Stock trend prediction using sentiment analysis

Qianyi Xiao [1], **Baha Ihnaini** [Corresp. 1]

[1] Department of Computer Science, Wenzhou Kean University, Wenzhou, Zhejiang, China

Corresponding Author: Baha Ihnaini
Email address: bihnaini@kean.edu

These days, the vast amount of data generated on the Internet is a new treasure trove for investors. They can utilize text mining and sentiment analysis techniques to reflect investors' confidence in specific stocks in order to make the most accurate decision. Most previous research just sums up the text sentiment score on each natural day and uses such aggregated score to predict various stock trends. However, the natural day aggregated score may not be useful in predicting different stock trends. Therefore, in this research, we designed two different time divisions: $0:00_t \sim 0:00_{t+1}$ and $9:30_t \sim 9:30_{t+1}$ to study how tweets and news from the different periods can predict the next-day stock trend. 260,000 tweets and 6,000 news from Service stocks (Amazon, Netflix) and Technology stocks (Apple, Microsoft) were selected to conduct the research. The experimental result shows that opening hours division ($9:30_t \sim 9:30_{t+1}$) outperformed natural hours division ($0:00_t \sim 0:00_{t+1}$).

1 # Stock trend prediction using sentiment analysis

2 Qianyi Xiao[1] and Baha Ihnaini[2]

3 [1] Department of Computer Science, Wenzhou-Kean University, Wenzhou, Zhejiang, China

4 [2] Department of Computer Science, Wenzhou-Kean University, Wenzhou, Zhejiang, China

6 Corresponding Author:

7 Baha Ihnaini[1]

8 88 Daxue Rd, Ouhai, Wenzhou, Zhejiang, 325060, China

9 Email address: bihnaini@kean.edu

25 # Stock trend prediction using sentiment analysis

26   28

27   29

30   Qianyi Xiao[1] and Baha Ihnaini[2]

31   [1] Department of Computer Science, Wenzhou-Kean University, Wenzhou, Zhejiang, China

32   [2] Department of Computer Science, Wenzhou-Kean University, Wenzhou, Zhejiang, China

Abstract

35   These days, the vast amount of data generated on the Internet is a new treasure trove for
36   investors. They can utilize text mining and sentiment analysis techniques to reflect investors'
37   confidence in specific stocks in order to make the most accurate decision. Most previous
38   research just sums up the text sentiment score on each natural day and uses such aggregated
39   score to predict various stock trends. However, the natural day aggregated score may not be
40   useful in predicting different stock trends. Therefore, in this research, we designed two different
41   time divisions: $0:00_t \sim 0:00_{t+1}$ and $9:30_t \sim 9:30_{t+1}$ to study how tweets and news from the
42   different periods can predict the next-day stock trend. 260,000 tweets and 6,000 news from
43   Service stocks (Amazon, Netflix) and Technology stocks (Apple, Microsoft) were selected to
44   conduct the research. The experimental result shows that opening hours division ($9:30_t \sim 9:30$
45    $_{t+1}$) outperformed natural hours division ($0:00_t \sim 0:00_{t+1}$) in specific trend predictions.

46

Introduction

48   For decades, stock trend prediction has been a popular topic due to its importance in the
49   economy and risk management [1-4]. However, its inborn complexity and uncertainty decide its
50   difficulty [5, 6]. For example, politics, wars, and many factors would sharply affect stock prices
51   [5, 7]. Thus, achieving the best result with the minimum required data is the goal [5]. In the past,
52   most research employed many financial or technical factors, which aimed to reflect investors'
53   interest and predict stock from a financial perspective [8, 9]. While a few years ago, with the
54   rapid expansion of social media, people could easily post and spread their emotions through
55   micro-blogging (e.g., Twitter and Reddit) [3, 10]. Therefore, researchers attempted to utilize
56   micro-blogging data to directly attain investors' moods regarding the stock market, as financial
57   decisions are significantly driven by emotion and mood [9, 11]. This provides the theoretical
58   foundation for connecting social media sentiments and stock fluctuations.

59   Twitter, one of the micro-blogging platforms, is a significant data source due to its popularity,
60   transparency, and timeliness [3]. Twitter provides cashtag symbol ($) search to obtain relevant
61   stock Twitter messages (tweets) [12]. For example, $AAPL is the stock topic only for Apple Inc.
62   Therefore, by harvesting tweets with $AAPL, we can build one dataset containing all AAPL
63   stock discussions.

64   Besides, news is also considered significant in stock prediction and has been widely used for
65   many years [13]. Expert's authoritative opinions, the description of one company's business and
66   many other factors hidden in the news can motivate investors to buy or sell stocks [14-18].
67   Among various financial media, Bloomberg, Forbes, and Reuters are some valuable financial
68   news sources [14, 16, 18]. In our research, we will use news attained from eight popular and
69   reputable websites or medias to conduct the stock trend prediction.

70   Natural language processing (NLP) is one computational technique to analyze and understand
71   human language [19, 20]. Sentiment analysis, a branch of the NLP, identifies and extracts
72   people's subjective attitudes, opinions, and emotions [21-23]. Sentiment analysis has been widely
73   used in checking online comments, and the main goal is to examine sentiment scores [22, 23].
74   This research will adopt sentiment analysis technology as the primary text analysis method.

75   In this paper, we harvest 2021 tweets and news mentioned Service stock AMZN, NFLX, and
76   Technology Stock AAPL, MSFT. We combine VADER sentiment analysis along with other
77   extracted tweets features to generate a novel weighted sentiment index $T_{weighted}$ for each tweet.
78   And we utilize FinBERT to analyze news titles and generate $N_{weighted}$ to represent news opinion
79   value. Besides, we also design two time divisions: $\textbf{0:00}_t \sim \textbf{0:00}_{t+1}$ and $\textbf{9:30}_t \sim \textbf{9:30}_{t+1}$ to
80   explore how tweets and news from different periods can predict the future stock trend. *Goal I:*
81   *Open$_{t+1}$ - Open$_t$* and *Goal II: Open$_{t+1}$ - Close$_t$* (t is given transaction date) are created to evaluate
82   their performance separately. Finally, six classifier algorithms (KNN, Tree, SVM, Random
83   Forest, Naïve Bayes, Logistic Regression) are applied to check the result with 10-fold cross-
84   validation.

