*مقاله پژوهشی*

# پیش‌بینی روند سهام با استفاده از شاخص احساسات و SVM بهبودیافته با تابع هزینه مبتنی بر آنتروپی احساسات

مهین یعقوب‌زاده، عباس ابراهیمی‌مقدم، مرتضی خادمی و هادی صدوقی یزدی

**چکیده: پیش‌بینی بازار سهام همیشه مورد توجه پژوهشگران بوده است. پیشرفت در زمینه هوش مصنوعی و الگوریتم‌های یادگیری ماشین باعث شده که بتوان از داده‌های متنی در کنار داده‌های عددی، جهت پیش‌بینی و عملکرد بهتر بازار بهره برد. در این پژوهش جهت پیش‌بینی روند شاخص بازار سهام نیویورک (NYSE) از داده‌های عددی، داده‌های متنی و یک مدل یادگیری ماشین استفاده شده است. ورودی مدل اولاً داده‌های عددی و ثانیاً نتایج تحلیل احساسات از متن‌های استخراج‌شده از شبکه X است. تحلیل احساسات با یک الگوریتم خاص مبتنی بر یادگیری ماشین (Fin-BERT) انجام شده است. همچنین برای بهبود نتایج پیش‌بینی، در طبقه‌بند پیشنهادی (SVM) دانش پیشینی که در مورد توزیع داده‌ها موجود است در تابع هزینه SVM وارد شده است. این دانش از طریق محاسبه آنتروپی احساسات به دست می‌آید. نتایج آزمایش‌ها نشان می‌دهند که با در نظر گرفتن آنتروپی احساسات در تابع هزینه مدل، نتایج پیش‌بینی بهبود می‌یابد.**

***کلیدواژه:* پیش‌بینی بازار سهام، تحلیل احساسات، Fin-BERT، SVM.**

## ۱- مقدمه

پیش‌بینی بازار سهام مدت‌هاست که توجه پژوهشگران و سرمایه‌گذاران را به خود جلب کرده است. در مطالعات و تحقیقات، نظریه‌هایی وجود دارد که بیان می‌کنند پیش‌بینی و عملکرد بهتر از بازار غیرممکن است [۱]. برخلاف این نظریه‌ها اقتصاددانان ادعا می‌کنند که سرمایه‌گذاران می‌توانند احساسی باشند، بنابراین رفتار آنها را می‌توان با استفاده از نظریه‌های روان‌شناسی توضیح داد. بر این اساس یک نظریه نسبتاً جدید به نام «بازار وفقی[1]» پیشنهاد شده است. این نظریه فرض می‌کند که بازارها را می‌توان با تجزیه و تحلیل رفتار سرمایه‌گذاران پیش‌بینی کرد. رفتار سرمایه‌گذاران را می‌توان از روی داده‌های متنی استخراج کرد. بنابراین استفاده از داده‌های متنی در کنار داده‌های عددی می‌تواند عملکرد یک مدل پیش‌بینی را بهبود بخشد [۲].

تحرکات بازار سهام تحت تأثیر اطلاعات عمومی و خصوصی است که از طریق مقالات خبری، گزارش شرکت‌ها و بحث‌های رسانه‌های اجتماعی به اشتراک گذاشته می‌شود. بنابراین تجزیه و تحلیل این منابع گسترده می‌تواند به سود شرکت‌کنندگان در بازار کمک کند [۳]. برای تحلیل تأثیر اطلاعات عمومی و خصوصی، محققان شروع به استخراج داده‌های متنی از این منابع کرده‌اند. با گذشت زمان تجزیه و تحلیل محتویات متن در تحقیقات بازار سهام به موضوعی جذاب تبدیل شده است. دو نوع استخراج اطلاعات از متن شامل ویژگی‌های سطحی (یا ساختاری متن) و ویژگی‌هایی که از تحلیل احساسات متن‌ها به دست می‌آیند وجود دارند [۴].

گام بعد از استخراج اطلاعات داده‌های متنی، تحلیل احساسات است. طبق فرهنگ لغت آکسفورد[2]، تحلیل احساسات به عنوان یک فرایند شناسایی، محاسباتی و دسته‌بندی نگرش‌های بیان‌شده در یک متن، در درجه اول برای تعیین اینکه آیا نگرش نویسنده نسبت به یک موضوع یا محصول خاص، مثبت، منفی یا خنثی است، تلقی می‌شود [۵]. تجزیه و تحلیل احساسات به شناسایی و استخراج احساسات انسانی از متن بدون ساختار با استفاده از ابزارهای یادگیری ماشین[3] و پردازش زبان طبیعی[4] مربوط می‌شود [۶]. تحلیل احساسات به خصوص از زمانی که مردم می‌توانند در شبکه‌های اجتماعی، وبگاه‌ها و دیگر پایگاه‌ها نظرات، عقاید و احساسات خود را با دیگران به اشتراک بگذارند، در زمینه‌های مختلفی کاربرد پیدا کرده است [۷]. مطالعات روان‌شناسانه نشان داده‌اند که علاوه بر اطلاعات، احساسات نیز بر تصمیمات بشر تأثیر می‌گذارد و ثابت شده است که تصمیمات مالی به طور قابل توجهی توسط احساسات هدایت می‌شوند [۸].

به منظور مطالعه چنین جنبه‌های عاطفی، الگوریتم‌های مختلف تجزیه و تحلیل احساسات با هدف پیش‌بینی تحرکات آتی بازار ارائه شده و روش‌های کشف رابطه بین خلق‌وخوی عمومی و عملکردهای مختلف بازارها پیشنهاد و به کار گرفته شده است. به طور مشخص از سال ۲۰۱۱ از تحلیل احساسات برای پیش‌بینی بازار استفاده می‌شود [۹]. با توجه به مطالعات انجام‌شده، یک چارچوب کلی مانند شکل ۱ برای پیش‌بینی بازار سهام با تحلیل احساسات البته در کنار داده‌های عددی می‌توان در نظر گرفت [۴]، [۱۰] و [۱۱].

شکل ۱ بلوک دیاگرام یک سامانه پیش‌بینی بازار سهام را که شامل دو بخش جداگانه برای پردازش داده‌های عددی و متنی است، نشان می‌دهد.

این مقاله در تاریخ ۲۶ خرداد ماه ۱۴۰۳ دریافت و در تاریخ ۵ مهر ماه ۱۴۰۳ بازنگری شد.

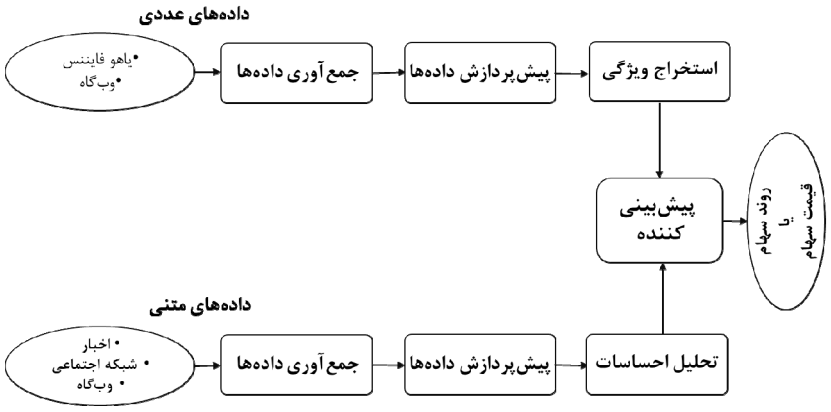
مهین یعقوب‌زاده، گروه برق، دانشکده مهندسی، دانشگاه فردوسی مشهد، مشهد، ایران، (email: mahin.yaqobzadeh@mail.um.ac.ir).

عباس ابراهیمی‌مقدم (نویسنده مسئول)، گروه برق، دانشکده مهندسی، دانشگاه فردوسی مشهد، مشهد، ایران، (email: a.ebrahimi@um.ac.ir).

مرتضی خادمی، گروه برق، دانشکده مهندسی، دانشگاه فردوسی مشهد، مشهد، ایران، (email: khademi@um.ac.ir).

هادی صدوقی یزدی، گروه کامپیوتر، دانشکده مهندسی، دانشگاه فردوسی مشهد، مشهد، ایران، (email: h-sadoghi@um.ac.ir).

1. Adaptive Market Hypothesis
2. Press Dictionary
3. Machine Learning
4. Natural Language Processing




شکل ۱: بلوک دیاگرام پیش‌بینی بازار سهام با داده‌های عددی و تحلیل احساسات داده‌های متنی.

