فصلنامه مهندسی مالی و مدیریت اوراق بهادار
دوره چهارده، شماره پنجاه و پنج، تابستان ۱۴۰۲
نوع مقاله : علمی پژوهشی
صفحات : ۸۷-۶۵

یادگیری عمیق برای پیش‌بینی بازار سهام با استفاده از اطلاعات عددی و متنی (رویکرد الگوریتم حافظه کوتاه مدت ماندگارLSTM)

سیده مژگان بهشتی مسئله گو[۱]
محمدعلی افشار کاظمی[۲]
جلال حقیقت منفرد[۳]
علی رضاییان[۴]

تاریخ دریافت مقاله : ۱۴۰۰/۱۲/۲۱ تاریخ پذیرش مقاله : ۱۴۰۱/۰۴/۲۶

چکیده

قیمت سهام تحت تأثیر عوامل بسیاری است،که کار پیش‌بینی را چالش‌برانگیز می‌کند. این پیش‌بینی اگر فقط داده‌های عددی یا اطلاعات متنی را در نظر بگیرد، اغلب بی‌اثر می‌شود. هدف این پژوهش ارائه یک روش پیش‌بینی قیمت روز آینده سهام بر اساس ساختار شبکه عصبی عمیق با استفاده از داده‌های قیمت، مجموعه‌ای از شاخص‌های فنی و سر تیتر اخبار به‌عنوان ورودی مدل است. برای این منظور از داده‌های سهام شاخص داوجونز و داده‌های خبری کانال ردیت استفاده شده است. از داده‌های سهام ویژگی‌های مبتنی بر شاخص فنی استخراج می‌شوند و داده‌های خبری توسط روش کوله‌کلمات به بردار ویژگی تبدیل می‌شوند و به شبکه حافظه کوتاه‌مدت ماندگار (LSTM) برای پیش‌بینی داده می‌شوند.از دقت به‌عنوان معیار ارزیابی عملکرد استفاده شده و آزمایش‌هایی بر روی دو مجموعه داده فقط عددی و فقط متنی برای ارزیابی استفاده همزمان دو منبع اطلاعاتی انجام پذیرفته است. همچنین ازسه شبکه ، SVM، MLP وRNN برای ارزیابی مدل استفاده شده است. نتایج نشان می‌دهد که مدل LSTM بالاترین دقت پیش‌بینی ۶۹,۱۹٪ را با استفاده از اخبار و داده‌های مالی به دست آورده است. داده‌های خبری با دقت۶۵,۶۲٪ و داده‌های عددی با دقت۵۱,۸۹٪ می‌باشند.همچنین مدلLSTM در مقایسه با شبکه‌های عصبی SVM و MLP و RNN از عملکرد بهتری برخوردار می‌باشد.

کلمات کلیدی

پیش‌بینی‌بازارسهام، پردازش‌زبان‌طبیعی، یادگیری‌عمیق، شاخص‌های‌فنی، حافظه‌کوتاه‌مدت‌ماندگار

۱-گروه فناوری اطلاعات، واحد تهران مرکزی، دانشگاه آزاد اسلامی، تهران، ایران. mojgan.bm1988@gmail.com
۲-گروه مدیریت صنعتی، واحد تهران مرکزی، دانشگاه آزاد اسلامی، تهران، ایران. (نویسنده مسئول) dr.mafshar@gmail.com
۳-گروه مدیریت صنعتی، واحد تهران مرکزی، دانشگاه آزاد اسلامی، تهران، ایران. jhm1847@gmail.com
۴-گروه مدیریت دولتی، دانشکده مدیریت و حسابداری، دانشگاه شهید بهشتی، تهران، ایران. a-rezaeian@sbu.ac.ir

مقدمه

بازارهای مالی به‌عنوان قلب اقتصاد جهان در نظر گرفته می‌شوند که در آن‌ها هرروز میلیاردها دلار دادوستد می‌شود. واضح است که پیش‌بینی خوب رفتار آینده بازارها در حوزه‌های مختلف بسیار ارزشمند خواهد بود. بازارهای بورس نقش مهمی در رشد اقتصادی ایفا می‌کنند بنابراین تجزیه‌وتحلیل رفتار آن‌ها و پیش‌بینی آینده آن‌ها می‌تواند در دستیابی به اهداف اقتصادی بسیار مفید باشد (بک و لوین[۱]، ۲۰۰۴). همچنین سرمایه‌گذاران برای تصمیم‌گیری صحیح (فروش یا نگهداری) سهامی که دارند باید روند قیمت سهام را به‌درستی پیش‌بینی کنند تا از این طریق سود قابل‌توجهی را تحقق بخشند (خان و همکاران[۲]، ۲۰۲۰). بااین‌حال، رفتار پیش‌بینی کار دشواری است زیرا قیمت سهام یک سری زمانی بسیار ناپایدار در حوزه مالی است (لی و یی[۳]، ۲۰۲۱). قیمت‌های سهام تحت تأثیر عوامل بسیاری از جمله نرخ بهره، نرخ ارز، سیاست پولی، اخبار، رسانه‌های اجتماعی، احساسات سرمایه‌گذار و ...قرار دارد. بنابراین مدل‌سازی رابطه بین قیمت سهام و این عوامل و پیش‌بینی روند قیمت سهام از جمله چالش‌های پیش روی پژوهشگران و سرمایه‌گذاران است (چن و همکاران[۴]، ۲۰۱۹).

سیستم‌های پیش‌بینی کننده مختلفی پیشنهاد شده است که از یک یا چند نوع داده استفاده می‌کنند. این سیستم‌ها اطلاعات مفیدی را برای سرمایه‌گذاران فراهم می‌کنند تا تصمیمات سرمایه‌گذاری را برای خرید یا فروش سهام بگیرند. اما استفاده از یک نوع داده ممکن است باعث افزایش دقت پیش‌بینی برای بازار سهام نشود. داده‌های قیمت تاریخی در رویکرد تجزیه‌وتحلیل فنی مورد استفاده قرار گرفته است که در آن از ریاضیات و روش‌های آماری سنتی مانند رگرسیون، میانگین نمایی، میانگین نمایی متحرک برای تجزیه‌وتحلیل داده‌ها و یافتن روندهای آینده بازار استفاده شده است. بااین‌حال روش‌های آماری اغلب فرض می‌کنند که سری‌های زمانی از یک فرایند خطی تشکیل شده‌اند و بنابراین در پیش‌بینی داده‌های غیرخطی سهام ضعیف عمل می‌کنند. بنابراین محققان از تکنیک‌های یادگیری ماشین مانند یادگیری عمیق بر روی داده‌های قیمت تاریخی سهام استفاده می‌کنند (چن و همکاران، ۲۰۱۹). اما در نظر گرفتن عوامل دیگر مهم است. منابع زیادی از داده‌های متنی مانند اخبار، توییت ها و گزارش‌های سالانه وجود دارد که می‌توانند اطلاعات مهم و معناداری را ارائه دهند که می‌تواند قیمت سهام را تحت تأثیر قرار دهد. داده‌های متنی به‌ویژه اخبار، منابع مناسبی از اطلاعات هستند زیرا امکان پیش‌بینی روندهای مالی با توجیه آن را فراهم می‌کنند (اسمانی و جوواد[۵]، ۲۰۲۱: چن و جیمز[۶]، ۲۰۱۱). به‌عنوان‌مثال یک مقاله خبری در مورد یک شرکت با کلمات یا عباراتی مانند استعفا، ریسک، عدم پرداخت بدهی به

