Realised Volatility Forecasting: Machine Learning via Financial Word Embedding

Eghbal Rahimikia
Alliance Manchester Business School; Oxford-Man Institute of Quantitative Finance
Stefan Zohren
Oxford-Man Institute of Quantitative Finance
Ser-Huang Poon
Alliance Manchester Business School

December 6, 2021

Overview

1. Financial textual analysis
2. FinText
3. Realised volatility (RV)
4. Forecsting performance
5. Explainable AI (XAI)
6. Conclusions

Financial textual analysis

Goal

- **Transforming financial textual data to numerical data.**

Challenges

- **Textual data is a high dimensional data.**
- **Approaches: From One Hot Encoding to Word Embedding.**
- **Dictionary (sentiment analysis) is the simplest, based on personal judgment but a transparent approach of natural language processing.**
- **Computational feasibility is still a big challenge.**

Challenges

In previous studies (main stream)...

- Loughran and McDonald (LM) Sentiment Word Lists (Loughran & McDonald, 2011, 2016) is a popular dictionary in finance for converting textual data to the sentiment.
- Based on personal judgment(s).
- Based on financial statements, not news.
- No link between the output of each study and the LM sentiment scores.

In previous studies (novel statistical approaches and ML)...

- Predicting Returns with Text Data (WP) (Ke et al., 2019) →Linking the output and text ✓– Transparency ✓– Nonlinearity ✗.
- Measuring news sentiment (J. Econom) (Shapiro et al., 2020) →Linking the output and text ✓– Transparency ✓– Nonlinearity ✗.
- The Role of Corporate Culture in Bad Times: Evidence from the COVID-19 Pandemic (Li et al., 2020) (JFQA) →Linking the output and text ✓– Transparency ✓– Nonlinearity ✗.

Our goals

- Developing FinText, a novel financial word embedding from Dow Jones Newswires Text News Feed Database.
- Many studies have attributed news as a major contributor to volatility (Engle & Ng, 1993; Engle & Martins, 2020; Conrad & Engle, 2021). →Realised Volatility Forecasting is chosen as an application for FinText in a Convolutional Neural Network (CNN) context.
- Explainable AI (XAI) methods (SHapley Additive exPlanations) & IG (integrated gradient) are applied for making ML models more transparent.

Linking the output and text ✓ Transparency ✓ Nonlinearity ✓

FinText

Word embedding is a type of word representation. In this representation, each word (token) is defined by a vector with size N from a corpus of text.

Table: An abstract representation of word embedding (Gentzkow et al., 2019).

Dimension	king	queen	prince	man	woman	child
Dimension 1 (Royalty)	0.99	0.99	0.95	0.01	0.02	0.01
Dimension 2 (Masculinity)	0.94	0.06	0.02	0.99	0.02	0.49
Dimension 3 (Age)	0.73	0.81	0.15	0.61	0.68	0.09
...						

- **Word2Vec** (Mikolov, Chen, et al., 2013), **Negative sampling** (Mikolov, Sutskever, et al., 2013), and **GloVe** (Pennington et al., 2014) are the main algorithms for training word embeddings.
- There are few pre-trained word embedding models available for download based on large text corpora (Wikipedia, Common Crawl, Twitter, etc.).

Steps & properties

- All types of news (viz. financial, political, weather, etc.) from Dow Jones Newswires Text News Feed from January 1, 2000, to September 14, 2015.
- Extensive text preprocessing of the news stories is applied to eliminate redundant characters, sentences, and structures (primary, begins with, ends with, genral, and Final checks categori).
- Bigram (two-word) phrases are detected and replaced with their bigram form.
- FinText consists of 2,733,035 unique tokens.
- Word2Vec algorithm (CBOW and skip-gram models) and FastText algorithm (CBOW and skip-gram models) are used for developing FinText financial word embedding.
- The dimension of FinText is 300.
- FinText is compared with pre-trained word embeddings from Google (Word2Vec algorithm) and Facebook (FastText algorithm).

General evaluation (Google analogy)

Table: Word embedding comparison (Google analogy)

Section	Word2Vec[a]			FastText		
	FinText[b] (CBOW)[c]	**FinText** (skip-gram)	**Google** (skip-gram)	**WikiNews** (skip-gram)	**FinText** (skip-gram)	**FinText** (CBOW)
capital-common-countries	77.27	85.50	83.60	**100**	85.93	47.40
capital-world	63.60	75.87	82.72	**98.78**	71.06	35.79
currency	22.49	36.69	**39.84**	25.00	32.54	10.65
city-in-state	19.93	60.48	74.64	**81.41**	58.20	15.83
family	63.46	70.51	90.06	**98.69**	58.97	59.62
gram1-adjective-to-adverb	27.47	33.00	32.27	70.46	50.59	**79.45**
gram2-opposite	33.33	32.50	50.53	**73.91**	50.83	71.67
gram3-comparative	77.65	75.04	91.89	**97.15**	77.06	87.39
gram4-superlative	61.67	55.00	88.03	**98.68**	62.14	90.71
gram5-present-participle	62.30	61.24	79.77	**97.53**	70.63	76.06
gram6-nationality-adjective	88.11	93.23	97.07	**99.12**	94.05	79.05
gram7-past-tense	42.02	39.92	66.53	**87.25**	37.98	31.09
gram8-plural	59.23	62.46	85.58	**98.69**	70.92	79.54
gram9-plural-verbs	53.26	54.53	68.95	**97.38**	61.59	79.17
overall	53.65	62.86	77.08	**91.44**	65.00	55.74

[a] For learning word embeddings from textual datasets, **Word2Vec** is developed by Mikolov, Chen, et al. (2013) and **FastText**, as an extension to Word2Vec algorithm, is developed by Bojanowski et al. (2017). [b] Developed word embedding on Dow Jones Newswires Text News Feed database (**FinText**); Publicly available word embedding trained on a part of Google news dataset with about 100 billion words (**Google**); Publicly available word embedding trained on Wikipedia 2017, UMBC webbase corpus and statmt.org news dataset (Mikolov et al., 2018) (**WikiNews**). [c] The continuous bag of words (CBOW) and Skip-gram are the proposed supervised learning models for learning distributed representations of tokens in Mikolov, Chen, et al. (2013).

General evaluation (gold-standard collections)

Table: Word embedding comparison (Gold-standard collections)

Benchmark	Word2Vec[a]			FastText		
	FinText[b] (CBOW)[c]	**FinText** (skip-gram)	**Google** (skip-gram)	**WikiNews** (skip-gram)	**FinText** (skip-gram)	**FinText** (CBOW)
WordSim-353[d] (relatedness)	0.3821	0.4993	**0.6096**	0.6018	0.4425	0.1677
WordSim-353 (similarity)	0.6126	0.6436	**0.7407**	0.6713	0.6393	0.4722
Simlex	0.2657	0.2650	0.3638	**0.3985**	0.2772	0.2574

[a] For learning word embeddings from textual datasets, **Word2Vec** is developed by Mikolov, Chen, et al. (2013) and **FastText**, as an extension to Word2Vec algorithm, is developed by Bojanowski et al. (2017). [b] Developed word embedding on Dow Jones Newswires Text News Feed database (**FinText**); Publicly available word embedding trained on a part of Google news dataset with about 100 billion words (**Google**); Publicly available word embedding trained on Wikipedia 2017, UMBC webbase corpus and statmt.org news dataset (Mikolov et al., 2018) (**WikiNews**). [c] The continuous bag of words (CBOW) and Skip-gram are the proposed supervised learning models for learning distributed representations of tokens in Mikolov, Chen, et al. (2013). [d] WordSim-353 (Agirre et al., 2009) is a gold-standard collection for measuring word relatedness and similarity, and Simlex (Hill et al., 2015) is another gold-standard collection tending to focus on similarity rather than relatedness or association.