در بخش پردازش داده‌های عددی، ابتدا داده‌های عددی که شامل قیمت و حجم معاملات و ... می‌باشند از منابع مختلف جمع‌آوری می‌شوند.

معمولاً روی داده‌های عددی عملیاتی همچون نرمال‌سازی، شناسایی روزهای تعطیل بازار و ... به عنوان پیش‌پردازش انجام می‌شود. سپس مجموعه‌ای از ویژگی‌ها از داده‌های پاکسازی‌شده استخراج می‌شوند که دقت پیش‌بینی به میزان زیادی به این ویژگی‌ها وابسته است. برای داده‌های عددی می‌توان شاخص‌های تکنیکال را محاسبه کرد یا قیمت‌های مربوط به سهام را مستقیم مورد استفاده قرار داد [۱۲] و [۱۳]. در این تحقیق نیز قیمت‌ها به طور مستقیم مورد استفاده قرار می‌گیرد.

در بخش داده‌های متنی، جمع‌آوری از منابع مختلف صورت می‌گیرد. داده‌های متنی در حوزه مالی را می‌توان به سه دسته اخبار، شبکه‌های اجتماعی و گزارش شرکت‌ها تقسیم کرد [۴]. شبکه‌های اجتماعی عموماً نظرات آزادانه کاربران را بیان می‌کنند و برای تحلیل احساسات بیشتر مورد استفاده قرار می‌گیرند [۱۴]. در مرحله پیش‌پردازش داده‌های متنی معمولاً روی متن‌ها عملیاتی همچون بخش‌بندی[1]، ریشه‌یابی[2]، حذف کلمات اضافی[3]، همسان‌سازی کلمات و حذف حروف بی‌ربط انجام می‌شود [۱۵]. برای استخراج ویژگی‌های احساسی از متن، تحلیل احساسات انجام می‌شود. دو رویکرد اصلی برای تحلیل احساسات، روش‌های مبتنی بر واژه‌نامه[4] و روش‌های مبتنی بر یادگیری ماشین[5] هستند که در این تحقیق از روش دوم برای تحلیل احساسات بهره گرفته شده است. رویکردهای مبتنی بر واژه‌نامه بر اساس استفاده از واژه‌نامه احساسات (یعنی لیستی از کلمات که هر کدام به درجه‌ای از احساسات نگاشته شده‌اند) است؛ در حالی که روش‌های یادگیری ماشین بر اساس طبقه‌بندهایی هستند که از نمونه متون یا جملات برچسب‌گذاری‌شده آموزش دیده‌اند [۱۰].

در نهایت از یک مدل پیش‌بینی‌کننده جهت پیش‌بینی بازار استفاده می‌شود. بسیاری از الگوریتم‌های یادگیری ماشین برای پیش‌بینی بازارهای سهام در مطالعات تحقیقاتی به کار گرفته شده‌اند. اساساً دو دسته مدل اصلی برای پیش‌بینی وجود دارد: مدل‌های طبقه‌بندی که سعی می‌کنند به سرمایه‌گذاران در فرایند تصمیم‌گیری خرید، فروش یا نگهداری سهام کمک کنند و مدل‌های رگرسیونی که سعی در پیش‌بینی قیمت سهام دارند؛ مانند قیمت بسته‌شدن. شبکه‌های عصبی حافظه‌دار[6]، شبکه‌های عصبی بازگشتی[7] و شبکه‌های عصبی پیچشی[8] از ابزارهای پرکاربرد در زمینه پیش‌بینی قیمت هستند [۶] و [۱۵]. مطالعات انجام‌شده نشان می‌دهند که SVM محبوب‌ترین تکنیک برای پیش‌بینی روند سهام است [۲] و [۱۶]؛ بنابراین در این تحقیق که پیش‌بینی روند سهام مورد نظر می‌باشد از تکنیک فوق به عنوان مدل پیش‌بینی‌کننده استفاده می‌شود.

در بخش دوم گزارشی از مطالعات انجام‌شده در زمینه پیش‌بینی بازار سهام مبتنی بر تحلیل احساسات آورده شده است. بخش سوم به معرفی روش پیشنهادی، بخش چهارم به ارزیابی، بخش پنجم به شبیه‌سازی، بخش ششم به نتیجه و بحث در مورد آزمایش‌های انجام‌شده و بخش هفتم به جمع‌بندی و نتیجه‌گیری از آزمایش‌های انجام‌شده می‌پردازد.

## ۲- پیشینه پژوهش

تحقیقات زیادی بر روی پیش‌بینی بازار سهام با استفاده از تحلیل احساسات انجام شده که در این قسمت به برخی از آنها اشاره می‌شود. نوسی و همکاران برای پیش‌بینی روند سهام مایکروسافت از داده‌های مربوط به قیمت سهام و تحلیل احساسات پیام‌های مربوط به این سهام که از شبکه توییتر و stockTweet جمع‌آوری شده، بهره برده‌اند. برای انجام تحلیل احساسات از روش مبتنی بر واژه‌نامه (با استفاده از کتابخانه‌های Vader و Textblob موجود در پایتون) استفاده کرده‌اند. میانگین احساسات روزانه به‌همراه قیمت‌ها به‌عنوان ورودی پیش‌بینی‌کننده در نظر گرفته شده‌اند. در این تحقیق SVM و Logestic regression برای پیش‌بینی روند سهام مورد استفاده قرار گرفته‌اند. نتایج نشان داده هنگامی که از کتابخانه Textblob برای تحلیل احساسات و از SVM برای پیش‌بینی استفاده شده، بهترین نتایج را گرفته‌اند [۱۷].

نهان کاچ دانگ و همکاران نیز برای پیش‌بینی قیمت سهام از تحلیل احساسات در کنار قیمت‌های مربوط به سهام بهره گرفته‌اند. مجموعه داده‌های متنی و عددی شامل توییت‌ها و قیمت‌های مربوط به ۱۶ سهام شناخته‌شده در یک بازه سه‌ساله است. تحلیل احساسات با دو مدل مبتنی بر واژه‌نامه (کتابخانه‌های Vader و Textblob) و یک مدل مبتنی بر یادگیری ماشین (Fin-BERT[9]) انجام شده است. پس از تحلیل احساسات نمرات روزانه احساسات استخراج شده و میانگین نمرات روزانه محاسبه می‌شوند. برای ارزیابی پیش‌بینی، ۳۰ الگوریتم مختلف (مانند

1. Tokenization
2. Lemmatization
3. Stop Word Removal
4. Lexicon Based
5. Machin Learning
6. LSTM
7. RNN
8. CNN
9. Financial Bidirectional Encoder Representations from Transformers Transformers

GRU، NN، CNN، RLL و ...) بررسی و مقایسه شده‌اند. آزمایش‌ها نشان می‌دهند که استفاده از تحلیل احساسات، عملکرد مدل‌های پیش‌بینی‌کننده را بهبود می‌دهد و نتیجه پیش‌بینی به روش انجام تحلیل احساسات و همچنین مدل پیش‌بینی‌کننده وابسته است [۱۸].

وانگ و همکاران، جهت پیش‌بینی روند سهام‌های مورد نظر (۵ سهام) از عناوین خبری و قیمت‌های مربوط به آنها در روش خود استفاده کرده‌اند. بر روی قیمت‌ها تحلیل تکنیکال انجام شده و چند شاخص تکنیکال از این داده‌ها استخراج شده است. بر روی عناوین خبری نیز پس از پیش‌پردازش متن‌ها، تحلیل احساسات با روش مبتنی بر واژه‌نامه (واژه‌نامه SenticNet) انجام شده است. نتایج تحلیل احساسات و تحلیل تکنیکال با یکدیگر ادغام شده و وارد پیش‌بینی‌کننده شده‌اند. در این تحقیق از شش مدل یادگیری ماشین ( Multi-Layer Perception, Support Vector Machine, Naïve Bayes, Random Forest, Logistic Regression و Extreme Gradient Boosting) جهت پیش‌بینی روند سهام استفاده و عملکرد آنها مقایسه شده است. این تحقیق نه‌تنها نقاط قوت و ضعف مدل‌های مختلف یادگیری ماشین را نشان می‌دهد، بلکه به این نتیجه می‌رسد که استفاده از تحلیل احساسات بهبود قابل توجهی در عملکرد پیش‌بینی روند سهام به همراه دارد. همچنین استفاده همزمان از شاخص‌های تکنیکال و تحلیل احساسات به افزایش دقت پیش‌بینی کمک کرده و از طرفی قابلیت هر مدل یادگیری ماشین در میان سهام مختلف متفاوت بوده است [۱۹].

