

Improving stock price prediction using technical indicators and sentiment analysis

Fatemeh Moodi, Amir Jahangard-Rafsanjani*, Sajad Zarifzadeh

Ph.D. Student, Department of computer engineering, Yazd University, Yazd, Iran ¹, Assistance Professor, Department of computer engineering, Yazd University, Yazd, Iran², Associated Professor, Department of computer engineering, Yazd University, Yazd, Iran³
E-mails: moodi@stu.yazd.ac.ir; jahangard@yazd.ac.ir, szarifzadeh@yazd.ac.ir

* means corresponding author

Abstract

With the increasing public use of social media platforms like Twitter, public opinion is influenced by posts shared by both influential individuals and the general public. These opinions and posts, particularly in the context of the stock market, guide traders in making buy or sell decisions. This study examines the impact of sentiment analysis on improving the accuracy of stock price prediction using six base estimators and five technical indicators. Experiments showed that selecting an appropriate time window (3, 7, 14, 30, 45, and 60 days) positively affects model performance. The prediction target was the stock price 7 days ahead. The results indicated that incorporating sentiment analysis improves prediction accuracy. The best-performing models were Linear Regression (LR), Multilayer Perceptron Regression (MLP), and Random Forest Regression (RF). Findings also revealed a correlation between tweet volume and trading volume, as well as a strong relationship between the average positive sentiment score in a day and the stock price. Specifically, based on the Mean Squared Error (MSE) evaluation metric, the best methods were LR, MLP, and RF. Stock price prediction improved by 17.37% using MLP and sentiment analysis, and by 34.78% using LR with sentiment analysis.

Keywords

Stock price prediction, Sentiment analysis, Technical indicator, Regression method, Big data.

Introduction

Social media are platforms where people publish their feelings and thoughts on various subjects. People or companies that are the subjects of these thoughts are affected by these posts. One of the main factors in predicting stock prices and understanding market behavior is sentiment analysis. Human factors such as public sentiments, the number of competitors, news, and media affect stock prices, and data on social networks such as Twitter can be used as a source for data analysis and price predictions. In this article, we demonstrate the power of social media and its impact on financial markets.

Proposed Work and Methodology

This article uses a heuristic method for estimating tweets based on VADER. The impact of different time window sizes has also been investigated. Basic regression methods are used for prediction. The results of these methods are compared with and without sentiment analysis, using six technical indicators: Decay, UO (Ultimate Oscillator), WCP (Weighted Closing Price), Willr (William's R%), and KDJ. We utilized several specific features based on the sentiment analysis performed by VADER. First, the number of positive, negative, and neutral comments was counted in the Tweets dataset each day. We also used the mean positive score, mean neutral score, and mean negative score for one day. Additionally, we used the tweet volume feature for one day and, finally, the ratio of positive tweets to negative tweets. The correlation between the variables created in the dataset was also examined.

Conclusion

As the digital currency market continues to evolve, the analysis of instrumental emotions will remain essential for understanding and predicting price movements. Here, a heuristic method was used to estimate the sentiment of tweets with VADER. In addition, we compared the effects of different window sizes on the performance of the model and showed through experiments that an appropriate time window size positively affects the model's performance. The time window sizes were set at 3, 7, 14, 30, 45, and 60 days, respectively, and the best time window was selected to predict stock prices. The goal was to predict stock prices after 7 days. The best models for predicting stock prices were linear regression, multilayer perceptron regression, and random forest regression. The results showed that the volume of tweets was associated with the volume of transactions, as well as the mean score of positive sentiment with stock prices.

بهبود پیش‌بینی قیمت سهام با استفاده از نشانگرهای فنی و تحلیل احساسات

فاطمه مودی

دانشجوی دکتری مهندسی کامپیوتر، دانشکده مهندسی کامپیوتر، دانشگاه یزد، یزد، ایران

امیر جهانگرد رفسنجانی

استادیار، دانشکده مهندسی کامپیوتر، دانشگاه یزد، یزد، ایران

سجاد ظریف‌زاده

دانشیار، دانشکده مهندسی کامپیوتر، دانشگاه یزد، یزد، ایران

چکیده

با افزایش استفاده عموم از رسانه‌های اجتماعی مثل توئیتر، افکار عمومی تحت تأثیر پست‌های منتشر شده توسط افراد مؤثر و عموم افراد قرار می‌گیرند. این افکار و پست‌های منتشر شده به خصوص در حوزه بازار سهام، به معامله‌گران که قصد خرید یا فروش سهام را دارند جهت می‌دهد. در این مقاله تأثیر تحلیل احساسات بر افزایش دقت پیش‌بینی قیمت سهام با استفاده از شش تخمین‌گر پایه و پنج نشانگر فنی بررسی شده است. آزمایش‌ها نشان دادند که اندازه پنجره زمانی مناسب (۳، ۷، ۱۴، ۳۰، ۴۵ و ۶۰ روزه) تأثیر مثبتی بر عملکرد مدل دارد. هدف پیش‌بینی قیمت سهام پس از ۷ روز بود. نتایج نشان‌دهنده بهبود پیش‌بینی با استفاده از تحلیل احساسات بود. بهترین مدل‌ها جهت پیش‌بینی قیمت سهام مدل‌های رگرسیون خطی، رگرسیون پرسپترون چندلایه و رگرسیون جنگل تصادفی بود. نتایج نشان داد که حجم توئیتهای با حجم معاملات در ارتباط است، همچنین میانگین امتیاز نظرات مثبت در یک روز، با قیمت سهام ارتباط بالایی داشت. به طور خاص، بهترین روش‌ها براساس MSE، روش‌های LR، MLP و RF هستند. پیش‌بینی قیمت سهام با روش MLP با تحلیل احساسات ۳۷/۱۷ درصد و با روش LR و با تحلیل احساسات ۳۴/۷۸ درصد بهبود یافته است.

کلمات کلیدی

پیش‌بینی قیمت سهام، تحلیل احساسات، نشانگر فنی، روش‌های رگرسیون، داده حجیم.

نام نویسنده مسئول: دکتر امیر جهانگرد رفسنجانی

ایمیل نویسنده مسئول: jahangard@yazd.ac.ir

تاریخ ارسال مقاله: ۱۴۰۲/۰۷/۲۴

تاریخ(های) اصلاح مقاله: ۱۴۰۴/۰۴/۰۴

تاریخ پذیرش مقاله: ۱۴۰۳/۰۴/۱۳

۱- مقدمه

تا خطرهای یک سرمایه‌گذاری را با تحلیل‌ها به شدت کاهش دهید. عدم اهمیت به آن، منجر به از دست‌دادن سرمایه خواهد شد. با توجه به سابقه طولانی پیش‌بینی سهام، اکثر مطالعات سعی در استفاده از نشانگرها برای قضاوت در مورد پیش‌بینی سهام [۳] داشته‌اند. در یک بررسی گسترده که ۵۷ مقاله منتشر شده در مجلات با تأثیر بالا را ترسیم کردند، نشان داده شد که استفاده از تکنیک‌های یادگیری ماشین برای پیش‌بینی قیمت سهام هنوز یک موضوع تحقیقاتی بسیار مورد بحث است [۴]. بررسی گسترده روش‌های رگرسیون با جنبه‌های مختلف استخراج شده از احساسات توئیتهای تاکنون انجام نشده است. در این مقاله، قدرت رسانه‌های اجتماعی و تأثیر آن در بازارهای مالی را نشان خواهیم داد. پیش‌بینی قیمت سهام با روش‌های مختلفی انجام می‌شود که به شرح زیر هستند:

- پیش‌بینی قیمت سهام با تحلیل فنی: یکی از راه‌های مؤثری که معامله‌گران برای پیش‌بینی قیمت سهام از آن‌ها استفاده می‌کنند، کمک گرفتن از تحلیل فنی (Technical) است [۵]. اساس این تحلیل روی قیمت و تاریخچه انواع دارایی‌های مالی تمرکز دارد. این تحلیل با کمک نمودارها می‌تواند پیش‌بینی خوبی از آینده‌ی یک دارایی مالی ارائه دهد. در تحلیل فنی قیمت‌های حال حاضر و انتظاری که از آن‌ها می‌رود را با گذشته آن قیمت بررسی می‌کنند. آن چیزی که در تحلیل فنی برای

رسانه‌های اجتماعی مکان‌هایی هستند که افراد احساسات و افکار خود را در مورد یک موضوع منتشر می‌کنند. مؤسسات، سازمان‌ها، افراد یا شرکت‌هایی که موضوع این افکار هستند، تحت تأثیر این پست‌ها قرار می‌گیرند. تعداد افرادی که از رسانه‌های اجتماعی استفاده می‌کنند، به طور قابل توجهی افزایش یافته است. تعداد کاربران رسانه‌های اجتماعی در سال ۲۰۰۴ کمتر از ۱ میلیون نفر بود و تعداد کاربران فیس‌بوک تا سال ۲۰۱۸ به ۲.۲۶ میلیارد رسیده است [۱]. بسیاری دیگر از پلتفرم‌های رسانه‌های اجتماعی برای ابراز وجود مردم ایجاد شده‌اند. در سال ۲۰۱۹ توئیتر به‌عنوان یکی از بزرگ‌ترین پلتفرم‌ها از نظر تعداد کاربران، بیش از ۳۳۰ میلیون کاربر داشت. در توئیتر، کاربران ایده‌هایی درباره هر نوع رویداد زندگی و همچنین برخی از مسائل مهم به اشتراک می‌گذارند.

پیش‌بینی بازار سهام کار بسیار پیچیده‌ای است و برای پیش‌بینی دقیق و کارآمدتر آینده بازار باید عوامل مختلفی در نظر گرفته شود. برخی از عوامل مؤثر بر وضعیت بازار جهت پیش‌بینی قیمت سهام عبارتند از: برآورد درآمدهای آتی، انتشار اخبار در مورد سود، اعلام سود سهام، تغییرات مدیریتی و غیره [۲]. روش متداولی که اکثر معامله‌گران برای پیش‌بینی بازار سهام استفاده می‌کنند، استفاده از نشانگرهای فنی است. پیش‌بینی قیمت سهام کمک می‌کند

- ارائه روشی برای تعیین امتیاز احساسات موجود در توییت‌ها با VADER.
- استفاده از جنبه‌های مختلف از تحلیل احساسات توییت‌ها، ویژگی‌هایی مانند: تعداد نظرات مثبت، تعداد نظرات منفی، میانگین امتیاز نظرات مثبت، میانگین امتیاز نظرات منفی، حجم توییت‌ها در یک روز و نسبت توییت-های مثبت به توییت‌های منفی.
- ارزیابی تأثیر پنجره‌های زمانی مختلف در پیش‌بینی قیمت سهام.
- بررسی ارتباط بین حجم توییت‌ها و حجم معاملات.
- بررسی رابطه بین میانگین تعداد توییت‌های مثبت و قیمت بسته‌شدن سهام.

۲- پیشینه تحقیق

در مقاله‌ای در سال ۲۰۱۵، به بررسی همبستگی بین داده‌های وب و بازارهای مالی پرداخته شده است [۸]. آن‌ها رابطه بین بازده قیمت سهام و احساسات بیان‌شده در توییت‌های مالی ارسال‌شده در تویتر را بررسی کردند. سه دسته اصلی از داده‌ها را در نظر گرفتند: اخبار وب، پرسش‌های موتور جستجو و رسانه‌های اجتماعی. اینجا به‌جای استفاده از احساسات اعلام سود، از احساسات جمع بیان‌شده در توییت‌های مالی استفاده شده است. ۳۰ سهم از شاخص DJIA تجزیه و تحلیل شده است. داده‌های سهام برای یک دوره ۱۵ ماهه بین سال‌های ۲۰۱۳ و ۲۰۱۴ جمع‌آوری شده‌اند. در این مقاله ابتدا رویدادهایی شناسایی شد که با افزایش فعالیت کاربران تویتر مشخص شده‌اند و سپس رفتار بازار در روزهای بعد از رویدادها مشاهده شد. با استفاده از همبستگی پیرسون و ولایت گرنجر (Granger)، وابستگی قابل توجهی بین احساسات تویتر و بازده غیرعادی در طول اوج حجم تویتر پیدا شد. این نه تنها برای پیک‌های مورد انتظار حجم تویتر (به عنوان مثال، اطلاعیه‌های فصلی)، بلکه برای پیک‌های مربوط به رویدادهای کمتر آشکار نیز معتبر است. طبق مطالعات آن‌ها، قطبیت احساسات پیک‌های تویتر، نشان‌دهنده‌ی جهت بازده‌های غیرعادی تجمعی است. مقدار بازده غیرعادی تجمعی نسبتاً کم است (حدود ۱ تا ۲ درصد)، اما این وابستگی از نظر آماری برای چندین روز پس از رویدادها معنی‌دار است.