Financial evaluation (analogy examples)

Table: Financial analogy examples

Analogy	Word embedding		
	Google	**WikiNews**	**FinText[a]**
debit:credit :: positive:X	positive	negative	negative
bullish:bearish :: rise:X	rises	rises	fall
apple:iphone :: microsoft:X	windows_xp	iphone	windows
us:uk :: djia:X	NONE[b]	NONE	ftse_100
microsoft:msft :: amazon:X	aapl	hmv	amzn
bid:ask :: buy:X	tell	ask-	sell
creditor:lend :: debtor:X	lends	lends	borrow
rent:short_term :: lease:X	NONE	NONE	long_term
growth_stock:overvalued :: value_stock:X	NONE	NONE	undervalued
us:uk :: nyse:X	nasdaq	hsbc	lse
call_option:put_option :: buy:X	NONE	NONE	sell

[a] *FinText* is the financial word embedding developed using Dow Jones Newswires Text News Feed database, Word2Vec algorithm and Skip-gram model. [b] Not in the vocabulary list.

Financial evaluation (2D visualization)

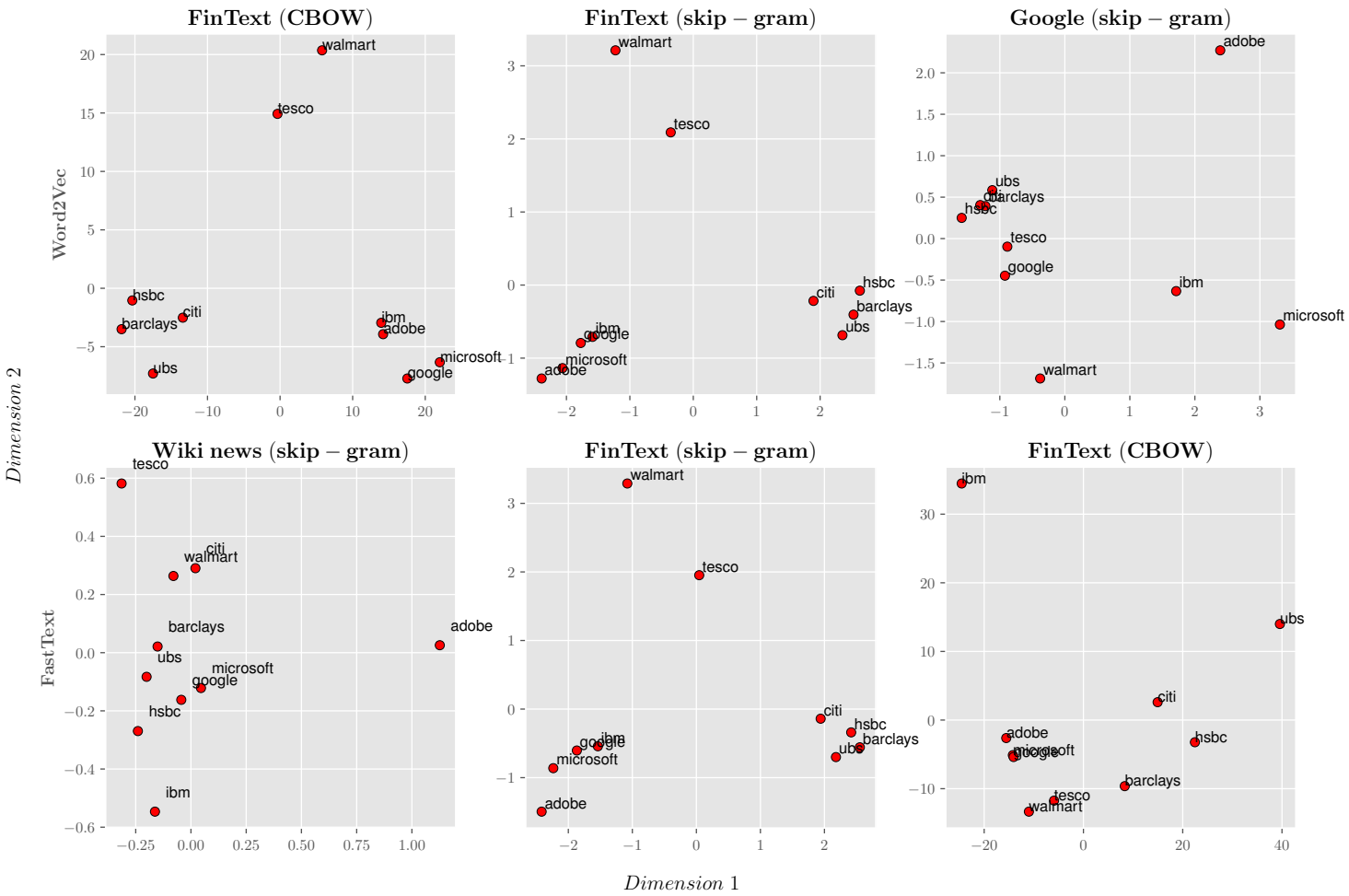


Figure: *2D visualization of word embeddings*

Realised Volatility (RV)

- **High-frequency financial data paves the way for calculation of realised volatility.**
- Andersen & Bollerslev (1998) and Barndorff-Nielsen & Shephard (2001) show that realised volatility has **lower measurement error and noise**.

RV definition

The integrated variance is not observable. **RV can be used as an approximation of integrated variance**:

$$RV_t \equiv \sum_{i=1}^{M} r_{t,i}^2, \tag{1}$$

where M is the sampling frequency and $r_{t,i} \equiv log(P_{t-1+i\delta}) - log(P_{t-1+(i-1)\delta})$. For $\delta \to 0$, RV_t is a consistent estimator for IV_t (Andersen & Bollerslev, 1998). Furthermore, **an increase in the sampling frequency of return series (M) causes a decrease in the measurement error and an increase in the microstructure noise.**

Forecasting structure

- Daily forecasting (Rolling window: 5 days/ Sampling frequency: 5 minutes).
- Out-of-sample data: 300 days (14 September 2015 to 18 November 2016).
- Training period: 2046 (27 July 2007 to 11 September 2015).
- Tickers: 23 stocks from NASDAQ market (based on highest liquidity and availability of data).
- Following Rahimikia & Poon (2020b), jump and normal days are selected based on Interquartile Range (IQR) ($Q_3 + 1.5IQR$, where $IRQ = Q_3 - Q_1$).
- We have purposely avoided parametric jump estimation as it is very sensitive to the assumption of the stock price dynamics, the bandwidth adopted data frequency and data reference period.
- Following White (2000), the Reality Check is implemented using the stationary bootstrap of Politis & Romano (1994) with 999 re-samplings and the average block length of 5 (each model against AR1, HAR, HAR-J, CHAR, SHAR, ARQ, HARQ, and HARQ-F).