مدل یادگیری ماشینی جدیدی (Gradient Boosting Classifier) را زمان و همکاران پیشنهاد داده‌اند که با ترکیب قیمت‌های مربوط به سهام‌ها با عوامل خارجی مانند احساسات شبکه‌های اجتماعی و اخبار مالی و روندهای نفت و طلا، دقت پیش‌بینی روند سهام را بهبود می‌بخشد. اثربخشی مدل پیشنهادی با ۷ سهام مختلف ارزیابی شده است. الگوریتم پیشنهادی نتایج ثابتی را تولید کرده و ترکیب داده‌های شبکه‌های اجتماعی با قیمت نفت دقت پیش‌بینی را افزایش داده است. این تحقیق همچنین تأثیر انتخاب ویژگی‌ها، گسترش ویژگی‌ها، کاهش توییت‌های هرز[1] و روندهای قیمت نفت را بر عملکرد الگوریتم‌ها بررسی کرده و نشان نشان داده که ادغام داده‌های مختلف می‌تواند درک جامع‌تری از روند سهام‌ها را فراهم کند [۲۰].

داراپنی و همکاران نیز از ادغام داده‌ها نظیر قیمت‌های روزانه سهام، ارزش طلا و نفت و ارزش دلار آمریکا به عنوان داده‌های عددی و اخبار مالی مربوط به سهام‌های مورد بررسی به عنوان داده‌های متنی، جهت پیش‌بینی قیمت سهام بهره برده‌اند. تحلیل احساسات اخبار با یک روش مبتنی بر واژه‌نامه انجام شده است. از نتایج تحلیل احساسات چند ویژگی استخراج شده و به همراه داده‌های عددی برای پیش‌بینی قیمت سهام‌ها وارد مدل Randoon Forest شده است. نتایج با پیش‌بینی‌کننده LSTM و با داده‌های مربوط به چهار سهام مختلف مقایسه شده است. نتایج نشان داده‌اند که استفاده از منابع مختلف، عملکرد مدل را می‌تواند بهبود بخشد. این تحقیق همچنین به تحلیل عملکرد دو مدل LSTM و Random Forest پرداخته و نشان داده که مدل LSTM به طور کلی دقت بیشتری دارد و نتایج بهتری را نسبت به مدل Random Forest ارائه می‌دهد. به‌ویژه، تحلیل احساسات در مدل Random Forest نتایج متفاوتی با LSTM به همراه داشته است. با این حال، تحلیل احساسات در مدل Random Forest در مورد برخی از سهام‌ها نتایج بهتری نسبت به مدل LSTM داشته است [۲۱].

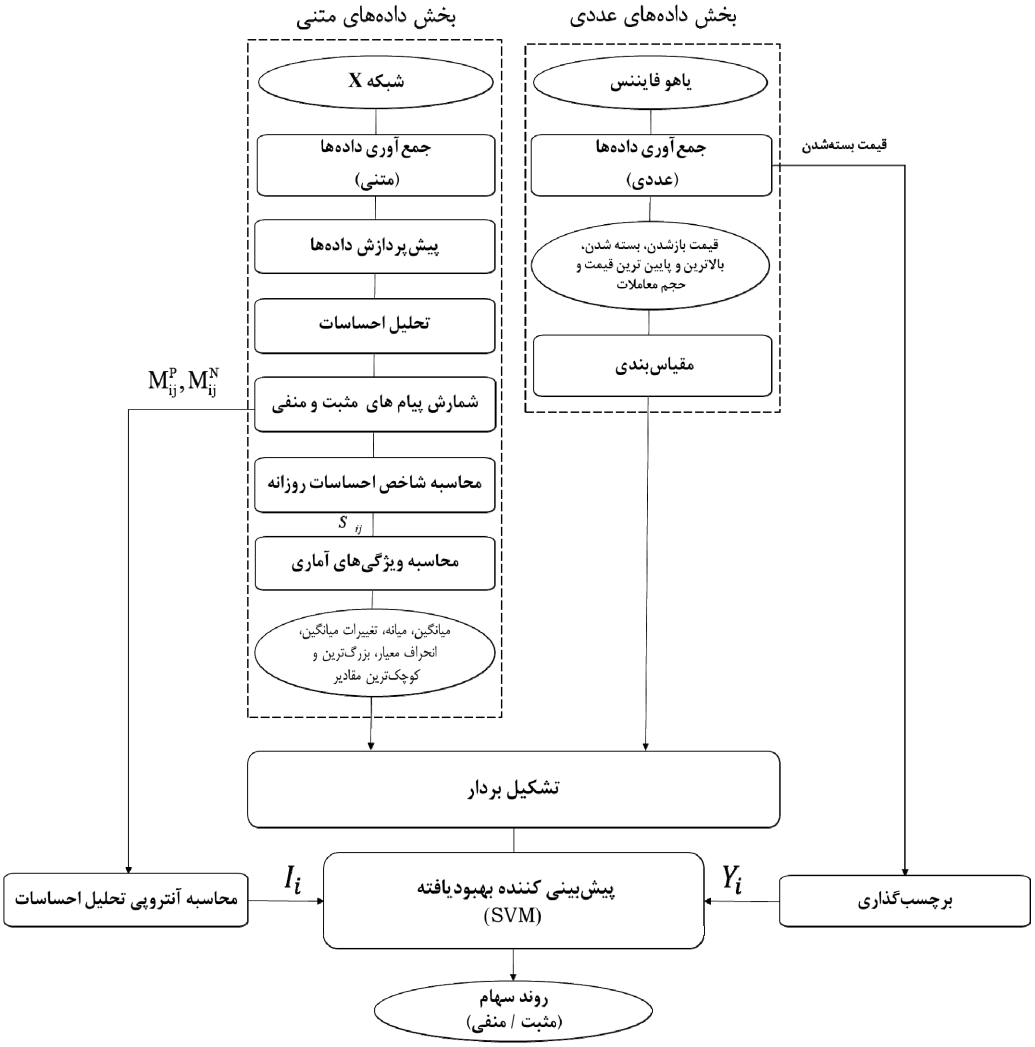
جی‌هوان‌کیم و همکاران برای پیش‌بینی شاخص سهام S&P ۵۰۰ از قیمت‌های بازار و خلاصه اخبار مربوطه استفاده کرده‌اند. تحلیل احساسات اخبار با مدل Fin-BERT انجام شده است. یک شبکه LSTM با این داده‌ها آموزش داده و سپس برای پیش‌بینی از آن بهره برده‌اند. آزمایش‌ها در دو حالت، بدون در نظر گرفتن تحلیل احساسات و با در نظر گرفتن آن انجام و مقایسه گردیده و نشان داده شده که استفاده از نتایج تحلیل احساسات در مدل، LSTM بهبود قابل توجهی در دقت پیش‌بینی ایجاد می‌کند. این تحقیق همچنین به بررسی استفاده از خلاصه‌های خبری به جای متون کامل پرداخت و نتایج به‌دست‌آمده از این تحقیق نشان داد که استفاده از داده‌های خلاصه‌شده و تحلیل احساسات آنها می‌تواند به کاهش زمان تحلیل و بهبود دقت پیش‌بینی کمک کند؛ بدون اینکه اطلاعات کلیدی از دست برود [۲۲].

شیائو و همکاران برای پیش‌بینی روند سهام از تحلیل احساسات توییت‌ها و اخبار منتشرشده بهره برده‌اند. در این کار از دو مدل مبتنی بر واژه‌نامه (واژه‌نامه Loughran-McDonalddictionary و کتابخانه Vader) و یک مدل مبتنی بر یادگیری ماشین (Fin-BERT) برای تحلیل احساسات داده‌ها استفاده شده است. از آنجا که ایده این بوده که زمان انتشار داده (توییت و خبر) روی پیش‌بینی اثرگذار است، برای محاسبه تحلیل احساسات در یک روز، دو بازه زمانی (بازه زمانی طبیعی ۰۰:۰۰~۰۰:۰۰ و بازه زمانی کاری ۰۹:۳۰~۰۹:۳۰) در نظر گرفته شده و بر اساس این بازه‌های زمانی، تحلیل‌ها انجام شده است. پژوهش‌های پیشین معمولاً از شاخص‌های روزانه جمع‌آوری‌شده برای تحلیل احساسات استفاده کرده‌اند، اما نتایج این تحقیق نشان می‌دهد که استفاده از تقسیمات زمانی، مانند ساعات بازشدن بازار می‌تواند به پیش‌بینی دقیق‌تر تغییرات قیمت سهام کمک کند. این روش جدید در مقایسه با نتایج تحقیقات قبلی، دقت بالاتری در پیش‌بینی تغییرات قیمت‌ها بر اساس احساسات کاربران شبکه‌های اجتماعی و اخبار مالی به دست آورده است [۲۳].

