دانشکده مهندسی کامپیوتر

بررسی ویژگی‌های متنی در حوزه خلقیات و اثر آن در بازارهای مالی با استفاده از روش‌های پردازش زبان‌های طبیعی

رشته مهندسی مهندسی کامپیوتر

نام دانشجو

علی سلطانی

استاد راهنما:

دکتر رضا انتظاری ملکی

شهریور ماه ۱۴۰۳


بسم الله الرحمن الرحيم

تأییدیه هیأت داوران جلسه دفاع از پایان نامه/رساله

نام دانشکده: دانشکده مهندسی کامپیوتر

نام دانشجو: علی سلطانی

عنوان پایان نامه یا رساله: بررسی ویژگی‌های متنی در حوزه خلقیات و اثر آن در بازارهای مالی با استفاده از روش‌های پردازش زبان‌های طبیعی

تاریخ دفاع:

رشته: مهندسی کامپیوتر

ردیف	سمت	نام و نام خانوادگی	مرتبه دانشگاهی	دانشگاه یا مؤسسه	امضا
1	استاد راهنما	دکتررضا انتظاری ملکی	استادیار	علم و صنعت ایران	
2	استاد مدعو داخلی	دکتر مرضیه ملکی مجد	استادیار	علم و صنعت ایران	

تأییدیه‌ی صحت و اصالت نتایج

باسمه تعالی

اینجانب علی سلطانی به شماره دانشجویی ۹۹۵۲۱۳۴۳ دانشجوی رشته مهندسی کامپیوتر مقطع تحصیلی کارشناسی تأیید مینمایم که کلیه ی نتایج این پایان نامه/رساله حاصل کار اینجانب و بدون هر گونه دخالت و تصرف است و موارد نسخه برداری شده از آثار دیگران را با ذکر کامل مشخصات منبع ذکر کرده ام.
در صورت اثبات خلاف مندرجات فوق، به تشخیص دانشگاه مطابق با ضوابط و مقررات حاکم (قانون حمایت از حقوق مؤلفان و مصنفان و قانون ترجمه و تکثیر کتب و نشریات و آثار صوتی، ضوابط و مقررات آموزشی، پژوهشی وانضباطی ...) با اینجانب رفتار خواهد شد و حق هر گونه اعتراض در خصوص احقاق حقوق مکتسب و تشخیص و تعیین تخلف و مجازات را از خویش سلب مینمایم. در ضمن، مسؤولیت هر گونه پاسخگویی به اشخاص اعم از حقیقی و حقوقی و مراجع ذیصلاح (اعم از اداری و قضایی) به عهده ی اینجانب خواهد بود و دانشگاه هیچ گونه مسؤولیتی در این خصوص نخواهد داشت.

نام و نام خانوادگی: علی سلطانی

امضا و تاریخ: ۱۳ / ۶ / ۱۴۰۳

مجوز بهره برداری از پایان نامه

بهره برداری از این پایان نامه در چهارچوب مقررات کتابخانه و با توجه به محدودیتی که توسط اساتید راهنما به شرح زیر تعیین میشود، بلامانع است:

- بهره برداری از این پایان نامه/ رساله برای همگان بلامانع است.
- بهره برداری از این پایان نامه/ رساله با اخذ مجوز از استاد راهنما، بلامانع است.
- بهره برداری از این پایان نامه/ رساله تا تاریخ ..... ممنوع است.

نام استاد یا اساتید راهنما:

تاریخ:

امضا:

چکیده

در عصر پیشرفت‌های سریع تکنولوژیکی، پیش‌بینی بازارهای مالی با سرعتی بی‌سابقه در حال تحول است. روش‌های سنتی که اغلب بر تحلیل‌های تک‌بعدی مانند تحلیل احساسات تمرکز دارند، دیگر برای فهم پیچیدگی‌های بازارهای مدرن کافی نیستند. این پژوهش با بهره‌گیری از مدل‌های پیشرفته زبانی و اطلاعات متنی، دقت پیش‌بینی‌ها را افزایش می‌دهد.