85

86   ## Related work

87   In the work of Bujari et al. [9], they used only tweets features and achieved 82% accuracy with
88   an ad-hoc model. However, no unified model fits all cases, and the prediction accuracy varies
89   greatly for different stocks with the same model. The limitation of this research is the
90   considerably short period, only 70 days (including non-transaction days) of data were collected.
91   In Checkley et al. [3] research, their dataset goes from the 17th February 2012 to 17th October
92   2014. They proposed a model to predict market volatility, volume, and returns (direction) by
93   using data granular to two-minute intervals. The evidence showed a causal link between all three
94   target and bull-bearish sentiment metrics. Among the five stocks investigated, market volatility
95   and volume are more predictable than market returns. A similar approach using more granular
96   data also appeared in Kinyua et al. [24] research, and they analyzed the impact of President
97   Trump's tweets on SPX and INDU in a 30-minute event window (15 min before tweet posted
98   and 15 min after tweet posted). To focus on the immediate influence of Trump's tweets, only
99   tweets in transaction hours were kept in this research. The model used Random forest, Decision
100   tree, and Logistic regression algorithm to predict how market response to Trump's tweets.
101   Machine learning algorithms showed that the inclusion of Trump's tweets significantly decreased
102   prediction RMSE. Both Trump's strong negative and positive sentiments resulted in an uptrend

103 of SPX and INDU indices. While, for all other sentiment categories, Trump's tweets caused a
104 downtrend.

105 Besides, Maqsood et al. [25] built a stock trend prediction model based on tweets related to local
106 or global events. The significant events from 2012 to 2016 in USA, Hong Kong, Turkey, and
107 Pakistan markets were investigated. For example, in USA market, the author examined how
108 tweets mentioned the US election 2012 (local event) and Brexit 2016 (global event) affected
109 Apple and Google stock fluctuations. The percentage of positive and negative tweets was
110 calculated per day for each event. The result revealed that not all major events severely impact
111 the stock market. However, more important events like the US election can strongly affect
112 algorithm performance.

113 Many researchers have tried to make predictions by using data other than tweets. Kabbani et al.
114 [13] used financial articles from 2016-01-01 to 2020-04-01 to predict the stock trend inside a
115 given day. Considering the rapid change of sentiments, only today and tomorrow trend is
116 predicted. After correlation analysis, highly correlated features with the article's sentiment scores
117 were selected in the final data set. The model used Linear Regression, Random Forest, Gradient
118 Boosting Machine Algorithm and hit 63.58% accuracy on average. Weng et al. [8] introduced a
119 model to predict the stock movement one day ahead. Different from sentiment analysis, the
120 author used Wikipedia traffic, Google news counts, some market data, and various technical
121 indicators to make predictions. The model just focused on Apple stock fluctuations from May
122 1st, 2012, to June 1st, 2015. Combining data from multiple sources, their expert system hit
123 85.8% accuracy, reflecting that the increase in data categories can boost the prediction result.

124 Andrius et al. [12] divided news and tweets sentiments into eight emotions (e.g. anger, fear).
125 Only a few sentiment emotions showed some correlation with the future stock movements. And
126 in most cases, the prediction based on technical factors achieves a better result without emotions
127 analysis. The same conclusion also got in Khedr et al. [21] research. They used N-gram and TF-
128 IDF method to generate the weight for each token and classify collected financial news into
129 positive and negative sentiment attributes. The sentiment attributes method achieved 59.18%
130 accuracy in the K-NN classifier, and sentiment & historical stock data method achieved 89.80%
131 accuracy. However, the author did not show the accuracy achieved with sole historical data. It
132 seems that technical factors are the determining factors in stock prediction while extract
133 sentiment attributes have harmful effects on the result, similar to Andrius's conclusion.

134 Except for traditional Machine learning methods, new technologies were also applied in stock
135 predictions. Nguyen et al. [26] designed a novel aspect-based sentiment method to calculate the
136 sentimental values for each topic in one sentence. Instead of extracting hidden topics together
137 with sentiments, this new model represents each message as a set of topics with corresponding
138 sentimental values. To examine the efficiency of the novel aspect-based sentiment method, the
139 authors applied SVM on 18 stocks data from July 2012 to July 2013. The aspect-based sentiment
140 feature method achieved 71.05% high accuracy for Amazon stock. Qiu et al. [27] created a new
141 sentiment analysis model based on Baidu AI Cloud, which provides an algorithm automatically
142 mines sentiment knowledge through unsupervised learning. Besides, a novel weighted sentiment
143 index considers the holiday effect. Sentiments in non-transaction were somehow allocated to the

144   next transaction day. This novel index increased seven of the eight algorithms' performance
145   compared to the sentiment index without the holiday effect. Cryptocurrency price forecasting is
146   also a hot topic outside of stock prediction. Parekh et al. [28] presented a hybrid model, *DL-*
147   *GuesS* in cryptocurrency price prediction considering historical price and twitter sentiments. *DL-*
148   *GuesS* model adopted LSTM and GRU deep learning to predict cryptocurrency prices
149   considering different window sizes. The proposed *DL-GuesS* model achieved better Bitcoin-
150   Cash performance than the model only using Bitcoin-Cash price.

151

Materials & Methods

153

154   **Figure1:** Flowchart of methodology

155

156   The proposed model presented in Figure 1 is designed to predict stock fluctuations and help
157   stock investors make proper investment decisions. For news and tweets data, they are first sent to
158   the sentiment analysis test section to select the most appropriate technology for applying
159   sentiment analysis respectively. Then both generated weighted scores will be summed up in
160   given intervals and apply holiday effect processing. Finally, summed news and tweets scores are
161   combined according to intervals as input features. While for stock data, we turn stock
162   fluctuations into binary changes and regard such change as the expected result in machine
163   learning.

Data Collection

165   **Tweets:** Compared with Twitter official API, the Twint library has many advantages and has
166   been used in research [29-31]. Twint has no restrictions in the official Twitter API and provides
167   many options to filter data (e.g. language selection). We use the cashtag symbol to harvest twitter
168   messages of Service stock AMZN, NFLX and Technology Stock AAPL, MSFT in 2021.

169   **News:** To avoid the bias of specific financial media, we scrape news from eight websites or
170   medias including CNBC, Forbes, The Street, Reuters, The Motley Fool, Business Insider and
171   Wall Street Journal, Bloomberg [14, 16, 18]. Finally, six thousand pieces of news are retained for
172   this study.