یوه‌کیو و همکاران به توسعه یک شاخص جدید احساسات برای پیش‌بینی روندهای سهام با استفاده از نظرات سرمایه‌گذاران و ناهنجاری‌های مالی مانند اثر روزهای هفته و تعطیلات پرداخته‌اند. داده‌های تاریخی دوساله شاخص SSE ۵۰ بورس شانگهای و نقدهای کاربران از Eastmoney.com جمع‌آوری و با استفاده از مدل تحلیل احساسات SKEP که بر پایه Baidu AI Cloud است، پردازش شده‌اند. هدف این شاخص جدید، بهبود دقت پیش‌بینی با وزن‌دهی به محتواهای متنی است. عملکرد این شاخص تعدیل‌شده با استفاده از مدل‌های یادگیری ماشین از جمله SVM، DT و KNN مورد ارزیابی قرار گرفت و نتایج نشان داد که شاخص تعدیل‌شده باعث بهبود دقت مدل‌ها می‌شود؛ به‌ویژه دقت مدل KNN به میزان ٪۱۲٫۲۵ و SVM به ٪۶۸٫۳۷ افزایش یافت. همچنین ترکیب شاخص احساسات با شاخص‌های بازار منجر به کاهش ریسک و افزایش سود برای سرمایه‌گذاران شد [۲۴].

لیو و همکاران نیز برای بهبود نتایج پیش‌بینی روند شاخص سهام S&P ۵۰۰ از شاخص تحلیل احساسات به همراه ویژگی‌های مربوط به قیمت استفاده کرده‌اند. برای بهبود نتایج تحلیل احساسات از مدل Fin-BERT بهره برده‌اند، سپس چهار ویژگی را از این نتایج استخراج کرده‌اند و برای پیش‌بینی در کنار ویژگی‌های مربوط به قیمت سهام به کار برده‌اند. همچنین برای بهبود نتایج از یک مدل Ensamble SVM استفاده شده است. نتایج نشان می‌دهند که استفاده از Fin-BERT دقت

1. Tweet Spam




شکل ۲: بلوک دیاگرام روش پیشنهادی.

پیش‌بینی را ۴ تا ۵ درصد افزایش می‌دهد. مدل Ensamble SVM نیز عملکرد بهتری نسبت به SVM اصلی داشته است [۲۵].

## ۳- روش پیشنهادی

طبق بررسی تحقیقات پیشین، داده‌های عددی و متنی، روش‌های تحلیل احساسات، نحوه در نظر گرفتن احساسات در پیش‌بینی و مدل‌های مختلف برای پیش‌بینی بازار سهام به‌طور گسترده مورد مطالعه قرار گرفته‌اند. هدف این تحقیق، بهبود دقت پیش‌بینی با ترکیب اطلاعات احساسات کاربران شبکه‌های اجتماعی و داده‌های مالی است. برخلاف رویکردهای معمول که احساسات را به‌عنوان یک شاخص ساده وارد مدل می‌کنند، این پژوهش از مفهومی به نام آنتروپی احساسات بهره می‌برد. این نوآوری به مدل اجازه می‌دهد تا در هنگام طبقه‌بندی، پیچیدگی و عدم قطعیت موجود در احساسات کاربران را نیز در نظر بگیرد. در واقع از این طریق به مدل اطلاعاتی در مورد پراکندگی احساسات پیام‌ها به مدل داده می‌شود. در این روش، تحلیل احساسات با مدل Fin-BERT انجام شده و برای پیش‌بینی روند شاخص سهام، یک مدل SVM با تابع هزینه‌ای جدید که بر اساس آنتروپی احساسات تنظیم شده است، پیشنهاد می‌شود. بلوک دیاگرام روش پیشنهادی در شکل ۲ نشان داده شده که شامل دو بخش اصلی برای پردازش داده‌های عددی و متنی است.

### ۳-۱ بخش داده‌های متنی

داده‌های متنی می‌توانند از منابع مختلف و با روش‌های گوناگون جمع‌آوری می‌شوند که در این کار، داده‌های متنی مد نظر با روش Web Crawel که الگوریتم آن را در محیط پایتون طراحی کرده‌ایم از شبکه X جمع‌آوری می‌شوند. پیام‌ها شامل حروف اضافه، شکلک، کلمات بدون ارتباط با موضوع و موارد دیگر هستند. برای تحلیل احساسات این داده‌ها نیاز است آنها را پاکسازی کرد تا آماده پردازش شوند. عملیات زیر برای پاکسازی داده‌های متنی جهت سهولت در پردازش بر روی پیام‌ها اعمال می‌شوند:

- حذف https، @، # و شکلک از توییت‌ها
- توکن‌سازی[۱] متن
- تبدیل حروف بزرگ به کوچک
- حذف کلمات غیرانگلیسی
- حذف ایست واژه‌ها[۲] از توییت‌ها

پس از پاکسازی، پیام‌ها آماده تحلیل احساسات هستند. تحلیل احساسات با مدل Fin-BERT انجام می‌شود[۳]. Fin-BERT یک مدل یادگیری بر

۱. Tokenization: فرایند تجزیه متن به کوچک‌ترین عبارات معنادار

۲. Stop Words: کلماتی که در زبان‌های طبیعی دارای سطح پایینی از اطلاعات هستند.

۳. توسط D. Araci در سال ۲۰۱۹ معرفی شد. برای اطلاعات بیشتر مطالعه شود.

پایه معماری BERT است که به‌طور خاص برای تحلیل احساسات در حوزه مالی آموزش دیده شده است. این مدل به‌صورت خاص بر روی داده‌های مالی شامل گزارش‌های سالانه، اخبار اقتصادی و توییت‌های مرتبط با بازار سهام آموزش دیده و می‌تواند احساسات مثبت، منفی و خنثی را از متن‌های مرتبط با امور مالی استخراج کند. با توجه به اینکه مدل‌های عمومی مانند BERT برای داده‌های مالی بهینه نشده‌اند، FinBERT توانسته است در تحلیل احساسات و پیش‌بینی‌های مرتبط با بازارهای مالی عملکرد بهتری داشته باشد. این مدل به‌ویژه برای تحلیل داده‌های شبکه‌های اجتماعی و اخبار مالی به کار می‌رود تا در پیش‌بینی روندهای بازار سهام نقش مؤثری ایفا کند [۲۵]. از این رو در سامانه پیشنهادی از این مدل برای تحلیل احساسات استفاده می‌شود. از آنجا که برای محاسبه شاخص احساسات، تعداد پیام‌های مثبت و منفی مورد نیاز است، پس از مشخص‌شدن درصد مثبت، منفی و خنثی‌بودن پیام‌ها، با قراردادن یک لایه Softmax در مدل Fin-BERT، پیام‌ها به دو دسته مثبت و منفی تقسیم می‌شوند. نهایتاً برای هر سهام در هر روز، مجموع تعداد پیام‌های مثبت و مجموع تعداد پیام‌های منفی به دست می‌آیند. پس از تحلیل احساسات و به دست آوردن تعداد پیام‌های مثبت و منفی برای هر سهام در هر روز، شاخص احساسات روزانه از (۱) به دست می‌آید. $S_{ij}$ شاخص احساسات روز $i$ام و سهام $j$ام، $M_{ij}^{P}$ تعداد پیام‌های مثبت و $M_{ij}^{N}$ تعداد پیام‌های منفی مربوط به سهام $j$ام در روز $i$ام می‌باشد. پس از محاسبه شاخص احساسات برای هر سهام در هر روز، ویژگی‌های میانگین، میانه، تغییرات میانگین، انحراف معیار، بزرگ‌ترین و کوچک‌ترین این شاخص‌ها برای هر روز محاسبه می‌شود [۲۴] و [۲۶]

$$S_{ij}=\begin{cases}\dfrac{2M_{ij}^{P}}{(M_{ij}^{P}+M_{ij}^{N})-1} & , M_{ij}^{P}>M_{ij}^{N}\\ 0 & , M_{ij}^{P}=M_{ij}^{N}\\ \dfrac{1-2M_{ij}^{N}}{M_{ij}^{P}+M_{ij}^{N}} & , M_{ij}^{P}<M_{ij}^{N}\end{cases} \quad (1)$$