در این مطالعه از تحلیل احساسات، به‌ویژه تحلیل احساسات استخراج شده از توییت‌ها، برای پیش‌بینی روندهای بازار استفاده شده است. احساسات موجود در توییت‌ها، که شامل احساسات مثبت، منفی و خنثی است، به عنوان عاملی تاثیرگذار بر حرکت بازارها در نظر گرفته شده‌اند. به کمک این تحلیل‌ها، سیستم قادر است روندهای کوتاه‌مدت بازار را با توجه به تغییرات لحظه‌ای احساسات کاربران تشخیص دهد. به عنوان مثال، توییت‌هایی با احساسات منفی معمولاً منجر به پیش‌بینی روند نزولی در بازار می‌شوند، در حالی که احساسات مثبت ممکن است نشانه‌ای از روند صعودی باشند.

در کنار این، مدل‌های پیشرفته یادگیری ماشین مانند شبکه‌های عصبی LSTM و اتوانکودرها نیز به کار گرفته شده‌اند. اتوانکودر ویژگی‌های پیچیده‌ای از داده‌ها را استخراج کرده و آن‌ها را به عنوان ورودی به مدل LSTM می‌دهد تا پیش‌بینی‌های دقیق‌تری انجام شود. این ترکیب از تحلیل احساسات توییت‌ها و پردازش دنباله‌ای مدل‌هایLSTM ، منجر به بهبود قابل توجهی در دقت پیش‌بینی‌ها شده است.

نتایج نشان می‌دهد که تحلیل احساسات و استفاده از توییت‌ها به همراه مدل‌های یادگیری عمیق، به طور مؤثری پیش‌بینی حرکات بازار را بهبود می‌بخشد. این روش، که هم احساسات بازار و هم ویژگی‌های پیچیده‌تر داده‌ها را در نظر می‌گیرد، نمایانگر گامی مهم در به‌کارگیری یادگیری ماشین و پردازش زبان طبیعی در حوزه مالی است.

واژه های کلیدی: پردازش زبان های طبیعی- تحلیل فضای مجازی- پیش بینی بازار مالی- تحلیل احساسات

فهرست مطالب

فهرست جدول‌ها

فهرست شکل ها

فصل ۱ : مقدمه

۱ – ۱ مقدمه

پیش‌بینی بازارهای مالی همواره موضوعی مهم و چالش‌برانگیز بوده است. نظریات سنتی مانند فرضیه بازار کارا[۱] بیان می‌کنند که بازارها به طور کامل کارا هستند و قیمت‌ها تمامی اطلاعات موجود را در خود منعکس می‌کنند. با این حال، فرضیه بازار تطبیقی[۲] این دیدگاه را به چالش کشیده و معتقد است که بازارها در طول زمان تکامل پیدا کرده و تحت تأثیر رفتار و روانشناسی فعالان بازار قرار می‌گیرند. این تغییر به سمت درک بهتر مالی رفتاری مسیرهای جدیدی برای پیش‌بینی بازارهای مالی باز کرده و بر اهمیت رفتار انسانی و احساسات در حرکت بازارها تأکید دارد. [۳۴] [۳۵]

یکی از روش‌های امیدوارکننده برای درک این جنبه‌های رفتاری استفاده از تحلیل متنی با پردازش زبان طبیعی (NLP)[۳]است. تکنیک‌های NLP در وظایف مختلفی مانند مدل‌سازی موضوعی، تحلیل رویداد محور اخبار، تحلیل احساسات عمومی در شبکه‌های اجتماعی، و تحلیل احساسات خبرها و پست‌های شبکه‌های اجتماعی برای پیش‌بینی روند بازارها به کار رفته‌اند. محققان از منابع متنوعی مانند اوراق سفید ارزهای دیجیتال و گزارش‌های مالی برای استخراج داده‌های ارزشمند استفاده کرده‌اند. اخیراً ظهور مدل‌های زبانی بزرگ مانندGPT و مدل‌های خاص حوزه مالی مانند FinBERT و CryptoBERT تحول عظیمی در تحلیل متنی برای پیش‌بینی بازارهای مالی ایجاد کرده است که امکان پیش‌بینی‌های دقیق‌تر و پیچیده‌تری را فراهم می‌کند. [۳][۹]