173   **Stock Data:** Yahoo Finance is an influential financial news and data website widely adopted in
174   stock research [32-34]. It provides historical stock price fluctuations [9]. Daily price data has six
175   main features: Open price, Close price (or Adjusted Close price), Highest price, Lowest price,
176   and Volume. We download data for four selected stocks from January 4, 2021 to December 31,
177   2021

178

Data preprocessing

180   Because the tweets dataset is built by using text scraped from Twitter, raw tweets with many
181   undesired stuff need to be cleaned. Emoticons, symbols, URLs, and some meaningless texts

182 (stopwords) are removed [30, 33]. Besides, tweets containing more than three cashtags are
183 discarded as meaningless messages, the same as Bujari's research [9]. For news preprocessing,
184 we only keep news that mentioned chosen stock name exactly in titles so that we can avoid the
185 disturb of much unrelated information. Moreover, we calculate the value *Goal I: $Open_{t+1}$ - $Open_t$*
186 and *Goal II: $Open_{t+1}$ - $Close_t$* (t is given transaction date) for each stock, then transform all
187 values into a binary variable, where positive values → 1 (uptrend) and negative values → 0
188 (downtrend) [8, 9]. The equation is shown below:

189

$$Goal\ I = \begin{cases} 1, & Open_t \leq Open_{t+1} \\ 0, & Open_t > Open_{t+1} \end{cases} \#(1)$$

191

$$Goal\ II = \begin{cases} 1, & Close_t \leq Open_{t+1} \\ 0, & Close_t > Open_{t+1} \end{cases} \#(2)$$

193 ## Sentiment analysis Test

194 *Testing datasets to find the most proper sentiment analysis technology*
195 **Financial PhraseBank:** Financial PhraseBank is one primary dataset for financial area
196 sentiment analysis [35, 36], which was created by Malo et al. [37]. Financial PhraseBank
197 contains 4845 news sentences found on the LexisNexis database and marked by 16 people with
198 finance backgrounds. Annotators were required to label the sentence as positive, negative, or
199 neutral market sentiment [37]. All 4845 sentences were kept with higher than 50% agreement. In
200 our study, we will use this dataset to test the performance of the lexicon or model on news text.

201 **Tweets_labelled_09042020_16072020:** It consists of 5000 tweets randomly selected out from
202 943,672 raw tweets [30]. The raw tweets were harvested by using Twitter hashtags and cashtags
203 like #SP500, $MSFT, $AAPL. Inside this file, 1300 were manually labeled with positive,
204 negative, or neutral market sentiment and reviewed by a second independent annotator. In our
205 study, we will use this dataset to test the performance of the lexicon or model on tweets.

206 *Sentiment analysis technologies*
207 ***A. VADER sentiment analysis***

208 In our model, VADER library is the first technology used to analyze text sentiment [38]. The
209 VADER library is a lexicon and rule-based framework that can detect the three dimensions of
210 sentiment in the text [38, 39]. The generated result consists of positive, neutral, negative, and
211 compound scores ranging from -1 to 1 for each tweet.

212 Here, we just apply VADER sentiment analysis and consider the compound score to represent
213 the general sentiment of each tweet. Then, to expand the influence of those popular tweets, we
214 create a new feature Tweets Weighted: $T_{weighted}$

$$T_{weighted} = compound\ score * (retweets + 1) \#(3)$$

216   where $T_{weighted}$ is the weighted sentiment compound score for each tweet.

217   ### B. Loughran-McDonald word dictionary

218   The Loughran-McDonald dictionary was widely adopted in financial area research [40], since it
219   was the first dictionary to explore the potential benefits by using specific financial domain words
220   [41]. The dictionary was initially developed using corporate 10-k reports between 1994 and 2008
221   [42]. We use version 2012, which contains 2349 negative and 354 positive words to help us
222   evaluate text sentiment [43].

223   ### C. FinBERT Model

224   FinBERT in a language classification model based on BERT to tackle NLP tasks in the financial
225   domain [44]. FinBERT was trained by 46,143 news documents TRC2-financial corpus.
226   According to Nguyen et al.'s paper [26], we generate a News opinion value, News Weighted:
227    $N_{weighted}$

228    $$N_{weighted} = \frac{pos_{score} - neg_{score}}{pos_{score} + neg_{score}} \#(4)$$

229   *Sentiment analysis test result*
230   Figure 2 shows the testing results of VADER, Loughran-McDonald dictionary and FinBERT
231   performance on dataset Tweets_labelled_09042020_16072020. VADER gets the best
232   classification accuracy 68%; Loughran-McDonald dictionary and FinBERT gain 56%, 53%
233   classification accuracy respectively. The results reflect that VADER has better classification
234   performance on messy Twitter messages, while the FinBERT trained on well-structured news
235   text gives the worst result in tweet classification. Besides, we can find that VADER far
236   outperforms other two methods in positive tweets classification and VADER also gives twice as
237   many positive labels as the others. VADER tends to label more tweets in the testing dataset as
238   positive than the Loughran-McDonald dictionary and FinBERT do.

239   Figure 3 contains the classification accuracy of VADER, Loughran-McDonald dictionary and
240   FinBERT on Financial PhraseBank, the result for FinBERT is gathered from Araci's paper [44],
241   while other two confusion matrix are calculated by using our own results. In news classification,
242   the situation is reversed, FinBERT gets 86% accuracy, but VADER is only 54%. FinBERT is
243   highly adaptive to standard financial text written by professionals compared with tweets
244   proposed by laymen. However, VADER is more suitable for general messages and cannot
245   identify domain-specific terms well.

246   According to the test results, we will apply VADER analysis on tweets dataset, and FinBERT
247   model on news data we gathered for prediction since they present better sentiment classification
248   results.

249

250

251

252

253

254 **Figure 2:** VADER, Loughran-McDonald dictionary and FinBERT performance on
255 Tweets_labelled_09042020_16072020

256

257

258

259 **Figure 3:** VADER, Loughran-McDonald dictionary and FinBERT performance on Financial
260 PhraseBank

261

262 Data manipulation
263 Because our idea is to explore how tweets and news from different periods can predict the future
264 stock trend, we also design two time divisions: Natural hours division **0:00$_t$ ~ 0:00$_{t+1}$** and
265 Opening hours division **9:30$_t$ ~ 9:30$_{t+1}$** (t is given natural date) and get Tweets in Natural hours:
266 TN_t, and Tweets in Opening hours: TO_t

267
$$TN_t = \sum_{i = 0:00_t}^{0:00_{t+1}} T_{weighted} \#(5)$$

268
$$TO_t = \sum_{i = 9:30_t}^{9:30_{t+1}} T_{weighted} \#(6)$$

269 News in Natural hours: NN_t, and News in Opening hours: NO_t are similar to above formular
270 above, except that $T_{weighted}$ is changed to $N_{weighted}$.

271 The new variables are the summation of sentiment values in each time division. For example, in
272 **9:30$_t$ ~ 9:30$_{t+1}$** time division on January 5th, $T_{weighted}$ from January 5th, 9:30 to January 6th, 9:30
273 are summed up to generate TO_t for January 5th.

