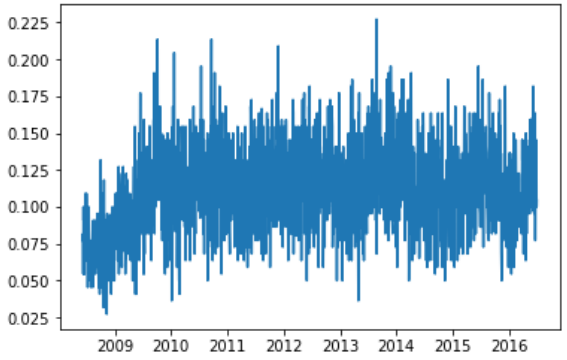
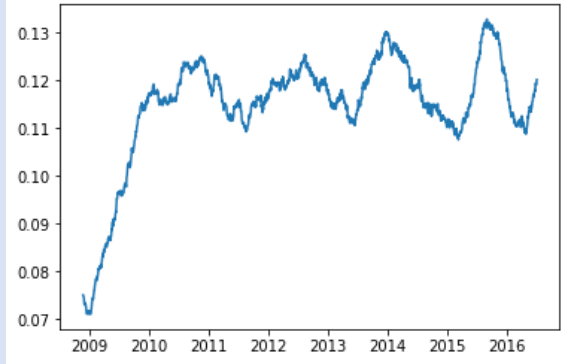
ABSTRACT

Efficient market hypothesis states that it's impossible to make consistent profits out of public information, due to it being already incorporated into prices. Hence, the only advantages comes from how fast we are able to distinguish the market sentiment and predict the correct change. In the present project we conduct natural language processing of news headlines in order to see whether incorporation of the sentiment embedded in them into the trading strategy leads to the improvement of its performance. We find that returns are significantly harder to predict than volatility and argue that using sentiment scores vs the unsupervised features selected leads to an improvement in prediction and, hence, the performance of our strategy.

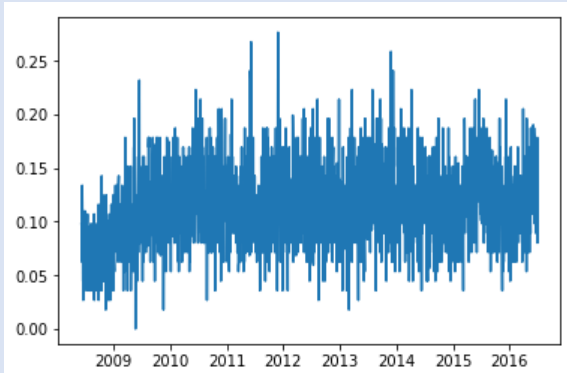
Motivation



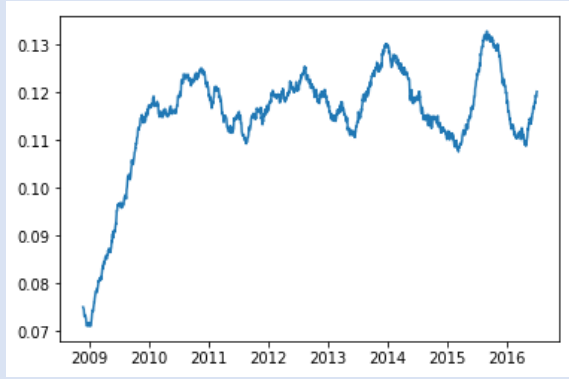
'Positive' score history on Reddit News for BOW with threshold = 50



'Positive' score history on Reddit News for BOW with threshold = 50 smoothed over 120 days