در سال ۲۰۲۰، دلان یا تأثیرگذاران احتمالی در تویتر مورد بررسی قرار گرفتند [۹]. شرکت‌های هدف در اینجا، گوگل، آمازون، اپل، تسلا و مایکروسافت که از بزرگترین شرکت‌های موجود در بازار بورس NASDAQ هستند، بودند. این مطالعه برای داده‌های ۵ سال (۲۰۱۵-۲۰۲۰) انجام شد. در مجموع ۳,۷۱۷,۹۶۴ تعداد توییت منحصربه‌فرد جمع‌آوری شد. در این مطالعه، تجزیه و تحلیل احساس با استفاده از فرهنگ لغت تجزیه و تحلیل احساسات لوگرن و مک دونالد^۱ استفاده شد. نتایج تحلیل احساسات برای مدل‌سازی الگوریتم‌های مختلف یادگیری ماشین استفاده شد. ۸ روش مختلف یادگیری ماشین شامل: SVMهای با هسته RBF، درخت تصمیم، AdaBoost، فرآیندهای گاوسی، نزدیک‌ترین همسایگان، تشخیص درجه دوم، جنگل تصادفی، پرسپترون چندلایه، ارزیابی شدند. فرآیندهای مدل‌های یادگیری ماشین با استفاده از روش جستجوی شبکه‌ای تنظیم شد. با مدل‌ها، افرادی که اثرات مثبت یا منفی بیش از حد زیادی به عنوان دلان یا تأثیرگذاران احتمالی داشتند، شناسایی شدند. نتیجه‌گیری شد که همبستگی بین تعداد توییت‌ها و حجم شرکت ۰/۴۴۸۷ است. به طور مشابه، همبستگی بین تعداد توییت و حجم شرکت برای GOOGL ۰/۵۹۰۱۶، برای GOOG ۰/۴۳۹۳۱۸، برای TSLA ۰/۵۴۴۶۰۵، برای MSFT ۰/۲۰۹۶۶۷، برای AMZN ۰/۲۹۱۱۱۴ است. براساس نتایج SVMهای هسته RBF نتایج بهتری از سایر تکنیک‌های طبقه‌بندی کننده می‌دهند.

پیش‌بینی بازار سهام اهمیت دارد، زمان است.

- پیش‌بینی قیمت سهام با تحلیل بنیادی: با تحلیل بنیادی (Fundamental) معامله‌گران می‌توانند به ارزش واقعی یک دارایی پی ببرند. البته کمتر از این روش استفاده می‌شود و معامله‌گران روی تحلیل فنی تمرکز بیشتری دارند. در تحلیل بنیادی برای تشخیص روند، عوامل مختلفی که روی قیمت یک دارایی مالی تأثیر می‌گذارند، بررسی می‌شوند. در یک بازار مالی در صورتی که ارزش واقعی یک دارایی از ارزش حال حاضر آن بیشتر باشد، نشان می‌دهد که قیمتی که در حال حاضر وجود دارد، پایین و ارزان است. اما وقتی ارزش ذاتی یک دارایی از ارزش حال حاضر آن کمتر باشد، نشان‌دهنده بالا بودن قیمت حال حاضر است. معامله‌گران برای پیش‌بینی بازار سهام در تحلیل بنیادی؛ به ارزش حقیقی یک دارایی توجه می‌کنند و خرید و فروش خود را انجام می‌دهند.
- پیش‌بینی قیمت سهام با تحلیل احساسات: یکی از عوامل اصلی برای پیش‌بینی قیمت سهام و درک رفتار بازار، تحلیل احساسات است. عوامل انسانی مانند احساسات عمومی، تعداد رقبای، اخبار و رسانه‌ها بر قیمت سهام تأثیرگذار هستند و می‌توان از داده‌های موجود در شبکه‌های اجتماعی مثل تویتر به عنوان منبع داده‌ای برای تحلیل احساسات (Sentiment Analysis) و پیش‌بینی قیمت‌ها استفاده کرد. با این حال، تحلیل احساسات نیازمند توسعه‌ی روش‌های مؤثر برای دستیابی به نتایج دقیق‌تر و قابل اطمینان است. بنابراین، تحقیقات جامع در زمینه تحلیل احساسات سهام برای بهبود دقت پیش‌بینی قیمت‌ها اجتناب‌ناپذیر است [۶]. تجزیه و تحلیل نظرات کاربران، که با عنوان تجزیه و تحلیل احساسات شناخته می‌شود [۷]، روشی برای اندازه‌گیری ابعاد مختلف احساسی موجود در بیان انسان است. کاربران در مورد بازارهای سهام در سراسر جهان در تویتر، توییت می‌کنند، یا ایده‌های سیاسی خود را به اشتراک می‌گذارند. تعداد کاربران و همچنین شرکت‌هایی که تحت تأثیر محتوای این توییت‌ها قرار گرفته‌اند نیز افزایش یافته است. برخی از کاربران بازار سهام که متوجه این تغییر شده‌اند شروع به دنبال کردن توییت‌ها در هنگام تصمیم‌گیری مالی کردند.
- نشانگرهای تحلیل فنی بازار سهام: یک نشانگر فنی اساساً یک نمایش ریاضی براساس مجموعه داده‌هایی مانند قیمت (بالا، پایین، باز شدن، بسته شدن و غیره) برای پیش‌بینی روند قیمت است. نشانگرهای فنی مجموعه‌ای از ابزارهایی هستند که در نمودار معاملاتی به کار می‌روند تا به شفاف‌تر کردن تحلیل بازار برای معامله‌گران کمک کنند. نشانگرها کارکردهای متفاوتی دارند. نشانگرهای فنی محاسبات ریاضی هستند که از قیمت و حجم گذشته برای تشخیص جهت و قدرت روندهای بازار استفاده می‌کنند، که به چهار دسته‌ی روند، حرکت (Momentum)، حجم (Volume) و نوسان (Volatility) تقسیم می‌شوند. برای مثال؛ نشانگر حرکت؛ سرعت و بزرگی حرکات جهتی قیمت را اندازه‌گیری می‌کند، نشانگر حجم، به قیمت و حجم مربوط می‌شود. نشانگر نوسانات، نشان‌دهنده درجه‌ی نوسان قیمت است.

در این مقاله سعی کردیم با استفاده از اندیکاتورهای فنی و جنبه‌های مختلف احساسات به کار رفته در توییت‌ها، قیمت سهام را پیش‌بینی کنیم. اینجا از روش‌های مختلف رگرسیون جهت پیش‌بینی قیمت سهام استفاده شده است.

نوآوری‌های این مقاله به شرح زیر می‌باشند:

¹ Loughran - McDonald (LMD)

در سال ۲۰۱۱، نویسندگان مقاله‌ای با پیش‌بینی تغییرات روزانه‌ی بسته- شدن شاخص دوجونز با دقت ۸۷/۶ درصد توجه بسیاری را به خود جلب کردند [۱۰]. علاوه بر این، بیش از ۹ میلیون توییت از حدود ۲/۷ میلیون کاربر را بین ۲۸ فوریه و ۱۹ دسامبر ۲۰۰۸ را تحلیل کردند. برای استخراج احساس توییت‌ها، از دو ابزار نرم‌افزاری رایگان موجود به نام‌های OF (Opinion Finder) و Google GOMS (Profile of Mood States) استفاده کردند. OF با استفاده از ۲۷۲۸ کلمه مثبت و ۴۹۱۲ کلمه منفی، یک واژه‌نامه احساس را با توییت‌ها مطابقت داد. درحالی‌که GOMS شش بعد مختلف حالت روح را شناسایی کرد؛ از جمله ساکن، هوشیار، مطمئن، حیاتی، مهربان و خوشحال. آن‌ها برای پیش‌بینی، هشت سری زمانی موردنظر را مدنظر قرار دادند: سری زمانی OF، شش سری زمانی GOMS و قیمت‌های گذشته‌ی شاخص دوجونز برای پیش‌بینی ارزش شاخص دوجونز، از یک شبکه عصبی برای پیش‌بینی ارزش شاخص دو جونی براساس ارزش‌های سه روز گذشته استفاده کردند. استفاده از ارزش‌های شاخص دو جونی سه روز گذشته، باعث دقت پیش‌بینی ۷۳/۳ درصد شد که به‌نظر می‌رسد برای تنها در نظر گرفتن ارزش‌های گذشته‌ی یک سری زمانی، مقدار نسبتاً بالایی است [۱۱]. ترکیب ارزش‌های شاخص دو جونی و مقادیر ساکن GOMS باعث دقت ۸۷/۶ درصد شد. در سال ۲۰۱۳، بولن جوهان و هوآینا مائو برای کار خود در این زمینه، یک پتنت آمریکایی دریافت کردند [۱۲].

به دلیل فعالیت‌های هک در حساب‌های توییتر و وب‌سایت‌ها، برخی از قیمت‌های سهام به شدت تحت تأثیر قرار گرفته است [۱۳]. همچنین مطالعات زیادی وجود دارد که تأثیر رسانه‌های اجتماعی را بر بازارهای بورس بررسی می‌کند. یکی از این مطالعات در سال ۲۰۱۸ توسط فیلیپس و همکاران در توییتر انجام شد [۱۴]. در اینجا تأثیر افکار عمومی برای یک نام تجاری در توییتر بر ارزش بازار آن بررسی شده است که با تجزیه و تحلیل احساسات به دقت ۸۷ درصد می‌رسد. همچنین بیان شده است که بدون تجزیه و تحلیل احساسات، دقت به ۶۸ درصد کاهش می‌یابد. تحقیق دیگری در مورد تأثیر مقالات اخبار مالی بر بازار سهام توسط لی و همکاران انجام شد [۱۵]. این مقالات برای ۶ سال از بورس هنگ‌کنگ جمع‌آوری شد. تجزیه و تحلیل احساسات انجام شد و سعی شد نتایج با حرکات بازار سهام مرتبط باشد. اینجا از واژه‌نامه تحلیل احساسات مالی استفاده شد. در این مطالعه، مدلی با ماشین‌های بردار پشتیبان با استفاده از تحلیل احساسات با استفاده از ۲ فرهنگ لغت مختلف تحلیل احساسات ایجاد شده است.

در تحلیل فنی از داده‌های گذشته بازار برای پیش‌بینی جهت قیمت‌ها استفاده می‌شود و برای شناسایی الگوها به روش‌های آماری وابسته است [۱۶، ۱۷]. برای حل مشکل پیش‌بینی قیمت بسته‌شدن سهام، دو مدل پیش‌بینی جدید مبتنی بر خوشه‌بندی براساس مدل‌های پیش‌بینی دومارحله‌ای پیشنهاد شد [۱۸]. در این مقاله از ده نشانگر فنی استفاده شده است. مدل‌های آماری، مدل‌های اقتصادسنجی، مدل‌های یادگیری ماشین، مدل‌های یادگیری عمیق (DL) و مدل‌های ترکیبی برای پیش‌بینی استفاده شده است. این کار قیمت بسته‌شدن سهام را برای ۱، ۵، ۱۰، ۲۰ و ۳۰ روز پیش‌بینی می‌کند. سپس الگوریتم خوشه‌بندی K-Means [۱۹] روی هر دو نشانگر فنی از ده نشانگر فنی بیان‌شده انجام شده است. بهترین و بدترین نتایج خوشه‌بندی؛ براساس ضریب Silhouette برای ارزیابی دو عامل، یعنی انسجام و جدایی خوشه‌بندی، برای تقسیم نمونه‌های داده اولیه سهام به دو خوشه مختلف انتخاب می‌شوند. از الگوریتم یادگیری گروهی Bagging [۲۰] برای یک مدل پیش‌بینی جدید به نام یادگیری گروهی-رگرسیون بردار پشتیبان و جنگل تصادفی (ESVR&RF) برای پیش‌بینی قیمت بسته‌شدن سهام در n روز قبل از

زمان‌بندی استفاده می‌شود. در نهایت، الگوریتم خوشه‌بندی K-Means و الگوریتم یادگیری گروهی برای بهبود بیشتر دقت پیش‌بینی قیمت سهام ترکیب شده‌اند. داده‌های مورد استفاده در اینجا؛ چهار سهام چینی، یعنی بانک توسعه پودونگ شانگهای (SH:600000)، اوراق بهادار (SH:600030)، CITIC، ZTE (SZ:000063) و LeTV (SZ:300104)، از ۱ ژانویه ۲۰۰۸ تا ۲۰ ژانویه ۲۰۱۹ می‌باشد. برای ارزیابی عملکرد مدل‌های پیش‌بینی از چهار معیار ارزیابی، از جمله میانگین درصد مطلق خطا (MAPE)، میانگین خطای مطلق (MAE)، ریشه نسبی میانگین مربعات خطا (rRMSE) و میانگین مربعات خطا (MSE)، استفاده شده است. نتایج تجربی نشان می‌دهد که مدل پیش‌بینی ترکیبی می‌تواند بهترین دقت پیش‌بینی کلی را برای پیش‌بینی سهام مالی به‌دست آورد. در سال ۲۰۲۰، یک مدل مرجع برای روند بازار سهام، روشی برای رهگیری پیش‌بینی قیمت سهام و همچنین مرجعی جهت پیش‌بینی بازار سهام و سرمایه‌گذاری سرمایه‌گذاران، ارائه شد. مدل پیش‌بینی در این مقاله پیش‌بینی براساس ARI-MA-LS-SVM است [۲۱]. برای شروع یک سهم خاص انتخاب می‌شود. سپس کاهش ویژگی روی داده‌ها اعمال می‌شود. سپس، مجموعه خشن (Rough Set) برای تعیین بردار ورودی نهایی مدل کاهش ویژگی استفاده می‌شود و مدل توسط نمونه‌های آموزشی، آموزش داده می‌شود. در نهایت از نمونه آزمایشی برای آزمایش دقت مدل استفاده می‌شود. ایده اصلی مدل پیش‌بینی ARI-MA-LS-SVM، استفاده از روش رگرسیون میانگین متحرک خودکار- گروهی برای کاهش ویژگی‌های شرطی در نسخه اصلی است. تابع هسته گاوسی شعاعی پایه^۲ انتخاب می‌شود. نتایج تجربی نشان می‌دهد که ARI-MA-LS-SVM دارای عملکرد پیش‌بینی عالی، دقت پیش‌بینی بالا و خطای کوچک است. این نشان می‌دهد که ARIMA-LS-SVM اثر پیش‌بینی بهتری نسبت به مدل LS-SVM و RS-SVM دارد.