تحلیل شبکه‌های اجتماعی به ویژه در سال‌های اخیر پیشرفت‌های زیادی داشته است. مطالعات نشان داده‌اند که هویت نویسنده می‌تواند عاملی مهم در دسته‌بندی توییت‌ها باشد. [۱۱][۱۰]
افراد تأثیرگذاری مانند ایلان ماسک و دونالد ترامپ تأثیر قابل توجهی بر بازارهای مالی دارند. همچنین، تحقیقات به بررسی رابطه علّی میان انتشار اطلاعات در پلتفرم‌هایی مانند توییتر و نوسانات بازار پرداخته‌اند، به‌ویژه در زمان‌هایی مانند همه‌گیری کرونا. علاوه بر این، استفاده از ویژگی‌هایی مانند بازتوییت‌ها، لایک‌ها و دیگر معیارهای تعاملی، به تحلیل‌های غنی‌تری از داده‌های شبکه‌های اجتماعی کمک کرده است. [۵][۴]

---

[۱] Efficient Market Hypothesis
[۲] Adaptive Market Hypothesis
[۳] Natural Language programming

بر اساس این پیشرفت‌ها، هدف از کار ما بهبود قدرت پیش‌بینی مدل‌های زبانی برای پیش‌بینی‌های کوتاه‌مدت بازار است. ما رویکردی نوین معرفی می‌کنیم که با تعیین مرزهای زمانی و حد سود/زیان، توییت‌ها را بر اساس حرکت‌های واقعی بازار طبقه‌بندی می‌کند. این روش به ما امکان می‌دهد که داده‌های ساختارمند و معنادارتری جمع‌آوری کنیم. [۲۵][۲]

در نهایت، این مطالعه از آخرین پیشرفت‌ها در زمینه NLPو تحلیل شبکه‌های اجتماعی استفاده می‌کند تا چارچوبی قوی برای پیش‌بینی بازارهای مالی ارائه دهد. با ادغام اطلاعات زمینه‌ای و به کارگیری تکنیک‌های نوآورانه مهندسی پرامپت، هدف ما ارائه تحلیل‌های دقیق‌تر و معنادارتر از محتوای مالی شبکه‌های اجتماعی است که در نهایت باعث تقویت قدرت پیش‌بینی مدل‌های زبانی در حوزه مالی می‌شود. [۲۲] [۱۹]

فصل ۲ : مروری بر منابع

۲ – ۱ مقدمه

تحقیقات در زمینه تحلیل احساسات بازارهای مالی و پیش‌بینی قیمت‌ها با استفاده از داده‌های شبکه‌های اجتماعی و داده‌های قیمتی در سال‌های اخیر توجه زیادی را به خود جلب کرده است. با افزایش داده‌های متنی مرتبط با بازارهای مالی در پلتفرم‌های شبکه‌های اجتماعی، تلاش‌های بسیاری برای به‌کارگیری روش‌های یادگیری ماشین و مدل‌های پیشرفته برای تحلیل و پیش‌بینی حرکات بازار صورت گرفته است. تحقیقات متعددی از تکنیک‌های مختلفی مانند پردازش زبان طبیعی ، شبکه‌های عصبی، و مدل‌های احتمالی برای استخراج اطلاعات ارزشمند از این داده‌ها استفاده کرده‌اند. این بخش به مرور مهم‌ترین پژوهش‌ها و روش‌های استفاده‌شده در این زمینه می‌پردازد و نقش آن‌ها را در پیشرفت روش‌های پیش‌بینی بازارهای مالی با تأکید بر تحلیل احساسات و روش‌های برچسب‌گذاری بررسی می‌کند. [۶][۲۲][۱۹]

در سال‌های اخیر، استفاده از پردازش زبان طبیعی (NLP) در حوزه مالی به یکی از ابزارهای کلیدی برای تحلیل داده‌های متنی و پیش‌بینی حرکات بازارهای مالی تبدیل شده است. با افزایش حجم اطلاعات متنی موجود در پلتفرم‌های شبکه‌های اجتماعی و اخبار اقتصادی، محققان به دنبال توسعه مدل‌هایی بوده‌اند که بتوانند از این داده‌ها برای بهبود دقت پیش‌بینی‌ها و تحلیل‌ها استفاده کنند. [۱۲][۶][۲۳]