274 In addition, calendar anomalies (holiday effect, day-of-the-week effect,) are common in financial
275 markets [45-47]. The holiday effect, which leads to abnormal stock price fluctuation on Monday,
276 has been extensively studied in the stock area [9, 45, 48]. The news released on weekends or
277 holidays is one reason that changes investors' behavior [27]. Inspired by Qiu et al. [27], we
278 generate an equation to consider holiday effect as shown below (n is days of vacation):

279
$$TO_{modified-t} = e^{-n}TO_{t-n} + \ldots + e^{-1}TO_{t-1} + TO_t \#(7)$$

280
$$TN_{modified-t} = e^{-n}TN_{t-n} + ... + e^{-1}TN_{t-1} + TN_t \#(8)$$

281 Modified score $NO_{modified-t}$ and $NN_{modified-t}$ are also similar to get, just to change Tweets
282 value to corresponding News value.

283 Take $TN_{modified-t}$ as an example, if n = 2 days, the TN_{t-2}, TN_{t-1} and TN_t can stand for
284 summed $T_{weighted}$ on t = Sunday, t-1 = Saturday, t-2 = Friday. As sentiments have a stronger
285 impact more recently, sentiment in the past decreasing exponentially [49]. Therefore, we
286 concentrate weekend sentiments on Sunday and use modified Sunday weighted score to predict
287 Monday's trend.

288 Finally, we will combine modified news and tweets score according to function Combine Natural
289 Hours CN and Combine Opening Hours CO:

290
$$CN = \alpha\, TN_{modified-t} + (1-\alpha)NN_{modified-t}\#(9)$$

291
$$CO = \alpha\, TO_{modified-t} + (1-\alpha)NO_{modified-t}\#(10)$$

292 In our experiment, result is the best when $\alpha = 0.25$. Thus, we apply 0.25 in the function to get CO
293 and CN.

294

Machine Learning
296 CO, CN and Goal I, Goal II are aligned in time series. Sentiment features are the input in KNN,
297 Tree [50, 51], SVM, Random Forest (RF), Naïve Bayes (NB), Logistic Regression (LR)
298 algorithms with 10-fold cross validation to check the Goal I, Goal II prediction accuracy
299 separately. This step was implemented by using Spyder and Orange from Anaconda. The
300 parameters of the Machine Learning models are shown below:

301 • Random Forest: n_estimators=10, max_depth=None, min_samples_split=5.
302 • Naïve Bayes: MultinominalNB, alpha=1.0, force_alpha=False, fit_prior=True,
303   class_prior=None.
304 • Logistic Regression: Penalty=L2 (Ridge), C=1.
305 • KNN: n_neighbors=5, metric='cosine'
306 • SVM: Kernel=Linear, Cost = 1, Regression loss epsilon = 0.1,
307 • Tree: Splitter = best, min_samples_leaf= 2, min_samples_split = 5, max_depth=100.

308

Evaluation Methodology
310 To evaluate the performance of different machine learning algorithms, we employ Classification
311 Accuracy (CA), Precision, Recall, and F1-score. Equations of these four are shown in the
312 following:

313

314
$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \#(11)$$

315
$$Precision = \frac{TP}{TP + FP} \#(12)$$

316
$$Recall = \frac{TP}{TP + FN} \#(13)$$

317
$$F1 - score = \frac{2 Precision * Recall}{Precision + Recall} \#(14)$$

318

319 TP stands for true positive classification, and FP is also positive classification but false. TN
320 stands for true negative classification, and FN is the positive result but wrongly classified as
321 negative.

322

323 **Results**

324 Tables 1~4 presents the six algorithms' performance for AAPL, AMZN, MSFT, NFLX in *Goal*
325 *I: Open$_{t+1}$ - Open$_t$* and Tables 5~8 are the performance for those companies in *Goal II: Open$_{t+1}$ -*
326 *Close$_t$*. As the result shows, Naïve Bayes is the best classification algorithm in our model, 6 out
327 of 8 best classification accuracy comes from Naïve Bayes.

328 In *Goal I: Open$_{t+1}$ - Open$_t$* prediction, all best Accuracy, F1-score, Precision, and Recall, except
329 AAPL's F1-Score, are generated by using CN. In contrast, all best results in *Goal II: Open$_{t+1}$ -*
330 *Close$_t$* prediction are achieved with CO.

331 Therefore, since Naive Bayes algorithm has proven to be the most effective, it demonstrates
332 strong performance in TD1 for Goal I (*Open$_{t+1}$ - Open$_t$*) and even better performance in TD2 for
333 Goal II (*Open$_{t+1}$ - Close$_t$*). To determine the overall performance, we calculate the average of
334 Naive Bayes accuracy from both Goal I in TD1 and Goal II in TD2.

335

336 **TD1: Natural hours' time division from 0:00$_t$ ~ 0:00$_{t+1}$**

337 **TD2: Opening hours' time division from 9:30$_t$ ~ 9:30$_{t+1}$**

338 **Bold values represent the best performances**

339   **Table 1**: Performance of Six ML classifiers on AAPL stock's Goal I

340

341   **Table 2**: Performance of Six ML classifiers on AMZN stock Goal I

342

343   **Table 3**: Performance of Six ML classifiers on MSFT stock Goal I

344

345   **Table 4**: Performance of Six ML classifiers on NFLX stock Goal I

346

347   **Table 5**: Performance of Six ML classifiers on AAPL stock Goal II

348

349   **Table 6**: Performance of Six ML classifiers on AMZN stock Goal II

350

351   **Table 7**: Performance of Six ML classifiers on MSFT stock Goal II

352

353   **Table 8**: Performance of Six ML classifiers on NFLX stock Goal II

354

Conclusion

355

356 In this paper, we develop a new weighted sentiment index $T_{weighted}$ by using Twitter retweets

357 counts and VADER sentiment score, and we utilize FinBERT to generate $N_{weighted}$ to represent

358 news opinion value. We propose a new time division, opening hours division, to study how

359 tweets and news released in different time periods can predict next-day stock movement. Based

360 on that, the summation of scores in different time divisions is calculated. Besides, the holiday

361 effect are also considered in this paper to construct modified score, which proved to be a more

362 reliable and realistic indicator [27]. Finally, we apply KNN, Tree, SVM, Random Forest, Naïve

363 Bayes, Logistic Regression algorithms on aligned series to evaluate the performance of

364 combined score CN and CO in predicting *Goal I: Open_{t+1} - Open_t* and *Goal II: Open_{t+1} - Close_t*.

365 The experimental result shows that Naïve Bayes is the best classification algorithm among six, 6

366 out of 8 best results are produced by Naïve Bayes. The possible reason is that the dataset might

367 be highly skewed, with one class being much more prevalent than the others. NB is known to

368 perform well in such scenarios. Furthermore, CN is better in predicting *Goal I: Open_{t+1} - Open_t*.

