEEE International Conference. Paris (2015).
9. Olaniyan R., Stamate D., Logofatu D.: Social web–based anxiety index's predictive information on S&P 500 revisited, Proceedings of the 3rd Intl. Symposium on Statistical Learning and Data Sciences, LNCS, Springer (2015).
10. Rechenthin M.D.: Machine–learning classification techniques for the analysis and prediction of high–frequency stock direction. Thesis. University of Iowa (2014).
11. Triacca U.: On the variance of the error associated to the squared return as proxy for volatility, Applied Financial Economics Letters, 3, pp. 255-27, In: Giles D E Some Properties of Absolute Returns as a Proxy for Volatility, Econometrics Working Paper EWP0706, University of Victoria, ISSN 1485-6441, pp. 3 (2007).
12. Vaiz J.S., Ramaswami M.A.: Study on Technical Indicators in Stock Price Movement Prediction Using Tree Algorithms. American Journal of Engineering Research– (AJER). Volume 5, Issue 12:207–212 (2016).