۲ – ۲ پیشینه استفاده از NLP در پیش‌بینی‌های مالی

تحقیقات انجام‌شده توسط *Xing et al. (۲۰۱۸)* به‌عنوان یکی از جامع‌ترین پژوهش‌های مروری در زمینه استفاده از تکنیک‌های NLP برای پیش‌بینی‌های مالی شناخته می‌شود. در این مقاله، روش‌های مختلفی که برای تحلیل متون مالی به‌کار گرفته شده‌اند، بررسی می‌شود و نتایج آن‌ها نشان می‌دهد که استفاده از این تکنیک‌ها می‌تواند به بهبود پیش‌بینی‌های مالی کمک شایانی کند. آن‌ها نشان دادند که مدل‌های مبتنی بر پردازش زبان طبیعی قادرند اطلاعات پنهان و پیچیده‌ای را از متون استخراج کرده و به مدل‌های پیش‌بینی وارد کنند که بهبود دقت پیش‌بینی‌ها را به دنبال دارد. [۱]

علاوه بر این، *Si et al. (۲۰۱۳)* نشان دادند که تجزیه و تحلیل احساسات مرتبط با موضوعات مختلف در شبکه‌های اجتماعی، به‌ویژه توییتر، می‌تواند نقش مهمی در پیش‌بینی نوسانات سهام ایفا کند. آن‌ها به‌طور خاص از تحلیل احساسات توییت‌ها برای پیش‌بینی قیمت سهام استفاده کردند و بهبود چشمگیری در نتایج

پیش‌بینی‌های خود مشاهده کردند. این تحقیق نشان داد که احساسات عمومی و واکنش‌های کاربران شبکه‌های اجتماعی به اخبار اقتصادی و مالی می‌تواند به‌عنوان یک شاخص مهم در پیش‌بینی قیمت سهام و بازارهای مالی استفاده شود. [۲]

۲ – ۲ – ۱ استفاده از یادگیری ماشین و تحلیل احساسات در پیش‌بینی‌های مالی

تحقیقات بسیاری نیز به استفاده از روش‌های یادگیری ماشین در ترکیب با تکنیک‌های پردازش زبان طبیعی برای پیش‌بینی‌های مالی پرداخته‌اند. یکی از پژوهش‌های برجسته در این زمینه، تحقیق ***Bollen et al.*** *(۲۰۱۱)*است که از تحلیل احساسات توییت‌ها و داده‌های عمومی شبکه‌های اجتماعی برای پیش‌بینی شاخص‌های بازارهای مالی استفاده کردند. نتایج این مطالعه نشان داد که احساسات عمومی که از توییت‌ها استخراج می‌شود، به‌طور قابل‌توجهی با نوسانات بازار ارتباط دارد و استفاده از این داده‌ها در مدل‌های پیش‌بینی می‌تواند به بهبود دقت پیش‌بینی‌های کوتاه‌مدت کمک کند. [۴]

به‌طور کلی، تحلیل احساسات[۴] یکی از تکنیک‌های پرکاربرد در استفاده از پردازش زبان طبیعی در بازارهای مالی است. پژوهش‌های متعددی نشان داده‌اند که با تحلیل احساسات استخراج‌شده از متن‌های منتشرشده در شبکه‌های اجتماعی، اخبار و رسانه‌های مالی، می‌توان اطلاعات ارزشمندی درباره روندهای بازار و نوسانات قیمتی به‌دست آورد ***Zhang et al.*** *(۲۰۲۰)* .نیز در مطالعه‌ای بر اهمیت این تکنیک تأکید کرده‌اند و نشان داده‌اند که با استفاده از تحلیل احساسات می‌توان تغییرات بازار را در مقیاس کوتاه‌مدت پیش‌بینی کرد. [۶]