369 Except for AAPL's F1-score, all best Accuracy, F1-score, Precision and Recall are generated by

370 using CN. In reverse, CO presents all best results in *Goal II: Open_{t+1} - Close_t* prediction. It

371 proves that non-natural time division is particularly useful in predicting specific goals. Based on

372 this result, we proposed our model for predicting *Goal I: Open_{t+1} - Open_t* and *Goal II: Open_{t+1} -*

373 *Close_t* separately in Figure 4 and Figure 5. Compared with Kabbani and Usta's result [13], which

374 gets 63.6% best accuracy by using seven features (including sentiment score and technical

375 factors). Our model gets the best 62.4% accuracy with only news and tweets sentiment features.

376

377

378

Figure 4: Proposed model for predicting *Goal I: Open$_{t+1}$ - Open$_t$*

380

381

Figure 5: Proposed model for predicting *Goal II: Open$_{t+1}$ - Close$_t$*

383

384

385

386

387

388

389

390

391

References

393

394   1.   Ou, J.A. and S.H. Penman, *Financial statement analysis and the prediction of stock returns.*
395   Journal of accounting and economics, 1989. **11**(4): p. 295-329.
396   2.   Dimson, E., *Risk measurement when shares are subject to infrequent trading.* Journal of Financial
397   Economics, 1979. **7**(2): p. 197-226.
398   3.   Checkley, M., D.A. Higón, and H. Alles, *The hasty wisdom of the mob: How market sentiment*
399   *predicts stock market behavior.* Expert Systems with applications, 2017. **77**: p. 256-263.
400   4.   Ranaldi, L., M. Gerardi, and F. Fallucchi, *Crypto Net: Using Auto-Regressive Multi-Layer Artificial*
401   *Neural Networks to Predict Financial Time Series.* Information, 2022. **13**(11): p. 524.
402   5.   Agrawal, J., V. Chourasia, and A. Mittra, *State-of-the-art in stock prediction techniques.*
403   International Journal of Advanced Research in Electrical, Electronics and Instrumentation
404   Engineering, 2013. **2**(4): p. 1360-1366.
405   6.   Holthausen, R.W. and D.F. Larcker, *The prediction of stock returns using financial statement*
406   *information.* Journal of accounting and economics, 1992. **15**(2-3): p. 373-411.
407   7.   Alzazah, F.S. and X. Cheng, *Recent Advances in Stock Market Prediction Using Text Mining: A*
408   *Survey.* E-Business-Higher Education and Intelligence Applications, 2020.
409   8.   Weng, B., M.A. Ahmed, and F.M. Megahed, *Stock market one-day ahead movement prediction*
410   *using disparate data sources.* Expert Systems with Applications, 2017. **79**: p. 153-163.
411   9.   Bujari, A., M. Furini, and N. Laina. *On using cashtags to predict companies stock trends.* in *2017*
412   *14th IEEE Annual Consumer Communications & Networking Conference (CCNC).* 2017. IEEE.

413 10. Lassen, D.S. and A.R. Brown, *Twitter: The electoral connection?* Social science computer review,
414   2011. **29**(4): p. 419-436.
415 11. Nofsinger, J.R., *Social mood and financial economics.* The Journal of Behavioral Finance, 2005.
416   **6**(3): p. 144-160.
417 12. Mudinas, A., D. Zhang, and M. Levene, *Market trend prediction using sentiment analysis: lessons*
418   *learned and paths forward.* arXiv preprint arXiv:1903.05440, 2019.
419 13. Kabbani, T. and F.E. Usta, *Predicting The Stock Trend Using News Sentiment Analysis and*
420   *Technical Indicators in Spark.* arXiv preprint arXiv:2201.12283, 2022.
421 14. García-Méndez, S., F. de Arriba-Pérez, A. Barros-Vila, and F.J. González-Castaño, *Detection of*
422   *temporality at discourse level on financial news by combining Natural Language Processing and*
423   *Machine Learning.* Expert Systems with Applications, 2022. **197**: p. 116648.
424 15. Hao, Z. and Y.-H.J. Chen-Burger, *An Investigation into Influences of Tweet Sentiments on Stock*
425   *Market Movements*, in *Agents and Multi-Agent Systems: Technologies and Applications 2022*.
426   2022, Springer. p. 87-97.
427 16. Li, Y. and Y. Pan, *A novel ensemble deep learning model for stock prediction based on stock prices*
428   *and news.* International Journal of Data Science and Analytics, 2022. **13**(2): p. 139-149.
429 17. Sharma, K. and R. Bhalla. *Stock Market Prediction Techniques: A Review Paper*. in *Second*
430   *International Conference on Sustainable Technologies for Computational Intelligence*. 2022.
431   Springer.
432 18. Srivastava, S., R. Tiwari, R. Bhardwaj, and D. Gupta. *Stock Price Prediction Using LSTM and News*
433   *Sentiment Analysis*. in *2022 6th International Conference on Trends in Electronics and*
434   *Informatics (ICOEI)*. 2022. IEEE.
435 19. Cambria, E. and B. White, *Jumping NLP Curves: A Review of Natural Language Processing*
436   *Research [Review Article].* IEEE Computational Intelligence Magazine, 2014. **9**(2): p. 48-57.
437 20. Hirschberg, J. and C.D. Manning, *Advances in natural language processing.* Science, 2015.
438   **349**(6245): p. 261-266.
439 21. Khedr, A.E. and N. Yaseen, *Predicting stock market behavior using data mining technique and*
440   *news sentiment analysis.* International Journal of Intelligent Systems and Applications, 2017.
441   **9**(7): p. 22.
442 22. Hussein, D.M.E.-D.M., *A survey on sentiment analysis challenges.* Journal of King Saud
443   University-Engineering Sciences, 2018. **30**(4): p. 330-338.
444 23. Chalothom, T. and J. Ellman, *Simple approaches of sentiment analysis via ensemble learning*, in
445   *information science and applications*. 2015, Springer. p. 631-639.
446 24. Kinyua, J.D., C. Mutigwe, D.J. Cushing, and M. Poggi, *An analysis of the impact of president*
447   *trump's tweets on the djia and S&P 500 using machine learning and sentiment analysis.* Journal
448   of behavioral and experimental finance, 2021. **29**: p. 100447.
449 25. Maqsood, H., I. Mehmood, M. Maqsood, M. Yasir, S. Afzal, F. Aadil, M.M. Selim, and K.
450   Muhammad, *A local and global event sentiment based efficient stock exchange forecasting using*
451   *deep learning.* International Journal of Information Management, 2020. **50**: p. 432-451.
452 26. Nguyen, T.H., K. Shirai, and J. Velcin, *Sentiment analysis on social media for stock movement*
453   *prediction.* Expert Systems with Applications, 2015. **42**(24): p. 9603-9611.
454 27. Qiu, Y., Z. Song, and Z. Chen, *Short-term stock trends prediction based on sentiment analysis and*
455   *machine learning.* Soft Computing, 2022: p. 1-16.
456 28. Parekh, R., N.P. Patel, N. Thakkar, R. Gupta, S. Tanwar, G. Sharma, I.E. Davidson, and R. Sharma,
457   *DL-GuesS: Deep Learning and Sentiment Analysis-based Cryptocurrency Price Prediction.* IEEE
458   Access, 2022: p. 1-1.