سرمایه‌گذار کمک می‌کند تا کاهش در قیمت سهام را پیش‌بینی کند (نصیر طوسی[۷] ،۲۰۱۵).علاوه بر این اخبار نامشخص مانند شوک‌های اقتصادی، جنگ، تروریسم، بلایای طبیعی می‌تواند بر روند بازار تأثیر بگذارد (اسمانی و جوواد،۲۰۲۱). در نتیجه کشف بهترین دانش از داده‌های متنی یکی از نیازهای اساسی است. بااین‌حال، چنین حجم عظیمی از رسانه‌های اجتماعی و داده‌های اخبار را نمی‌توان به‌طور کامل توسط سرمایه‌گذاران ارزیابی کرد؛ بنابراین، یک سیستم پشتیبانی تصمیم‌گیری خودکار برای سرمایه‌گذاران ضروری است، زیرا این سیستم روند سهام را به‌طور خودکار با استفاده از چنین مقادیر زیادی از داده‌ها ارزیابی خواهد کرد .این سیستم خودکار می‌تواند با استفاده از الگوریتم‌های یادگیری ماشین ساخته شود .پیدا کردن الگوریتم‌هایی که در پیش‌بینی روند بازار سهام با استفاده از داده‌های خارجی، مانند اخبار مالی، مؤثرتر هستند، این سیستم‌ها می‌توانند با استفاده از الگوریتم‌های یادگیری ماشین ساخته شوند. پیدا کردن الگوریتم‌هایی که در پیش‌بینی روند بازار سهام با استفاده از داده‌های متنی و عددی مؤثر هستند بسیار مهم است (خان و همکاران ،۲۰۲۰). یادگیری عمیق یکی از مدل‌های یادگیری ماشین است و مجموعه‌ای از الگوریتم‌ها است که سعی در مدل‌سازی مفاهیم مفهومی سطح بالا با استفاده از یادگیری در سطوح و لایه‌های مختلف دارد که باعث دقت بالا در نتایج خواهد شد. با طراحی الگوریتم‌های یادگیری عمیق می‌توان حجم عظیمی از داده‌های غیرخطی را آموزش داد. در مقایسه با الگوریتم‌های قبلی یادگیری ماشین این الگوریتم‌ها را می‌توان برای حل مسائل غیرخطی به کار برد (یو و ین[۸]،۲۰۲۰:هوآنگ و همکاران[۹] ،۲۰۱۷).

از انجاییکه قیمت سهام با بسیاری از عوامل مرتبط است و استفاده از یک نوع داده ممکن است باعث افزایش دقت پیش‌بینی نشود بنابراین بیش‌ترین اطلاعات مفید و مرتبط تا جایی که ممکن است پیش‌بینی را تضمین می‌کند چگونه می‌توان از طریق اخبار و قیمت‌های تاریخی موجب افزایش دقت پیش‌بینی سرمایه‌گذاری شد؟ استفاده از شبکه‌های عصبی عمیق چه میزان مؤثر و کارآمد است؟ این پژوهش قصد دارد با طراحی یک مدل شبکه عصبی عمیق مبتنی بر داده‌های متنی و عددی به‌دقت پیش‌بینی بازار سهام کمک کند. هدف این مدل این است که در عین توجه به اخبار به شاخص‌های فنی بازار برای کاهش ریسک توجه شود. آنچه در ادامه خواهد آمد ابتدا بررسی مروری پژوهش‌های پیشین و سپس طراحی و توسعه مدل خواهد آمد. در انتها پس از احصاء نتایج حاصل از متدلوژی، در مورد عواید استفاده از این مدل بحث و بررسی خواهد شد.

ادبیات و مبانی نظری

تجزیه‌وتحلیل بازار سهام که شامل جمع‌آوری، تفسیر و یکپارچه‌سازی اطلاعات مرتبط است به درک و پیش‌بینی روند قیمت سهام و اتخاذ تصمیمات سرمایه‌گذاری برای کاهش خطرات و به دست آوردن مزایای بالاتر کمک خواهد کرد. پیش‌بینی قیمت سهام توجه بسیاری را از سوی سرمایه‌گذاران به‌منظور دستیابی به بازده‌های بالاتر جلب کرده است بااین‌حال قیمت سهام تحت تأثیر عوامل زیادی ازجمله عوامل سیاسی-اقتصادی و بازار و همچنین تکنولوژی و رفتار سرمایه‌گذار قرار دارد که کار پیش‌بینی را چالش‌برانگیز می‌کند.(جی و همکاران ،۲۰۲۰)

در حال حاضر بسیاری از تکنیک‌ها و مدل‌های مختلف برای پیش‌بینی بازار سهام به کار گرفته شده است مانند مدل‌های آماری سنتی، روش‌های یادگیری ماشین، شبکه‌های عصبی مصنوعی و غیره (هاو و جاو،۲۰۲۰). مدل‌های متداول آماری سنتی شامل اتو رگرسیون، میانگین متحرک و اتورگرسیون میانگین متحرک است. تمام این رویکردها به‌طور عمده بر خود سری‌های زمانی تمرکز دارند درحالی‌که بسیاری از عوامل تأثیرگذار مانند اطلاعات زمینه را نادیده می‌گیرند. به‌طور خاص آن‌ها داده‌های قبلی و داده‌های بعدی را به‌عنوان متغیر مستقل و وابسته با هدف به دست آوردن رابطه کمی بین آن‌ها در نظر می‌گیرند. همچنین این روش‌ها اغلب به برخی فرضیات و پیش‌آگاهی نیاز دارند. ازآنجاکه داده‌های سهام ماهیتاً غیرثابت یا غیرخطی هستند روش‌های تحلیلی مرسوم را بی‌اثر می‌کند. علاوه بر این مقدار اطلاعات پردازش‌شده توسط مدل‌سازی و پیش‌بینی بازار سهام اغلب بسیار بزرگ است که چالش‌های زیادی را برای طراحی الگوریتم ایجاد می‌کند. به‌منظور غلبه بر محدودیت‌های مدل‌های آماری مدل‌های مبتنی بر یادگیری ماشین اتخاذ شده‌اند. این مدل‌ها شامل ماشین بردار پشتیبان، رگرسیون بردار پشتیبان و الگوریتم‌های یادگیری عمیق می‌باشند (سو و همکاران،۲۰۲۰). امروزه یادگیری عمیق به دلیل توانایی عالی آن در ترسیم روابط غیرخطی و اتخاذ دانش زمینه‌ای محدود به روشی جدید در یادگیری ماشین تبدیل‌شده است. یادگیری عمیق قابلیت پردازش داده قدرتمندی دارد که می‌تواند مشکلات ناشی از پیچیدگی سری زمانی را حل کند. با توسعه یادگیری عمیق روش‌های زیادی مبتنی بر آن‌ها برای پیش‌بینی سهام مورداستفاده قرار گرفته است و برخی از نتایج اساسی را به دست آورده‌اند. یادگیری عمیق برای اولین بار در سال ۲۰۰۵ معرفی شد و از سال ۲۰۱۲ به‌طورجدی در نظر گرفته شده است. در واقع یادگیری عمیق به معنی بررسی روش‌های جدید برای شبکه‌های عصبی مصنوعی است. شبکه‌های عصبی یک‌لایه پنهان داخلی دارند و یک شبکه با چندین لایه پنهان داخلی یک شبکه عصبی عمیق

نامیده می‌شود. مدل‌های شبکه عصبی عمیق شامل شبکه‌های عصبی رمزگذار خودکار (AE)، شبکه‌های باور عمیق، شبکه‌های پیچشی (CNN) و شبکه‌های بازگشتی (RNN) می‌باشند. شبکه‌های پیچشی برای داده‌های بصری منحصربه‌فرد هستند و کاربرد گسترده‌ای در تشخیص تصویر و ویدئو دارند. علاوه بر این شبکه‌های بازگشتی برای داده‌های سری زمانی مناسب هستند. RNN شبکه‌های عصبی بازگشتی با یک یا چند حلقه بازگشت هستند. از لحاظ تئوری این شبکه‌ها می‌توانند داده‌ها در یک دنباله طولانی را ثبت و بهره‌برداری کنند (غلام‌زاده و باقرزاده ۲۰۱۹). حافظه کوتاه‌مدت ماندگار (LSTM) یک نوع شبکه بازگشتی است که از توسعه RNN ها پدیدار شده و توسط جرس و همکاران بهبود یافته است. حافظه بلندمدت به وزن‌های آموخته‌شده و حافظه کوتاه‌مدت به حالت‌های درونی سلول‌ها اشاره دارد. ویژگی اصلی این شبکه امکان یادگیری وابستگی بلندمدت است که با RNN غیرممکن است. یک شبکه LSTM معمولی از بلوک‌های حافظه مختلفی به نام سلول تشکیل شده است. دو حالت وجود دارد که به سلول بعدی منتقل می‌شوند: حالت سلول و حالت پنهان. بلوک حافظه مسئول به خاطر سپردن چیزهاست و دست‌کاری در این حافظه از طریق سه مکانیزم عمده به نام گیت‌ها انجام می‌شود. گیت ورودی که یادگیری را کنترل می‌کند، گیت فراموشی که آنچه را که باید فراموش شود کنترل می‌کند و گیت خروجی که مقدار محتوا را برای تغییر کنترل می‌کند.

پیشینه پژوهش

در این بخش تحقیقات مرتبط با پیش‌بینی سهام با استفاده از شبکه‌های عصبی عمیق بر اساس قیمت سهام و اخبار بررسی می‌شوند.