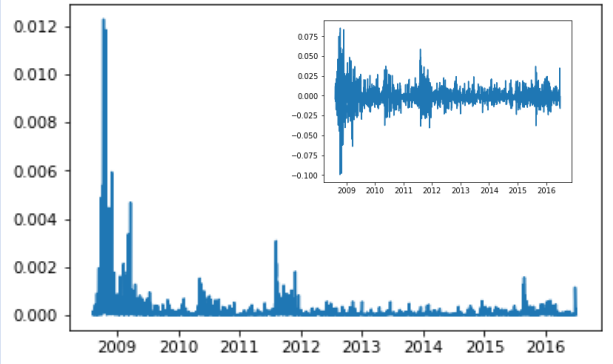


'Positive' score history on Reddit News for BOW with threshold = 50

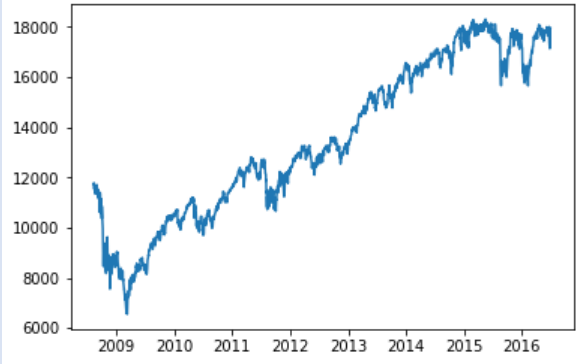


'Positive' score history on Reddit News for BOW with threshold = 500 smoothed over 120 days

Observations:
- Positive sentiment has been increasing since 2008
- Picture doesn't significantly differ for Bag of Words with 50 and 500 words as threshold



Realized Variance history of Dow Jones Index



Price history of Dow Jones Index

Linear Regression on the whole dataset (for Bag-of-Words with threshold of 500) for the volatility and returns gives the following results:

Variance

VAR	COEF	P-VALUE
Vol(-1)	0.1743	0.000***
Return(-1)	-0.0066	0.000***
'Positive'	-0.0008	0.085*
'Surprise'	0.0021	0.049**

Return

VAR	COEF	P-VALUE
Vol(-1)	0.9990	0.045**
Return(-1)	-0.1047	0.000***
'Anger'	0.0314	0.067*
'Surprise'	-0.0436	0.068*

- Volatility is sticky
- Returns reverse
- Sentiments matter

Feature Selection and Model Fitting

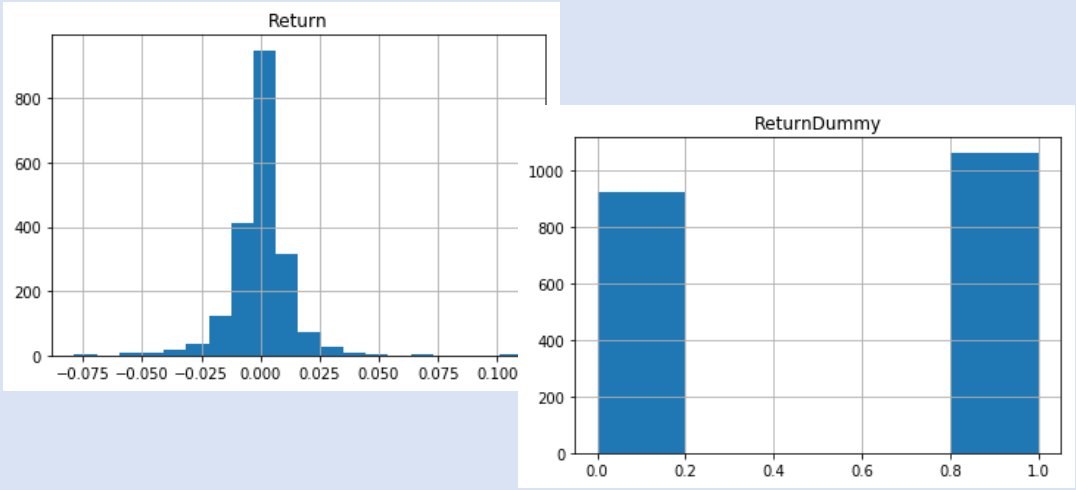
Variables for Predictions:
1. Next Day Return
2. Next Day Realized Volatility

Models for Feature Selection:

1. PCA
2. Factor Analysis
3. LDA
4. *Sentiment Scoring: Positive, Negative, 'Anger', 'Anticipation', 'Disgust', 'Fear', 'Joy', 'Sadness', 'Surprise', 'Trust'*

Models for Prediction:
1. OLS
2. Ridge
3. Lasso
4. Random Forest

Well-distributed data enables **balanced** predictions:



TASK: MINIMIZE PREDICTION ERROR (MSE)
HOW THE TASK LEADS TO BETTER RESULT: *BETTER PREDICTION ENABLES BETTER PROFITS*
RESULT: TRADING STRATEGY