Loss functions (cardinal and complementary measures)

Patton (2011) shows that many lost functions do not work well in the presence of noise in the volatility proxy.

$$MSE(RV_t, \widehat{RV_t}) \equiv \frac{1}{N}\sum_{t=1}^{N}(RV_t - \widehat{RV_t})^2, \tag{2}$$

$$QLIKE(RV_t, \widehat{RV_t}) \equiv \frac{1}{N}\sum_{t=1}^{N}\left(\frac{RV_t}{\widehat{RV_t}} - log(\frac{RV_t}{\widehat{RV_t}}) - 1\right), \tag{3}$$

$$MDA(RV_t, RV_{t-1}, \widehat{RV_t}) \equiv \frac{1}{N}\sum_{t=1}^{N} 1_{sign(RV_t - RV_{t-1}) == sign(\widehat{RV_t} - RV_{t-1})}, \tag{4}$$

where RV_t is the true RV at time t, $\widehat{RV_t}$ is the forecast RV at time t, N is the number of days in the out-of-sample period, and $sign(\cdot)$ and 1 are the sign and indicator functions.

RV descriptive statistics

Table: RV descriptive statistics (from 27 July 2007 to 18 November 2016)

Ticker[a]	Min	Max	1st quantile	Median	3rd quantile	Mean	STD	Kurtosis	Skewness
AAPL	0.102	198.574	1.021	1.879	3.987	4.953	12.664	83.884	8.028
MSFT	0.067	133.679	0.935	1.577	2.918	3.289	7.268	104.476	8.786
INTC	0.146	130.167	1.165	1.920	3.709	4.065	7.596	71.609	6.875
CMCSA	0.148	153.376	0.951	1.827	3.751	4.014	8.419	95.178	8.139
QCOM	0.122	373.543	0.959	1.872	3.873	4.673	13.778	280.384	13.730
CSCO	0.163	343.946	1.030	1.791	3.438	4.348	12.995	262.347	13.440
EBAY	0.215	252.608	1.461	2.592	4.946	5.525	12.785	139.670	9.850
GILD	0.222	259.489	1.383	2.179	3.900	4.719	13.706	173.238	11.815
TXN	0.183	287.897	1.111	1.999	3.987	4.107	9.249	398.325	15.651
AMZN	0.206	547.030	1.683	2.882	5.720	7.562	23.925	185.115	11.593
SBUX	0.164	192.629	1.086	1.968	4.308	4.714	11.255	114.155	9.331
NVDA	0.317	1104.351	2.180	4.382	9.414	9.591	29.432	837.584	24.558
MU	0.563	359.620	4.204	7.137	13.584	14.355	26.204	61.344	6.711
AMAT	0.292	114.376	1.715	2.812	5.150	5.153	8.149	61.231	6.438
NTAP	0.257	290.647	1.594	2.903	5.743	6.283	14.419	149.830	10.163
ADBE	0.216	569.720	1.153	2.081	3.952	5.059	15.730	693.479	21.309
XLNX	0.229	251.383	1.224	2.184	4.258	4.359	9.382	265.118	12.977
AMGN	0.159	214.156	1.006	1.727	3.126	3.468	9.764	221.110	13.295
VOD	0.134	219.033	0.780	1.487	3.369	4.252	11.788	115.204	9.471
CTSH	0.246	485.894	1.214	2.162	5.266	6.103	17.479	315.162	14.555
KLAC	0.154	499.808	1.278	2.514	5.126	5.689	17.915	395.464	17.684
PCAR	0.214	389.930	1.285	2.563	5.800	6.014	13.514	294.091	12.810
ADSK	0.358	693.772	1.637	2.823	5.256	6.833	24.263	413.773	17.814

[a] Tickers are ranked according to their liquidity (high to low).

An abstract representation of NLP-ML

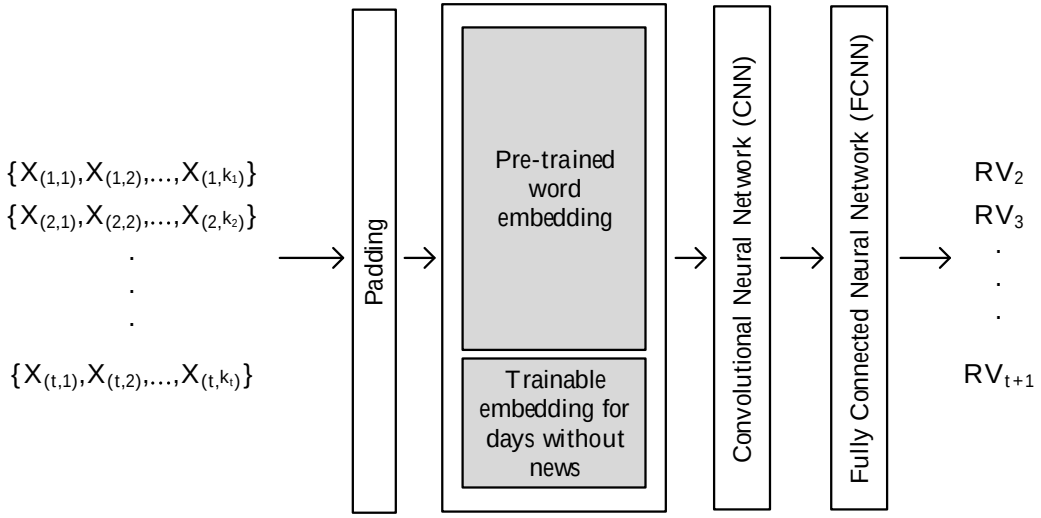


Figure: An abstract representation of NLP-ML model

Notes: $\{X_{(t,1)}, X_{(t,2)}, \ldots, X_{(t,k_t)}\}$ consists of news headlines of day t and $X_{(t,k_t)}$ is the kth token of input t. Also, RV_{t+1} is the RV of day $t+1$ (next day RV). Padding with the maximum length of 500 is adopted to ensure that all inputs of the neural network have the same length. The word embedding block consists of two different word embeddings. To capture days without any news, a trainable word embedding is used.

News selection criteria

- Stock-related and hot political news are used for this analysis.
- The tag ('about') is used for extracting news
- 'About' denotes a story about a ticker but of no particularly significant.