بازارهای بورس هر روز حجم زیادی از داده‌های معامله را تولید می‌کنند که شبکه‌های عصبی عمیقی را با حجم زیادی از داده‌ها برای آموزش و بهبود توانایی‌های پیش‌بینی آن‌ها فراهم می‌کند. بسیاری از تحقیقات دریافته‌اند که شبکه عصبی عمیق به دلیل توانایی حافظه، توانایی یادگیری بهتری برای داده‌های سری زمانی نسبت به سایر روش‌های یادگیری ماشین دارد. ژانگ و تن (۲۰۱۸)، از داده‌های قیمت تاریخی برای پیش‌بینی و رتبه‌بندی بازده آینده سهام از طریق یک مدل انتخاب سهام جدید بر اساس یک شبکه عصبی عمیق استفاده کردند .چن، ژو (۲۰۱۵)، از یک شبکه عصبی عمیق آموزش‌دیده توسط داده‌های قیمت برای پیش‌بینی نوسانات روزانه سهام در بازار سهام چین استفاده کردند .نتایج تحقیقات آن‌ها نشان می‌دهد که نرمال‌سازی داده‌ها برای بهبود دقت بسیار مفید است و دقت پیش‌بینی به‌طور قابل‌توجهی با افزایش ابعاد داده قیمت بهبود می‌یابد .آن‌ها همچنین اجرای پیش‌بینی‌ها برای انواع

مختلف سهام به‌طور جداگانه را پیشنهاد می‌کنند که می‌تواند دقت را بیشتر بهبود بخشد .لی، کایو و پن (۲۰۱۹)، سیستمی را می‌سازند که از یک معماری یادگیری عمیق برای بهبود نمایش‌های ویژگی بهره می‌برد و از ماشین یادگیری افراطی برای پیش‌بینی اثرات بازار استفاده می‌کند .آن‌ها نتیجه گرفتند که نمایش ویژگی‌های عمیق آموخته‌شده همراه با ماشین یادگیری افراطی می‌تواند دقت پیش‌بینی بازار را بهتر کند. نلسون و همکاران[۱۰] (۲۰۱۷)، روند آینده سهام در بورس اوراق بهادار برزیل را بر اساس قیمت تاریخی و شاخص‌های فنی پیش‌بینی می‌کنند .آن‌ها نشان می‌دهند که شبکه عصبی عمیق از نظر دقت پیش‌بینی سهام در مقایسه با دیگر روش‌های یادگیری ماشین، دستاوردهای قابل‌توجهی دارد .لی و لیائو (۲۰۱۸)، نیز توانایی یادگیری بین یادگیری عمیق و روش‌های یادگیری ماشینی کلاسیک را در بازار سهام چین مقایسه می‌کنند. آن‌ها مدل‌ها را با داده‌های قیمت و شاخص‌های فنی آموزش می‌دهند. نتایج طبقه‌بندی نشان می‌دهد که شبکه‌های عصبی عمیق دقت بالاتری دارند. فیشر و کراوس (۲۰۱۸)، تائید می‌کنند که شبکه عصبی عمیق بهتر از طبقه‌بندی کننده‌های بدون حافظه برای شاخص S&P عمل می‌کنند. جاو و همکاران (۲۰۱۶)، بازار سهام را با استفاده از شبکه عصبی بازگشتی پیش‌بینی کردند. این مطالعه با هدف بررسی امکان‌سنجی و کارایی LSTM در پیش‌بینی بازار سهام انجام شده است. بر اساس نتایج ،میانگین دقت مدل LSTM در پیش‌بینی شش سهم ۵۴,۸۳٪ بود که بالاترین و پایین‌ترین دقت به ترتیب ۵۹,۵٪ و ۴۹,۷۵٪ است. نتایج نشان داد که شبکه مورداستفاده عملکرد بهتری به نسبت سایر شبکه‌های عمیق موردبررسی در تحقیق است. چن و همکاران (۲۰۱۸)، یک مدل پیش‌بینی شاخص سهام مبتنی بر یادگیری عمیق را ارائه دادند. آن‌ها از داده‌های قیمت سهام CSI300 از بازار سهام چین استفاده کردند و شبکه پیشنهادی خود را با سه شبکه عصبی مصنوعی سنتی باهم مقایسه کردند. آن‌ها ادعا کردند که روش یادگیری عمیق از سه شبکه عصبی مصنوعی سنتی بهتر عمل می‌کند. آن‌ها همچنان متوجه شدند که افزایش مقدار داده عملکرد پیش‌بینی را افزایش می‌دهد که این نشان می‌دهد که یادگیری عمیق ویژگی‌های غیرخطی داده‌ها را در برمی‌گیرد. لی و پن (۲۰۲۱)، یک رویکرد عمیق را برای پیش‌بینی حرکت آینده سهام برای شاخص S&P پیشنهاد کردند. این مدل از یک روش یادگیری ترکیبی برای ترکیب دو شبکه عصبی بازگشتی و به دنبال آن یک شبکه عصبی کاملاً متصل استفاده می‌کند. تحقیقات نشان داد که مدل ترکیبی عمیق آن‌ها به‌طور قابل‌توجهی از بهترین مدل پیش‌بینی موجود عمل می‌کند و میانگین مربعات خطا را از ۴۳۸,۹۴ به ۱۸۶,۳۲ کاهش می‌دهد. آن‌ها ادعا کردند که روش‌های یادگیری عمیق می‌توانند واقعاً روند قیمت سهام آینده را به‌طور مؤثرتری پیش‌بینی کنند.

اخبار به‌عنوان یکی از عوامل محرک در بازار سهام، تأثیرات عمده‌ای بر تکامل قیمت در جنبه‌های مختلف ازجمله اقتصاد داخلی / جهانی و موقعیت‌های مالی و همچنین احساسات سرمایه‌گذاران می‌گذارد. ما و همکاران (۲۰۱۶)، مطالعه‌ای باهدف تسهیل عملکرد پیش‌بینی‌های بازار سهام مبتنی بر اخبار، یک مدل بازنمایی پراکنده جدید از اخبار(DRNEWS)را ارائه داده که اخبار را به‌عنوان بردارهای پیوسته که اطلاعات جمعی و همچنین دانش بین متنی را در مقالات خبری مختلف توصیف می‌کنند، قرار می‌دهد. توجه به متون اخیر در مورد بازنمایی‌های متنی مقالات خبری و کاربردهای آن‌ها در شبکه‌های عمیق به پیش‌بینی حرکات سهام جلب شده است .به‌عنوان‌مثال، لی (۲۰۱۶)، مدل بردار پاراگراف برای آموزش تعبیه اخبار پیشنهاد کرده است که فرآیند تصمیم‌گیری استراتژی‌های تجارت سهام را تسهیل می‌کند.علاوه بر این، دینگ و همکاران (۲۰۱۵)، استخراج رویداد را پیشنهاد می‌کنند .آن‌ها از عناوین خبری برای نگاشت رویداد خبری و پیش‌بینی سهام مبتنی بر رویداد مربوطه استفاده کردند، درحالی‌که هو و همکاران (۲۰۱۸)، از یک چارچوب یادگیری عمیق برای پیش‌بینی سهام مبتنی بر رویداد استفاده می‌کنند .پیش‌بینی روند سهام اخبار محور با نمایش اخبار به‌عنوان میانگین بردار کلمات انجام می‌شود همچنین آن‌ها با ساخت یک مدل پیش‌بینی قیمت سهام با استفاده از یک شبکه عصبی به‌عنوان رمزگذار اخبار و کلمات در تی‌تر به‌عنوان ورودی استفاده کردند . آکیتاوهمکاران[11] (۲۰۱۶) ، مقالات خبری را با استفاده از بردارها ارائه می‌دهند .آن‌ها یک شبکه عصبی عمیق را برای پیش‌بینی قیمت بسته شدن ۵۰ سهم در بورس سهام توکیو آموزش می‌دهند.