459   29.   Dutta, U., R. Hanscom, J.S. Zhang, R. Han, T. Lehman, Q. Lv, and S. Mishra, *Analyzing Twitter*
460   *Users' Behavior Before and After Contact by the Russia's Internet Research Agency.* Proceedings
461   of the ACM on Human-Computer Interaction, 2021. **5**(CSCW1): p. 1-24.
462   30.   Bruno Taborda, A.d.A., José Carlos Dias, Fernando Batista, Ricardo Ribeiro, *Stock Market Tweets*
463   *Data*, in *2021 IEEE 7th International Conference on Computing, Engineering and Design (ICCED)*.
464   2021, IEEE: IEEE Dataport.
465   31.   Yohapriyaa, M. and M. Uma, *Multi-variant Classification of Depression Severity Using Social*
466   *Media Networks Based on Time Stamp*, in *Intelligent Data Communication Technologies and*
467   *Internet of Things*. 2022, Springer. p. 553-564.
468   32.   Ahmar, A.S. and E.B. Del Val, *SutteARIMA: Short-term forecasting method, a case: Covid-19 and*
469   *stock market in Spain.* Science of the Total Environment, 2020. **729**: p. 138883.
470   33.   Budiharto, W., *Data science approach to stock prices forecasting in Indonesia during Covid-19*
471   *using Long Short-Term Memory (LSTM).* Journal of Big Data, 2021. **8**(1).
472   34.   Topcu, M. and O.S. Gulal, *The impact of COVID-19 on emerging stock markets.* Finance Research
473   Letters, 2020. **36**: p. 101691.
474   35.   Ding, N., Y. Qin, G. Yang, F. Wei, Z. Yang, Y. Su, S. Hu, Y. Chen, C.-M. Chan, and W. Chen, *Delta*
475   *tuning: A comprehensive study of parameter efficient methods for pre-trained language models.*
476   arXiv preprint arXiv:2203.06904, 2022.
477   36.   Ye, Q., B.Y. Lin, and X. Ren, *Crossfit: A few-shot learning challenge for cross-task generalization*
478   *in nlp.* arXiv preprint arXiv:2104.08835, 2021.
479   37.   Malo, P., A. Sinha, P. Korhonen, J. Wallenius, and P. Takala, *Good debt or bad debt: Detecting*
480   *semantic orientations in economic texts.* Journal of the Association for Information Science and
481   Technology, 2014. **65**(4): p. 782-796.
482   38.   Hutto, C.J. *vaderSentiment*. 2016; Available from: https://github.com/cjhutto/vaderSentiment.
483   39.   Pano, T. and R. Kashef, *A Complete VADER-Based Sentiment Analysis of Bitcoin (BTC) Tweets*
484   *during the Era of COVID-19.* Big Data and Cognitive Computing, 2020. **4**(4): p. 33.
485   40.   Wang, T., C. Yuan, and C. Wang, *Does applying deep learning in financial sentiment analysis lead*
486   *to better classification performance?* 2020.
487   41.   Karalevicius, V., N. Degrande, and J. De Weerdt, *Using sentiment analysis to predict interday*
488   *Bitcoin price movements.* The Journal of Risk Finance, 2018.
489   42.   Loughran, T. and B. McDonald, *When is a liability not a liability? Textual analysis, dictionaries,*
490   *and 10-Ks.* The Journal of finance, 2011. **66**(1): p. 35-65.
491   43.   Loughran, T. and B. McDonald, *Textual analysis in accounting and finance: A survey.* Journal of
492   Accounting Research, 2016. **54**(4): p. 1187-1230.
493   44.   Araci, D., *Finbert: Financial sentiment analysis with pre-trained language models.* arXiv preprint
494   arXiv:1908.10063, 2019.
495   45.   Berument, H. and H. Kiymaz, *The day of the week effect on stock market volatility.* Journal of
496   economics and finance, 2001. **25**(2): p. 181-193.
497   46.   Jacobs, B.I. and K.N. Levy, *Calendar anomalies: Abnormal returns at calendar turning points.*
498   Financial Analysts Journal, 1988. **44**(6): p. 28-39.
499   47.   Kiymaz, H. and H. Berument, *The day of the week effect on stock market volatility and volume:*
500   *International evidence.* Review of financial economics, 2003. **12**(4): p. 363-380.
501   48.   Shiller, R.J., *From efficient markets theory to behavioral finance.* Journal of economic
502   perspectives, 2003. **17**(1): p. 83-104.
503   49.   Barberis, N., R. Greenwood, L. Jin, and A. Shleifer, *X-CAPM: An extrapolative capital asset pricing*
504   *model.* Journal of financial economics, 2015. **115**(1): p. 1-24.
505   50.   scikit-learn. *1.10. Decision Trees*. 2023; Available from: https://scikit-
506   learn.org/stable/modules/tree.html.

507   51.   Gérard BIAU, G.C., Ludovic KOEHLER, and Pierre-Henri WATTELLE, *Random Forests and Decision*
508   *Trees: A Comparison.* Journal of Machine Learning Research, 2018. **19**: p. 1-25.

509

Performance of Six ML classifiers on AAPL stock's Goal I

Model	Accuracy		F1 – score		Precision		Recall	
	TD1	TD2	TD1	TD2	TD1	TD2	TD1	TD2
KNN	0.532	0.569	0.528	0.569	0.529	0.569	0.532	0.569
LR	0.600	0.502	0.551	0.351	**0.635**	0.270	0.600	0.502
NB	**0.604**	0.545	0.577	0.543	0.616	0.544	**0.604**	0.545
RF	0.552	0.584	0.551	**0.584**	0.551	0.584	0.552	0.584
SVM	0.484	0.471	0.373	0.304	0.547	0.224	0.484	0.471
Tree	0.496	0.541	0.488	0.538	0.489	0.539	0.496	0.541

1   **Table 3**: Performance of Six ML classifiers on AAPL stock's Goal I

Performance of Six ML classifiers on AMZN stock Goal I

Model	Accuracy		F1 – score		Precision		Recall	
	TD1	TD2	TD1	TD2	TD1	TD2	TD1	TD2
KNN	0.600	0.514	0.597	0.510	0.597	0.509	0.600	0.514
LR	0.600	0.545	0.582	0.385	0.599	0.297	0.600	0.545
NB	**0.624**	0.510	**0.600**	0.421	**0.633**	0.446	**0.624**	0.510
RF	0.540	0.561	0.540	0.561	0.540	0.561	0.540	0.561
SVM	0.476	0.475	0.428	0.387	0.507	0.529	0.476	0.475
Tree	0.540	0.510	0.537	0.505	0.536	0.505	0.540	0.510