۲ – ۲ – ۲ ترکیب داده‌های متنی و قیمتی

یکی دیگر از رویکردهای پرکاربرد در استفاده از NLP در پیش‌بینی‌های مالی، ترکیب داده‌های متنی با داده‌های قیمتی است ***Zou et al.*** *(۲۰۱۹)* .در پژوهش خود به بررسی روش‌هایی پرداختند که داده‌های قیمتی و متنی را برای ایجاد مدل‌های پیش‌بینی ترکیب می‌کنند. این تحقیق نشان داد که ترکیب این دو نوع داده می‌تواند اطلاعات دقیق‌تر و جامع‌تری را برای پیش‌بینی روندهای آینده بازار فراهم کند. با استفاده از روش‌هایی مانند برچسب‌گذاری سه‌گانه[۵](TBL)، این مدل‌ها توانستند به تحلیل دقیق‌تری از حرکات بازار دست یابند. [۲۸] [۱۸]

---

[۴] Sentiment Analysis

[۵] Triple Barrier Labeling

۲ – ۲ – ۳ تکنیک‌های برچسب‌گذاری پیشرفته در تحلیل بازارهای مالی

روش‌های برچسب‌گذاری داده‌ها، مانند TBL که توسط *Lopez de Prado* معرفی شده است، از جمله تکنیک‌های پیشرفته‌ای هستند که به بهبود دقت مدل‌های پیش‌بینی کمک می‌کنند. این تکنیک به‌طور خاص برای تحلیل حرکات کوتاه‌مدت بازار طراحی شده است و قادر است با استفاده از برچسب‌های دقیق، ارتباط بین احساسات و نوسانات قیمتی را بهتر نشان دهد. مطالعاتی مانند تحقیق *Lopez de Prado* (۲۰۱۸) نشان داده‌اند که استفاده از این روش‌های برچسب‌گذاری باعث افزایش دقت پیش‌بینی در مقایسه با روش‌های سنتی تحلیل احساسات می‌شود. [۱۷]

۲ – ۳ تحلیل احساسات در پیش‌بینی بازارهای مالی

یکی از پژوهش‌های پیشگام در این حوزه، تحقیق *Bollen et al. (۲۰۱۱)* است که نشان داد حالت روحی کاربران توییتر می‌تواند شاخص‌های بازار سهام را پیش‌بینی کند. در این مطالعه، احساسات عمومی استخراج شده از پست‌های توییتر به‌عنوان یک شاخص مستقل برای پیش‌بینی نوسانات شاخص‌های بازار مورد استفاده قرار گرفت. نتایج این تحقیق نشان داد که تغییرات در حالت روحی کاربران می‌تواند تا حدودی با نوسانات بازار همبستگی داشته باشد و از آن برای بهبود پیش‌بینی‌های مالی استفاده شود. [۴]

علاوه بر این، تحقیق *Tetlock (۲۰۰۷)* به تأثیر رسانه‌ها در شکل‌گیری احساسات سرمایه‌گذاران پرداخته است. در این مطالعه، تحلیل احساسات مندرج در مقالات خبری و تأثیر آن بر رفتار سرمایه‌گذاران و روندهای بازار بررسی شد. نتایج نشان داد که رسانه‌ها می‌توانند احساسات منفی و مثبت را در میان سرمایه‌گذاران ایجاد کرده و به‌طور مستقیم بر تصمیم‌گیری‌های سرمایه‌گذاری تأثیرگذار باشند. [۵]

تحقیق دیگری که بر روی تحلیل احساسات در فضای دارایی‌های دیجیتال تمرکز دارد، مطالعه *Kulakowski et al. (۲۰۲۳)* است. این پژوهش به طبقه‌بندی احساسات پست‌های مرتبط با رمزارزها در شبکه‌های اجتماعی پرداخته و نشان داده است که تحلیل احساسات در حوزه دارایی‌های دیجیتال نیز می‌تواند به‌عنوان یک ابزار مهم برای پیش‌بینی نوسانات قیمت این دارایی‌ها استفاده شود. این تحقیق به‌طور خاص نشان داد که احساسات کاربران شبکه‌های اجتماعی در ارتباط با رمزارزها می‌تواند به‌طور مستقیم بر قیمت و نوسانات بازار رمزارزها تأثیر بگذارد. [۳]

۲ – ۳ – ۱ کاربردهای تحلیل احساسات در بازارهای مالی

تحقیقات انجام‌شده در این زمینه نشان می‌دهد که تحلیل احساسات از دو جنبه مهم در پیش‌بینی بازارهای مالی نقش دارد. از یک سو، بررسی احساسات کاربران شبکه‌های اجتماعی و تحلیل تأثیر آن‌ها بر شاخص‌های

بازار، به سرمایه‌گذاران کمک می‌کند تا درک بهتری از روندهای بازار به‌دست آورند. از سوی دیگر، تحلیل احساسات منتشرشده در رسانه‌های مالی و خبری می‌تواند اطلاعات مهمی درباره وضعیت عمومی بازار و تمایلات سرمایه‌گذاران ارائه دهد.