X: ***Today's Volatility & Sentiment***
→ **Reduced Bag of Words *OR* Sentiment Scores**

Y: Volatility & Return Next Day

Dataset: 2008 - 2016

Parameters to Tune for each Combination of Prediction and Feature Selection Models:

1. Rolling Window of Sentiment Features
2. The Size of Reduced Features

Initial dataset:

X

Y

TRAIN	TEST
7 years	1 year

Hyperparameters choice:

Reduced X of Size N

Y

	VALIDATION
	1 year

Rolling averaging window of size M

MSE for Validation Set

Grid Search of Best Combination of Reduced Features Size and Rolling Averaging Window Size

N/M	1	2	..	5	10	20	60	..	250
3									
5									
10									
20									
50									
...									
300									

1. **Choose the grid cell** with the **lowest MSE on the Validation Set**: select corresponding Feature Size and Rolling Averaging Window Size
2. **Refit** chosen parameters on the whole **Train Set** once
3. **Predict Test Set** performance

Assumptions:

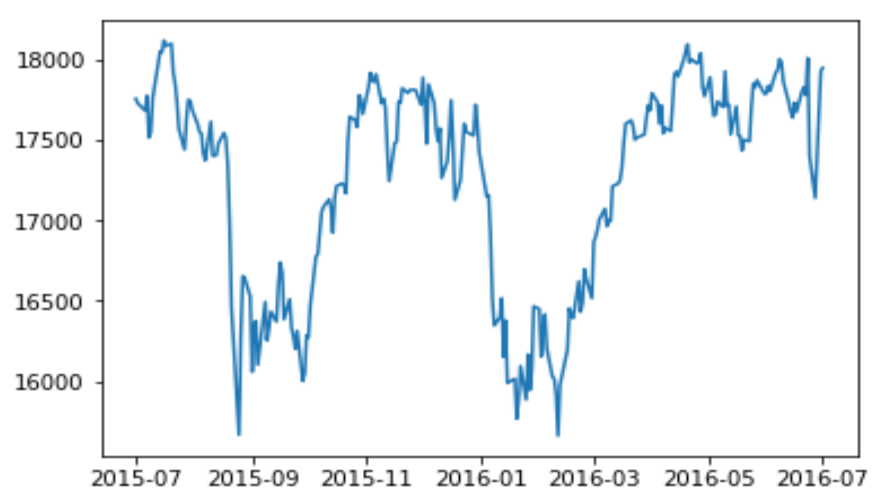
1. Reaction to the same words in the news stays the same over time

Observations:

1. Rolling windows of 1-10 days are typically chosen for models
2. Number of components of 10-20 are typically chosen for models

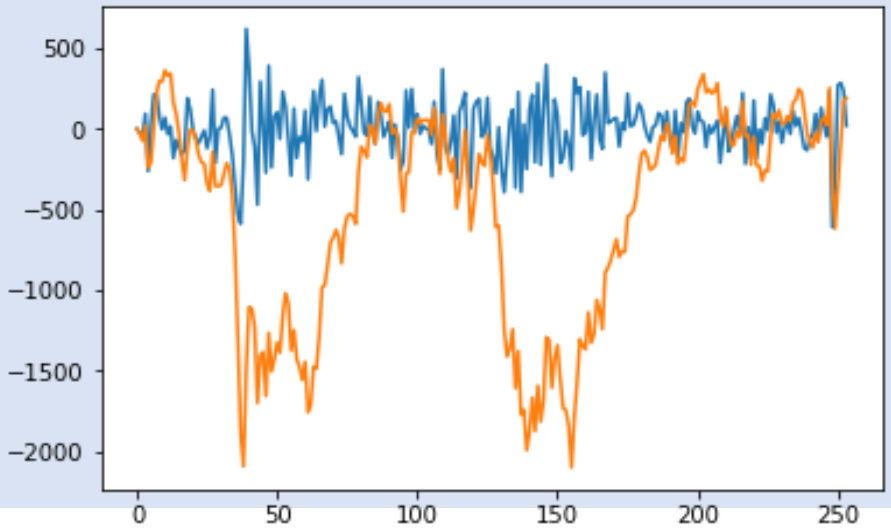
Benchmark Models Comparison

Dow Jones Index on the Test Set:



ARMA – GARCH Sharpe-Ratio Based Prediction P&L:

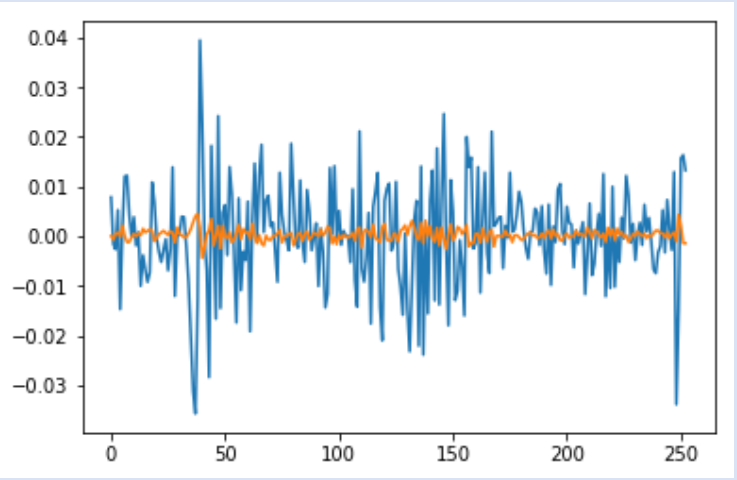
Buy and Hold P&L:



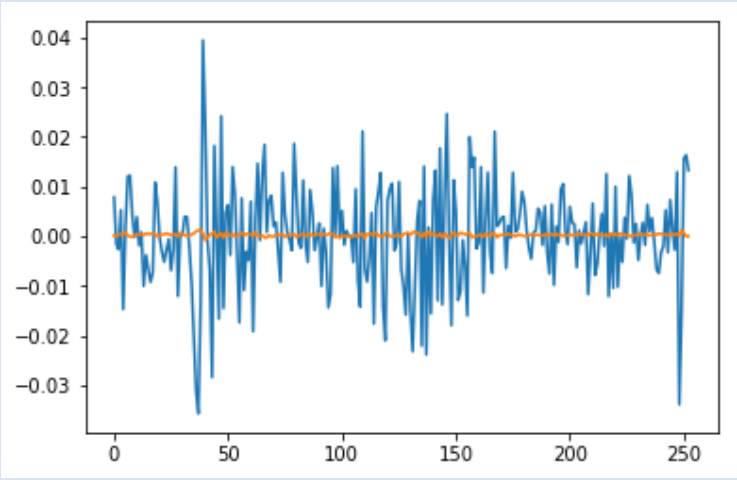
- Comparison with benchmarks will allow to draw further, more educated, conclusions on unsupervised vs. supervised feature selection in volatility and return predictions
- Big volatility in the selected test period seems to be in need of better forecasts in the process of portfolio construction

Return and Volatility Predictions

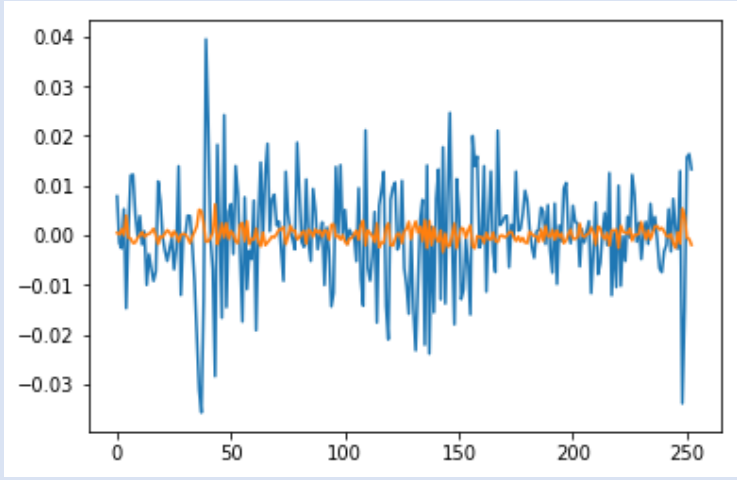
PCA OLS Return Prediction vs Actual Returns:



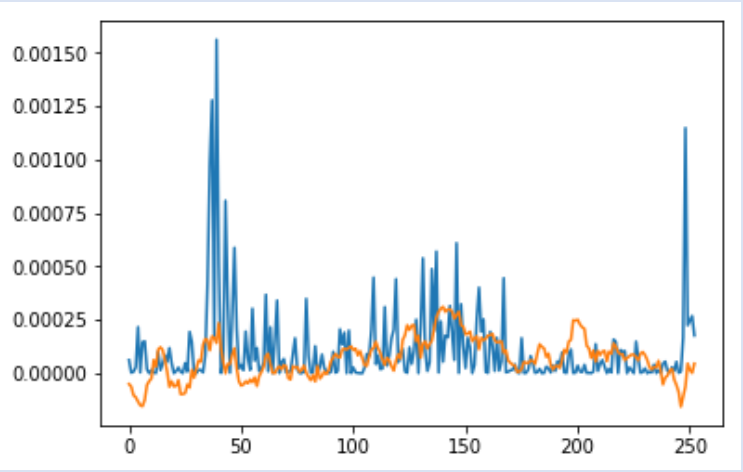
PCA Ridge Return Prediction vs Actual Returns:



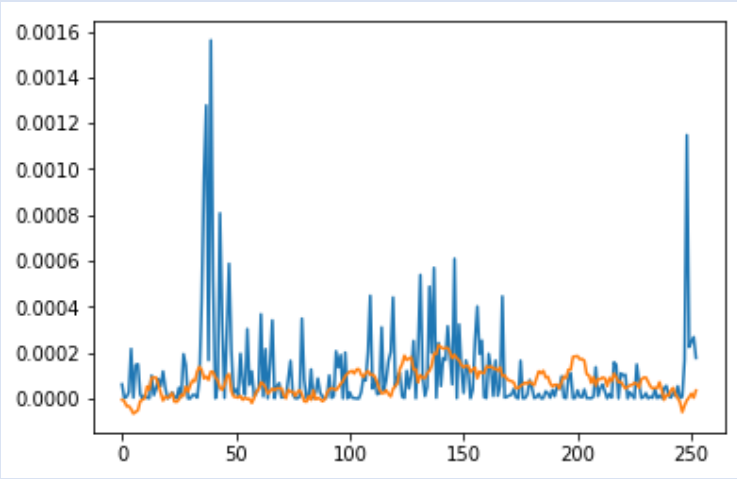
Sentiment Score Return Prediction vs Actual Returns (Averaged RW = 1, OLS) :



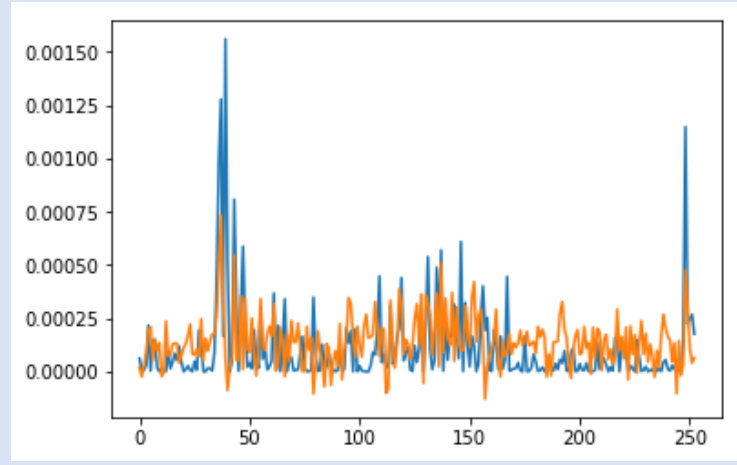
PCA OLS RV Prediction vs Actual RV:



PCA Ridge RV Prediction vs Actual RV:



Sentiment Score Return Prediction vs Actual Returns (Averaged RW = 1, OLS) :



Limitations and Further Research Ideas

- Negative volatility prediction disable us from the optimal portfolio construction, so we need to adjust for that.
- Expansion of the data subset can lead to significant improvement in results – the more headlines are in the pool – the better results we are supposed to have.
- Combination of supervised and un-supervised methods of feature selection could shed new light into the sentiment analysis for return and volatility predictions.