استفاده از منابع اطلاعاتی بیشتر، مانند متون خبری و قیمت‌های سهام، به روندی در پیش‌بینی سهام تبدیل شده است. چن و همکاران (۲۰۲۰)، یک مدل پیش‌بینی بازار سهام با استفاده از اطلاعات دوگانه (متن و عدد) ارائه کردند. آن‌ها در مرحله پیش‌پردازش از روش CNN برای استخراج ویژگی‌های دوگانه استفاده کردند که نشان‌دهنده روند بلندمدت داده‌های تاریخی و کوتاه‌مدت بازار است. در مرحله مدل‌سازی سری زمانی از یک شبکه LSTM-AE استفاده کردند. نتایج آن‌ها نشان داد که مدل پیشنهادی در مقایسه با سایر مدل‌ها از جمله SVR و CNN عملکرد بهتری دارد بطوریکه خطا را تا ۱۳,۴۷٪ و ۱۷,۶۳٪ کاهش می‌دهد. هو، لیو، (۲۰۱۸)، نوسانات روزانه سهام را با آموزش یک مدل GRU دو طرفه با بردارهای خبری و داده‌های قیمت پیش‌بینی می‌کنند. دینگ و همکاران (۲۰۱۶)، وارگاس، دی لیما، اوسوکوف (۲۰۱۷) ، چن و همکاران(۲۰۲۰)، دنگ و همکاران (۲۰۱۹)، لی و وانگ (۲۰۲۰) قدرت یادگیری عمیق برای پیش‌بینی سهام را همراه با داده‌های متنی و عددی ارائه داده‌اند.آن ها به این

نتیجه رسیده‌اند که شبکه‌های عمیق عملکرد بهتری برای داده‌هایی با حجم بالا به نسبت سایر روش‌های یادگیری ماشین ایجاد می‌کنند.

سؤال‌های پژوهش

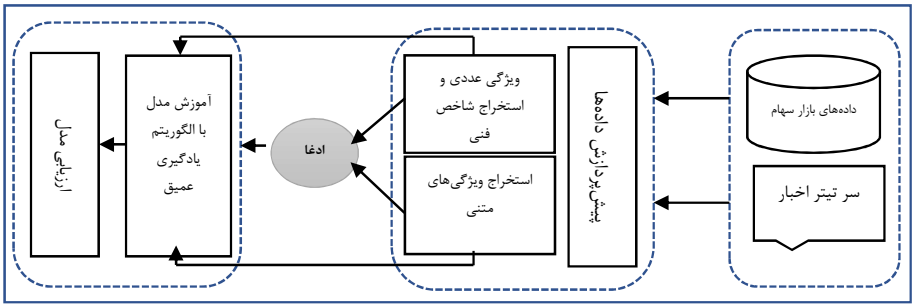
آیا استفاده از هر دو داده متنی و عددی عملکرد پیش‌بینی را بهبود می‌بخشد؟

آیا استفاده از شبکه LSTM باعث افزایش قدرت پیش‌بینی می‌شود؟

عملکرد مدل LSTM در مقایسه با مدل SVM و RNNوMLP چگونه است؟

روش‌شناسی

این بخش گام‌های اجراشده در چارچوب پیشنهادی برای پیش‌بینی سهام را توصیف می‌کند. شکل ۱ رویکرد این تحقیق را برای پیش‌بینی قیمت سهام با ادغام اطلاعات متنی و عددی با استفاده از شبکه عصبی عمیق (LSTM) نشان می‌دهد.



شکل ۱- مراحل اجرای تحقیق منبع: یافته‌های تحقیق

جمع‌آوری داده‌ها

این بخش به توصیف فرایند جمع‌آوری داده‌ها، منابع داده‌های جمع‌آوری‌شده و ساختار داده‌های جمع‌آوری‌شده می‌پردازد. در تحقیق حاضر از داده‌های قیمت شاخص[۲۲] Dow Jones در بازه زمانی ۲۰۰۸/۰۸/۱۱ تا ۲۰۱۶/۰۶/۲۸ بهره گرفته شده است، همچنین برای مجموعه داده متن‌کاوی از سر تیتر اخبار در بازه زمانی سال‌های ۲۰۰۰ تا ۲۰۱۶ استخراجی از کانال خبری Reddit استفاده شده است. جهت تعیین برچسب این داده‌ها نیز از داده‌های سری زمانی شاخص Dow Jones در همین بازه زمانی استفاده شده است.

متغیرهای عددی

داده‌های قیمت تاریخی سهام در سایت یاهو در دسترس هستند و داده‌های قیمت سهام برای دوره زمانی انتخابی در فرمت فایل csv جمع آوری شده است. قیمت‌ها شامل ۷ ویژگی می‌باشند که عبارت‌اند از تاریخ معاملات سهام، قیمت آزاد سهام، حداکثر قیمت معاملات سهام، حداقل قیمت معاملات سهام، قیمت بستن سهام، تعداد سهام مبادله شده و قیمت بستن سهام در صورت پرداخت سود سهام به سرمایه‌گذاران را نشان می‌دهد .

متغیرهای متنی

دومین داده مهم برای این تحقیق، داده‌های اخبار است. محققان از منابع مختلفی مانند Reuters [۲۳]، FINET[۲۴] و Reddit [۲۵] استفاده کرده‌اند. برای این تحقیق از کانال خبری Reddit برای جمع‌آوری اخبار استفاده شده است. فایل اخبار خام چهار ویژگی دارد که شامل منبع اخبار، لینک خبری، سر تیتر اخبار و تاریخ انتشار. استفاده از عناوین خبری کوتاه به‌جای مقالات خبری باعث منحصربه‌فرد شدن این مرحله می‌شود که استفاده از متن کوتاه را برای تجزیه‌وتحلیل احساسی اخبار ممکن می‌سازد .

نمایش اطلاعات متنی (پردازش زبان طبیعی)

پیش‌پردازش

داده‌های اخبار به‌صورت خام هستند و باید قبل از اعمال به الگوریتم‌های یادگیری ماشین پیش‌پردازش شوند. در این پژوهش در نخستین مرحله پیش‌پردازش سرتیترهای اخبار در کنار یکدیگر قرار گرفته است. ازآنجایی‌که همه کلمات حاوی اطلاعات در محتوای متنی نیستند کلمات ایست (stop words) که واژه‌هایی بااهمیت کم برای بازیابی اطلاعات هستند حذف شده است. سپس برای تبدیل داده‌های بدون ساختار در ویژگی‌ها به ویژگی‌های ساختاریافته، متن Tokenization شده است به این معنی که اسناد متنی به کلمات ساده با فضاهای خالی و نقطه‌گذاری مورداستفاده برای تشخیص و جداسازی کلمات تجزیه شده و در نهایت ریشه‌یابی واژه‌ها (Lemmatization) انجام می‌شود با استفاده از جایگزین کردن یک کلمه با ریشه آن.

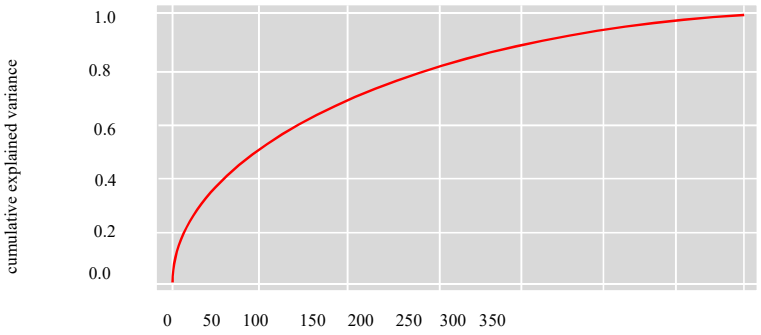
استخراج ویژگی‌های متنی

ویژگی‌های متن در این تحقیق ازسرتیتر اخبار استخراج می‌شوند و می‌توانند نگرش‌های

سرمایه‌گذاران نسبت به سهام، روند کلی نظرات عمومی و اطلاعات تصمیم‌گیری را تحت تأثیر قرار دهند. برای این منظور از روش کیسه کلمات استفاده شده است.