این تحقیقات نشان داده‌اند که ترکیب تحلیل احساسات با داده‌های مالی می‌تواند ابزار قدرتمندی برای بهبود دقت پیش‌بینی‌های بازار فراهم کند. به‌عنوان مثال، استفاده از تحلیل احساسات همراه با مدل‌های مالی سنتی می‌تواند به شناسایی فرصت‌های جدید سرمایه‌گذاری کمک کند و به سرمایه‌گذاران این امکان را بدهد که با دقت بیشتری تصمیم‌گیری کنند.

۲ – ۴ پیش‌بینی قیمت رمزارزها

در سال‌های اخیر، پیش‌بینی قیمت رمزارزها توجه بسیاری از پژوهشگران و سرمایه‌گذاران را به خود جلب کرده است. این حوزه پژوهشی به دلیل ویژگی‌های منحصربه‌فرد بازار رمزارزها و نوسانات بالای قیمت‌ها با چالش‌ها و فرصت‌های جدیدی همراه است. تحقیقات متعددی در این زمینه به بررسی منابع مختلف داده و روش‌های پیش‌بینی پرداخته‌اند.

یکی از پژوهش‌های مهم در این حوزه، مطالعه *Mohapatra et al. (۲۰۲۰)* است که پلتفرم KryptoOracle را معرفی کرده‌اند. این پلتفرم به‌طور بلادرنگ از احساسات توییتر برای پیش‌بینی قیمت رمزارزها استفاده می‌کند. در این تحقیق، تحلیل احساسات کاربران توییتر به‌عنوان یک منبع داده مهم برای پیش‌بینی قیمت رمزارزها مورد استفاده قرار گرفته و نتایج نشان داده است که احساسات کاربران می‌تواند به‌طور مؤثری با تغییرات قیمت رمزارزها همبستگی داشته باشد. [۷]

همچنین *Critien et al. (۲۰۲۲)* ترکیبی از تحلیل احساسات توییتر و حجم داده‌ها را برای پیش‌بینی تغییرات قیمت بیت‌کوین به‌کار گرفته‌اند. این تحقیق نشان داد که هم احساسات کاربران و هم حجم تعاملات شبکه‌های اجتماعی می‌تواند به‌عنوان شاخص‌های مهمی برای شناسایی روندهای قیمت بیت‌کوین استفاده شود. این مطالعه تأکید بر اهمیت حجم داده‌ها در کنار تحلیل احساسات دارد و نشان می‌دهد که افزایش حجم تعاملات، به‌ویژه در شبکه‌های اجتماعی، می‌تواند به‌عنوان نشانه‌ای از نوسانات قیمت عمل کند. [۱۳]

در تحقیق دیگری، *Kraaijeveld and Smedt (۲۰۲۰)* قدرت پیش‌بینی احساسات عمومی کاربران توییتر را برای پیش‌بینی قیمت رمزارزها بررسی کرده‌اند. این پژوهش نشان داد که داده‌های شبکه‌های اجتماعی، به‌ویژه احساسات عمومی، می‌توانند به‌طور قابل توجهی در پیش‌بینی قیمت رمزارزها مؤثر باشند. نتایج این تحقیق به‌طور خاص بر تأثیر احساسات کاربران توییتر در تعیین روندهای قیمت رمزارزها تأکید دارد. [۱۵]

این تحقیقات نشان می‌دهند که تحلیل داده‌های شبکه‌های اجتماعی به‌ویژه احساسات کاربران، می‌تواند ابزار قدرتمندی برای پیش‌بینی قیمت رمزارزها باشد. با توجه به نوسانات شدید بازار رمزارزها و نقش کلیدی احساسات در تصمیم‌گیری‌های سرمایه‌گذاری، استفاده از روش‌های ترکیبی که از داده‌های اجتماعی و مالی بهره می‌برند، می‌تواند منجر به پیش‌بینی‌های دقیق‌تری شود. این مطالعات همچنین چالش‌های منحصربه‌فرد پیش‌بینی قیمت رمزارزها را برجسته می‌کنند و نشان می‌دهند که روش‌های سنتی به تنهایی قادر به پیش‌بینی دقیق این بازار پیچیده نیستند.