در مطالعه‌ای از ۳۳ ترکیب مختلف نشانگرهای فنی برای پیش‌بینی سهام برای معاملات کوتاه‌مدت استفاده شد [۲۲]. از داده‌های بورس ملی هند (NSE)؛ دو سهام بانک ICICI و بانک دولتی هند، از سال ۲۰۰۸ تا ۲۰۱۸ استفاده شده است. در ابتدا با استفاده از تکنیک انتخاب ویژگی بوروتا (Boruta) ویژگی‌های مؤثر انتخاب شدند و سپس برای پیش‌بینی قیمت سهام، مدل پیش‌بینی رگرسیون ANN (شبکه عصبی مصنوعی) استفاده شد. از ANN سه لایه به همراه تابع فعال‌سازی تابع سیگموئید استفاده شد. همچنین مقدار آستانه ۰/۵ در نظر گرفته شد. پارامترهای گرادینانزولی مونتوم^۳ برای تعیین وزن و کاهش حداقل جهانی^۴ بررسی شد. در این مقاله از معیارهای MAE و RMSE برای ارزیابی استفاده شد و براساس نتایج ۱۲ درصد بهتر از روش موجود است.

در سال ۲۰۲۲ روشی برای بهبود نشانگرهای فنی مبتنی بر نوپرزادایی با انواع توابع پایه موجک و همچنین برای انتخاب زیرمجموعه ویژگی بهینه از مجموعه ویژگی‌ها ارائه کردند [۲۳]. در این مقاله از اطلاعات قیمت سهام بدون نویز برای محاسبه نشانگرهای فنی مؤثرتر استفاده شد. داده‌های این مقاله شامل چهار بازار سهام، SSEC بازار سهام چین، HSI بازار سهام هنگ‌کنگ، S&P 500 و DJI از بازار سهام ایالات متحده است. ۱۸ نشانگر فنی در اینجا بررسی شده است. از پنجره‌های زمانی با اندازه متغیر استفاده شده است. اندازه پنجره زمانی در این مطالعه به ترتیب ۳، ۵، ۱۰، ۱۵، ۳۰، ۴۵ و ۶۰ تنظیم شده است و هدف پیش‌بینی جهت حرکت قیمت سهام پس از سه روز بود. نتایج نشان داد که دقت مدل پس از حذف نویز موجک بهبود یافته است.

در سال ۲۰۲۱، در مقاله‌ای چهل و چهار نشانگر فنی را برای داده‌های روزانه سهام هشتاد و هشت سهم محاسبه کردند [۲۴]. آن‌ها پیش‌بینی حرکت سهام را به عنوان یک مسئله طبقه‌بندی در نظر گرفتند. اهمیت نشانگرهای فنی

^۴ Global minimum^۲ Gaussian Radial Basis Kernel Function^۳ Gradient descent momentum

بررسی اجمالی کارهای گذشته براساس مدل یا روش استفاده شده، مجموعه داده و معیارهای ارزیابی در جدول ۱ آمده است.

را با آموزش مستقل مدل رگرسیون لجستیک، ماشین بردار پشتیبان و جنگل تصادفی و به عنوان ویژگی های ورودی مدل مورد استفاده قرار دادند. از دقت طبقه بندی و ضریب همبستگی ماتیوس (MCC) به عنوان معیارهای ارزیابی استفاده شد. نتایج نشان داد که مدل آن ها به خوبی کار می کند.

جدول ۱- مقایسه کارهای گذشته

معیار ارزیابی	مجموعه داده	مدل / روش	مرجع
Accuracy	گوگل، آمازون، اپل، تسلا و میکروسافت	SVM ها با هسته RBF، درخت تصمیم، AdaBoost، فرآیندهای گاوسی، نزدیک ترین همسایگان، تشخیص درجه دوم، جنگل تصادفی، پرسپترون چندلایه	دوغان و همکاران [۹] (۲۰۲۰)
MAPE MAE RMSE MSE	بانک توسعه شانگهای پودونگ (SH: 600000)، اوراق بهادار Citic (SH: 600030)، ZTE (SZ: 000063) و LETV (SZ: 300104)	رگرسیون بردار پشتیبان گروهی و جنگل تصادفی (ESVR & RF)، مدل های آماری، مدل های اقتصادسنجی، مدل های یادگیری ماشین، مدل های یادگیری عمیق (DL) و مدل های ترکیبی	ایکس یو و همکاران [۲۰] (۲۰۲۰)
MAE RMSE	داده های بورس اوراق بهادار ملی هند (NSS)، دو سهام بانک ICICI و بانک دولتی هند	ANN	نایک و همکاران [۲۲] (۲۰۱۹)
F1-Score	SSEC بازار سهام چین، HSI بازار سهام هنگ کنگ، S&P 500 و DJI از بازار سهام ایالات متحده	جنگل تصادفی	جی و همکاران [۲۳] (۲۰۲۲)
MAPE	داده های قیمت های تاریخی سهام، چندین شاخص فنی؛ شمارش و نمرات احساسات مقالات خبری منتشر شده برای یک سهام مشخص؛ روند در جستجوهای Google برای نشانگر سهام داده شده؛ و تعداد بازدیدکنندگان منحصربه فرد برای صفحات مربوط به ویکی پدیا	رگرسیون شبکه عصبی، رگرسیون بردار پشتیبان، رگرسیون درخت تقویت شده و رگرسیون جنگل تصادفی	ونگ و همکاران [۲۵] (۲۰۱۸)
Accuracy	داده های تویتر و قیمت سهام NASDAQ	طبقه بندی کننده ی ساده بیز، درخت تصمیم و ماشین بردار پشتیبان	پاتل و همکاران [۲۶] ۲۰۱۶
RMSE R MAPE	S&P index	MLP, LSTM, CNN, attention-based Neural Networks	گو و همکاران [۲۷] (۲۰۲۰)
%absolute error R2	Not stated	Multiple regression PCA	تانه و همکاران [۲۸] (۲۰۱۸)

(درخت تصمیم)، مبتنی بر نزدیک ترین همسایه (رگرسیون K-نزدیک ترین همسایه)، مبتنی بر شبکه عصبی (رگرسیون پرسپترون چندلایه)، مبتنی بر ماشین بردار پشتیبان (رگرسیون بردار پشتیبان) و مبتنی بر گروه (جنگل تصادفی) [۳۳].

۳- مبانی و روش ها

۳-۱- روش پیشنهادی

(۱) رگرسیون خطی (LR): رگرسیون خطی نوعی تابع پیش بینی کننده ی خطی است که در آن متغیر وابسته به صورت ترکیبی خطی از متغیرهای مستقل پیش بینی می شود.

(۲) رگرسیون درخت تصمیم (DTR): درخت تصمیم یک تابع درختی متشکل از چندین گره تصمیم گیری است. هر گره غیر برگ یک ویژگی است. در درخت تصمیم، ویژگی داده ها در گره های داخلی شاخه ها و نتیجه آن ها در برگ هر شاخه نشان داده می شود.

(۳) رگرسیون K-نزدیک ترین همسایه (KNN): روش K-نزدیک ترین همسایه یکی از معروفترین روش ها برای تخمین است [۳۴]. رگرسیون مبتنی بر همسایگان را در مواردی که متغیرها پیوسته هستند می توان استفاده کرد. رگرسیون K-نزدیک ترین همسایه، یادگیری را براساس (K) نزدیک ترین همسایه های هر نقطه پیاده سازی می کند و از وزن های یکنواخت استفاده می کند [۳۲].

(۴) رگرسیون پرسپترون چندلایه (MLP): MLP شامل یک لایه ورودی، لایه های پنهان و یک لایه خروجی است. این مدل با الگوریتم پس انتشار آموزش می بیند [۳۲]. MLP خطای مجذور (Squared Error) را با استفاده از LBFGS

بلوک دیگرام پیشنهادی در شکل ۱ نشان داده شده است. مطابق با بلوک دیگرام پیشنهادی در شکل ۱، ابتدا: داده های سهام شرکت میکروسافت، از یاهو فایننس جمع آوری می شوند. سپس نشانگرهای فنی محاسبه می شود و مجموعه داده جدیدی شامل نشانگرهای فنی ایجاد می شود. در مرحله بعد پیش پردازش داده های جمع آوری شده، جهت حذف مقادیر از دست رفته و جایگزینی آن با میانگین داده های موجود، انجام می شود. اینجا هر ویژگی در مجموعه داده با مقیاس کننده حداقل-حداکثر (Min-Max) نرمال سازی شده است؛ تا داده ها به یک مقیاس مشترک تبدیل شوند. آماده سازی داده های توییت از مجموعه داده ی توییت نیز انجام می شود و بعد از آن، تحلیل احساسات توییت ها انجام می شود. سپس ویژگی های مورد نظر از توییت ها محاسبه و مجموعه داده جدیدی براساس مجموعه داده سهام و توییت ها ایجاد می گردد. در ادامه، با انواع روش های رگرسیون، پیش بینی قیمت سهام انجام خواهد شد.

۳-۲- روش های رگرسیون

در این مقاله شش گروه از تخمین گر ها از کتابخانه scikit-learn [۳۲] بررسی شده اند که شامل طیف وسیعی از تخمین گر ها هستند، تخمین گر هایی مثل؛ تخمین گر خطی ساده (رگرسیون خطی)، مبتنی بر درخت تصمیم

(۵) حجم (۶) قیمت بسته شدن تعدیل شده. در اینجا، قیمت بسته شدن به عنوان متغیر وابسته و مابقی اطلاعات (ستون حجم برای مقایسه با حجم توییت‌ها استفاده شد) فقط جهت محاسبه نشانگرهای فنی استفاده و سپس از مجموعه داده حذف شده است. در این مقاله برخی از نشانگرهای فنی مؤثر بر اساس مقاله [۲۹] انتخاب شده است. نشانگرهای استفاده شده در اینجا به شرح زیر می‌باشند: WCP (Weighted Closing Price)، UO (Ultimate Oscillator)، Decay، KDJ، WILLR (William's R%)

جدول ۲- فرایارامترهای استفاده شده

مدل	فرایارامترهای تنظیم شده
LR	copy_X= True, fit_intercept= True, n_jobs= None, positive= False
	mean: 0.008834, stdev: 0.00387
DTR	Criterion= 'squared_error', max_depth= 9, min_samples_leaf= 2
	mean: -0.007852, stdev: 0.000825
KNN	leaf_size= 10, metric= 'manhattan', n_neighbors=2, weights='distance'
	mean: -0.002565, stdev: 0.000457
MLP	Activation='logistic', alpha= 1, hidden_layer_sizes= 50, learning_rate= 'constant', learning_rate_init= 0.0001, max_iter= 200, random_state= 1, solver= 'lbfgs'
	mean: 0.008714, stdev: 0.000445
SVR	C= 0.1, gamma= 0.1, kernel= 'rbf'
	mean: -0.00356, stdev: 0.000862
RFR	max_features= 20, n_estimators= 200, Criterion= 'squared_error', min_samples_split= 2
	mean: -0.008155, stdev: 0.000662

یا نزول گرادین تصادفی^۵ بهینه می‌کند. توابع فعال سازی برای لایه پنهان شامل توابع: Exponential, Relu, Sigmoid, Softplus, Softsign, Tanh, Logistic, LeakyReLU, Identity, Elu, Selu است.

(۵) رگرسیون بردار پشتیبان (SVR): ماشین بردار پشتیبان (SVM) یک طبقه بندی کننده خطی تعمیم یافته است (یادگیری بانظارت) که داده ها را به دسته های باینری طبقه بندی می کند [۳۳]. در سال ۱۹۹۶، واپنیک و همکارانش، نسخه ای از SVM را پیشنهاد دادند که به جای طبقه بندی، عمل رگرسیون را انجام می دهد. این مورد به SVR معروف است [۳۵].

(۶) رگرسیون جنگل تصادفی (RFR): جنگل تصادفی یک مدل گروهی با درختان تصمیم گیری پایه ایجاد می کند. به طور تصادفی از هر نمونه نمونه برداری می کند تا درخت های تصمیم گیری مختلف را آموزش دهد.