کیسه کلمات

یکی از رایج‌ترین نوع نمایش ویژگی در داده‌های متنی، کیسه لغات[۱۲] (BOW) است که ابتدا توسط هریس[۱۳] (۱۹۵۴) به آن اشاره شده است و هنوز تکنیک غالب است (ژای و ماسونگ[۱۴]، ۲۰۱۶). کیسه لغات اساساً یک ماتریس است که در آن هر سند به‌عنوان یک ردیف برداری و ویژگی‌های (معمولاً کلمات) به‌عنوان ستون‌های این ماتریس نشان داده شده است. ستون‌های این ماتریس نه‌تنها شامل شرایط موجود در نوشتار، بلکه تمام اصطلاحات موجود در پیکره زبانی است. مقادیر ویژگی را می‌توان به‌صورت باینری (وجود و عدم وجود یک ویژگی) و مقادیر عددی نشان داد. مقادیر عددی می‌توانند شامل هر عدد صحیح یا مقدار پیوسته استخراج‌شده از محتوای متنی یا برخی سنجش‌ها وزن دهی می‌شود. در این تحقیق پس از پیش‌پردازش داده‌های متنی با استفاده از تابع CountVectorizer کتابخانه Sklearn واژه‌های جملات هر داده به بردار ویژگی عددی تبدیل شده است. این تابع مدل BOW (Bag of Words) را که یکی از روش‌های معمول پردازش زبان طبیعی برای استخراج بردار ویژگی از متن است پیاده‌سازی می‌کند. خروجی این تابع یک ماتریس خلوت (Sparse Matrix) است. که نوع داده بهینه‌شده برای ماتریس‌هایی است که مقادیر غیر صفر آن‌ها بسیار کم است و تنها مقادیر غیر صفر را نگهداری می‌کند.به این دلیل که روش BOW معمولاً یک بردار ویژگی با بعد بزرگ تولید می‌کند(بردار ویژگی معمولاً صدها یا حتی هزاران بعد دارد). بااین‌حال، روش پیشنهادی تنها ۱۲بعد از ویژگی‌های مالی دارد. بعد نامتعادل دو نوع ویژگی منجر به دو مشکل جدی می‌شود. ازیک‌طرف، این عدم تعادل اهمیت ویژگی مالی در مدل پیش‌بینی را تضعیف خواهد کرد. از سوی دیگر، ابعاد بزرگ بردار ویژگی متن بر سرعت آموزش روش پیشنهادی تأثیر خواهد گذاشت، بنابراین بعد بردارهای ویژگی متن کاهش داده شده است. برای این منظور با بهره‌گیری از روش PCA (Principal Component Analysis) کاهش بعد انجام شده است.



شکل۲- توزیع تجمعی واریانس روش PCA منبع : محاسبات تحقیق

نمایش اطلاعات عددی

پیش پردازش

به دلیل اینکه داده‌های عددی در یک بازه مشخص قرار داشته باشند باید آن‌ها را نرمال کرد بدین معنی که مقادیر یک ویژگی بدون از دست رفتن اختلاف مقادیر به مقیاسی واحد نگاشت می‌شوند. بدین منظور از روش MIN-MAX بر اساس رابطه زیر استفاده شده تا تمامی مقادیر مجموعه داده در بازه (۰و۱) قرار بگیرند. در این رابطه x بردار ویژگی و x'بردار ویژگی نرمالیزه شده است.

$$x' = \frac{(x - \min(x))}{\max(x) - \min(x)} \qquad (۱)$$

شاخص های فنی

شاخص‌های فنی مالی به شاخص‌های محاسبه‌شده بر اساس داده‌های تجارت سهام، ازجملهRSI ، MACD وWilliams %R گفته می‌شود .شاخص‌هایی که در ادامه معرفی می‌شوند به دلیل پذیرش گسترده آن‌ها در مدل پیشنهادی مورداستفاده قرار گرفته‌اند. در این تحقیق به‌منظور استخراج شاخص‌های فنی با استفاده از داده‌های قیمتی از کتابخانه Stockstate از توابع کتابخانه‌ای keras با استفاده از زبان برنامه‌نویسی پایتون در محیط Anaconda استفاده شده است.

تغییرات پیوسته قیمت به بالا یا پایین در سه روز

در مرحله اول الگوریتم ۱، روند هر روز محاسبه می‌شود. بدین منظور قیمت امروز را از قیمت دیروز کم می‌شود، اگر مقدار به‌دست‌آمده عدد غیر منفی باشد، در این صورت روند مثبت و در غیر این صورت

روند منفی است. در مرحله بعد در الگوریتم ۲ پیوستگی مداوم موردبررسی قرار می‌گیرد. بالا و پایین بودن پیوسته با در نظر گرفتن روند سه روز گذشته محاسبه می‌شود

```
Step 1-trend calculation
Data: Close price vector (cV)
Result: Trend on each day(t)
cV = difference(cV);
Append 0 to cV;
j=0
for i in cV do
    if i> 0 then
        t[j] =1;
    else
        t[j]=0;
    end
    increment j;
end

Algorithm to Calculate Trend on Each Day
```

الگوریتم ۱-تشخیص روند قیمت

```
Step 2- identification of continuous five days up/down
Data: Date vector along with trend vector(d)
Result: Vector containing relationship between trend and volume
traded(t)
i=0;
j=0;
for i,i+1,i+2,i+3 in d do
    for j,j+1,j+2,j+3 in t do
        if j = = j+1 && j = = j+2 && j = = j+3 then
            cD=1;
        else
            cD = 0
        end
    end
end

Algorithm to check continuous days up/down
```

الگوریتم ۲- تشخیص روند قیمت متوالی

تغییرات حجمحجم معاملات در هر روز با روند در همان روز مقایسه می‌شود تا الگوی تغییر حجم را مانند الگوریتم ۳ به دست آورد .

```
Date: vector containing volume traded on each day (vV) and vector containing trend on each day (t)
Result: vector containing relationship between trend and volume of stock traded(vD)
vV= difference(volume);
append 0 to Vv ;
for i in Vv do
    for j in t do
        if i = = - 1 then
            vD =0;
        else if  i>i-1 && j= = 1 then
            if i> (2* (i-1) then
                vD = 1;
            else
                vD= 0.5;
            end
        else if i< i-1&& j = = 0 then
            if 2* I < (i-1) then
                vD=-1;
            else
                vD=-0.5;
            end
        else
            vD= -1
        end
    end
end
    Algorithm to Find the Relationship Between Trend and Volume of Stock Traded
```

الگوریتم۳- تشخیص رابطه روند قیمت و حجم

سه مورد دیگر از شاخص‌های فنی متداول در زمینه معاملات بورس RSI ، MACD و Williams %Rهستند. این‌ها ویژگی‌هایی است که به دلیل پذیرش گسترده آن‌ها می‌توانند در مدل عصبی مورداستفاده قرار می‌گیرند. در ادامه به‌طور خلاصه این شاخص‌ها توضیح داده شده است:

شاخص نسبی قدرت(RSI)

یک شاخص حرکت فنی است که قدرت تاریخی یا ضعف قیمت سهام را نشان می‌دهد .همچنین در یک بازه زمانی مشخص، ضرر و سود را به شرح زیر مقایسه می‌کند.

$$\mathrm{RSI} = 100\frac{100}{(1+\mathrm{RS})}$$

$$RS = \frac{AverageGain}{AverageLoss} \qquad (۲)$$

همگرایی یا واگرایی میانگین حرکت (Moving Average Convergence and Divergence)MACD

یک شاخص فنی است که روند قیمت سهام را نشان می‌دهد .برابر است با اختلاف میانگین ۱۲ روزه و ۲۶ روزه میانگین حرکت‌نمایی(EMA:Exponential Moving Average)

$$MACD = (12DaysEMA - 26DaysEMA) \qquad (۳)$$

شاخص Williams %R

شاخص فنی مبتنی بر گشتاور است که شرایط افراط در خرید یا فروش را نشان می‌دهد:

$$\%R = \frac{(HighestHigh - currentClose)}{(HighestHigh - LowestLow)}x - 100 \qquad (۴)$$

یادگیری ماشین

به‌منظور انجام یادگیری ماشین باید داده‌ها جمع‌آوری شود تا آن را به‌عنوان یک دانش پایه برای سیستم مورد استفاده قرار گیرد. بدین منظور از الگوریتم‌های یادگیری عمیق شبکه عصبی بازگشتی استفاده شده است.

شبکه عصبی بازگشتی(LSTM)

شبکه حافظه بلندمدت کوتاه‌مدت (LSTM) نسخه اصلاح‌شده شبکه‌های عصبی بازگشتی (RNN) هستند که وابستگی‌های بلندمدت در داده‌ها را به شیوه‌ای کارآمد به یاد می‌آورند. هسته یک شبکه یک

واحد حافظه (یا سلول) است که در شکل۳ نشان داده شده است. یک سلول از سه سیگموید و یک لایه تانژانت تشکیل شده که سه ورودی تشکیل می‌دهند که اطلاعات را در خارج و داخل سلول سازمان‌دهی می‌کند. ورودی و خروجی به ترتیب اطلاعات ورودی و خروجی را در واحد حافظه کنترل می‌کنند. دروازه فراموشی می‌تواند واحد حافظه را با یک تابع سیگموید بازگردانی کند.