این ویژگی‌های آماری که در بازه $(-1,1)$ قرار دارند، در نهایت به عنوان ویژگی‌های داده‌های متنی به مدل پیش‌بینی‌کننده، داده می‌شوند [۲۴] و [۲۶]. برای تنظیم سامانه پیش‌بینی‌کننده پیشنهادی (SVM)، نیاز به محاسبه آنتروپی نتایج تحلیل احساسات است. این پارامتر از تعداد پیام‌های مثبت و منفی که قبلاً به دست آمده‌اند محاسبه می‌شود. آنتروپی احساسات کل پیام‌ها در روز $i$ام است و از رابطه زیر به دست می‌آید

$$I_i=-(P_i^{P}\log P_i^{P}+P_i^{N}\log P_i^{N}) \quad (2)$$

که $P_i^{P}$ به فراوانی نسبی پیام‌های مثبت و $P_i^{N}$ به فراوانی نسبی پیام‌های منفی هر روز اشاره دارند و این احتمال‌ها از (۳) به دست می‌آیند

$$P_i^{P}=\frac{\sum_{i=1}^{N_D}\sum_{j=1}^{N_S}p_{ij}^{p}}{N_s}$$
$$P_i^{N}=\frac{\sum_{i=1}^{N_D}\sum_{j=1}^{N_S}p_{ij}^{n}}{N_s} \quad (3)$$

که در آن $p_{ij}^{p}$ احتمال مثبت‌بودن پیام‌ها، $p_{ij}^{n}$ احتمال منفی‌بودن پیام‌ها برای سهام $j$ام در روز $i$ام می‌باشند. $N_D$ تعداد روزها و $N_S$ تعداد سهام‌ها را در بر می‌گیرد. مقادیر $p_{ij}^{p}$ و $p_{ij}^{n}$ طبق (۴) محاسبه می‌شوند

$$p_{ij}^{p}=\frac{M_{ij}^{P}}{M_{ij}^{P}+M_{ij}^{N}}$$
$$p_{ij}^{n}=\frac{M_{ij}^{N}}{M_{ij}^{P}+M_{ij}^{N}} \quad (4)$$

چنانچه قبلاً نیز گفته شد $M_{ij}^{P}$ تعداد پیام‌های مثبت و $M_{ij}^{N}$ تعداد پیام‌های منفی مربوط به سهام $j$ام در روز $i$ام را نشان می‌دهند.

## ۳-۲ بخش داده‌های عددی

داده‌های عددی که در این تحقیق مورد استفاده قرار گرفته است، ۵ مورد شامل قیمت‌های بسته‌شدن و بازشدن، بالاترین و پایین‌ترین قیمت و حجم معاملات انجام‌شده مربوط به شاخص یا سهام مورد نظر می‌باشد [۲۴] و [۲۶]. این داده‌ها از سایت yahoofinance جمع‌آوری می‌شود.

پیش‌پردازش این دسته داده‌ها شامل مقیاس‌بندی و برچسب‌گذاری می‌شود. از آنجا که ویژگی‌های برگرفته از داده‌های متنی (۶ ویژگی) در بازه $(-1,1)$ قرار دارند، مقادیر مربوط به داده‌های عددی طوری مقیاس‌بندی می‌شوند که با آنها در یک مقیاس قرار بگیرند. در نهایت این پنج ویژگی از داده‌های عددی به همراه شش ویژگی از داده‌های متنی، تشکیل بردار داده و به‌عنوان ورودی به مدل پیش‌بینی‌کننده داده می‌شوند.

از طرفی، از آنجا که ابزار پیش‌بینی‌کننده پیشنهادی (SVM) بر اساس یادگیری تحت نظارت است، نیاز است بردار ورودی برچسب‌گذاری شود که این عمل با روش بازگشتی صورت می‌گیرد. رابطه زیر نحوه برچسب‌گذاری بردار ورودی را که با مقایسه قیمت بسته‌شدن شاخص در دو روز متوالی انجام می‌شود، نشان می‌دهد

$$Y_i=\begin{cases}+1 & , close_{i-1}<close_i\\ -1 & , close_{i-1}>close_i\end{cases} \quad (5)$$

## ۳-۳ پیش‌بینی‌کننده بهبودیافته

در این بخش SVM با هسته گوسی به عنوان پیش‌بینی‌کننده پیشنهادی استفاده می‌شود که برای پیاده‌سازی آن در محیط پایتون از کتابخانه Sckitlearn استفاده شده است. برای تنظیم پارامترهای مدل از جست‌وجوی شبکه‌ای[1] بر روی پارامترهای SVM بر اساس کمینه‌کردن تابع هزینه پیشنهادی استفاده می‌شود. نقش تابع هزینه در SVM این است که با استفاده از آن مرز تصمیم‌گیری بین دو دسته تعیین می‌شود. تابع هزینه اجازه می‌دهد که مرز تصمیم‌گیری، بهتر انتخاب شده و در نتیجه دقت پیش‌بینی‌کننده افزایش یابد. تابع هزینه SVM از کمینه‌کردن رابطه زیر به دست می‌آید

$$J=E\{L(e_i)\}+l \quad (6)$$

که $e_i$ مقدار خطای مربوط به هر داده است که از حاصل‌ضرب برچسب واقعی داده $(Y_i)$ در برچسب یا کلاس پیش‌بینی‌شده $\hat{Y}_i$ توسط مدل به دست می‌آید. $L(e_i)$ تابع زیان نمونه‌هاست که در این تحقیق تابع Hing-Loss در نظر گرفته می‌شود و $E\{L(e_i)\}$ امید ریاضی تابع زیان نمونه‌هاست که از طریق (۷) به دست می‌آید

$$E\{L(e_i)\}=\sum_i f_i\max(0;1-Y_i\hat{Y}_i) \quad (7)$$

1. Grid Search

جدول ۱: لیست سهام‌ها.

| ردیف | نماد سهام | توضیح |
|---|---|---|
| ۱ | AAPL | Apple Inc |
| ۲ | AMC | AMC Entertainment Holdings, Inc |
| ۳ | AMD | Advanced Micro Devices, Inc. |
| ۴ | ATER | Aterian, Inc |
| ۵ | AMZN | Amazon.com Inc |
| ۶ | BABA | Alibaba Group Holding Limited |
| ۷ | DIS | The Walt Disney Company |
| ۸ | FB | Meta Platforms/Facebook, Inc |
| ۹ | GME | GameStop Corp |
| ۱۰ | MSFT | Microsoft Corporation |
| ۱۱ | NFLX | Netflix, Inc |
| ۱۲ | NIO | NIO Inc |
| ۱۳ | NVDA | NVIDIA Corporation |
| ۱۴ | QQQ | Invesco QQQ Trust |
| ۱۵ | SPY | SPDR S&P ۵۰۰ ETF Trust |
| ۱۶ | TSLA | Tesla, Inc |
| ۱۷ | UBER | Uber Technologies, Inc |

$$E\{L(e_i)\} = \sum_i f_i \max(۰;۱-Y_i\hat{Y}_i) \quad (۷)$$

در این رابطه به $f_i$ دانش پیشین می‌گویند که معمولاً برای ساده‌کردن محاسبات آن را در نظر نمی‌گیرند؛ ولی در این پژوهش برای بهبود مدل SVM از $f_i = ۱-I_i$ برای این منظور استفاده شده که $I_i$ همان آنتروپی در (۲) است. در نتیجه تابع هزینه مدل به صورت رابطه زیر بازنویسی می‌شود

$$J = l + \sum_i (۱-I_i)\max(۰;۱-Y_i\hat{Y}_i) \quad (۸)$$

## ٤- ارزیابی

ارزیابی شامل معیارهای دقت، حساسیت و صحت می‌شود که طبق روابط زیر محاسبه می‌شوند

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (۹)$$

$$Recall = \frac{TP}{TP+FN} \quad (۱۰)$$

$$Precision = \frac{TP}{TP+FP} \quad (۱۱)$$

در این تعریف $TP$ به تعداد نمونه‌های مثبت و $TN$ به تعداد نمونه‌های منفی که به درستی تشخیص داده شده‌اند، اشاره دارند. همچنین $FP$ تعداد نمونه‌های مثبت و $FN$ تعداد نمونه‌های منفی هستند که در کلاس اشتباه قرار گرفته‌اند.