A detailed representation of NLP-ML model

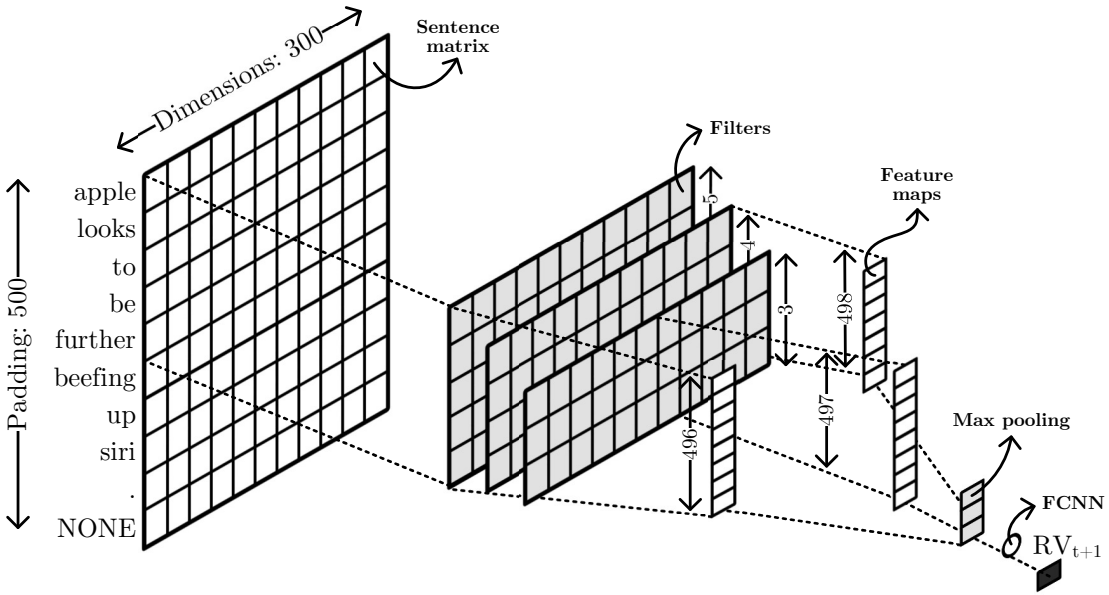


Figure: A detailed representation of NLP-ML model

Notes: The sentence matrix is a 500×300 matrix with a maximum length of padding of 500 and word embedding dimensions of 300. In this matrix, each token is defined by a vector of 300 values. This structure contains three filters of different sizes. The filters with the size of 3, 4, and 5 generate feature maps with the size of 498, 497, and 496, respectively. Max pooling and a fully connected neural network (FCNN) are applied then as the next steps. The output of this network is the RV of the next day (RV_{t+1}).

RV forecasting performance evaluation (MSE)

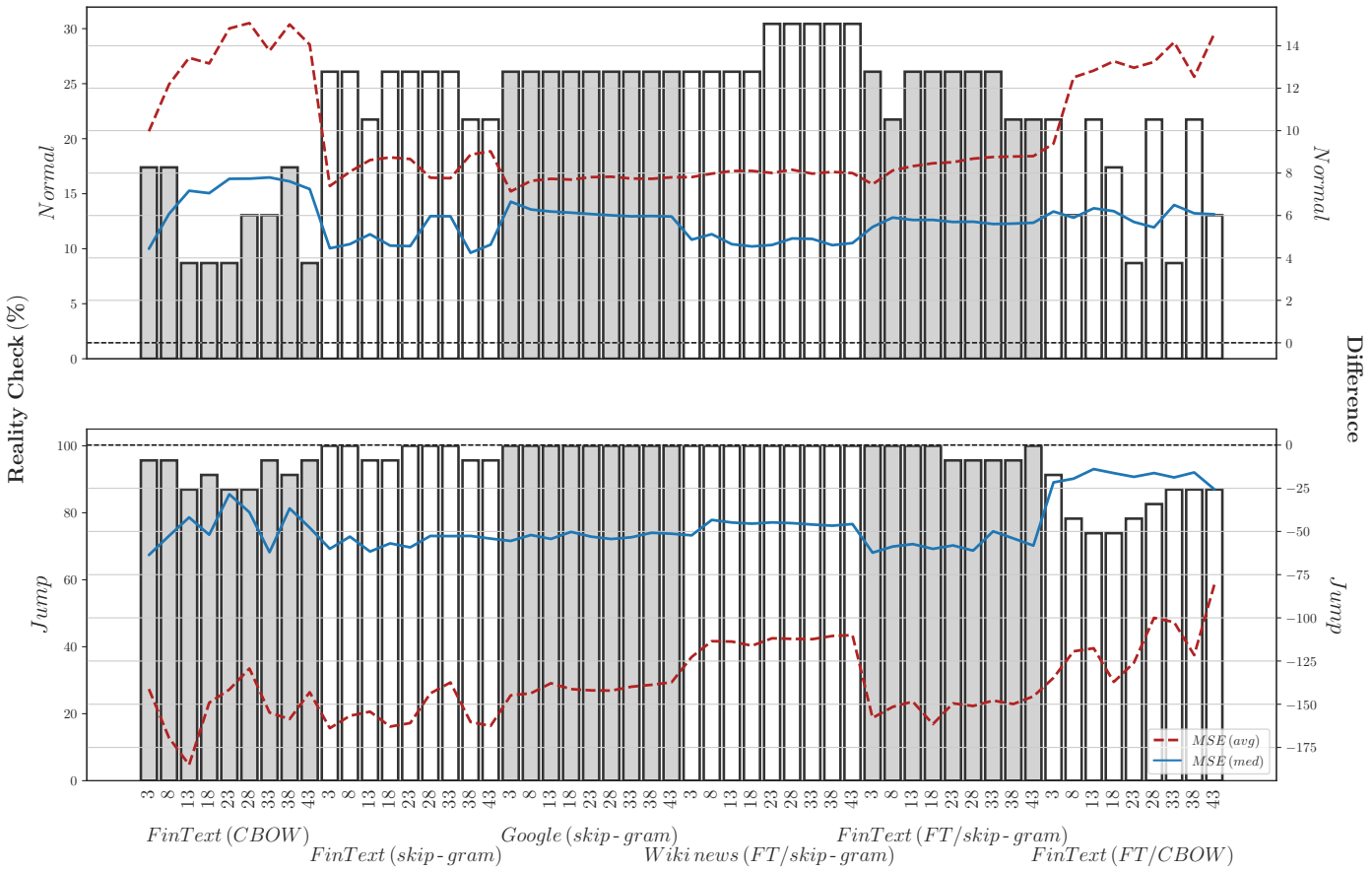


Figure: *Forecast evaluation under MSE*

RV forecasting performance evaluation (QLIKE)