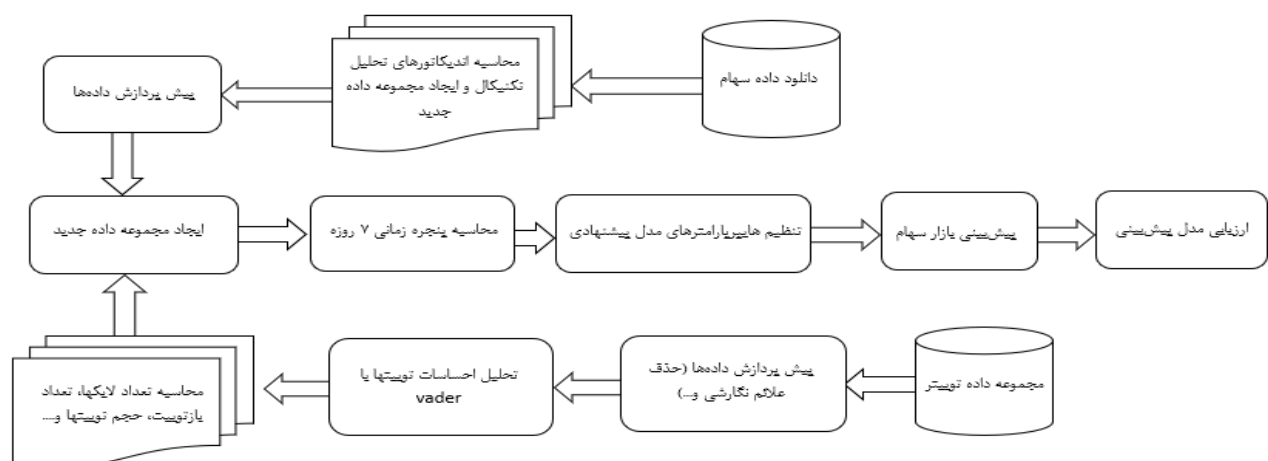
در این پژوهش از جستجوی شبکه ای برای پیدا کردن بهترین فرایارامتر استفاده شده است. ارزیابی هر مجموعه از فرایارامترها با استفاده از اعتبارسنجی - متقابل k-fold انجام می شود که مجموعه داده آموزشی را به k برابر تقسیم می کند. در این مقاله تنظیم فرایارامترها با اعتبارسنجی متقابل k-fold مکرر؛ با k=10 و ۳ بار تکرار انجام شده است. فرایارامترهای استفاده شده در این مقاله در جدول ۲ آمده است.

در اینجا از نسبت تقسیم ۷۰ به ۳۰ برای تقسیم مجموعه داده به مجموعه داده آموزش و آزمایش استفاده می کنیم. همچنین از اعتبارسنجی متقابل k-fold برای جلوگیری از برازش بیش از حد استفاده می شود. اعتبارسنجی متقابل چندین مدل را در زیر مجموعه های متعدد مجموعه داده ایجاد و ارزیابی می کند که جمعیتی از معیارهای عملکرد را فراهم می کند.

پنجره زمانی، روشی برای تبدیل مسئله پیش بینی سری زمانی به یک مشکل یادگیری با نظارت است. در اینجا مانند مقاله [۲۹] پنجره زمانی برای پیش بینی با روش های رگرسیون، با اندازه متغیر برای مجموعه داده در نظر گرفته شده و مجموعه داده جدیدی ایجاد شده است.

۳-۳- مجموعه داده سهام

در این مقاله از داده های سهام شرکت مایکروسافت از تاریخ ۲۰۱۵/۰۱/۰۱ تا ۲۰۱۹/۱۲/۳۱ استفاده شده است. داده های خام از شش ستون تشکیل شده است: (۱) قیمت باز شدن (۲) قیمت بالا (۳) قیمت پایین (۴) قیمت بسته شدن



شکل ۱- بلوک دیاگرام پیشنهادی

⁵ Stochastic Gradient Descent

۳-۴- مجموعه داده‌ی توییت و آماده‌سازی داده‌های توییت

مجموعه داده حجیم استفاده شده برای تحلیل احساسات، توییت‌هایی درباره شرکت‌های برتر از ۲۰۱۵/۰۱/۰۱ تا ۲۰۱۹/۱۲/۳۱، شامل توییت‌هایی درباره شرکت‌های آمازون، اپل، گوگل، مایکروسافت و تسلا است. مجموعه داده توییت در مجموع شامل ۳,۷۱۷,۹۶۴ تعداد توییت منحصر به فرد است. تعداد ۳۷۵,۷۱۱ توییت در مورد مایکروسافت هستند. توییت‌ها همگی به زبان انگلیسی هستند و متن توییت‌ها بدون پیوند و علامت‌های تیک‌دار در طول پردازش اولیه هستند [۹]. محتوای این مجموعه داده شامل توییت‌های منحصر به فرد با ویژگی‌هایی مانند شناسه توییت، نویسنده توییت، تاریخ ارسال، متن توییت، تعداد نظرات، لایک‌ها و باز توییت‌های توییت‌های مطابق با شرکت مرتبط است. مجموعه داده توییت‌های انتخاب شده شامل نماد MSFT می‌باشند.

پیش‌پردازش داده‌های توییت شامل: حذف علائم نگارشی و عناصر غیرضروری مانند URL‌ها و نشانه‌های تبدیل است. همچنین، کلمات توییت‌ها به شیوه‌ای پردازش می‌شوند تا از واژه‌های توقف (Stop Words) و علائمی که اطلاعات ارزشمندی ندارند، خالی شوند. این کارها باعث کاهش حجم داده و افزایش سرعت پردازش می‌شود. علاوه بر این، استخراج اطلاعات ارتباطی از توییت‌ها نیز بسیار حائز اهمیت است. تشخیص و تفکیک کاربران، منشن‌ها (Mentions)، هشتک‌ها و اموجی‌ها (Emojis) به عنوان عواملی که نقش مهمی در توییت‌ها ایفا می‌کنند، باعث می‌شود تا اطلاعات مرتبط با شبکه‌ها و موضوعات مورد نظر استخراج شوند. به طور کلی، نیاز به پیش‌پردازش توییت‌ها به دلیل پیچیدگی ساختار اطلاعات درون آن‌ها است. با انجام مراحل پیش‌پردازش مانند حذف اطلاعات غیرضروری و استخراج ویژگی‌های مهم، داده‌های توییت آماده‌تر و منظم‌تر برای تحلیل‌های بعدی می‌شوند.

۳-۵- تخمین احساس هر توییت

تحلیل احساسات به صورت دستی یک فرآیند کاری حساس و گاهی اوقات مستعد خطا است. بنابراین جای تعجب نیست که بسیاری از محققان NLP به شدت به فرهنگ لغت‌های موجود به عنوان منابع اولیه تکیه می‌کنند. می‌توان احساسات سهام را با استفاده از توییت‌های مربوط به یک شرکت اندازه‌گیری کرد. اینجا از تحلیل‌گر احساسات داخلی و از پیش آموزش‌دیده NLTK به نام VADER^۶ [۳۰] استفاده شده است. VADER از رویکردی مبتنی بر واژگان استفاده می‌کند تا با در نظر گرفتن زمینه و شدت احساس، امتیازات احساسی را به کلمات در یک متن اختصاص دهد، که درک دقیق‌تری از احساسات در پست‌های رسانه‌های اجتماعی را فراهم می‌کند و آن را به ابزاری برای تجزیه و تحلیل احساسات توییت‌های مرتبط با بازار سهام تبدیل می‌کند [۳۱]. VADER به فرهنگ لغتی متکی است که کلمات و دیگر ویژگی‌های واژگانی متعدد مشترک برای بیان احساسات در میکرو بلاگ‌ها را ترسیم می‌کند. این ویژگی‌ها عبارتند از: فهرست کامل شکلک‌های سبک غربی^۷ (به عنوان مثال D: و P:) کلمات اختصاری مربوط به احساسات (مثلاً LOL و ROFL) کلمات عامیانه رایج با ارزش احساسی (مثلاً meh و Nah) و ایجاد یک فرهنگ لغت کامل. یکی از مزیت‌های این روش، توانایی تعیین امتیازات مثبت و منفی برای هر توییت است و درک دقیق‌تری از احساسات کاربران در توییت‌ها به دست می‌آید. برای مثال، اگر توییتی حاوی کلمات مثبتی مانند "عالی" باشد، VADER این احساسات را به عنوان احساسات مثبت تشخیص می‌دهد. به همین ترتیب، اگر کلمات منفی مانند "بد" یا "ناامیدکننده" در توییت‌ها وجود داشته باشد، این احساسات را به عنوان منفی تشخیص می‌دهد. در ادامه می‌توان از نتایج تحلیل احساسات

به دست آمده با استفاده از VADER، در مورد میزان مثبت یا منفی بودن یک احساس خاص حاکم بر توییت‌ها، به عنوان نتیجه‌گیری و ارزیابی کلی، استفاده کرد.

چهار ستون جدید در مجموعه داده، بعد از تجزیه و تحلیل احساسات ایجاد می‌شود. ستون امتیاز ترکیبی^۸، ستون احساسات خنثی^۹، ستون احساسات مثبت و ستون احساسات منفی.

- امتیاز ترکیبی، معیاری است که مجموع نمرات واژگان را محاسبه می‌کند و در بازه‌ی بین ۱- (بسیار منفی) تا ۱ (بسیار مثبت) نرمال شده است.
 - احساسات مثبت: هر توییتی که امتیاز ترکیبی آن بین ۰/۵ و ۱ باشد، به عنوان احساس مثبت شناخته می‌شود.
 - احساسات خنثی: هر توییتی که امتیاز ترکیبی آن بین ۰/۵- و ۰/۵ قرار بگیرد، به عنوان احساس خنثی شناخته می‌شود.
 - احساسات منفی: هر توییتی که امتیاز ترکیبی آن بین ۱- و ۰/۵- باشد، به عنوان احساس منفی شناخته می‌شود.
- به عبارت دیگر، اگر امتیاز ترکیبی توییت، بیشتر از ۰/۵ باشد، آن توییت به عنوان احساس مثبت در نظر گرفته می‌شود؛ اگر بین ۰/۵- و ۰/۵ باشد، به عنوان احساس خنثی شناخته می‌شود و اگر بین ۱- و ۰/۵- باشد، به عنوان احساس منفی شناخته می‌شود.
- از این روش برای تبدیل احساسات موجود در توییت‌ها به صورت نمایش عددی و مقایسه پذیر استفاده می‌شود. این امتیاز ترکیبی به محققان و تحلیل‌گران اجازه می‌دهد که با توجه به امتیازات مثبت، منفی و خنثی، توییت‌ها را در دسته‌بندی‌های مختلف مورد بررسی قرار دهند و از احساسات کاربران برای تحلیل‌های مختلف بهره‌برداری کنند.

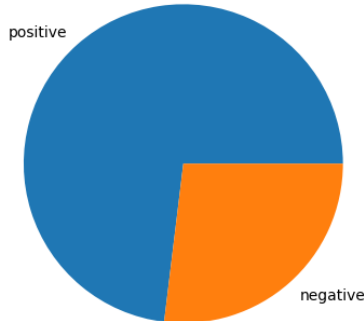
۳-۶- روش پیشنهادی برای تخمین احساس توییت‌ها

در این مطالعه ما برای استفاده از امتیاز ترکیبی از روش پیشنهادی زیر برای طبقه‌بندی استفاده می‌کنیم:

- اگر امتیاز ترکیبی بین ۱- و ۰/۶- باشد؛ احساس توییت، منفی است.
 - اگر امتیاز ترکیبی بین ۰/۶- و ۰/۳۲ باشد؛ احساس توییت، خنثی است.
 - اگر امتیاز ترکیبی بین ۰/۳۲ و ۱+ باشد؛ احساس توییت، مثبت است.
- مجموعه توییت‌ها برای شرکت مایکروسافت شامل ۱۸۲۶ روز منحصر به فرد است. از آنجایی که ممکن است در یک تاریخ به دلیل تعطیلات و غیره داده مربوط به سهام موجود نباشد، ممکن است توییت‌های پست شده بیشتر از داده‌های سهام شرکت مربوطه باشند، بنابراین در این تاریخ‌ها هیچ مقداری برای سهام نداریم. ما از مقادیر بسته شدن سهام در روز/روزهای قبل (به جای میانگین-گیری قیمت بسته شدن) استفاده کردیم و آن را در مقدار خالی (روز/روزهای بعد) قرار دادیم تا همه روزها در مجموعه داده‌ی توییت‌ها حاوی مقدار قیمت بسته شدن باشند. دلیل استفاده از داده‌های روز قبل برای مقادیر بدون قیمت، این است که ممکن است سهام در بازه تاریخی طولانی (۲۰۲۰-۲۰۱۵) خیز بسیار زیادی داشته باشد و میانگین‌گیری باعث افزایش قیمت سهام در یک روز خاص شود؛ در حالی که مقادیر قیمت سهام در روز قبل و بعد از آن تاریخ بسیار کمتر از این میانگین باشند و این باعث کاهش دقت در نتیجه‌گیری خواهد شد. در اینجا از چندین ویژگی خاص براساس تحلیل احساسات انجام شده توسط VADER استفاده کردیم. ابتدا تعداد نظرات مثبت و منفی و خنثی در هر روز

^۸ Compound^۹ Neutral^۶ Valence Aware Dictionary and sEntiment Reasoner^۷ Western-style emoticons

است. همانطور که مشهود است، قیمت سهام شرکت مایکروسافت و تحلیل احساسات توپیت‌ها دارای نوسانات بسیاری است که باید تأثیر این احساسات بر قیمت سهام شرکت مایکروسافت در بازه‌ی کوچکتري بررسی شود. در شکل ۳، درصد توپیت‌های مثبت و منفی مربوط به سهام شرکت مایکروسافت نشان داده شده است، براساس شکل ۳، توپیت‌های مثبت ۳۸/۴۱ درصد و توپیت‌های منفی ۱۴/۴۷ درصد است.