1 **Table 4**: Performance of Six ML classifiers on AMZN stock Goal I

Performance of Six ML classifiers on MSFT stock Goal I

Model	Accuracy		F1 – score		Precision		Recall	
	TD1	TD2	TD1	TD2	TD1	TD2	TD1	TD2
KNN	0.520	0.451	0.515	0.443	0.515	0.441	0.520	0.451
LR	0.536	0.533	0.380	0.383	0.294	0.406	0.536	0.533
NB	**0.600**	0.482	**0.573**	0.417	**0.604**	0.428	**0.600**	0.482
RF	0.464	0.463	0.459	0.459	0.457	0.458	0.464	0.463
SVM	0.488	0.502	0.417	0.433	0.546	0.570	0.488	0.502
Tree	0.472	0.494	0.467	0.484	0.465	0.484	0.472	0.494

1   **Table 5**: Performance of Six ML classifiers on MSFT stock Goal I

Performance of Six ML classifiers on NFLX stock Goal I

Model	Accuracy		F1 – score		Precision		Recall	
	TD1	TD2	TD1	TD2	TD1	TD2	TD1	TD2
KNN	0.500	0.443	0.498	0.442	0.499	0.442	0.500	0.443
LR	**0.588**	0.494	**0.588**	0.364	**0.589**	0.409	**0.588**	0.494
NB	0.512	0.478	0.507	0.468	0.511	0.473	0.512	0.478
RF	0.468	0.478	0.468	0.478	0.468	0.480	0.468	0.478
SVM	0.496	0.502	0.409	0.479	0.505	0.510	0.496	0.502
Tree	0.512	0.502	0.508	0.502	0.511	0.502	0.512	0.502

1 **Table 6**: Performance of Six ML classifiers on NFLX stock Goal I

Performance of Six ML classifiers on AAPL stock Goal II

Model	Accuracy		F1 – score		Precision		Recall	
	TD1	TD2	TD1	TD2	TD1	TD2	TD1	TD2
KNN	0.488	0.570	0.482	0.569	0.480	0.567	0.488	0.570
LR	0.568	0.570	0.412	0.417	0.323	0.329	0.568	0.570
NB	0.524	**0.609**	0.419	**0.585**	0.415	**0.600**	0.524	**0.609**
RF	0.540	0.516	0.539	0.515	0.537	0.514	0.540	0.516
SVM	0.532	0.484	0.464	0.476	0.481	0.518	0.532	0.484
Tree	0.544	0.539	0.543	0.537	0.542	0.536	0.544	0.539

1   **Table 7**: Performance of Six ML classifiers on AAPL stock Goal II

Performance of Six ML classifiers on AMZN stock Goal II

Model	Accuracy		F1 − score		Precision		Recall	
	TD1	TD2	TD1	TD2	TD1	TD2	TD1	TD2
KNN	0.532	0.551	0.524	0.542	0.520	0.537	0.532	0.551
LR	0.616	**0.621**	0.470	0.476	0.379	0.386	0.616	**0.621**
NB	0.616	**0.621**	0.470	0.476	0.379	0.386	0.616	**0.621**
RF	0.488	0.520	0.486	0.518	0.484	0.517	0.488	0.520
SVM	0.604	0.547	0.471	0.533	0.454	0.526	0.604	0.547
Tree	0.488	0.594	0.483	**0.582**	0.480	**0.578**	0.488	0.594

1   **Table 8**: Performance of Six ML classifiers on AMZN stock Goal II

Performance of Six ML classifiers on MSFT stock Goal II

Model	Accuracy		F1 – score		Precision		Recall	
	TD1	TD2	TD1	TD2	TD1	TD2	TD1	TD2
KNN	0.492	**0.590**	0.488	**0.587**	0.485	**0.586**	0.492	**0.590**
LR	0.560	0.578	0.411	0.424	0.324	0.334	0.560	0.578
NB	0.560	0.551	0.430	0.501	0.468	0.514	0.560	0.551
RF	0.496	0.512	0.493	0.513	0.491	0.516	0.496	0.512
SVM	0.444	0.508	0.397	0.506	0.487	0.504	0.444	0.508
Tree	0.528	0.523	0.515	0.514	0.513	0.511	0.528	0.523

1   **Table 9**: Performance of Six ML classifiers on MSFT stock Goal II

Performance of Six ML classifiers on NFLX stock Goal II

Model	Accuracy		F1 – score		Precision		Recall	
	TD1	TD2	TD1	TD2	TD1	TD2	TD1	TD2
KNN	0.516	0.555	0.514	0.547	0.512	0.546	0.516	0.555
LR	0.556	0.563	0.403	0.408	0.316	0.320	0.556	0.563
NB	0.560	**0.594**	0.459	**0.568**	0.526	**0.584**	0.560	**0.594**
RF	0.488	0.566	0.489	0.566	0.491	0.566	0.488	0.566
SVM	0.484	0.531	0.442	0.482	0.540	0.494	0.484	0.531
Tree	0.492	0.551	0.486	0.546	0.484	0.544	0.492	0.551