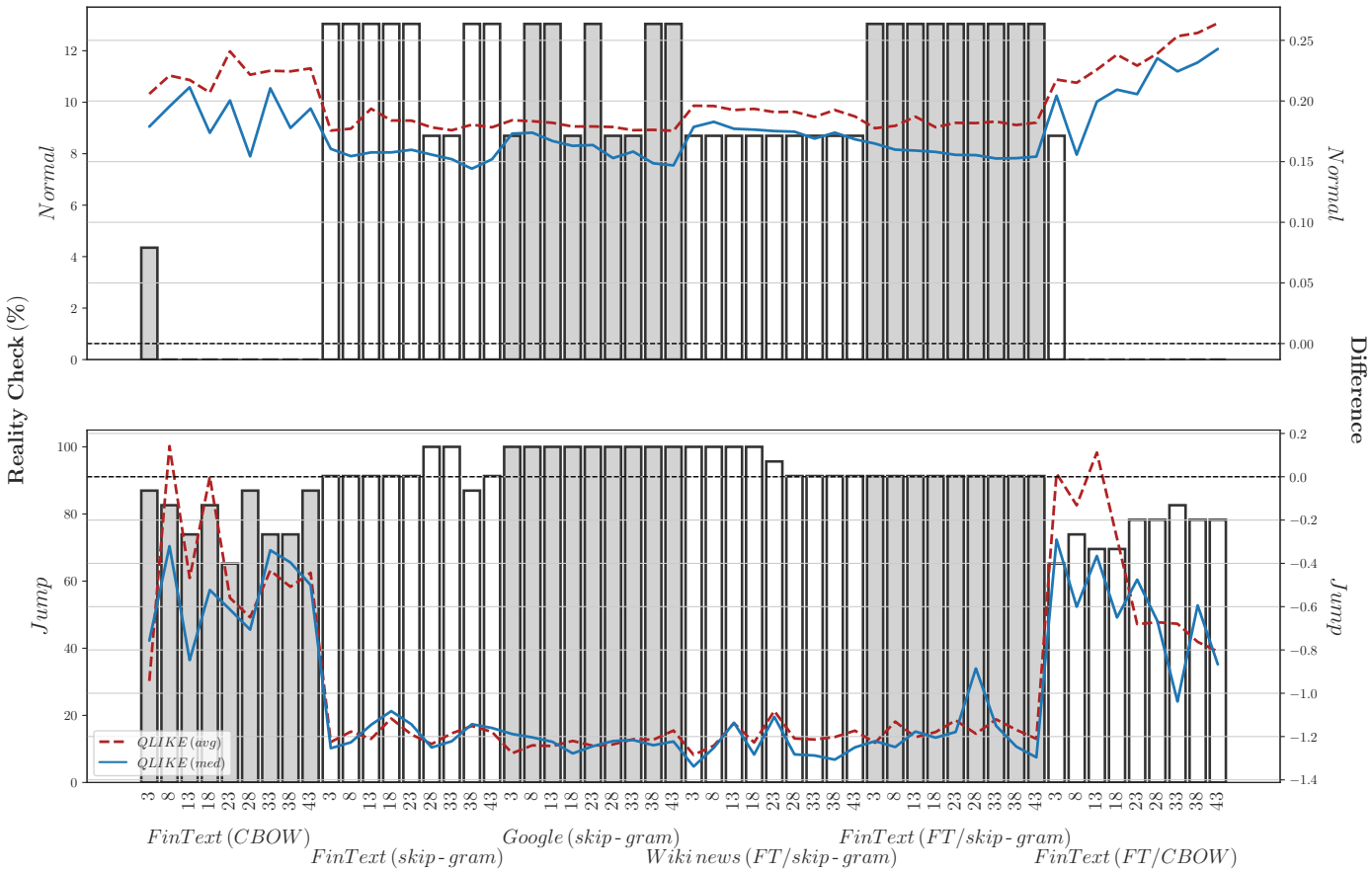


Figure: *Forecast evaluation under QLIKE*

A horse race

- NLP-ML is compared with CHARx (extended CHAR) in Rahimikia & Poon (2020a) ML group of models in Rahimikia & Poon (2020b).

Table: CHARx, ML group, and NLP-ML structure comparison

	Model group		
Criterion	**CHARx**	**ML group**	**NLP-ML**
Variable type	news(LM[a]) + financial	news(LM) + financial	news
News type	headline + body	headline + body	headline
Historical information	last 23 days	last 23 days	last day
Training sequence	daily	daily	every 5 days

[a] LM is the Loughran-McDonald dictionary approach (Loughran & McDonald, 2011).