## ٥- شبیه‌سازی

جهت پیاده‌سازی روش پیشنهادی به دو دسته داده نیاز است که این داده‌ها توسط مؤلفان جمع‌آوری شده است. داده‌های عددی مربوط به شاخص بازار سهام نیویورک که شامل قیمت‌های بسته‌شدن و بازشدن، بالاترین و پایین‌ترین قیمت و حجم معاملات انجام گردیده است که به‌طور مستقیم از وبگاه یاهو فایننس در بازه زمانی ۲۰۱۸/۰۱/۱۰ تا ۲۰۲۰/۰۱/۰۱ جمع‌آوری شده است[1]. در قسمت داده‌های متنی، پیام‌های مربوط به هفده سهام بازار بورس آمریکا با روش WebCrawel از شبکه X[2] در بازه زمانی گفته‌شده نیز جمع‌آوری می‌شود که پیام‌های مربوط به روزهای تعطیل در نظر گرفته نمی‌شوند. جدول ۱ اطلاعات مربوط به سهام‌هایی را که پیام‌های مرتبط با آنها جمع‌آوری شده، نمایش می‌دهد. ستون دوم و سوم این جدول به ترتیب نماد سهام و شرکت مرتبط را نشان می‌دهد. طبق این مجموعه داده‌ها پارامترهای روابط بخش چهارم برابر مقادیر زیر قرار می‌گیرند

$$N_D = ۴۹۶$$
$$N_S = ۱۷$$
$$i = \{۱,۲,\ldots,۴۹۶\}$$
$$j = \{۱,۲,\ldots,۱۷\}$$

پس از پاکسازی و تحلیل احساسات پیام‌ها، تعداد پیام‌های مثبت و منفی به دست می‌آید که در جدول ۲ قسمتی از نتایج را می‌توان مشاهده کرد. طبق این جدول، سهام‌های "fb"، "aapl" و "spy" در مجموع بیشترین تعداد پیام‌ها را داشته‌اند. این امر نشان می‌دهد که این سهام‌ها در طول دوره تحلیل از محبوبیت یا توجه بیشتری در شبکه اجتماعی برخوردار بوده‌اند. جدول ۳ نیز قسمتی از نتایج محاسبه شاخص احساسات روزانه را نشان می‌دهد. همان طور که مشاهده می‌شود برای هر سهام در هر روز یک عدد به‌عنوان شاخص احساسات روزانه معرفی شده که مقداری بین $(-۱,۱)$ است. هرچه تعداد پیام‌های مثبت نسبت به پیام‌های منفی یک سهام بیشتر باشد، مقدار این شاخص به عدد یک نزدیک‌تر است. همان طور که می‌توان دید سهام‌های "tsla"، "msft" و "aapl" بالاترین شاخص مثبت و سهام‌های "spy"، "qqq" و "aapl" پایین‌ترین شاخص منفی را دارند.

## ٦- نتیجه و بحث

در این تحقیق سامانه پیش‌بینی‌کننده پیشنهادی با یک SVM معمولی و همچنین با تحقیق مشابه [۲۶] و با داده‌های یکسان مقایسه شده است. برای آموزش و آزمایش مدل پیش‌بینی‌کننده پیشنهادی از اعتبارسنجی متقابل پنج‌بخشی[3] استفاده شده است. تمامی داده‌ها به پنج بخش مساوی تقسیم شده و در هر مرحله یک بخش به عنوان داده آزمایشی و چهار بخش دیگر به عنوان داده‌های آموزشی در نظر گرفته و ارزیابی می‌شوند. این روال پنج بار تکرار شده تا همه بخش‌ها یک بار برای آموزش و یک بار برای آزمایش به کار روند. در نهایت میانگین این پنج مرحله به عنوان نتیجه نهایی ارزیابی پیش‌بینی‌کننده در نظر گرفته شده است. از طرفی پارامترهای مدل (c, gamma) با استفاده از جست‌وجوی شبکه‌ای در حین آموزش انتخاب می‌شوند. جدول ۴ نتایج شبیه‌سازی را نشان می‌دهد که در آن مقادیر ارزیابی و پارامترهای مدل (c, gamma) سامانه پیشنهادی و دو روش دیگر آورده شده است.

با مقایسه روش پیشنهادی با SVM می‌توان دید که مقادیر دقت، صحت و مقدار حساسیت به میزان ۳٪ بهبود یافته است. همان طور که انتظار می‌رفت با در نظر گرفتن تابع هزینه پیشنهادی در SVM، عملکرد کلی آن بهبود یافته است.

1. https://finance.yahoo.com/quote/%5ENYA/history
2. https://x.com
3. 5-Fold Cross Validation

جدول ۲: تعداد پیام‌های مثبت و منفی سهام‌ها.

| سهام | ۲۰۱۸/۰۱/۱۰ | ۲۰۱۸/۰۱/۱۱ | ۲۰۱۸/۰۱/۱۲ | ۲۰۱۸/۰۱/۱۳ | ۲۰۱۸/۰۱/۱۴ |
|---|---|---|---|---|---|
| aapl | {'مثبت':۱۲۱،'منفی':۵۸} | {'مثبت':۹۹،'منفی':۵۱} | {'مثبت':۷۲،'منفی':۵۸} | {'مثبت':۵۴،'منفی':۲۳} | {'مثبت':۲۹،'منفی':۲۱} |
| amc | {'مثبت':۱،'منفی':۰} | {'مثبت':۰،'منفی':۰} | {'مثبت':۰،'منفی':۰} | {'مثبت':۰،'منفی':۰} | {'مثبت':۰،'منفی':۰} |
| amd | {'مثبت':۴،'منفی':۶} | {'مثبت':۷،'منفی':۴} | {'مثبت':۴،'منفی':۰} | {'مثبت':۲،'منفی':۲} | {'مثبت':۶،'منفی':۱} |
| amzn | {'مثبت':۲۹،'منفی':۱۳} | {'مثبت':۲۹،'منفی':۱۸} | {'مثبت':۲۳،'منفی':۲۲} | {'مثبت':۱۲،'منفی':۸} | {'مثبت':۱۳،'منفی':۹} |
| baba | {'مثبت':۷،'منفی':۶} | {'مثبت':۸،'منفی':۵} | {'مثبت':۹،'منفی':۲} | {'مثبت':۲،'منفی':۲} | {'مثبت':۳،'منفی':۳} |
| dis | {'مثبت':۵،'منفی':۳} | {'مثبت':۴،'منفی':۶} | {'مثبت':۳،'منفی':۲} | {'مثبت':۱،'منفی':۲} | {'مثبت':۱،'منفی':2} |
| fb | {'مثبت':۵۷۷،'منفی':۴۹۸} | {'مثبت':۸۶۹،'منفی':۷۳۹} | {'مثبت':۱۳۵۴،'منفی':۸۷۴} | {'مثبت':۷۹۱،'منفی':۶۲۰} | {'مثبت':۷۶۶،'منفی':۶۱۲} |
| gme | {'مثبت':۰،'منفی':۰} | {'مثبت':۰،'منفی':۰} | {'مثبت':۴،'منفی':۱} | {'مثبت':۰،'منفی':۰} | {'مثبت':۰،'منفی':۰} |
| infy | {'مثبت':۰،'منفی':۰} | {'مثبت':۰،'منفی':۰} | {'مثبت':۰،'منفی':۰} | {'مثبت':۰،'منفی':۰} | {'مثبت':۰،'منفی':۰} |
| msft | {'مثبت':۱۱،'منفی':۸} | {'مثبت':۵،'منفی':۳} | {'مثبت':۳،'منفی':۴} | {'مثبت':۵،'منفی':۱} | {'مثبت':۴،'منفی':۳} |
| nflx | {'مثبت':۱۴،'منفی':۷} | {'مثبت':۲۲،'منفی':۱۷} | {'مثبت':۱۴،'منفی':۱۷} | {'مثبت':۱۰،'منفی':۵} | {'مثبت':۱۱،'منفی':۹} |
| nio | {'مثبت':۰،'منفی':۰} | {'مثبت':۰،'منفی':۰} | {'مثبت':۰،'منفی':۰} | {'مثبت':۰،'منفی':۰} | {'مثبت':۰،'منفی':۱} |
| nvda | {'مثبت':۱۱،'منفی':۷} | {'مثبت':۱۵،'منفی':۱۳} | {'مثبت':۱۱،'منفی':۶} | {'مثبت':۸،'منفی':۶} | {'مثبت':۵،'منفی':۱۰} |
| qqq | {'مثبت':۲۵،'منفی':۱۳} | {'مثبت':۳۱،'منفی':۸} | {'مثبت':۱۷،'منفی':۱۱} | {'مثبت':۵،'منفی':۳} | {'مثبت':۴،'منفی':۱} |
| spy | {'مثبت':۵۳۵،'منفی':۲۰۷} | {'مثبت':۷۳۳،'منفی':۲۲۷} | {'مثبت':۴۴۵،'منفی':۲۶۰} | {'مثبت':۲۰۲،'منفی':۱۱۴} | {'مثبت':۲۰۱،'منفی':۸۰} |
| tsla | {'مثبت':۳۸،'منفی':۵۸} | {'مثبت':۷۳،'منفی':۷۵} | {'مثبت':۳۷،'منفی':۳۷} | {'مثبت':۲۱،'منفی':۲۳} | {'مثبت':۳۰،'منفی':۲۱} |
| uber | {'مثبت':۰،'منفی':۰} | {'مثبت':۰،'منفی':۰} | {'مثبت':۰،'منفی':۰} | {'مثبت':۱،'منفی':۰} | {'مثبت':۰،'منفی':۱} |

مرجع [۲۶] از روش مبتنی بر واژه‌نامه برای تحلیل احساسات و از یک مدل SVM پایه جهت پیش‌بینی روند بازار استفاده کرده است. با مقایسه نتایج این کارها با سامانه پیشنهادی می‌توان دید که مقادیر دقت و صحت به میزان ٪۶ و مقدار حساسیت به میزان ٪۳ افزایش یافته است. این بهبود حاصل تغییر تابع هزینه پیشنهادی و همچنین انجام روش تحلیل احساسات در سامانه پیشنهادی می‌باشد.