با توجه به اطلاعات x_t جریان اطلاعات در یک سلول می‌تواند به‌صورت زیر فرموله بندی شود:

$$f_t = \sigma\left(w_f[h_{t-1}, x_t] + b_f\right) \quad (۱)$$

$$i_t = \sigma(w_i \cdot [h_{t-1}, x_t] + b_i) \quad (۲)$$

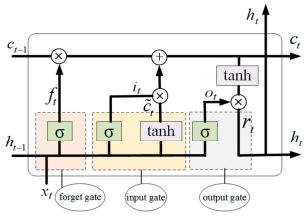
$$\tilde{c} = tanh(W_c \cdot [h_{t-1}, x_t] + b_c \quad (۳)$$

$$C_t = f_t * C_t - 1 + i_t * \tilde{C}_t \quad (۴)$$

$$o_t = \sigma(w_o[h_{t-1}, x_t] + b_o) \quad (۵)$$

$$h_t = o_{t^*}\, tanh(c_t) \quad (۶)$$

که در آن f_t، i_t و o_t به ترتیب دروازه‌های فراموشی،ورودی و خروجی را نشان را در زمان t نشان می‌دهند.C_t بردار حالت سلول است که در مرحله ۴ به‌روزرسانی می‌شود و h_tحالت پنهان را در زمان فعلی t نشان می‌دهد. $tanh$ تابع تانژانت هایپربولیک است که در ۴ و ۶ عملگر ضرب نقطه به نقطه را نشان می‌دهد. حالت سلولی جدید C_t که اطلاعات جدید را نشان می‌دهد، با میزان تصمیم‌گیری ما برای به‌روزرسانی هر ارزش حالت مقیاس بندی می‌شود.



شکل ۳-ساختار شبکه LSTM منبع : چن و همکاران (۲۰۱۸)

ادغام داده های متنی و عددی

برچسب‌گذاری داده‌های متنی

جهت تعیین برچسب داده‌های متنی از داده‌های سری زمانی شاخص Dow Jones در بازه زمانی

سال‌های ۲۰۰۰ تا ۲۰۱۶ استفاده شده است. بدین ترتیب که اگر قیمت روز جاری بیشتر از قیمت روز قبل باشد برچسب داده ۱ و در غیر این صورت برچسب داده ۰ خواهد بود.

بردار ورودی x_t از LSTM از ترکیب بردار گروه متن(خبر) p_t و بردار قیمت سهم n_t بدست می‌آید.بردار ورودی x_t از متصل کردن یا ادغام P_t و N_t ساخته می‌شود که $N_t\ P_t\ x_t$ به صورت زیر محاسبه می شود:

$$P_t = W_p p_t - bp$$
$$N_t = W_n n_t - b_n$$
$$\boldsymbol{x_t = [P_t N_t]}$$

ارزیابی عملکرد

عملکرد مدل ارائه شده با استفاده از معیار ارزیابی سنجیده می‌شود.ازاینجایی‌که مسئله یک مسئله طبقه‌بندی است از معیار نرخ دقت استفاده شده است.

نتایج پژوهش

تنظیم پارامترهای شبکه LSTM جهت انتخاب معماری بهینه

الگوریتم بهینه‌ساز Adam الگوریتمی محبوب در حوزه یادگیری عمیق به‌حساب می‌آید چراکه با استفاده از آن می‌توان خیلی سریع به نتایج بهینه و مطلوب دست پیدا کرد. در این تحقیق نیز از آن به‌عنوان الگوریتم بهینه‌ساز شبکه LSTM استفاده شده است. نرخ یادگیری بهینه‌ساز آدام برای هریک از وزن‌های شبکه حفظ می‌شود و این نرخ با شروع فرایند یادگیری به‌صورت جداگانه تطبیق داده می‌شود. از تابع زیان آنتروپی متقاطع دودویی نیز استفاده شده است. Batch-size یا همان اندازه ورودی و Epoch یا همان تعداد دوره‌ها برای شبکه LSTM بر روی ۱۰ و حداکثر تعداد دوره‌های آموزشی ۶۰ تنظیم شده است. زمانی که زیان داده‌های آموزشی را نمی‌توان بهینه کرد یا پس از چند دوره تکرار بزرگ‌تر می‌شود، آموزش بعدی مدل ضروری نیست. مکانیزم Early-Stop می‌تواند فرایند آموزش را متوقف کند و زمان آموزش شبکه عصبی را ذخیره کند. این مکانیزم در این تحقیق برای داده‌های آموزشی اعمال‌شده که به‌طور خودکار آموزش را متوقف می‌کند زمانی که زیان دیگر در ۲۰ دوره کاهش نمی‌یابد و مدلی ذخیره می‌شود که کمترین زیان را دارد. همچنین ازآنجایی‌که مدل‌های شبکه عصبی عمیق در معرض بیش برازش هستند چون تعداد لایه‌های اضافه‌شده این امکان را ایجاد می‌کند که وابستگی نایاب

در داده‌های آموزشی مدل ایجاد شود. در این تحقیق از روش DropOut استفاده شده است تا مشکل بیش برازش حل شود. همچنین جهت انتخاب معماری بهینه LSTM مدل با پارامترهای مختلف اجرا شده و در نهایت بهترین نتایج حاصل از معماری مقایسه گردیده است بنابراین مدل پیشنهادی حاصل از اجرای الگوریتم با پارامترهای متفاوت است. از ۲ پارامتر تعداد نورون‌های متفاوت در لایه‌های پنهان با اندازه ۶۴ ، ۱۲۸ و ۲۵۶ و از دو تابع فعال‌ساز Tanh و Relu استفاده شده است که نتایج حاصل از آن در جدول ۱ آورده شده است.

جدول ۱- ارزیابی شبکه LSTM با پارامترهای مختلف

پارامتر	شبکه LSTM با تابع فعال‌سازی Relu		
تعداد نورون لایه مخفی	۶۴	۱۲۸	۲۵۶
روش سنجش خطا (دقت)	٪۶۸,۶۵	٪۶۹,۱۹	٪۶۶,۴۹
پارامتر	شبکه LSTM با تابع فعال‌سازی Tanh		
تعداد نورون لایه مخفی	۶۴	۱۲۸	۲۵۶
روش سنجش خطا (دقت)	٪۶۴,۳۲	٪۶۴,۰۲	٪۶۰,۸۹

منبع: محاسبات تحقیق

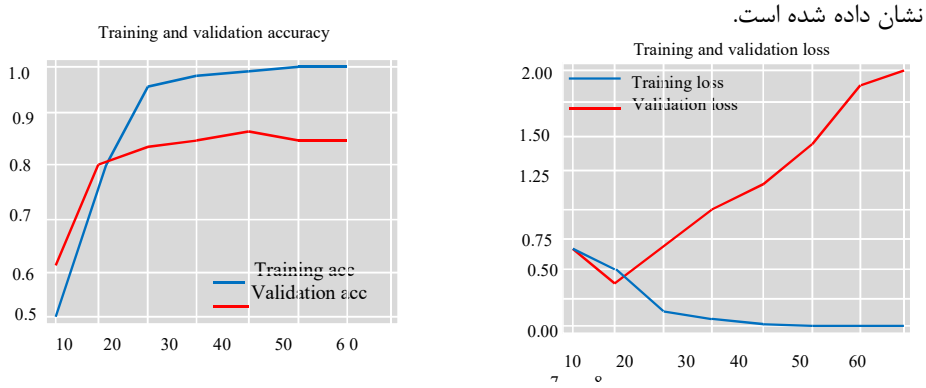
همان‌طور که در جدول ۱ مشخص است بهترین نتیجه برای تعداد نورون ۱۲۸ و تابع فعال‌ساز Reluاست.