RV forecasting performance evaluation (MSE) comparison

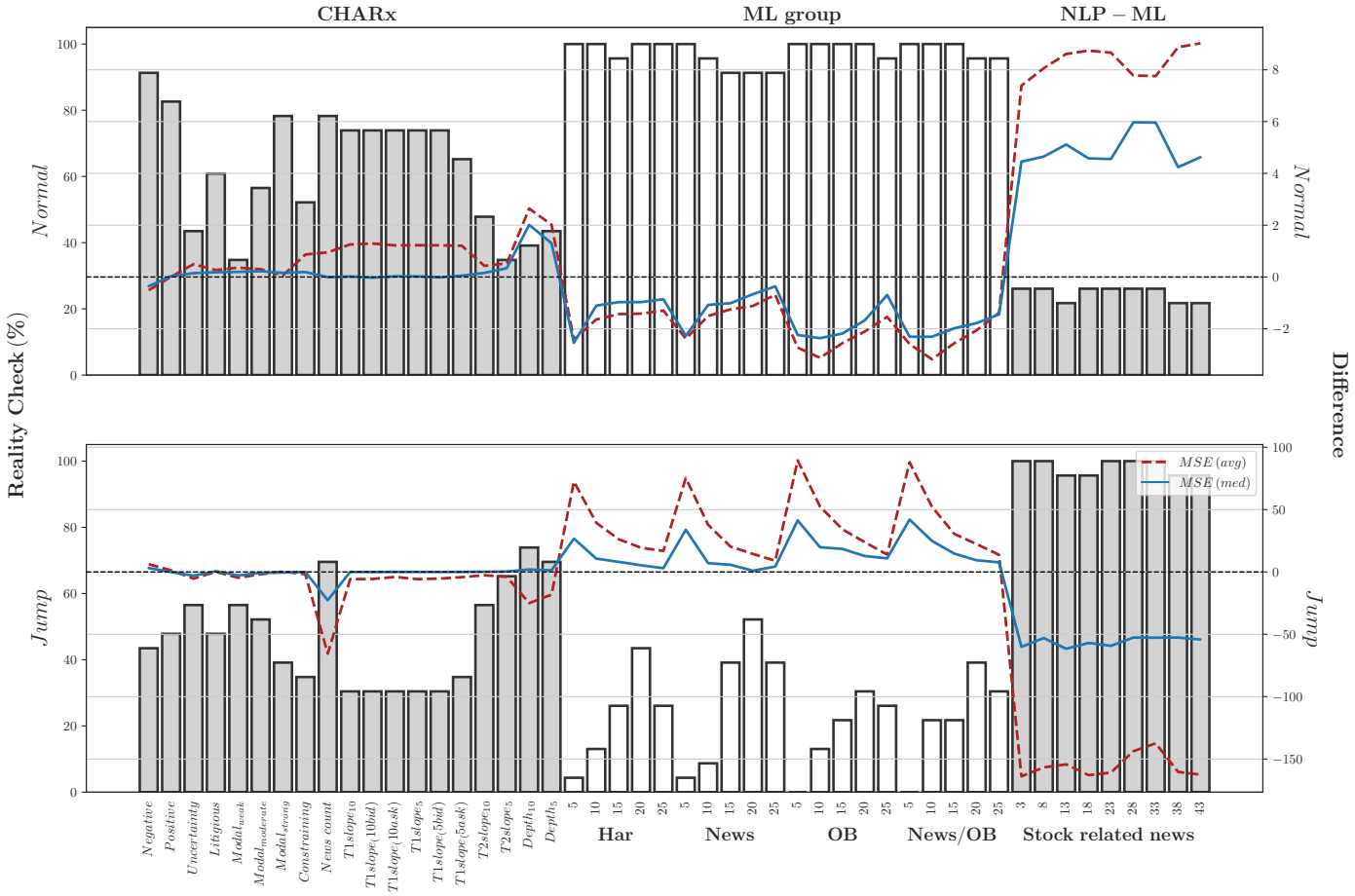


Figure: *ML, extended CHARx, and NLP-ML models comparison (MSE)*

RV forecasting performance evaluation (QLIKE) - comparison

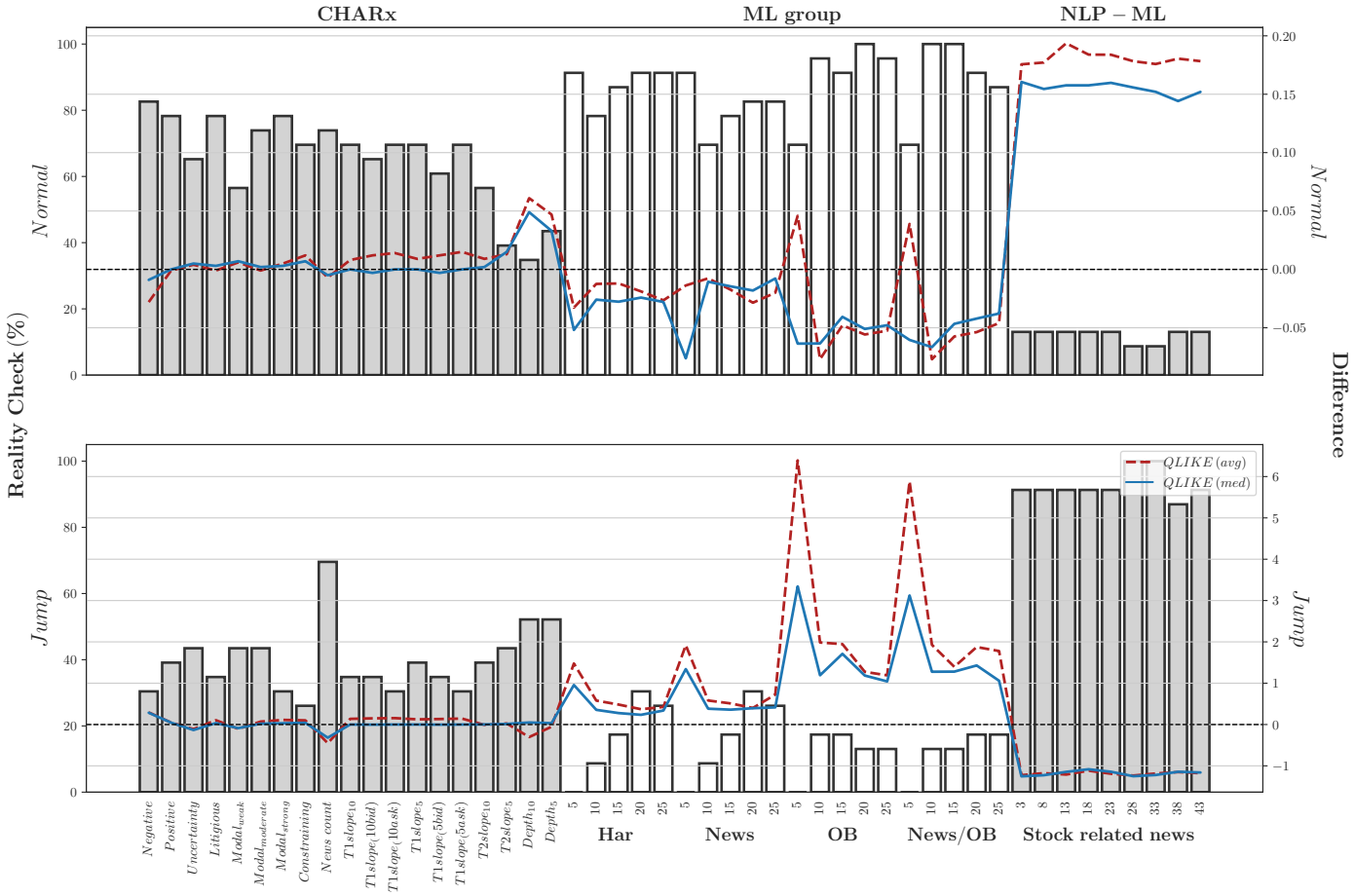
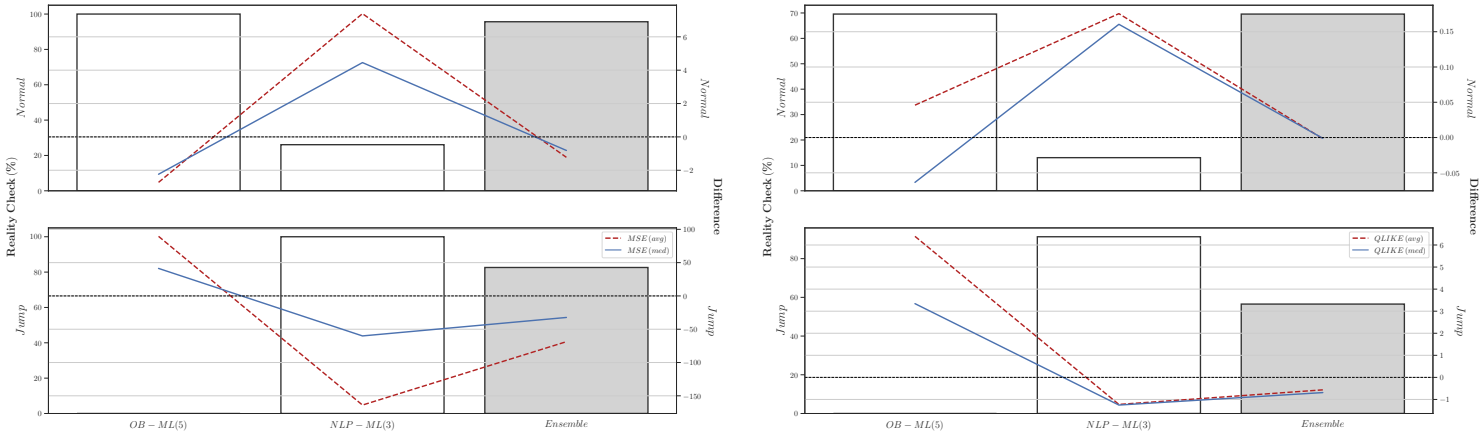