با توجه به نتایج این تحقیقات، مشخص است که بازار رمزارزها از پیچیدگی‌های خاصی برخوردار است که نیازمند رویکردهای نوآورانه و استفاده از منابع متنوع داده است. ترکیب تحلیل‌های احساسات با داده‌های بازار می‌تواند به بهبود عملکرد مدل‌های پیش‌بینی و کاهش ریسک سرمایه‌گذاری در این حوزه کمک کند.

۲ – ۵ تکنیک‌های پیشرفته یادگیری ماشین در پیش‌بینی بازارهای مالی

در سال‌های اخیر، تکنیک‌های پیشرفته یادگیری ماشین به‌طور گسترده‌ای برای بهبود پیش‌بینی بازارهای مالی به کار گرفته شده‌اند. این تکنیک‌ها با استفاده از مدل‌های پیچیده و چندلایه توانسته‌اند الگوهای پنهان و پیچیده‌ای را از داده‌های بازار استخراج کنند و دقت و استحکام پیش‌بینی‌ها را بهبود بخشند.

یکی از رویکردهای نوآورانه، مدل پیشنهادی *Zhou et al. (۲۰۲۰)* است که از یک شبکه توجه عصبی چندوجهی با سازگاری دامنه‌ای[۶] برای پیش‌بینی مالی استفاده می‌کند. این مدل، با ترکیب اطلاعات از منابع مختلف داده و توجه به ویژگی‌های خاص هر دامنه، توانسته است بهبود قابل توجهی در دقت پیش‌بینی‌ها در مقایسه با مدل‌های سنتی به دست آورد. این تکنیک با تمرکز بر ترکیب داده‌های مالی و اطلاعات متنی، به شناسایی بهتر نوسانات و روندهای بازار کمک می‌کند. [۶]

همچنین، *Huang et al. (۲۰۲۲)* مدلی به نام *FinBERT* را معرفی کرده‌اند که یک مدل زبان بزرگ تخصصی برای تحلیل متون مالی است FinBERT .با استفاده از روش‌های پردازش زبان طبیعی و مدل‌های یادگیری عمیق، توانسته است اطلاعات مالی مهمی را از متون تخصصی استخراج کند. این مدل در حوزه تحلیل احساسات و پیش‌بینی وقایع مالی بسیار مؤثر بوده است FinBERT .با توجه به ساختار و محتوای متون مالی، به سرمایه‌گذاران و تحلیل‌گران کمک می‌کند تا اطلاعات کلیدی را از اخبار و گزارش‌های مالی به‌دست آورند و تصمیم‌گیری‌های بهتری انجام دهند. [۹]

---

[۶] Multi-Modality Neural Attention Network Domain Adaptive

در یک پژوهش دیگر، *Ding et al. (۲۰۱۵)* از مدل‌های یادگیری عمیق برای پیش‌بینی سهام بر اساس وقایع خبری استفاده کرده‌اند. این پژوهش نشان داد که استفاده از تکنیک‌های یادگیری عمیق برای تحلیل داده‌های خبری می‌تواند به‌طور موثری منجر به بهبود دقت پیش‌بینی‌ها در بازارهای مالی شود. مدل‌های یادگیری عمیق با توانایی پردازش حجم زیادی از داده‌ها و استخراج الگوهای پنهان، ابزار قدرتمندی برای پیش‌بینی نوسانات بازار فراهم کرده‌اند. [۸]