شکل ۳- درصد توپیت‌های مثبت و منفی مربوط به سهام شرکت مایکروسافت

۴-۱- بررسی رابطه بین حجم توپیت و حجم معاملات

در اینجا رابطه بین «حجم توپیت» و «حجم معاملات» را جهت بررسی همبستگی بررسی می‌کنیم. حجم توپیت، تعداد توپیت‌های منتشرشده در یک روز خاص است. شکل ۴، حجم توپیت‌های منتشرشده و همچنین حجم معاملات انجام‌شده در تاریخ مورد بررسی این مقاله است.

باتوجه به شکل ۴، به نظر می‌رسد بین حجم توپیت و حجم معاملات رابطه وجود دارد. به عنوان مثال، در فاصله بین سال‌های ۲۰۱۶ تا ۲۰۱۷ افزایش چشم‌گیری در حجم توپیت‌ها و حجم معاملات مشاهده می‌شود. همچنین نمودار همبستگی را برای اندازه‌گیری رابطه متغیرها ایجاد می‌کنیم، از نمودارهای بالا، انتظار همبستگی مثبت بین حجم توپیت و حجم معاملات را داریم.

در شکل ۵، نمودار همبستگی میان ویژگی‌ها آمده است. با توجه به همبستگی میان ویژگی‌ها در شکل ۵؛ همبستگی میان میانگین امتیاز احساسات مثبت و قیمت بسته‌شدن سهام ۰/۵۵ است که نشان‌دهنده همبستگی و ارتباط بسیار زیاد این دو می‌باشد. برای نشان‌دادن این همبستگی به‌صورت بصری؛ شکل ۱۰، نمودار قیمت بسته‌شدن سهام را در کنار نمودار میانگین توپیت‌های مثبت در بازه زمانی ۲۰۱۹/۰۱/۰۱ تا ۲۰۱۹/۱۲/۳۱ نشان می‌دهد. همچنین نشانگرهای WCP و DECAP همبستگی ۱۰۰ درصدی با قیمت بسته‌شدن سهام دارند. همچنین حجم معاملات با تعداد توپیت‌های مثبت همبستگی بیشتری (۰/۱۶) نسبت به تعداد توپیت‌های منفی (۰/۱۳) دارد. همبستگی میان میانگین امتیاز احساسات مثبت با نشانگرهای WCP و DECAP نیز بسیار زیاد (۰/۵۵) است. همچنین میانگین امتیاز احساسات مثبت با نشانگر KDJ از ۰/۱۲ و با D، ۰/۱۵ همبستگی دارد.

۴-۲- بررسی رابطه بین میانگین تعداد توپیت‌های مثبت و قیمت بسته‌شدن سهام

در شکل ۶، رابطه بین میانگین تعداد توپیت‌های مثبت و قیمت بسته‌شدن سهام نشان داده شده است. با توجه به شکل ۶، به دلیل اینکه میانگین امتیاز نظرات مثبت با قیمت بسته‌شدن سهام در ارتباط هست، می‌توان نتیجه گرفت که امتیاز نظرات مثبت در هرروز به پیش‌بینی قیمت همان روز کمک می‌کند. در این مقاله از روش‌های LR، DTR، KNN، MLP، SVR و RFR استفاده

از مجموعه داده توپیت‌ها شمارش شد و در متغیرهای "تعداد نظرات مثبت" (count_Positive)، "تعداد نظرات خنثی" (count_Neutral) و "تعداد نظرات منفی" (count_Negative) قرار داده‌شد. سپس برای بررسی بیشتر، از "میانگین امتیاز نظرات مثبت"، "میانگین امتیاز نظرات خنثی" و "میانگین امتیاز نظرات منفی" در یک روز نیز استفاده کردیم و به‌ترتیب در متغیرهای score_mean_Positive، score_mean_Neutral و score_mean_Negative قرار دادیم. همچنین از ویژگی "حجم توپیت‌ها" در یک روز به عنوان متغیر tweet_vol و در نهایت از متغیر "نسبت توپیت‌های مثبت به توپیت‌های منفی" با نام count_ratio_NegPos استفاده کردیم.

۳-۷- مشخصات سیستمی

مدل‌های پیشنهادی در پایتون ۳.۹.۱۳ با کتابخانه Scikit-Learn نسخه ۱.۳.۰ آموزش و تجزیه و تحلیل شدند. در Numpy نسخه ۱.۲۲.۴، Pandas نسخه ۱.۴.۴ و nltk نسخه ۳.۷.

تمام آزمایشات با استفاده از یک سیستم مجهز به پردازنده Intel corei7، NVIDIA GEFORCE و رم ۱۲ گیگابایتی انجام شد. زمان اجرای مدل‌های پیشنهادی با تحلیل احساسات و بدون تحلیل احساسات در جدول ۸ آمده است.

۳-۸- معیارهای ارزیابی

در مسائل رگرسیون، پیش‌بینی یک یا چند ویژگی هدف براساس ویژگی‌های ورودی به مدل انجام می‌شود. ویژگی هدف برخلاف مسائل طبقه‌بندی، پیوسته است. معیارهای ارزیابی از سه جهت کاربرد داشته و حائز اهمیت است: (۱) برای آموزش مدل‌ها می‌توانند به عنوان تابع هزینه استفاده شوند، (۲) برای مقایسه مدل‌ها با یکدیگر می‌توانند استفاده شوند، (۳) برای پیش‌بینی خطای مدل در مواجه با داده‌های جدید استفاده شوند. در این مقاله از معیارهای ارزیابی MAE، MAPE، RMSE و MSE مطابق با جدول ۳ استفاده شده است.

جدول ۳- فرمول معیارهای ارزیابی

فرمول	معیار ارزیابی
$MSE = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2$	Mean Square Error (MSE)
$RMSE = \sqrt{MSE(y, \hat{y})}$	Root Mean Squared Error (RMSE)
$MAE = \frac{1}{N} \sum_{i=1}^N y_i - \hat{y}_i $	Mean Absolute Error (MAE)
$MAPE = \frac{100}{N} \sum_{i=1}^N \left \frac{\hat{y}_i - y_i}{y_i} \right $	Mean Absolute Percentage Error (MAPE)

۴- نتایج و بحث

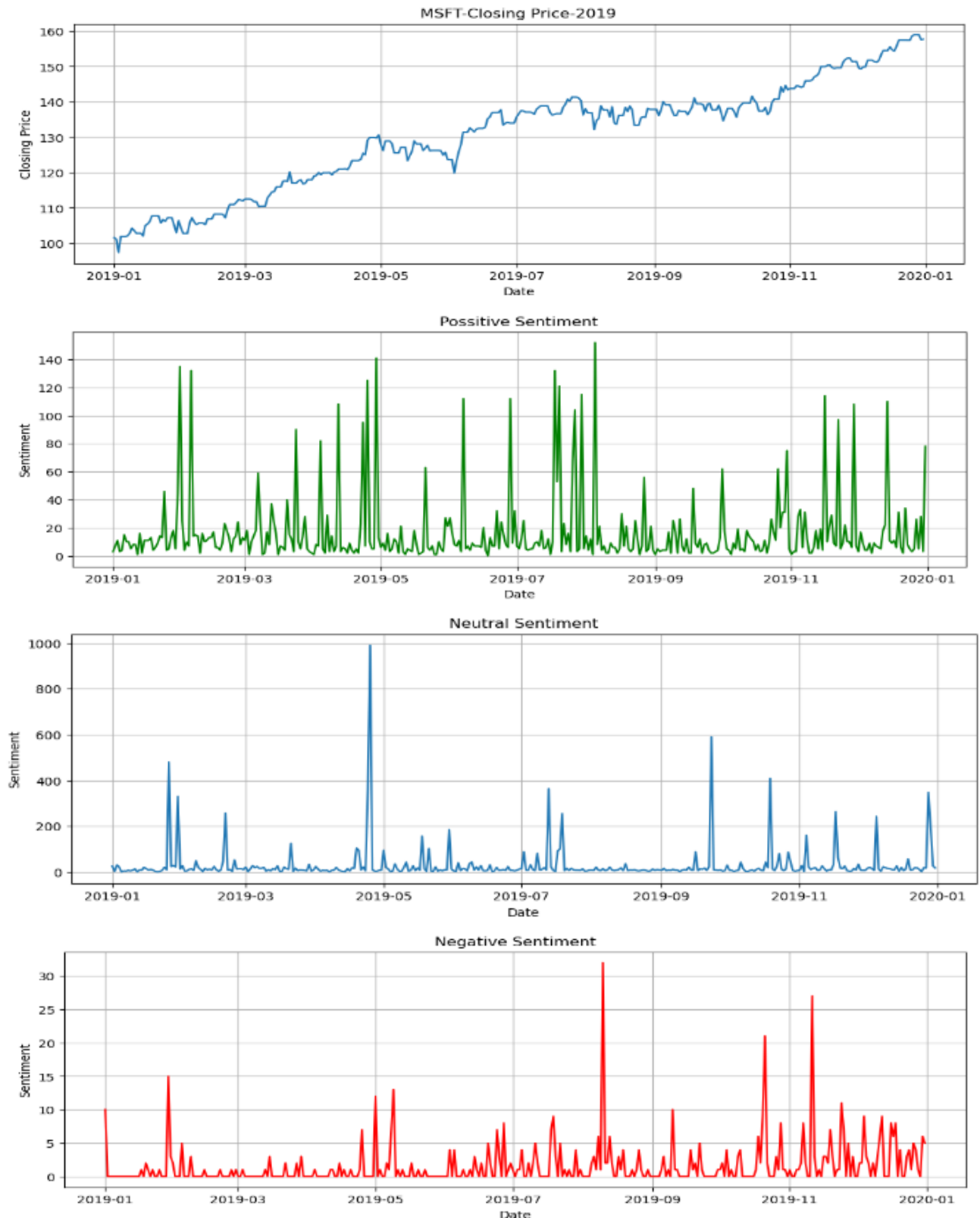
همانطور که در بخش قبل گفته شد، هر توپیت به عنوان مثبت، خنثی یا منفی براساس امتیاز ترکیبی پیشنهادی دسته‌بندی می‌شود، سپس برای تعیین اینکه تغییر روند قیمت از روی احساسات قابل تشخیص است یا خیر؛ نمودار قیمت بسته‌شدن سهام در کنار نمودارهای تعداد نظرات مثبت، تعداد نظرات منفی و تعداد نظرات خنثی ترسیم شده است. در بازه زمانی مورد بررسی این مقاله، با وجود بالا و پایین بودن قیمت سهام، احساسات مثبت توپیت‌ها همچنان ادامه داشت. بنابراین بازه کوچکتري را جهت بررسی بیشتر در نظر می‌گیریم. تغییرات قیمت سهام شرکت مایکروسافت در تاریخ ۲۰۱۹/۰۱/۰۱ تا ۲۰۱۹/۱۲/۳۱ به‌همراه تحلیل احساسات توپیت‌ها، در شکل ۲ نشان داده شده

مختلف اندازه‌گیری شده، آورده شده است.