1   **Table 10**: Performance of Six ML classifiers on NFLX stock Goal II

Figure 1

Flowchart of methodology

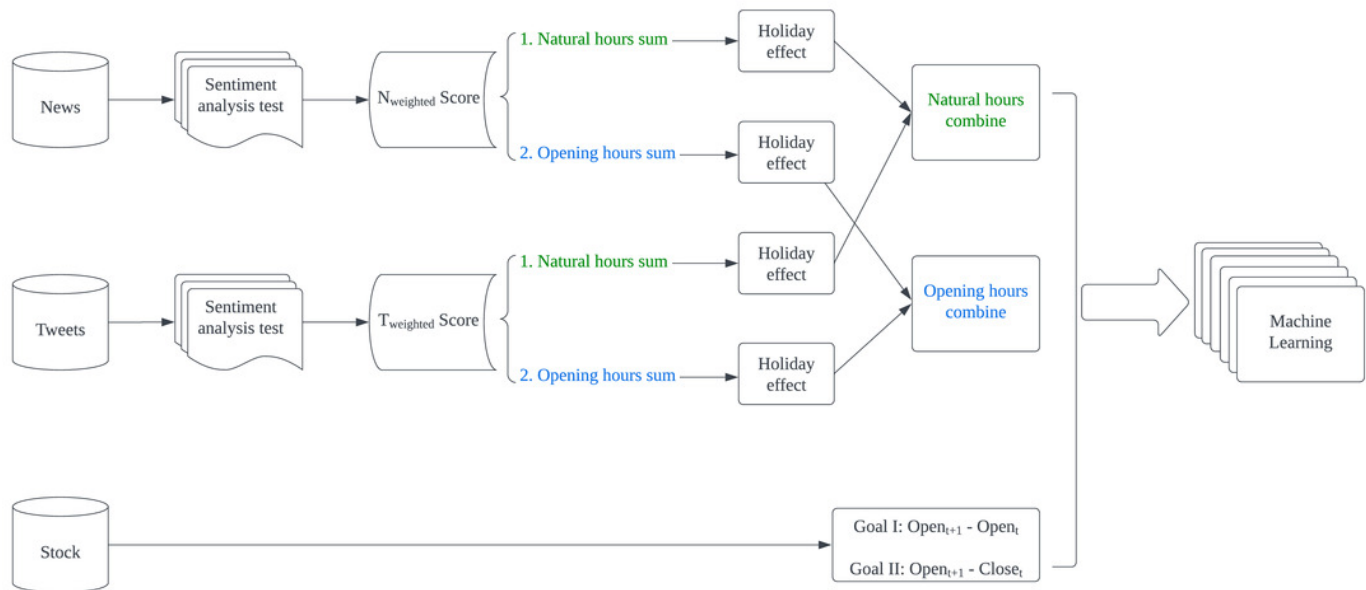



Figure 2

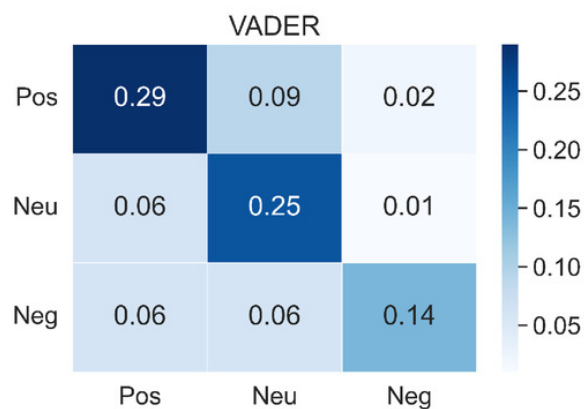
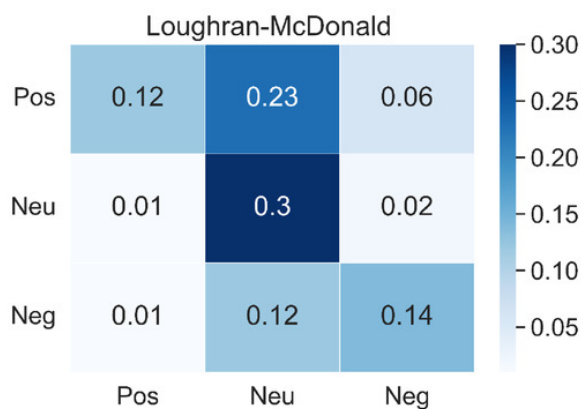
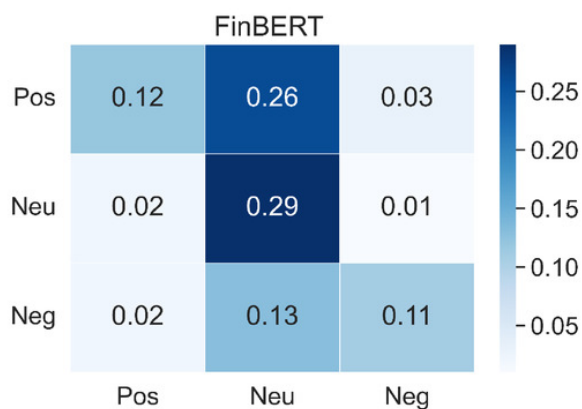
VADER, Loughran-McDonald dictionary and FinBERT performance on Tweets_labelled_09042020_16072020

(A) VADER on tweets

(B) Loughran-McDonald dictionary on tweets

(C) FinBERT on tweets

Figure 3

VADER, Loughran-McDonald dictionary and FinBERT performance on Financial
PhraseBank

(A) VADER on news

(B) Loughran-McDonald dictionary on news

(C) FinBERT on news

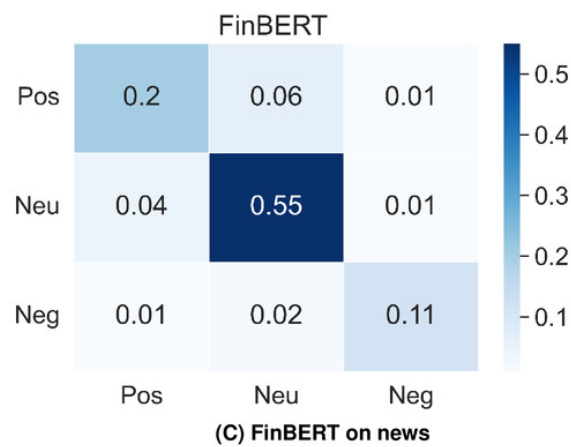
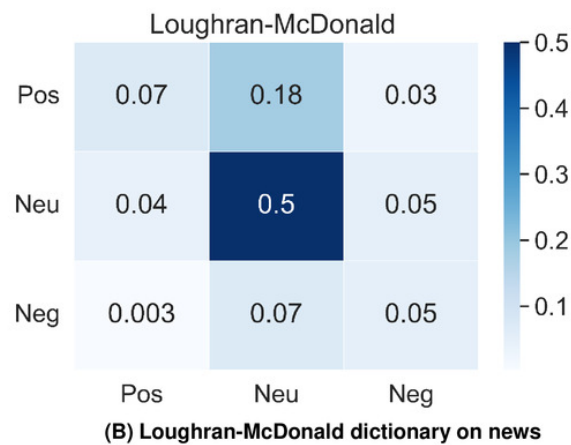
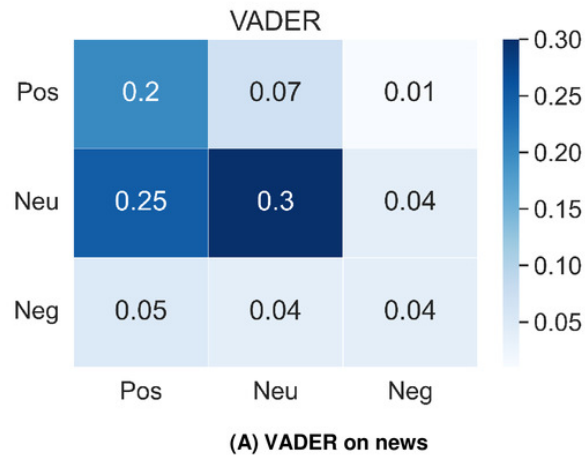



Figure 4

Proposed model for predicting *Goal I: Open$_{t+1}$ - Open$_t$*

Figure 5

Proposed model for predicting *Goal II: Open$_{t+1}$ - Close$_t$*