Figure: *ML, extended CHARx, and NLP-ML models comparison (QLIKE)*

Ensemble model

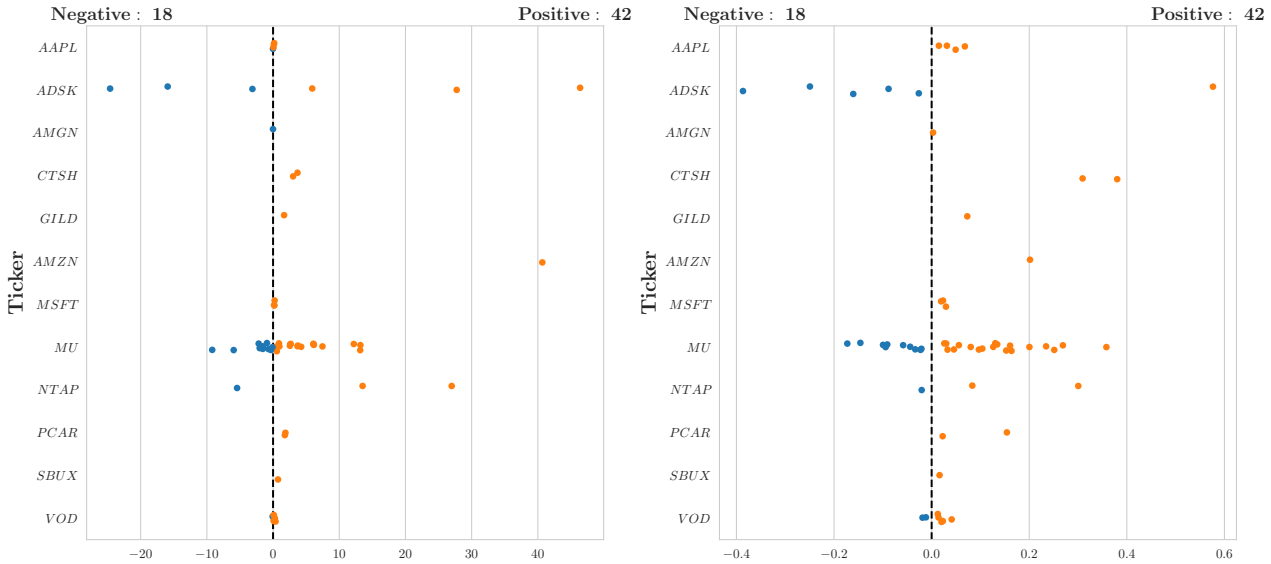


(a) Forecast evaluation under MSE

(b) Forecast evaluation under QLIKE

Figure: Performance of ensemble model, against OB-ML(5) and NLP-ML(3)

XAI: ‘Loss’ in Stock-Related News



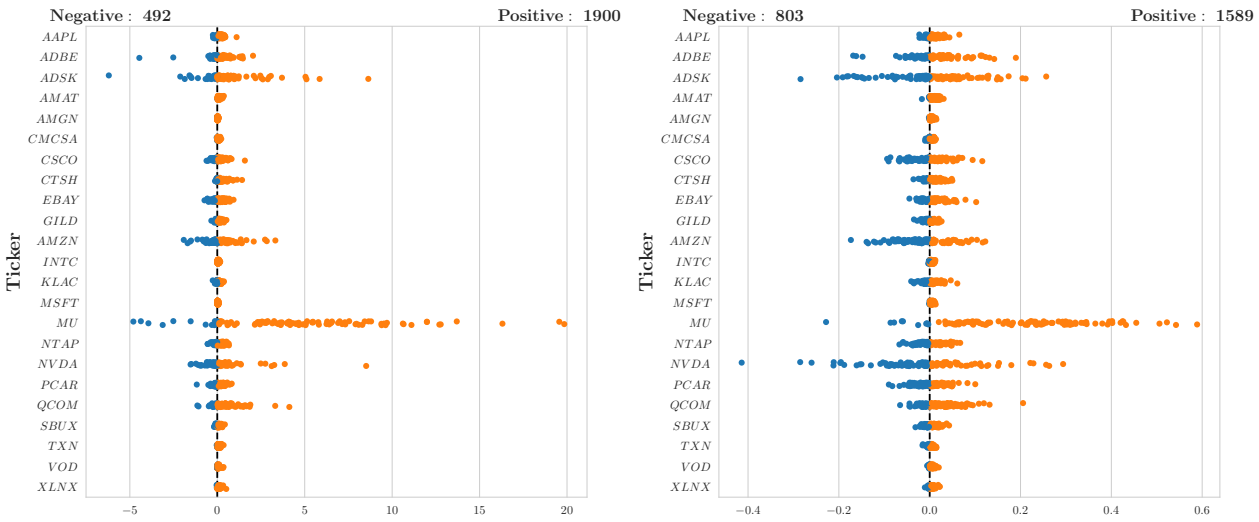
(a) IG attribution values

(b) SHAP values

Figure: Explainer results for ‘loss’ (ticker-related news)

Notes: The left (right) figure presents the IG attribution (SHAP) values for ‘loss’ as the chosen token considering the stock-related news published during the out-of-sample period. The x-axis is the reported explainer value, and the y-axis is the ticker name.

XAI: ‘donald_trump’ in hot political news



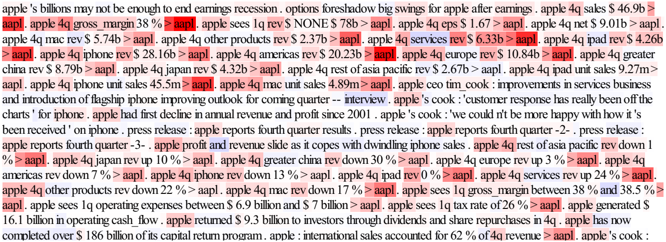
(a) IG attribution values *(b) SHAP values*

Figure: Explainer results for ‘donald_trump’ (hot political news)

Notes: The left (right) figure presents the IG attribution (SHAP) values for ‘donald_trump’ as the chosen token considering the hot political news published during the out-of-sample period. The x-axis is the reported explainer value, and the y-axis is the ticker name.

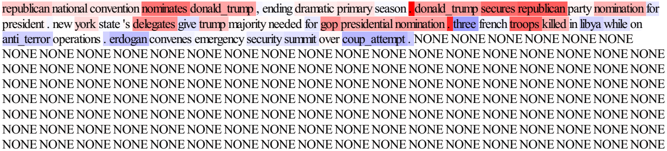
XAI: SHAP and IG explainer representations

Financial textual analysis
Challenges
Our goals
FinText
Steps & properties
General evaluation (Google analogy)
General evaluation (gold-standard collections)
Financial evaluation (analogy examples)
Financial evaluation (2D visualization)
Realised volatility (RV)
Forecasting structure and variables
RV descriptive statistics
An abstract representation of NLP-ML
A detailed representation of NLP-ML
Forecsting performance
RV forecasting performance evaluation (QLIKE)
A horse race
RV forecasting performance evaluation (MSE) - comparison
RV forecasting performance evaluation (QLIKE) - comparison
Ensemble model
Explainable AI (XAI)
XAI: ‘donald_trump’ in hot political news
XAI: SHAP and IG explainer representations
Conclusions
References