برای ارزیابی نقش روش تحلیل احساسات می‌توان از مقایسه [۲۶] با SVM استفاده کرد؛ زیرا هر دو از یک SVM برای پیش‌بینی استفاده کرده‌اند و تفاوت آنها صرفاً در روش تحلیل احساسات می‌باشد. همان طور که مشاهده می‌شود، مقادیر دقت و صحت ٪۲ بهبود یافته‌اند که بیانگر برتری تحلیل احساسات مبتنی بر یادگیری ماشین نسبت به روش مبتنی بر واژه‌نامه در [۲۶] است.

## ۷- جمع‌بندی و نتیجه‌گیری

در این تحقیق اثر تحلیل احساسات برگرفته‌شده از توییت‌ها در پیش‌بینی روند شاخص بازار سهام نیویورک با در نظر گرفتن آن در تابع هزینه مدل پیش‌بینی‌کننده و در نظر گرفتن آن به عنوان ورودی مدل پیش‌بینی‌کننده بررسی شد. تحلیل احساسات پیام‌ها با یک مدل مبتنی بر یادگیری ماشین (Fin-BERT) انجام شد. سپس با استفاده از نتایج این تحلیل شاخص احساسات روزانه محاسبه گردید. شش ویژگی از نتایج شاخص احساسات محاسبه شد و در کنار داده‌های مالی به مدل پیش‌بینی‌کننده بهبودیافته تزریق شد. در SVM نیز آنتروپی احساسات که از نتایج تحلیل احساسات محاسبه شد، در تابع هزینه مدل به عنوان دانش پیشین در نظر گرفته شد. در نهایت با روش‌های ارزیابی مختلف عملکرد روش پیشنهادی بررسی شد. از نتایج ارزیابی‌ها و مقایسه با تحقیقات دیگر می‌توان به این نتیجه رسید که پیش‌بینی‌کننده بهبودیافته، عملکرد پیش‌بینی را بهبود می‌دهد. همچنین تحلیل احساسات مبتنی بر یادگیری ماشین نسبت به روش مبتنی بر واژه‌نامه عملکرد بهتری داشته است.

## ۸- کارهای آینده و پیشنهادها

برای بهبود دقت پیش‌بینی در تحقیقات آینده پیشنهاد می‌شود که منابع متنی بیشتری نظیر اخبار و نظرات شبکه‌های اجتماعی به‌طور گسترده‌تری مورد استفاده قرار گیرند. همچنین شاخص‌های تکنیکال و سایر داده‌های عددی می‌توانند به‌عنوان ورودی‌های مدل در نظر گرفته شوند. بررسی زمان‌بندی تأثیر داده‌های متنی و بهبود روش‌های پیش‌پردازش برای شناسایی و تحلیل دقیق‌تر احساسات مرتبط نیز می‌تواند به بهبود دقت مدل‌ها کمک کند. علاوه بر این، آزمایش مدل‌های یادگیری ماشین پیشرفته مانند XGBoost و Random Forest با تابع هزینه پیشنهادی می‌تواند به تحلیل دقیق‌تری از عملکرد این مدل‌ها و بهبود پیش‌بینی‌ها منجر شود. مقایسه نتایج این مدل‌ها با مدل‌های موجود به شناسایی نقاط قوت و ضعف هر یک کمک کرده و می‌تواند به ارتقای عملکرد کلی پیش‌بینی بازار سهام منجر شود.

جدول ۳: شاخص احساسات روزانه سهام‌ها.

| سهام | ۲۰۱۸/۰۱/۱۰ | ۲۰۱۸/۰۱/۱۱ | ۲۰۱۸/۰۱/۱۲ | ۲۰۱۸/۰۱/۱۳ | ۲۰۱۸/۰۱/۱۴ |
|---|---|---|---|---|---|
| aapl | −۰٫۳۵۲۰ | −۰٫۳۲۰۰ | −۰٫۱۰۷۷ | −۰٫۴۰۲۶ | −۰٫۱۶۰۰ |
| amc | −۱٫۰۰۰۰ | ۰٫۰۰۰۰ | ۰٫۰۰۰۰ | ۰٫۰۰۰۰ | ۰٫۰۰۰۰ |
| amd | ۰٫۲۰۰۰ | −۰٫۲۷۲۷ | −۱٫۰۰۰۰ | ۰٫۰۰۰۰ | −۰٫۷۱۴۳ |
| amzn | −۰٫۳۸۱۰ | −۰٫۲۳۴۰ | −۰٫۰۲۲۲ | −۰٫۲۰۰۰ | −۰٫۱۸۱۸ |
| baba | −۰٫۰۷۶۹ | −۰٫۲۳۰۸ | −۰٫۶۳۶۴ | ۰٫۰۰۰۰ | ۰٫۰۰۰۰ |
| dis | −۰٫۲۵۰۰ | ۰٫۲۰۰۰ | ۰٫۲۰۰۰ | ۰٫۳۳۳۳ | ۰٫۳۳۳۳ |
| fb | −۰٫۰۷۳۵ | −۰٫۰۸۰۸ | −۰٫۲۱۵۴ | −۰٫۱۲۱۲ | −۰٫۱۱۱۸ |
| gme | ۰٫۰۰۰۰ | ۰٫۰۰۰۰ | −۰٫۶۰۰۰ | ۰٫۰۰۰۰ | ۰٫۰۰۰۰ |
| infy | ۰٫۰۰۰۰ | ۰٫۰۰۰۰ | ۰٫۰۰۰۰ | ۰٫۰۰۰۰ | ۰٫۰۰۰۰ |
| msft | −۰٫۱۵۷۹ | −۰٫۲۵۰۰ | ۰٫۱۴۲۹ | −۰٫۶۶۶۷ | −۰٫۱۴۲۹ |
| nflx | −۰٫۳۳۳۳ | −۰٫۱۲۸۲ | ۰٫۰۹۶۸ | −۰٫۳۳۳۳ | −۰٫۱۰۰۰ |
| nio | ۰٫۰۰۰۰ | ۰٫۰۰۰۰ | ۰٫۰۰۰۰ | ۰٫۰۰۰۰ | ۱٫۰۰۰۰ |
| nvda | −۰٫۱۷۶۵ | −۰٫۰۷۱۴ | −۰٫۲۹۴۱ | −۰٫۱۴۲۹ | ۰٫۳۳۳۳ |
| qqq | −۰٫۳۱۵۸ | −۰٫۵۸۹۷ | −۰٫۲۱۴۳ | −۰٫۲۵۰۰ | ۰٫۶۰۰۰ |
| spy | −۰٫۴۴۲۰ | −۰٫۵۲۷۱ | −۰٫۲۶۲۴ | −۰٫۲۷۸۵ | ۰٫۴۳۰۶ |
| tsla | −۰٫۰۰۸۳ | −۰٫۰۱۳۵ | ۰٫۰۰۰۰ | ۰٫۰۲۲۲ | −۰٫۱۷۶۵ |
| uber | ۰٫۰۰۰۰ | ۰٫۰۰۰۰ | ۰٫۰۰۰۰ | −۱٫۰۰۰۰ | ۱٫۰۰۰۰ |

جدول ۴: نتایج شبیه‌سازی.

| روش‌ها | ACC | Percision | Recall | C | Gamma |
|---|---|---|---|---|---|
| [۲۶] | ۰٫۸۲ | ۰٫۸۶ | ۰٫۸۳ | ۰٫۰۰۰۱۱۲ | ۹۴۷۳٫۶۸۴ |
| SVM | ۰٫۸۴ | ۰٫۸۸ | ۰٫۸۳ | ۰٫۰۰۰۴۸۳ | ۶۳۱۵٫۷۸۹ |
| روش پیشنهادی | ۰٫۸۸ | ۰٫۹۲ | ۰٫۸۶ | ۰٫۰۰۰۴۸۳ | ۶۳۱۵٫۷۸۹ |

# مراجع


[1] S. Usmani and J. A. Shamsi, "News sensitive stock market prediction: literature review and suggestions," *Peer J. Comput. Sci.*, vol. 7, no. 4, Article ID: e490, 2021.