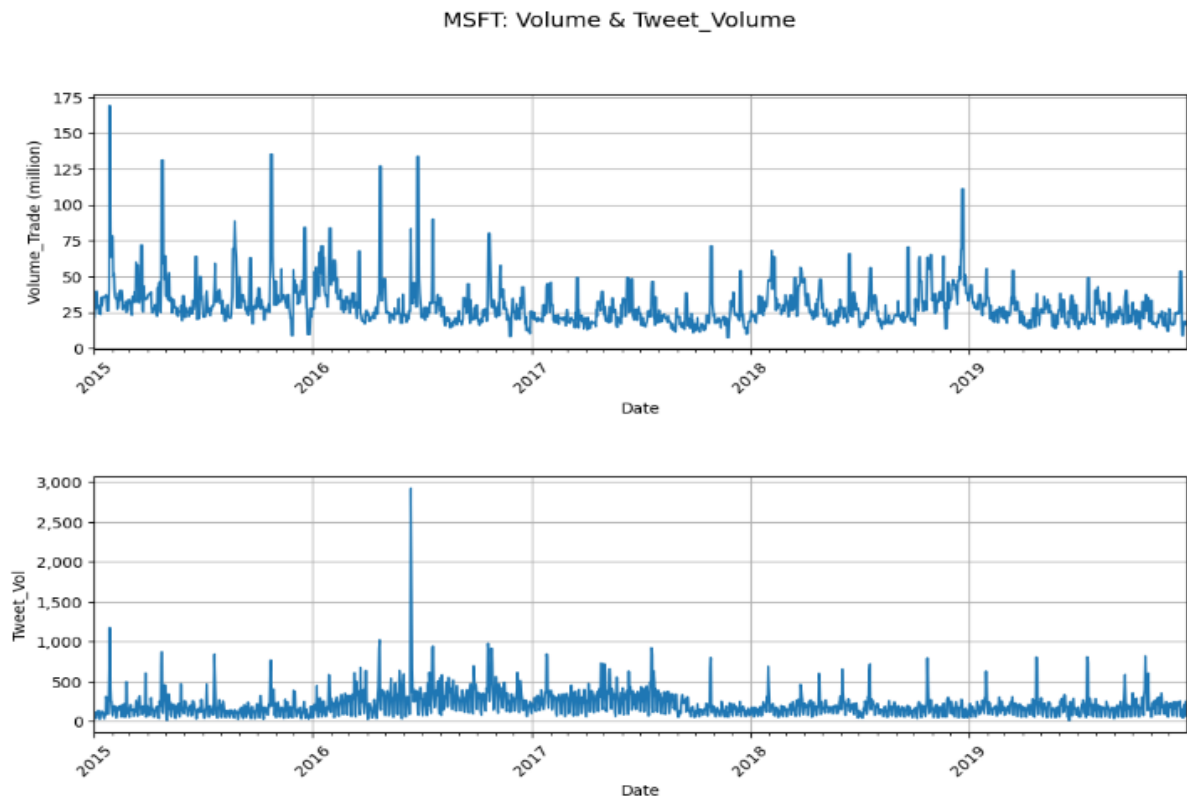
براساس نتایج جداول ۴ و ۵ بهترین روش‌ها جهت پیش‌بینی قیمت سهام با استفاده از تحلیل احساسات و براساس معیار ارزیابی MSE، روش‌های LR، MLP و RF هستند. پیش‌بینی قیمت سهام با روش رگرسیون MLP با استفاده از تحلیل احساسات ۳۷/۱۷ درصد بهبود یافته است. همچنین پیش‌بینی قیمت سهام با روش رگرسیون LR نیز با استفاده از تحلیل احساسات ۳۴/۷۸ درصد

شده است. نتایج روش‌ها بدون استفاده از تحلیل احساسات در جدول ۳ و نتایج با استفاده از تحلیل احساسات در جدول ۴ آورده شده است.

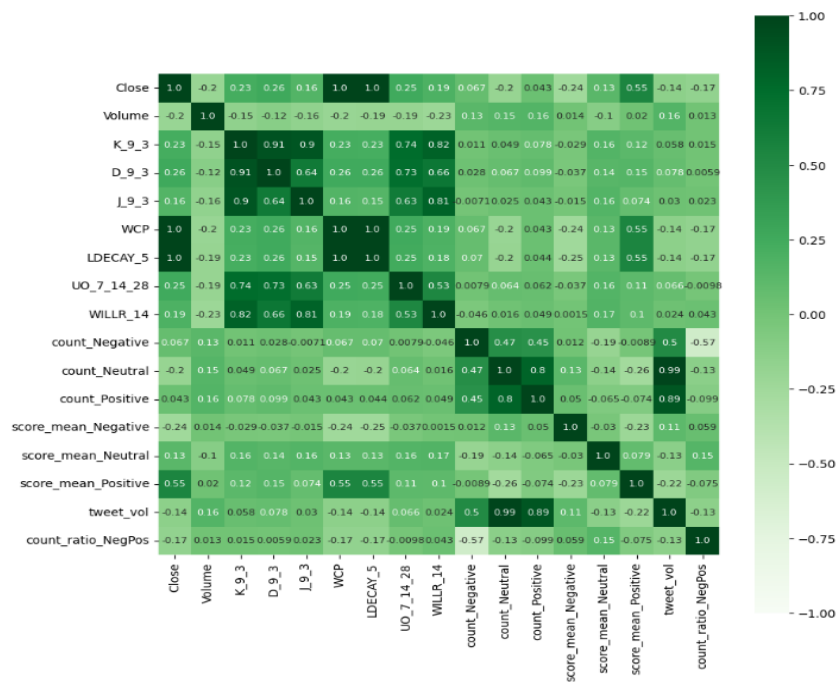
برای بررسی تأثیر تحلیل احساسات با قیمت سهام، پنجره زمانی ۷ روزه در این مقاله در نظر گرفته شد و پیش‌بینی قیمت سهام پس از یک هفته انجام می‌شود. در جداول ۴ و ۵ نتایج پیش‌بینی قیمت سهام توسط روش‌های مختلف براساس مجموعه داده آزمایشی (۳۰ درصد داده‌ها) که توسط معیارهای ارزیابی



شکل ۲- تغییرات قیمت سهام شرکت مایکروسافت به همراه تحلیل احساسات توئیته‌ها



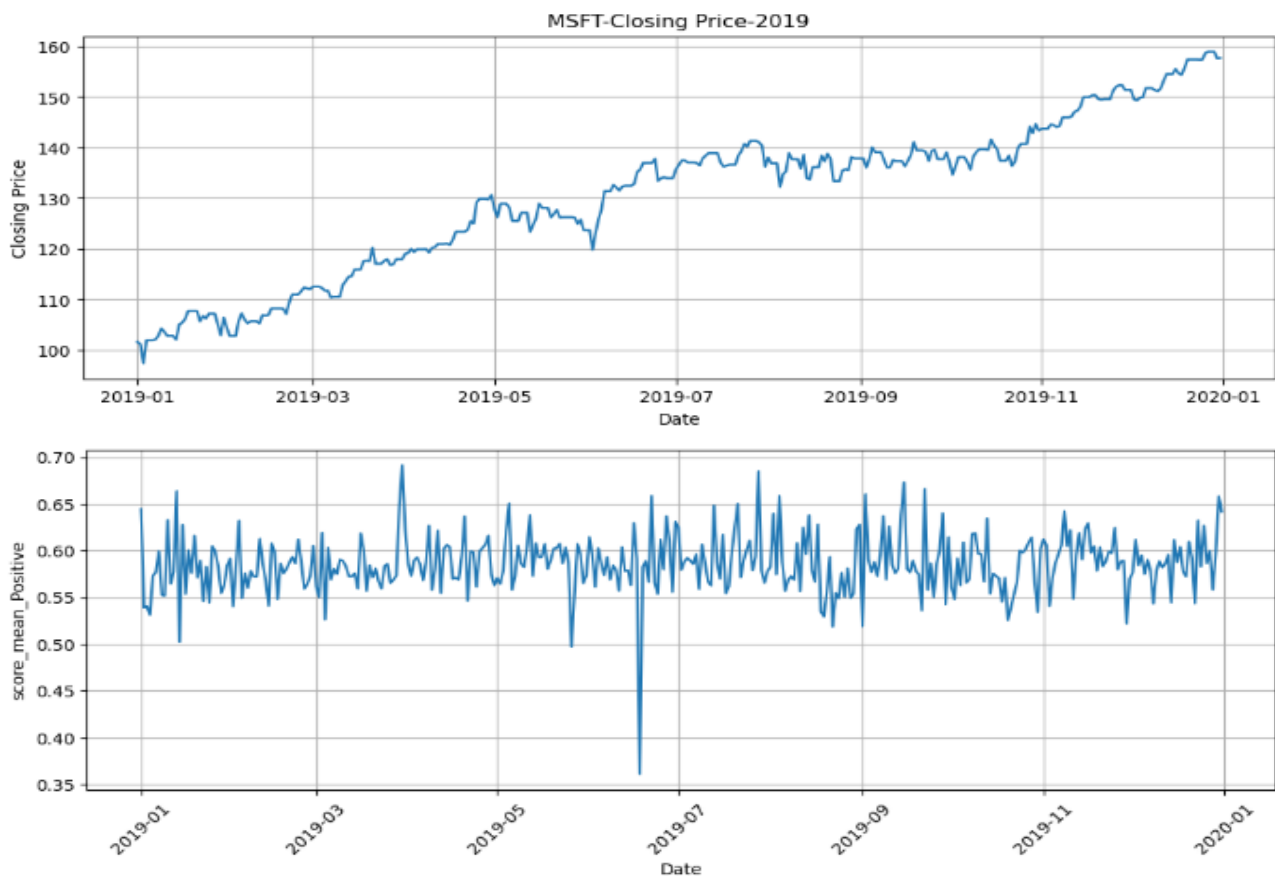
شکل ۴-ارتباط بین حجم توییت‌های منتشر شده و حجم معاملات



شکل ۵- نمودار همبستگی میان ویژگی‌ها

بهبود یافته است. پیش‌بینی قیمت سهام با رگرسیون RF و تحلیل احساسات تا ۱۰، پیش‌بینی قیمت سهام با استفاده از تحلیل احساسات بهبود چشم‌گیری نیز ۱ درصد بهبود داشت.

در شکل‌های ۷ تا ۱۰ نتایج پیش‌بینی توسط بهترین روش‌ها با پنجره زمانی ۷ روزه، یعنی دو روش LR و MLP نشان داده شده است. براساس شکل‌های ۷

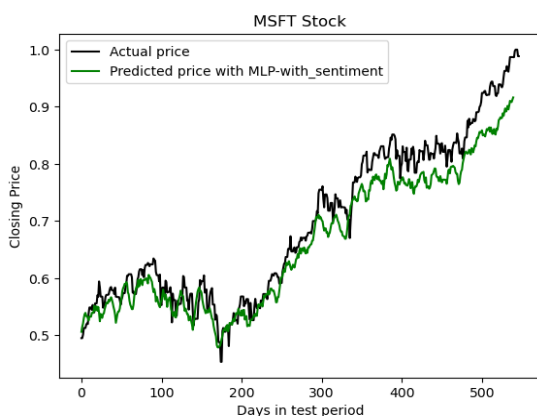


شکل ۶- رابطه بین میانگین تعداد توییت‌های مثبت و قیمت بسته‌شدن سهام

۴-۳- حذف نویز از سری زمانی

عناصر نویز در یک سری زمانی می‌تواند مشکلات قابل توجهی ایجاد کند و حذف نویز قبل از ساخت هر مدل بسیار توصیه می‌شود. به فرآیند به حداقل رساندن نویز، نویززدایی (denoising) گفته می‌شود. یکی از روش‌هایی که معمولاً برای از بین بردن نویز از یک سری زمانی استفاده می‌شود؛ استفاده از میانگین‌های لغزان (Rolling means) است.

میانگین‌های لغزان، میانگین پنجره مشاهدات قبلی است، جایی که پنجره دنباله‌ای از مقادیر داده‌های سری زمانی است. میانگین برای هر اندازه پنجره، محاسبه می‌شود. این می‌تواند به حداقل رساندن نویز در داده‌های سری زمانی کمک کند. در جدول ۶، ویژگی‌های مجموعه داده‌ی ورودی بدون نویز برای پیش‌بینی خروجی بدون نویز آورده شده است.



شکل ۷- پیش‌بینی سهام توسط روش MLP با تحلیل احساسات

جدول ۴- پیش‌بینی بازار سهام توسط روش‌های مختلف بدون

تحلیل احساسات

مدل / معیار ارزیابی	MSE	RMSE	MAE	MAPE
LR	۰/۰۰۰۲۲	۰/۰۱۵۱۳	۰/۰۱۱۲۲	۰/۰۱۷۱۰
DTR	۰/۰۵۲۸۷	۰/۲۲۹۹۳	۰/۱۸۶۷۲	۰/۲۴۰۴۱
KNN	۰/۰۶۲۸۸	۰/۲۵۰۷۶	۰/۲۱۳۲۸	۰/۲۸۱۲۵
MLP	۰/۰۰۳۱۲	۰/۰۵۵۸۵	۰/۰۵۰۶۸	۰/۰۶۹۹۵
SVR	۰/۰۶۹۳۱	۰/۲۶۳۲۷	۰/۲۲۸۴۷	۰/۳۲۴۴
RF	۰/۰۵۴۵۶	۰/۲۳۳۵۹	۰/۱۹۱۰۲	۰/۲۴۶۶۸

جدول ۵- پیش‌بینی بازار سهام توسط روش‌های مختلف با تحلیل

احساسات

مدل / معیار ارزیابی	MSE	RMSE	MAE	MAPE
LR	۰/۰۰۰۱۵	۰/۰۱۲۳۴	۰/۰۰۸۳۹	۰/۰۱۲۷۶
DTR	۰/۰۵۳۴۲	۰/۲۳۱۱۳	۰/۱۸۶۰۳	۰/۲۳۸۵۷
KNN	۰/۰۶۹۶۷	۰/۲۶۳۹۵	۰/۲۲۵۵۴	۰/۲۹۸۱۷
MLP	۰/۰۰۱۹۶	۰/۰۴۴۳۱	۰/۰۳۸۱۵	۰/۰۵۱۹۰
SVR	۰/۰۷۱۱۹	۰/۲۶۶۸۱	۰/۲۴۱۳۴	۰/۳۲۸۴۲
RF	۰/۰۵۴۰۹	۰/۲۳۲۵۶	۰/۱۸۸۶۶	۰/۲۴۲۷۶