apple 's billions may not be enough to end earnings recession . options foreshadow big swings for apple after earnings . apple 4q sales $ 46.9b > aapl . apple 4q gross_margin 38 % > aapl . apple sees 1q rev $ NONE $ 78b > aapl . apple 4q eps $ 1.67 > aapl . apple 4q net $ 9.01b > aapl . apple 4q mac rev $ 5.74b > aapl . apple 4q other products rev $ 2.37b > aapl . apple 4q services rev $ 6.33b > aapl . apple 4q ipad rev $ 4.26b > aapl . apple 4q iphone rev $ 28.16b > aapl . apple 4q americas rev $ 20.23b > aapl . apple 4q europe rev $ 10.84b > aapl . apple 4q greater china rev $ 8.79b > aapl . apple 4q japan rev $ 4.32b > aapl . apple 4q rest of asia pacific rev $ 2.67b > aapl . apple 4q ipad unit sales 9.27m > aapl . apple 4q iphone unit sales 45.5m > aapl . apple 4q mac unit sales 4.89m > aapl . apple ceo tim_cook : improvements in services business and introduction of flagship iphone improving outlook for coming quarter -- interview . apple 's cook : 'customer response has really been off the charts ' for iphone . apple had first decline in annual revenue and profit since 2001 . apple 's cook : 'we could n't be more happy with how it 's been received ' on iphone . press release : apple reports fourth quarter results . press release : apple reports fourth quarter -2- . press release : apple reports fourth quarter -3- . apple profit and revenue slide as it copes with dwindling iphone sales . apple 4q rest of asia pacific rev down 1 % > aapl . apple 4q japan rev up 10 % > aapl . apple 4q greater china rev down 30 % > aapl . apple 4q europe rev up 3 % > aapl . apple 4q americas rev down 7 % > aapl . apple 4q iphone rev down 13 % > aapl . apple 4q ipad rev 0 % > aapl . apple 4q services rev up 24 % > aapl . apple 4q other products rev down 22 % > aapl . apple 4q mac rev down 17 % > aapl . apple sees 1q gross_margin between 38 % and 38.5 % > aapl . apple sees 1q operating expenses between $ 6.9 billion and $ 7 billion > aapl . apple sees 1q tax rate of 26 % > aapl . apple generated $ 16.1 billion in operating cash_flow . apple returned $ 9.3 billion to investors through dividends and share repurchases in 4q . apple has now completed over $ 186 billion of its capital return program . apple : international sales accounted for 62 % of 4q revenue > aapl . apple 's cook :

(a) SHAP explainer representation (stock-related news) (2016-10-26)



republican national convention nominates donald_trump , ending dramatic primary season . donald_trump secures republican party nomination for president . new york state 's delegates give trump majority needed for gop presidential nomination . three french troops killed in libya while on anti_terror operations . erdogan convenes emergency security summit over coup_attempt . NONE NONE NONE NONE NONE NONE
NONE NONE NONE NONE NONE NONE NONE NONE NONE NONE NONE NONE NONE NONE NONE NONE NONE NONE
NONE NONE NONE NONE NONE NONE NONE NONE NONE NONE NONE NONE NONE NONE NONE NONE NONE NONE
NONE NONE NONE NONE NONE NONE NONE NONE NONE NONE NONE NONE NONE NONE NONE NONE NONE NONE
NONE NONE NONE NONE NONE NONE NONE NONE NONE NONE NONE NONE NONE NONE NONE NONE NONE NONE
NONE NONE NONE NONE NONE NONE NONE NONE NONE NONE NONE NONE NONE NONE NONE NONE NONE NONE
NONE NONE NONE NONE NONE NONE NONE NONE NONE NONE NONE NONE NONE NONE NONE NONE NONE NONE
NONE NONE NONE NONE NONE NONE NONE NONE NONE NONE NONE NONE NONE NONE NONE NONE NONE NONE

(b) IG explainer representation (hot political news) (2016-07-20)

Figure: SHAP and IG news visualization (AAPL)

Conclusions

- FinText word embedding works better and more sensitive in detecting financial jargon.
- Headlines of stock-related news $\rightarrow$substantial improvement on forecasting RV jump days (beating all HAR-family of models).
- Headlines of hot political news $\rightarrow$to a lesser extent, is crucial for improving RV forecasting performance*.
- A simple ensemble model combining textual and financial data (LOB) $\rightarrow$dominates all HAR-family of models on both normal jump volatlity days.
- Explainable AI methods help to measure the impact of given tokens on realised volatility forecasts.

Thank You

MANCHESTER 1824
The University of Manchester

UNIVERSITY OF OXFORD

OXFORD-MAN INSTITUTE

Agirre, E., Alfonseca, E., Hall, K., Kravalova, J., Pasca, M., & Soroa, A. (2009). A study on similarity and relatedness using distributional and wordnet-based approaches.

Andersen, T. G., & Bollerslev, T. (1998). Answering the skeptics: Yes, standard volatility models do provide accurate forecasts. *International economic review*, 885–905.

Barndorff-Nielsen, O. E., & Shephard, N. (2001). Non-gaussian ornstein–uhlenbeck-based models and some of their uses in financial economics. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *63*(2), 167–241.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, *5*, 135–146.

Conrad, C., & Engle, R. F. (2021). Modelling volatility cycles: the (mf)ˆ 2 garch model. *Available at SSRN*.

Engle, R. F., & Martins, S. (2020). Measuring and hedging geopolitical risk.

Engle, R. F., & Ng, V. K. (1993). Measuring and testing the impact of news on volatility. *The journal of finance*, *48*(5), 1749–1778.

Gentzkow, M., Kelly, B., & Taddy, M. (2019). Text as data. *Journal of Economic Literature*, *57*(3), 535–74.

Hill, F., Reichart, R., & Korhonen, A. (2015). Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, *41*(4), 665–695.

Ke, Z. T., Kelly, B. T., & Xiu, D. (2019). *Predicting returns with text data* (Tech. Rep.). National Bureau of Economic Research.

Li, K., Liu, X., Mai, F., & Zhang, T. (2020). The role of corporate culture in bad times: Evidence from the covid-19 pandemic. *Journal of Financial and Quantitative Analysis*, 1–68.

Loughran, T., & McDonald, B. (2011). When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of Finance*, *66*(1), 35–65.

Loughran, T., & McDonald, B. (2016). Textual analysis in accounting and finance: A survey. *Journal of Accounting Research*, *54*(4), 1187–1230.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C., & Joulin, A. (2018). Advances in pre-training distributed word representations. In *Proceedings of the international conference on language resources and evaluation (lrec 2018)*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, *26*, 3111–3119.

Patton, A. J. (2011). Volatility forecast comparison using imperfect volatility proxies. *Journal of Econometrics*, *160*(1), 246–256.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (emnlp)* (pp. 1532–1543).

Politis, D. N., & Romano, J. P. (1994). The stationary bootstrap. *Journal of the American Statistical association*, *89*(428), 1303–1313.

Rahimikia, E., & Poon, S.-H. (2020a). Big data approach to realised volatility forecasting using har model augmented with limit order book and news. *Available at SSRN 3684040*.

Rahimikia, E., & Poon, S.-H. (2020b). Machine learning for realised volatility forecasting. *Available at SSRN 3707796*.

Shapiro, A. H., Sudhof, M., & Wilson, D. J. (2020). Measuring news sentiment. *Journal of Econometrics*.

White, H. (2000). A reality check for data snooping. *Econometrica*, *68*(5), 1097–1126.