مجموعه داده‌ها بر اساس تاریخ آن‌ها به ۲ بخش تقسیم می‌شود داده‌های عددی از تاریخ ۲۰۰۸/۰۸/۱۱ تا۲۰۱۵/۰۸/۲۸ حدود ۹۰٪ از کل داده‌ها به‌عنوان مجموعه آموزشی استفاده می‌شود و از تاریخ ۲۰۱۵/۰۸/۲۹ تا ۲۰۱۶/۰۶/۲۸ حدود ۱۰٪ از داده‌ها برای آزمون استفاده خواهد شد و برای داده های متنی ازتاریخ ۲۰۰۰ تا ۲۰۱۵ حدود ۹۰٪ برای داده‌های آزمایش و از تاریخ ۲۰۱۵ تا ۲۰۱۶ برای داده‌های آزمون استفاده می‌شود .اعتبار سنجی متقابل بر روی مجموعه آموزشی برای انتخاب بهینه مدل انجام می‌شود که در این بخش برای تقسیم داده‌ها به دو مجموعه آموزش و اعتبارسنجی از روش تصادفی Hold upاستفاده شده است. مدل برای سه دوره آموزش داده شده و سپس میانگین همه صحت‌ها محاسبه و از آن به‌عنوان صحت مدل استفاده می‌شود. در این تحقیق به منظور اثر بخشی ادغام داده های عددی و متنی دو روش استفاده شده است اولین آزمایش تنها از داده های عددی به عنوان ورودی مدل LSTM در نظر گرفته شده است و دومین آزمایش تنها از اطلاعات متنی (سر تیتر خبر) استفاده

می‌شود. و همچنین برای ارزیابی، مدلLSTM از سه شبکه SVM و MLP و RNNکه مدل‌های پرکاربردی در شبکه عصبی هستند به‌منظور ارزیابی مدل مورد مقایسه قرار می‌گیرند.

نتایج اجرا در فاز آزمایش

امتیازات دقت در اعتبار سنجی متقابل برای شبکه LSTM که از هر دو داده عددی و اخبار، LSTM برای فقط داده‌های عددی و فقط داده‌های متنی و همچنین شبکه‌های SVM و MLP و RNN در جدول ۲ نشان داده شده است. همچنین نشان دادن برخی از منحنی‌های اتلاف و دقت LSTM به‌منظور چگونگی تعمیم مدل به داده‌های اعتبار سنجی با افزایش تعداد دوره‌ها ارزشمند است که در شکل ۴ نشان داده شده است.



شکل ۴-دقت و هدر رفت شبکه LSTM با ترکیب ویژگی‌های متن و قیمت منبع: محاسبات تحقیق

شکل ۴- فرایند آموزش و اعتبار سنجی متقابل LSTM با استفاده از هر دو داده متن و عدد را نشان می‌دهد. زیر گراف سمت راست منحنی‌های اتلاف را در مجموعه آموزشی و اعتبار سنجی نشان می‌دهد. می‌توان اشاره کرد که اتلاف مجموعه اعتبار سنجی بعد از ۲۰ دوره کاهش و به حداقل خود می‌رسد اما به‌شدت در تکرارهای آموزشی بعدی افزایش می‌یابد. به‌منظور دستابی به بهترین عملکرد برای داده‌های تست فرایند آموزش در ۲۰ امین دوره متوقف می‌شود زیر گراف سمت چپ منحنی دقت را در مجموعه آموزشی و اعتبار سنجی نشان می‌دهد که در آن دقت مجموعه اعتبار سنجی افزایش اما پس از چند دوره کاهش می‌یابد.

جدول ۲- دقت مدل‌ها در مجموعه اعتبار سنجی

پارامتر	RNN	SVM	MLP	قیمتLSTM	خبرLSTM	قیمت و خبرLSTM
دقت دسته‌بندی	٪۶۵	٪۵۱	٪۵۷	٪۵۳	٪۷۱	٪۸۴

منبع:محاسبات تحقیق

توجه به جدول۲ و نتایج به‌دست‌آمده از آن می‌توان بیان کرد، که $LSTM_{\text{قیمت و خبر}}$ که هر دو داده متنی و عددی در نظر گرفته عملکرد مناسب‌تری را در اعتبار سنجی متقابل نسب به سایر مدل‌ها داشته است.

نتایج اجرا در فاز آزمون

اثربخشی ادغام داده های عددی و متنی

در این تحقیق به منظور اثربخشی استفاده از اطلاعات متنی و عددی به صورت همزمان دو روش استفاده شده است.اولین آزمایش تنها از ازطلاعات متنی (سرتیتر خبر)به عنوان ورودی به مدل LSTM استفاده شده است و دومین آزمایش تنها از اطلاعات عددی(قیمت و شاخص فنی) به عنوان ورودی به مدل LSTM استفاده کرده است.نتایج در جدول ۳ نشان داده شده است.

جدول ۳ -دقت مدل‌ها برای فاز آزمون برای اثربخشی همزمان داده های قیمت و خبر

روش	قیمتLSTM	خبرLSTM	قیمت و خبرLSTM
دقت داده‌های تستی	%۵۱/۸۹	%۶۵/۶۲	%۶۹/۱۹

منبع:محاسبات تحقیق

می‌توان مشاهده کرد $LSTM_{\text{قیمت و خبر}}$ که از هر دو داده عدد و متن استفاده کرده است عملکرد بهتری را با دقت ۶۹,۱۹٪ داشته است بعد از آن مدل $LSTM_{\text{خبر}}$ با دقت ۶۵,۶۲٪ که فقط از داده‌های متنی استفاده کرده است و مدل $LSTM_{\text{عدد}}$ با دقت ۵۱,۸۹٪ قرار گرفته است.

اثربخشی مدل LSTM

برای ارزیابی شبکه LSTM پیشنهادی در این تحقیق عملکرد مدل با 3 طبقه‌بندی کننده پایه ارزیابی شده است .اولین طبقه‌بندی کننده پایه رگرسیون‌بردارپشتیبان(SVM) با تابع شعاعی(RBF)است.دومین طبقه‌بندی کننده خط پایه شبکه پرسپترون چند لایه (MLP) با دولایه مخفی و ۲۵۰ نورون در هرلایه.

و سومین مدل شبکه بازگشتی ساده (Simple RNN) برای ارزیابی درنظر گرفته شده است. نتایج در جدول ۴ نشان داده شده است.

جدول ۴- مقایسه دقت مدل‌ها در فازآزمون برای اثربخشی شبکه LSTM

روش	Simple RNN	SVM	MLP	LSTM
دقت داده‌های تستی	%۶۱/۸۹	%۴۸/۲۴	%۵۰/۷۸	%۶۹/۱۹

منبع:محاسبات تحقیق

همان‌طور که مشخص است مدل پیشنهاد شده این تحقیق دارای بهترین عملکرد با میزان دقت ۶۹,۱۹٪ است.بعد از آن مدل RNN با میزان دقت ۶۱,۸۹٪ و شبکه MLP با دقت ۵۰,۷۸٪ و SVM با دقت ۴۸,۲۴٪ قرارگرفته است.

جدول ۵ -=مقایسه دقت مدل‌ها برای همه نوع داده

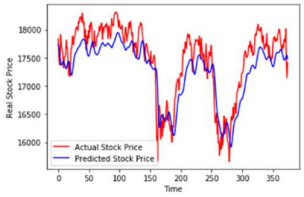
پارامتر	RNN	SVM	MLP	قیمتLSTM	خبرLSTM	قیمت و خبرLSTM
دقت دسته‌بندی	%۶۱/۸۹	%۵۲/۲۴	%۵۵/۷۸	%۵۱/۸۹	%۶۵/۶۲	%۶۹/۱۹

منبع : محاسبات تحقیق

بعد از به دست آمدن همه نتایج بر اساس جدول ۵همان‌طور که مشخص است مدل LSTM $_{\text{قیمت و خبر}}$ که از هر دو داده عددی و متنی استفاده کرده است دارای بهترین عملکرد نسبت به سایر مدل‌ها با میزان دقت ۶۹,۱۹٪ است. مدل LSTM که از داده‌های متنی استفاده کرده است به نسبت مدل RNN و MLP و SVM عملکرد بهتری با دقت ۶۵,۶۲٪ دارد که نشان‌دهنده عملکرد خوب مدل LSTM در ارزیابی سری‌زمانی که از رویدادهای خبری استفاده کرده است.اما عملکرد مدل LSTM که فقط از داده‌های عددی استفاده کرده در مقایسه با مدل RNN عملکرد ضعیف‌تری دارد ولی به نسبت عملکرد دو شبکه SVM و MLP توانسته است نتایج بهتری را در پیش‌بینی کسب کند.