جدول ۶- بررسی پنجره‌های زمانی با اندازه متغیر با انواع روش -

های رگرسیون روی مجموعه داده بدون نویز با MSE

مدل	روزه ۳	روزه ۷	روزه ۱۴	روزه ۳۰	روزه ۴۵	روزه ۶۰
LR	۰/۰۰۰۰۳	۰/۰۰۰۰۱	۰/۰۰۰۰۲	۰/۰۰۰۰۰	۰/۰۰۰۰۰	۰/۰۰۰۰۰
DTR	۰/۰۵۲۵۲	۰/۰۵۴۲۲	۰/۰۵۴۴۳	۰/۰۵۵۱۳	۰/۰۵۷۳۲	۰/۰۶۵۱۳
KNN	۰/۰۷۱۳۹	۰/۰۷۰۶۳	۰/۰۶۹۴۵	۰/۰۷۲۵۴	۰/۰۶۹۳۵	۰/۰۶۹۸۷
MLP	۰/۰۰۲۸۶	۰/۰۰۱۴۴	۰/۰۰۲۰۴	۰/۰۰۱۹۶	۰/۰۰۲۳۰	۰/۰۰۳۶۹
SVR	۰/۰۵۲۹۵	۰/۰۷۳۱۰	۰/۱۰۴۷۱	۰/۱۴۶۹۶	۰/۱۶۵۱۱	۰/۱۷۷۷۶
RF	۰/۰۵۳۱۶	۰/۰۵۷۱۲	۰/۰۵۶۰۳	۰/۰۵۷۹۴	۰/۰۵۹۵۸	۰/۰۶۴۰۹

با توجه به جدول ۶، نتایج بررسی روش‌های رگرسیون روی مجموعه داده بدون نویز تغییر محسوسی با نتایج روی مجموعه داده با حضور نویز در جدول ۷ ندارد و نشان‌دهنده استحکام روش‌ها در برابر نویز می‌باشد.

۴-۴- ارزیابی تأثیر پنجره‌های زمانی مختلف در پیش‌بینی قیمت

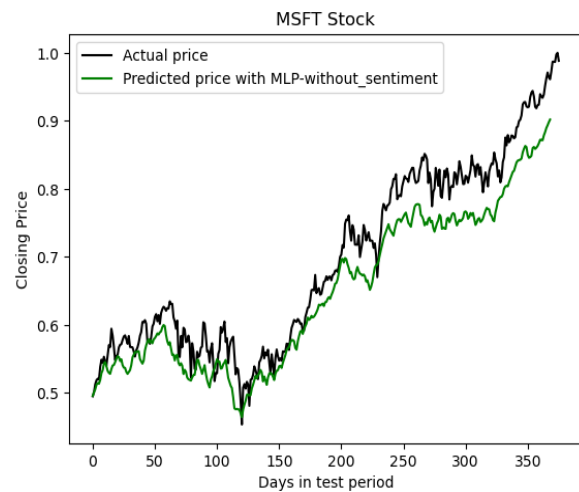
سهام

این مقاله تلاش می‌کند تا مجموعه‌های ویژگی‌هایی با ابعاد زمانی مختلف را با توجه به پنجره‌های زمانی با اندازه متفاوت ایجاد کند، تا بهترین پنجره زمانی جهت پیش‌بینی قیمت سهام شناسایی شود. در اینجا پنجره‌های زمانی مختلفی جهت بررسی نتایج پیش‌بینی اعمال شد. پنجره‌های زمانی بررسی شده در این مقاله ۳، ۷، ۱۴، ۳۰، ۴۵ و ۶۰ روزه است که در جدول ۷ آمده است. براساس جدول ۷، در همه موارد نتایج پیش‌بینی با پنجره زمانی ۷ روزه، بهتر بود. وقتی پنجره زمانی روی ۳، ۷، ۱۴، ۳۰، ۴۵ و ۶۰ تنظیم شد، با افزایش اندازه پنجره زمانی، تعداد ویژگی‌ها به سرعت افزایش یافت. این منجر به بسیاری از ویژگی‌های اضافی در مجموعه ویژگی‌های ایجاد شده با توجه به پنجره زمانی شد که حجم کار محاسباتی را به میزان قابل توجهی افزایش می‌دهد. نتایج نشان می‌دهد که هر چه این فاصله طولانی‌تر باشد، تأثیر نشانگرهای فنی تاریخی کمتر است.

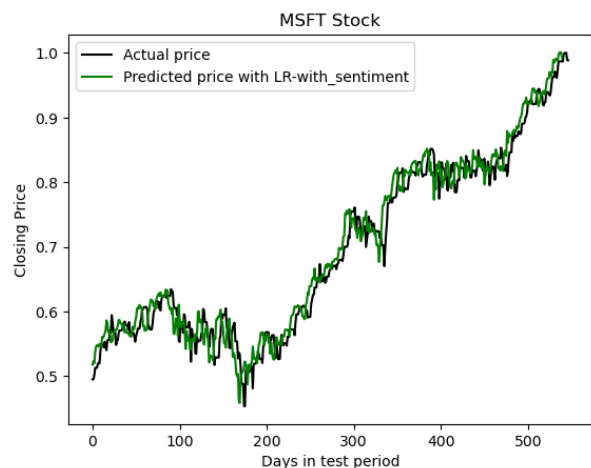
همانطور که از جدول ۷ مشاهده می‌شود، با افزایش پنجره زمانی، عملکرد مدل کاهش می‌یابد و این نشان‌دهنده تأثیر مثبت فاصله زمانی کوتاه برای پیش‌بینی است. پس می‌توان نتیجه گرفت که، با افزایش پنجره زمانی تأثیر نشانگرهای فنی و همچنین تأثیر احساسات توپیت‌ها کم می‌شود و عملکرد مدل کاهش می‌یابد. براساس نتایج جدول ۷ بهترین مدل‌ها جهت پیش‌بینی قیمت سهام، به ترتیب مدل‌های LR و MLP هستند.

در جدول ۸، نتایج روش‌های مورد بررسی با پنجره زمانی ۳ روزه با معیار ارزیابی F-measure آورده شده است. در جدول ۸، بهترین مدل‌ها براساس معیار F-measure، مدل‌های LR و MLP هستند.

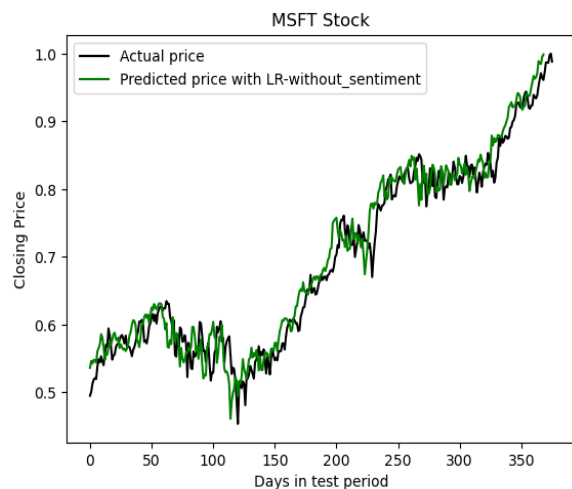
همچنین در جدول ۹، زمان اجرای روش‌های رگرسیون با تحلیل احساسات و بدون تحلیل احساسات برحسب ثانیه درج شده است. براساس جدول ۹، کمترین زمان مربوط به روش RF و سپس روش LR همراه با تحلیل احساسات بود. کمترین زمان برای روش‌های بدون تحلیل احساسات، مربوط به روش DTR و سپس روش LR بود.



شکل ۸- پیش‌بینی سهام توسط روش MLP بدون تحلیل احساسات



شکل ۹- پیش‌بینی سهام توسط روش LR با تحلیل احساسات



شکل ۱۰- پیش‌بینی سهام توسط روش LR بدون تحلیل احساسات

جدول ۷- بررسی پنجره‌های زمانی با اندازه متغیر با انواع روش‌های

رگرسیون

مدل	معیار ارزیابی	پنجره ۳ روزه	پنجره ۷ روزه	پنجره ۱۴ روزه	پنجره ۳۰ روزه	پنجره ۴۵ روزه	پنجره ۶۰ روزه
LR	MSE	۰/۰۰۰۱۵	۰/۰۰۰۱۵	۰/۰۰۰۱۶	۰/۰۰۰۱۹	۰/۰۰۰۲۴	۰/۰۰۰۳۲
	RMSE	۰/۰۱۳۰۹	۰/۰۱۳۳۴	۰/۰۱۳۵۴	۰/۰۱۳۸۸	۰/۰۱۵۴۸	۰/۰۱۷۷۸
	MAE	۰/۰۰۰۸۱۷	۰/۰۰۰۸۳۹	۰/۰۰۰۸۵۸	۰/۰۰۰۹۹۶	۰/۰۰۱۱۶۶	۰/۰۰۱۳۶۰
	MAPE	۰/۰۱۲۴۵	۰/۰۱۲۷۶	۰/۰۱۲۹۷	۰/۰۱۴۸۹	۰/۰۱۷۰۵	۰/۰۲۰۳۷
DTR	MSE	۰/۰۵۲۹۰	۰/۰۵۳۴۲	۰/۰۵۳۷۹	۰/۰۵۵۴۵	۰/۰۵۸۲۹	۰/۰۵۹۷۸
	RMSE	۰/۲۳۰۰۰	۰/۲۳۱۱۳	۰/۲۳۱۹۳	۰/۲۳۵۴۸	۰/۲۴۱۴۳	۰/۲۴۴۵۱
	MAE	۰/۱۸۴۸۶	۰/۱۸۶۰۳	۰/۱۸۷۰۸	۰/۱۹۱۷۱	۰/۱۹۹۶۲	۰/۲۰۳۰۵
	MAPE	۰/۲۳۷۲۸	۰/۲۳۸۵۷	۰/۲۳۹۴۶	۰/۲۴۴۷۷	۰/۲۵۵۰۸	۰/۲۵۸۲۴
KNN	MSE	۰/۰۷۱۳۴	۰/۰۶۹۶۷	۰/۰۶۶۲۹	۰/۰۶۹۱۷	۰/۰۶۸۲۰	۰/۰۶۸۹۳
	RMSE	۰/۲۶۷۱۰	۰/۲۶۳۹۵	۰/۲۶۵۴۶	۰/۲۶۳۰۰	۰/۲۶۱۱۵	۰/۲۶۲۵۵
	MAE	۰/۲۲۹۶۶	۰/۲۲۵۵۴	۰/۲۲۵۹۲	۰/۲۲۳۶۷	۰/۲۲۴۷۲	۰/۲۲۳۵۹
	MAPE	۰/۳۰۵۴۰	۰/۲۹۸۱۷	۰/۲۸۸۸۶	۰/۲۹۲۲۶	۰/۲۹۳۱۲	۰/۲۹۹۰۶
MLP	MSE	۰/۰۰۳۰۱	۰/۰۰۱۹۶	۰/۰۰۲۲۲	۰/۰۰۵۵۲	۰/۰۰۳۵۷	۰/۰۰۶۵۵
	RMSE	۰/۰۵۴۸۶	۰/۰۴۴۳۱	۰/۰۵۶۷۵	۰/۰۷۴۳۰	۰/۰۵۹۷۸	۰/۰۸۰۹۳
	MAE	۰/۰۵۰۱۹	۰/۰۳۸۱۵	۰/۰۴۸۹۶	۰/۰۶۳۷۸	۰/۰۵۰۱۳	۰/۰۶۸۷۰
	MAPE	۰/۰۶۹۵۶	۰/۰۵۱۹۰	۰/۰۶۵۷۹	۰/۰۸۴۱۹	۰/۰۶۵۵۲	۰/۰۸۹۰۳
SVR	MSE	۰/۰۵۱۷۴	۰/۰۷۱۱۹	۰/۱۱۲۸۶	۰/۱۷۸۶۹	۰/۳۱۴۸۹	۰/۳۳۷۷۷
	RMSE	۰/۲۲۷۴۷	۰/۲۶۶۸۱	۰/۳۳۵۹۵	۰/۴۲۲۷۲	۰/۴۶۳۵۷	۰/۵۸۷۶۲
	MAE	۰/۲۱۰۲۴	۰/۲۴۱۳۴	۰/۳۰۱۵۳	۰/۳۸۵۳۴	۰/۴۳۰۷۴	۰/۵۵۹۶۶
	MAPE	۰/۲۹۰۰۵	۰/۳۲۸۴۲	۰/۴۰۷۲۲	۰/۵۲۰۸۵	۰/۵۸۴۸۸	۰/۶۲۵۵۷
RF	MSE	۰/۰۵۳۲۹	۰/۰۵۴۰۹	۰/۰۵۷۰۴	۰/۰۵۹۶۰	۰/۰۶۱۸۴	۰/۰۶۵۰۸
	RMSE	۰/۲۳۰۸۴	۰/۲۳۳۵۶	۰/۲۳۸۸۲	۰/۲۴۴۱۳	۰/۲۴۸۶۸	۰/۲۵۵۱۱
	MAE	۰/۱۸۵۵۵	۰/۱۸۸۶۶	۰/۱۹۶۱۰	۰/۲۰۳۵۳	۰/۲۰۸۲۹	۰/۲۱۶۲۲
	MAPE	۰/۲۳۸۴۷	۰/۲۴۲۷۶	۰/۲۵۲۹۷	۰/۲۶۲۶۷	۰/۲۶۷۷۱	۰/۲۷۷۶۷