[2] F. G. Monteiro and D. R. Ferreira, "How much does stock Prediction improve with sentiment analysis?" in *Proc. 5th ECML PKDD Workshop*, pp. 16-31, Ghent, Belgium, 18-18 Sept. 2020.

[3] O. Bustos, "Stock market movement forecast: a systematic review," *Expert Syst. Appl.*, vol. 156, Article ID: 113464, Oct. 2020.

[4] K. Fataliyev, A. Chivukula, M. Prasad, and W. Liu, *Stock Market Analysis with Text Data: A Review*, arXiv preprint arXiv:2106.12985.

[5] M. Rodríguez-Ibánez, A. Casánez-Ventura, F. Castejón-Mateos, and P.-M. Cuenca-Jiménez, "A review on sentiment analysis from social media platforms," *Expert Systems with Applications,* vol. 223, Article ID: 119862, 2021.

[6] P. Sudhir and V. D. Suresh, "Comparative study of various approaches, applications and classifiers for sentiment analysis," in *Glob. Transitions Proc.*, vol. 2, no. 2, pp. 205-211, Nov. 2021.

[7] C. Michele, *et al.*, "Machine learning sentiment analysis, COVID-19 news and stock market reactions," *Research in International Business and Finance*, vol. 64, Article ID: 101881, Jan. 2023.

[8] D. Kelvin, *et al.*, "Financial sentiment analysis: techniques and applications," *ACM Computing Surveys*, vol. 56, no. 9, Article ID: 220, Sept. 2024.

[9] D. Nabanita, *et al.*, "Effect of public sentiment on stock market movement prediction during the COVID-19 outbreak," *Social Network Analysis and Mining*, vol. 12, no. 1, Article ID: 92, 2022.

[10] N. Rouf, *et al.*, "Stock market prediction using machine learning techniques: a decade survey on methodologies, recent developments, and future directions," *Electron.*, vol. 10, no. 21, Article ID: 2127, 2021.

[11] N. Jing, Z. Wu, and H. Wang, "A hybrid model integrating deep learning with investor sentiment analysis for stock price prediction," *Expert Syst. Appl.*, vol. 178, Article ID: 115019, Sept. 2021.

[12] D. Kumar, P. K. Sarangi, and R. Verma, "A systematic review of stock market prediction using machine learning and statistical techniques," in *Mater. Today Proc.*, vol. 49, pt. 8., pp. 3187-3191, 2022.

[13] C. Chou, J. Park, and E. Chou, "Predicting stock closing price after COVID-19 based on sentiment analysis and LSTM," in *Proc. IEEE Adv. Inf. Technol. Electron. Autom. Control Conf.*, pp. 2752-2756, Chongqing, China, 12-14 Mar. 2021.

[14] M. Peji, Text Mining for Big Data Analysis in Financial Sector: A Literature Review, 2019.

[15] K. Mishev, A. Gjorgjevikj, I. Vodenska, L. T. Chitkushev, and D. Trajanov, "Evaluation of sentiment analysis in finance: from lexicons to transformers," *IEEE Access*, vol. 8, pp. 131662-131682, 2020.

[16] Y. Qi, "Stock prediction under COVID-19 based on LSTM," in *Proc. IEEE Asia-Pacific Conference on Image Processing, Electronics and Computers,* pp. 93-98, Dalian, China, 4-16 Apr. 2021.

[17] N. Christina and C. Tjortjis, "A methodology for stock movement prediction using sentiment analysis on Twitter and stocktwits data," in *Proc. 6th South-East Europe Design Automation, Computer Engineering, Computer Networks and Social Media Conf.,* vol. 7 pp., Preveza, Greece, 24-26 Sept.2021.

[18] C. Liapis, A. Karanikola, and S. Kotsiantis, "Investigating deep stock market forecasting with sentiment analysis," *Entropy*, vol. 25, no. 2, Article ID: 219, Feb. 2023.

[19] W. Zhaoxia, *et al.*, "Learning-based stock trending prediction by incorporating technical indicators and social media sentiment," *Cognitive Computation*, vol. 15, no. 3, pp. 1092-1102, 2023.

[20] N. Zaman, *et al.*, "Stock market prediction based on machine learning and social sentiment analysis," *TechRxiv.* Mar. 27, 2023., DOI: 10.36227/techrxiv.22315069.v1.

[21] D. Narayana, *et al.*, "Stock price prediction using sentiment analysis and deep learning for Indian markets," arXiv preprint arXiv:2204.05783, 2022.

[22] K. Jihwan, H. Kim, and S. Y. Choi, "Forecasting the S&P 500 index using mathematical-based sentiment analysis and deep learning models: a FinBERT transformer model and LSTM," *Axioms*, vol. 12, no. 9, Article ID: 835, Sept. 2023.

[23] Q. Xiao and B. Ihnaini, "Stock trend prediction using sentiment analysis," *Peer J. Computer Science*, vol. 9, Article ID; e1293, 2023.

[24] Y. Qiu, Z. Song, and Z. Chen, "Short-term stock trends prediction based on sentiment analysis and machine learning," *Soft Comput.*, vol. 26, no. 5, pp. 2209-2224, Mar. 2022.

[25] J. Liu, J. Leu, and S. Holst, "Stock price movement prediction based on stocktwits investor sentiment using FinBERT and ensemble SVM," *Peer J. Computer Science*, vol. 9, Article ID; e1403, 2023.

[26] R. Ren, D. D. Wu, and D. D. Wu, "Forecasting stock market movement direction using sentiment analysis and support vector machine," *IEEE Syst. J.*, vol. 13, no. 1, pp. 760-770, Mar. 2019.



**مهین یعقوب‌زاده** در سال ۱۳۹۷ مدرک کارشناسی مهندسی برق خود را از دانشگاه دولتی بیرجند دریافت نمود. ایشان در سال ۱۳۹۸ وارد دوره کارشناسی ارشد مهندسی برق گرایش مخابرات سیستم در دانشگاه فردوسی مشهد گردید و در سال ۱۴۰۳ موفق به اخذ مدرک کارشناسی ارشد از دانشگاه مذکور شد. از سال ۱۴۰۱، ایشان به عنوان مهندس ناظر تأسیسات الکتریکی مشغول به فعالیت است. زمینه‌های علمی مورد علاقه نام‌برده شامل پردازش سیگنال، هوش مصنوعی، اینترنت اشیا و هوشمندسازی ساختمان می‌باشد.

**دکتر مرتضی خادمی درخ** در سال ۱۳۷۵ مدرک دکتری مهندسی برق (با تخصص در فشرده‌سازی تصویر) خود را از دانشگاه ولنگنگ در کشور استرالیا دریافت نمود. نام‌برده هم‌اکنون با مرتبه علمی استاد در گروه مهندسی برق دانشگاه فردوسی مشهد مشغول به فعالیت می‌باشد. زمینه‌های علمی مورد علاقه ایشان شامل مخابرات ویدیویی، پردازش سیگنال‌های تصویری و ویدیویی و پردازش سیگنال‌های پزشکی می‌باشد.

**عباس ابراهیمی مقدم** مدرک دکتری مهندسی برق (مخابرات) خود را از دانشگاه مک‌مستر کانادا در سال ۱۳۸۹ دریافت نمودد. ایشان هم‌اکنون با مرتبه علمی استادیار در گروه مهندسی برق دانشگاه فردوسی مشهد مشغول به فعالیت است. زمینه‌های علمی مورد علاقه نام‌برده شامل پردازش سیگنال‌های تصویری و ویدیویی، پردازش سیگنال‌های حیاتی و پردازش تصاویر هوایی می‌باشد.

**هادی صدوقی یزدی** در سال ۱۳۸۴ مدرک دکتری مهندسی برق (الکترونیک) خود را از دانشگاه تربیت مدرس ایران دریافت نمودند. نام‌برده هم‌اکنون با مرتبه علمی استاد در گروه مهندسی کامپیوتر دانشکده مهندسی دانشگاه فردوسی مشهد مشغول به فعالیت است. زمینه‌های علمی مورد علاقه ایشان شامل یادگیری ماشین و شبکه‌های عصبی می‌باشد.