ازآنجایی‌که مدل LSTM $_{\text{قیمت و خبر}}$ دارای بهترین عملکرد هم در فاز آزمایش و هم در فاز آزمون (بر اساس جدول ۱ و جدول ۲) در بین سایر مدل‌ها را دارد؛ بنابراین به‌عنوان مدل منتخب در نظر گرفته شده است که نشان از برازش مناسب مدل دارد. در ادامه به‌منظور بررسی رابطه بین مقدار پیش‌بینی‌شده و مقدار واقعی بسته شدن شاخص داوجونز از ضریب همبستگی پیرسون استفاده شده است که مقدار آن ۰,۹۰۲ و سطح معناداری ۰,۰۰ است و ازآنجایی‌که سطح معناداری مدل منتخب کمتر از ۰,۰۵ وجود

همبستگی بین مقدار واقعی و مقدار پیش‌بینی ازلحاظ آماری مورد تائید است. شکل ۵ نمودار قیمت واقعی و پیش‌بینی‌شده مدل LSTM قیمت و خبر برای قیمت بسته شدن سهم در داده‌های آزمون را نشان می‌دهد.



شکل ۵- مقدار واقعی و پیش‌بینی بسته شدن سهم برای قیمت و خبرLSTMدر فاز آزمون را نشان می‌دهد.

بحث و نتیجه‌گیری

این تحقیق یک مدل یادگیری عمیق که پیش‌بینی روزانه حرکت قیمت سهام را با استفاده از ترکیب قیمت، شاخص‌های فنی و عناوین خبری می‌باشد، ارائه می‌دهد. برای این منظور یک سیستم پیش‌بینی سهام ایجاد و قیمت‌های تاریخی و شاخص‌های فنی و عناوین خبری به‌عنوان ورودی به شبکه حافظه کوتاه‌مدت ماندگار(LSTM) داده شده است. برای ارزیابی اثربخشی ادغام داده‌های عددی و متنی مدل با دو مجموعه داده فقط داده‌های متنی که شامل سرتیتر اخبار و فقط داده‌های عددی که شامل قیمت‌ها و شاخص‌های فنی استخراج شده است نیز آموزش داده شده است. همچنین برای ارزیابی اثربخشی مدل LSTM از سه شبکه پایه SVM و MLP وRNN استفاده شده است. نتایج نشان می‌دهد: در مدل‌هایی که از هر دو منبع متن و عدد استفاده شده است عملکرد شبکه LSTM خبر و قیمت با میزان دقت(٪۶۹,۱۹) عملکرد بهتری را به نسبت مدل LSTM که فقط از داده‌های متنی با دقت ٪۶۵,۶۲و LSTM که فقط از داده‌های عددی با دقت ٪۵۱,۸۹داشته است. نتایج ارزیابی برای اثربخشی شبکه LSTM نشان داد که مدل پیشنهادی عملکرد بهتری نسبت به مدل MLP با دقت ٪۵۵,۷۸و مدل SVM با دقت٪۵۲,۲۴ و مدل RNN با دقت ٪۶۱,۸۹داشته است.که این نشان‌دهنده عملکرد خوب شبکه LSTM در پیش‌بینی سری‌های زمانی که از هر دو منبع استفاده می‌کنند دارد.برای کارهای آینده پیشنهاد می‌گردد تغییر در نحوه‌ی جمع‌آوری اخبار (استفاده از متن کل خبر به‌جای سر تیترهای خبری)و روش‌های دیگری که

متن را استخراج می‌کنند و همچنین استفاده از روش‌های دیگر کاهش بعد که کمک شایانی برای تعادل بین ویژگی‌های متن و خبر است تا بتوان پیش‌بینی دقیق‌تری صورت پذیرد

منابع

1) Akita, Ryo, et al. (2016). "Deep learning for stock prediction using numerical and textual information.IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS).
2) Beck, Thorsten, and Ross Levine. (2004)."Stock markets, banks, and growth: Panel evidence." Journal of Banking & Finance 28.3: 423-442.
3) Chan, Samuel WK, and James Franklin. (2011)"A text-based decision support system for financial sequence prediction." Decision Support Systems 52.1: 189-198.
4) Chen, Deli, et al. (2019)"Incorporating fine-grained events in stock movement prediction." arXiv preprint arXiv:1910.05078.
5) Chen, K., Zhou, Y., & Dai, F. (2015, October). A LSTM-based method for stock returns prediction: A case study of China stock market. In 2015 IEEE international conference on big data (big data) (pp. 2823-2824). IEEE.
6) Chen, Yingxuan, Weiwei Lin, and James Z. Wang. (2019). "A dual-attention-based stock price trend prediction model with dual features." IEEE Access 7 ,148047-148058.
7) Deng, Shumin, et al. (2019). "Knowledge-driven stock trend prediction and explanation via temporal convolutional network." Companion Proceedings of The World Wide Web Conference.
8) Ding, Xiao, et al. (2016). "Knowledge-driven event embedding for stock prediction." Proceedings of coling, the 26th international conference on computational linguistics: Technical papers.
9) Fischer, T., & Krauss, C. (2018). Deep learning with long short-term memory networks for financial market predictions. European journal of operational research, 270(2), 654-669.
10) Hao, Y., & Gao, Q. (2020). Predicting the trend of stock market index using the hybrid neural network based on multiple time scale feature learning. Applied Sciences, 10(11), 3961.
11) Hu, Z., Liu, W., Bian, J., Liu, X., & Liu, T. Y. (2018, February). Listening to chaotic whispers: A deep learning framework for news-oriented stock trend prediction. In Proceedings of the eleventh ACM international conference on web search and data mining (pp. 261-269).

12) Huang, Yuxuan, Luiz Fernando Capretz, and Danny Ho. (2019)."Neural network models for stock selection based on fundamental analysis." IEEE Canadian Conference of Electrical and Computer Engineering (CCECE).
13) Ji, X., Wang, J., & Yan, Z. (2021). A stock price prediction method based on deep learning technology. International Journal of Crowd Science.
14) Khan, Wasiat, et al. (2020). "Stock market prediction using machine learning classifiers and social media, news." Journal of Ambient Intelligence and Humanized Computing, 1-24.
15) Li, X., Cao, J., & Pan, Z. (2019). Market impact analysis via deep learned architectures. Neural Computing and Applications, 31(10), 5989-6000.
16) Li, Yang, and Yi Pan. (2021). "A novel ensemble deep learning model for stock prediction based on stock prices and news." International Journal of Data Science and Analytics ,1-11.
17) Ma, Y., Zong, L., & Wang, P. (2020). A novel distributed representation of news (drnews) for stock market predictions. arXiv preprint arXiv:2005.11706.
18) Nassirtoussi, Arman Khadjeh, et al. (2015)."Text mining of news-headlines for FOREX market prediction: A Multi-layer Dimension Reduction Algorithm with semantics and sentiment." Expert Systems with Applications 42.1: 306-324.
19) Nelson, David MQ, Adriano CM Pereira, and Renato A. De Oliveira. (2017.) "Stock market's price movement prediction with LSTM neural networks." International joint conference on neural networks (IJCNN).
20) Nikou, M., Mansourfar, G., & Bagherzadeh, J. (2019). Stock price prediction using DEEP learning algorithm and its comparison with machine learning algorithms. Intelligent Systems in Accounting, Finance and Management, 26(4), 164-174.
21) Usmani, Shazia, and Jawwad A. Shamsi. (2021)."News sensitive stock market prediction: literature review and suggestions." PeerJ Computer Science 7: e490.
22) Vargas, Manuel R., Beatriz SLP De Lima, and Alexandre G. Evsukoff. (2017)."Deep learning for stock market prediction from financial news articles." international conference on computational intelligence and virtual environments for measurement systems and applications.
23) Xu, Y., Chhim, L., Zheng, B., & Nojima, Y. (2020, July). Stacked deep learning structure with bidirectional long-short term memory for stock market prediction. In International Conference on Neural Computing for Advanced Applications (pp. 447-460). Springer, Singapore.
24) Yu, Pengfei, and Xuesong Yan. (2020)."Stock price prediction based on deep neural networks." Neural Computing and Applications 32.6: 1609-1628.

25) Zhai, ChengXiang, and Sean Massung.(2016). Text data management and analysis: a practical introduction to information retrieval and text mining. Morgan & Claypool,.
26) Zhang, X., & Tan, Y. (2018, June). Deep stock ranker: A LSTM neural network model for stock selection. In International conference on data mining and big data (pp. 614-623). Springer, Cham.

یادداشت‌ها:

1 Beck&levine

1 Dang
1 Huang
1 Bag of Words
1 Hariss
1 Zhai&massung
2 Khan et.al
3 Li&Yi
4 Chen et.al
5 Usmani &Jawwad
6 Chan&james
7 Nasirtoussi
8 Yu&Yan
9 Huang

10 Nelson et al

11 Akita
12 Bag of Words
13 Hariss
14 Zhai&massung