محققان توانسته‌اند به‌طور دقیق جهت و میزان حرکت قیمت را پیش‌بینی کنند. با این حال، توجه به این نکته مهم است که پیش‌بینی قیمت سهام کار پیچیده‌ای است که تحت تأثیر عوامل متعددی قرار دارد. درحالی‌که تجزیه و تحلیل احساسات بینش‌هایی را ارائه می‌دهد، باید در ارتباط با سایر تکنیک‌های تحلیل بنیادی و تکنیکی برای تصمیم‌گیری آگاهانه سرمایه‌گذاری استفاده شود. قدرت احساسات توپیت در توانایی آن‌ها برای انعکاس احساسات و نظرات جمعی بازار نهفته است، اما آن‌ها نباید تنها مبنایی برای استراتژی‌های سرمایه‌گذاری باشند. با بهبود مستمر و اصلاح مدل‌های یادگیری ماشین و یادگیری عمیق، محققان می‌توانند دقت پیش‌بینی را افزایش دهند و بینش‌هایی را برای سرمایه‌گذاران و معامله‌گران در این بازار پویا و سریع ارائه دهند. از یک روش اکتشافی برای تخمین احساسات توپیت‌ها با VADER استفاده شد. علاوه بر این، ما اثرات پنجره‌های زمانی اندازه متفاوت را بر عملکرد مدل مقایسه کردیم و از طریق آزمایش‌ها نشان دادیم که اندازه پنجره زمانی مناسب تأثیر مثبتی بر عملکرد مدل دارد. در این مقاله اندازه پنجره زمانی به ترتیب ۳، ۷، ۱۴، ۳۰، ۴۵ و ۶۰ روزه تنظیم شد و بهترین پنجره زمانی برای پیش‌بینی قیمت سهام انتخاب شد. هدف پیش‌بینی قیمت سهام پس از ۷ روز بود. بهترین مدل‌ها جهت پیش‌بینی قیمت سهام مدل‌های رگرسیون خطی، رگرسیون پرسپترون چندلایه و رگرسیون جنگل تصادفی بود. براساس تحلیل احساسات انجام شده؛ نتایج نشان داد که حجم توپیت‌ها با حجم معاملات در ارتباط است، همچنین میانگین امتیاز نظرات مثبت با قیمت سهام ارتباط بالایی داشت. به طور کلی، استفاده از تحلیل احساسات به طور قابل ملاحظه‌ای دقت پیش‌بینی را بهبود می‌بخشد.

کارهای آینده

بررسی این مقاله با دارایی‌های مالی مختلف و همچنین بررسی با نشانگرهای دیگر برای کارهای آینده قابل انجام است. علاوه بر این، همچنین پیش‌بینی قیمت و حرکت سهام با سایر شبکه‌های عصبی و مدل‌های یادگیری عمیق مانند مدل‌های LSTM و شبکه‌های عصبی کانولوشنال که ابزار قدرتمندی در ثبت رابطه بین احساسات توپیت و پیش‌بینی قیمت هستند، در آینده انجام خواهد شد.

مراجع

- [1] E. Ortiz-Ospina and M. Roser, "The rise of social media," *Our World in Data*, 2023. Available: <https://ourworldindata.org/rise-of-social-media>.
- [2] M. M. P. Bharambe and S. C. Dharmadhikari, "Stock market analysis based on artificial neural network with big data," in *Proc. 8th Post Graduate Conf. for Information Technology*, 2017.
- [3] J. J. Murphy, *Technical analysis of the financial markets: A comprehensive guide to trading methods and applications*, Penguin, 1999.
- [4] B. M. Henrique, V. A. Sobreiro, and H. Kimura, "Literature review: Machine learning techniques applied to financial market prediction," *Expert Syst. Appl.*, vol. 124, pp. 226–251, 2019.
- [5] L. Z. Zhen, Y. H. Choo, A. K. Muda, and A. Abraham, "Forecasting FTSE Bursa Malaysia KLCI trend with hybrid particle swarm optimization and support vector machine technique," in *Proc. 2013 World Congr. Nature Biol. Inspired Comput.*, IEEE, pp. 169–174, Aug. 2013.
- [6] O. Kraaijeveld and J. De Smedt, "The predictive power of public Twitter sentiment for forecasting cryptocurrency prices," *J. Int. Financial Markets Inst. Money*, vol. 65, 101188, 2020.
- [7] A. Mohammadi, M. Yazdian Dehkordi, M. A. Nemat Bakhsh, (2019). Implicit feature detection using Persian language syntactic rules and feature clustering. *Journal of Electrical Engineering*, University of Tabriz. , Volume 50, Number 3, pp. 1395–1404, 2019.
- [8] G. Ranco, D. Aleksovski, G. Caldarelli, M. Grčar, and I. Mozetič, "The effects of Twitter sentiment on stock price returns," *PLoS One*, vol. 10, no. 9, e0138441, 2015.

جدول ۸- بررسی انواع روش‌های رگرسیون با معیار ارزیابی F-measure

مدل / معیار ارزیابی	LR	DTR	KNN	MLP	SVR	RF
F-measure	۰/۷۱۶	۰/۵۰۰	۰/۴۸۱	۰/۶۶۳	۰/۵۷۷	۰/۵۴۸

جدول ۹- بررسی زمان اجرای روش‌های رگرسیون با و بدون

تحلیل احساسات (Second)

مدل / معیار ارزیابی	LR	DTR	KNN	MLP	SVR	RF
با تحلیل احساسات	۲/۴۲	۲/۸۹	۲/۷۴	۳/۵۱	۳/۶۰	۲/۴۰
بدون تحلیل احساسات	۱/۶۹	۱/۱۶	۱/۹۵	۲/۵۳	۱/۹۳	۲/۲۹

در جدول ۱۰ مقایسه مدل پیشنهادی با کارهای گذشته آمده است، بهترین مدل در مقایسه با کارهای گذشته، مدل MLP بود.

روش / معیار ارزیابی	MAE	MSE
C-E-SVR&RF [۱۹]	۰/۳۲۲۳	۰/۱۷۴۳
E-SVR&RF [۱۹]	۰/۲۳۶۹	۰/۱۸۴۰
SVR	۰/۲۱۰۲۴	۰/۰۵۱۷۴
RF	۰/۱۸۵۸۵	۰/۰۵۳۲۹
ANN [۲۲]	۱۵/۱۲۲۱	-
MLP	۰/۰۳۸۱۵	۰/۰۰۱۹۶

۵- نتیجه‌گیری

تجزیه و تحلیل احساسات، به‌ویژه در رسانه‌های اجتماعی مانند توپیت، نقش مهمی در پیش‌بینی افزایش قیمت سهام ایفا می‌کند. با تجزیه و تحلیل احساسات توپیت‌ها و استفاده از روش‌های یادگیری ماشین و یادگیری عمیق،

- Proc. 2nd Int. Conf. Emerging Trends Comput. Electr. Eng. (ICETCE)*, Jaipur, India, pp. 261–268, Feb. 2019.
- [23] G. Ji, J. Yu, K. Hu, J. Xie, and X. Ji, "An adaptive feature selection schema using improved technical indicators for predicting stock price movements," *Expert Syst. Appl.*, vol. 200, 116941, 2022.
- [24] A. U. Haq, A. Zeb, Z. Lei, and D. Zhang, "Forecasting daily stock trend using multi-filter feature selection and deep learning," *Expert Syst. Appl.*, vol. 168, 114444, 2021.
- [25] B. Weng, L. Lu, X. Wang, F. M. Megahed, and W. Martinez, "Predicting short-term stock prices using ensemble methods and online data sources," *Expert Syst. Appl.*, vol. 112, pp. 258–273, 2018.
- [26] U. Patel, "Twitter data predicting stock price using data mining techniques," M.S. thesis, Univ. of Bridgeport, 2016.
- [27] P. Gao, R. Zhang, and X. Yang, "The application of stock index price prediction with neural network," *Math. Comput. Appl.*, vol. 25, no. 3, 53, 2020. Available: <http://dx.doi.org/10.3390/mca25030053>
- [28] D. V. Thanh, M. H. N. Minh, and D. D. Hieu, "Building unconditional forecast model of stock market indexes using combined leading indicators and principal components: Application to Vietnamese stock market," *Indian J. Sci. Technol.*, vol. 11, no. 2, pp. 1–13, 2018. <https://doi.org/10.17485/ijst/2018/v11i2/104908>
- [29] F. Moodi, A. Jahangard-Rafsanjani, & S. Zarifzadeh, "Feature selection and regression methods for stock price prediction using technical indicators," *arXiv:2310.09903*. <https://doi.org/10.48550/arXiv.2310.09903>.
- [30] C. Hutto and E. Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text," in *Proc. Int. AAAI Conf. Web Social Media*, vol. 8, no. 1, pp. 216–225, May 2014.
- [31] Y. Sovbetov, "Factors influencing cryptocurrency prices: Evidence from bitcoin, ethereum, dash, bitcoin, and monero," *J. Econ. Financial Anal.*, vol. 2, no. 2, pp. 1–27, 2018.
- [32] <https://scikit-learn.org>
- [33] F. Li, W. Lu, J. W. Keung, X. Yu, L. Gong, and J. Li, "The impact of feature selection techniques on effort-aware defect prediction: An empirical study," *IET Softw.*, vol. 17, no. 2, pp. 168–193, 2023.
- [34] A. M. Sefidian, N. Daneshpour, (2018). Applying Regression Models to High Correlation Subsets to Improve the Placement of Numerical Embedded Values. *Journal of Electrical Engineering*, University of Tabriz. Vol. 48, No. 3, pp. 1187–1200, 2018.
- [35] R. K. Dash, T. N. Nguyen, K. Cengiz, and A. Sharma, "Fine-tuned support vector regression model for stock predictions," *Neural Comput. Appl.*, pp. 1–15, 2021. <https://doi.org/10.1007/s00521-021-05842-w>.
- [9] M. Doğan, Ö. Metin, E. Tek, S. Yumuşak, and K. Öztoprak, "Speculator and influencer evaluation in stock market by using social media," in *Proc. 2020 IEEE Int. Conf. Big Data*, pp. 4559–4566, Dec. 2020.
- [10] J. L. Bollen and H. Mao, *U.S. Patent No. 8,380,607*, Washington, DC: U.S. Patent and Trademark Office, 2013.
- [11] X.-J. Zhang, "Twitter mood predicts the stock market," *J. Comput. Sci.*, vol. 2, no. 1, pp. 1–8, 2011.
- [12] P. C. Tetlock, M. Saar-Tsechansky, and S. Macskassy, "More than words: Quantifying language to measure firms' fundamentals," *J. Finance*, vol. 63, no. 3, pp. 1437–1467, 2008.
- [13] T. C. Lin, "The New Market Manipulation," *Emory Law J.*, vol. 66, 2017.
- [14] L. Phillips, C. Dowling, K. Shaffer, N. Hodas, and S. Volkova, "Using social media to predict the future: A systematic literature review," *arXiv preprint arXiv:1706.06134*, 2017.
- [15] X. Li, H. Xie, L. Chen, J. Wang, and X. Deng, "News impact on stock price return via sentiment analysis," *Knowl.-Based Syst.*, vol. 69, pp. 14–23, 2014.
- [16] C. D. Kirkpatrick II and J. A. Dahlquist, *Technical analysis: The complete resource for financial market technicians*, FT Press, 2010.
- [17] E. K. Sahin and S. Demir, "Greedy-AutoML: A novel greedy-based stacking ensemble learning framework for assessing soil liquefaction potential," *Eng. Appl. Artif. Intell.*, vol. 119, 105732, 2023.
- [18] Y. Xu, C. Yang, S. Peng, and Y. Nojima, "A hybrid two-stage financial stock forecasting algorithm based on clustering and ensemble learning," *Appl. Intell.*, vol. 50, pp. 3852–3867, 2020.
- [19] F. Moodi and H. Saadatfar, "An improved K-means algorithm for big data," *IET Softw.*, vol. 16, pp. 48–59, 2021.
- [20] M. M. Kumbure, C. Lohrmann, P. Luukka, and J. Porras, "Machine learning techniques and data for stock market forecasting: A literature review," *Expert Syst. Appl.*, vol. 197, 116659, 2022.
- [21] C. Xiao, W. Xia, and J. Jiang, "Stock price forecast based on combined model of ARIMA-LS-SVM," *Neural Comput. Appl.*, vol. 32, pp. 5379–5388, 2020.
- [22] N. Naik and B. R. Mohan, "Optimal feature selection of technical indicator and stock prediction using machine learning technique," in