علاوه بر این، رویکردهای ترکیبی مانند استفاده از شبکه‌های عصبی عمیق با تکنیک‌های دیگر مانند شبکه‌های عصبی بازگشتی (LSTM) و مکانیزم‌های توجه، توانسته‌اند به بهبود پیش‌بینی‌های مالی کمک کنند. این مدل‌ها با تحلیل داده‌های سری زمانی و استفاده از اطلاعات گذشته، توانسته‌اند نوسانات و تغییرات کوتاه‌مدت و بلندمدت در بازارهای مالی را پیش‌بینی کنند. همچنین، به کارگیری روش‌های یادگیری نظارت‌نشده و شبکه‌های خودرمزگذار[۷] نیز به تحلیل بهتر و شناسایی الگوهای پنهان در داده‌های مالی کمک کرده است. [۳۶]

نتایج این تحقیقات نشان می‌دهند که تکنیک‌های پیشرفته یادگیری ماشین، به‌ویژه مدل‌های مبتنی بر یادگیری عمیق، می‌توانند ابزار قدرتمندی برای تحلیل و پیش‌بینی دقیق بازارهای مالی باشند. این تکنیک‌ها با ترکیب داده‌های چندمنبعی و تحلیل خودکار اطلاعات مالی، به سرمایه‌گذاران و تحلیل‌گران کمک می‌کنند تا تصمیمات بهتری بر اساس داده‌های مالی و نوسانات بازار اتخاذ کنند. از آنجایی که بازارهای مالی پیچیدگی‌های زیادی دارند و عوامل مختلفی بر آنها تأثیرگذار است، استفاده از مدل‌های پیشرفته و چندلایه می‌تواند بهبود چشم‌گیری در پیش‌بینی دقیق و کاهش ریسک‌های مرتبط با سرمایه‌گذاری داشته باشد.

این مطالعات به وضوح نشان می‌دهند که روش‌های نوین یادگیری ماشین و یادگیری عمیق می‌توانند راهکارهای مؤثری برای تحلیل و پیش‌بینی نوسانات بازارهای مالی فراهم کنند. از آنجا که تحلیل‌های سنتی گاهی از پیچیدگی‌های بازار و داده‌های مالی ناتوان هستند، استفاده از مدل‌های پیشرفته یادگیری ماشین می‌تواند نتایج دقیق‌تر و بهتری ارائه دهد و به بهبود دقت پیش‌بینی‌های مالی کمک کند.

۲ – ۶ یکپارچه‌سازی اطلاعات زمینه‌ای در پیش‌بینی‌های مالی

در سال‌های اخیر، بسیاری از مطالعات به بررسی نحوه‌ی ادغام اطلاعات زمینه‌ای پرداخته‌اند تا قدرت پیش‌بینی مدل‌های مالی را افزایش دهند. این رویکرد نوآورانه با هدف ارتقای دقت و کاربردی‌تر کردن پیش‌بینی‌ها صورت می‌گیرد و نشان می‌دهد که افزودن اطلاعات زمینه‌ای می‌تواند به مدل‌ها کمک کند تا نوسانات قیمت‌ها و روندهای بازار را بهتر شناسایی کنند.

---

[۷] Autoencoders

مطالعه *Zou and Herremans (۲۰۲۳)* یکی از نمونه‌های برجسته در این زمینه است که مدل *PreBit* را معرفی می‌کند. این مدل چندوجهی با استفاده از تعبیه‌های[۸] مدل Twitter FinBERT ، توانسته است پیش‌بینی‌های دقیقی از تغییرات شدید قیمت بیت‌کوین ارائه دهد. یکی از ویژگی‌های کلیدی این مدل، استفاده از اطلاعات زمینه‌ای متنوع است که به پیش‌بینی‌های بهتر کمک می‌کند. در این مدل، داده‌های توییتر و تحلیل‌های زبان طبیعی با هم ترکیب می‌شوند تا حرکت‌های شدید قیمت را شناسایی کنند *PreBit* .نشان می‌دهد که چگونه ترکیب داده‌های متنی و مالی با استفاده از تکنیک‌های پیشرفته می‌تواند به بهبود نتایج پیش‌بینی کمک کند. [۸][۹][۱۸]

این رویکرد با تمرکز بر اطلاعات زمینه‌ای که ممکن است در پیش‌بینی‌های مالی سنتی نادیده گرفته شوند، دقت پیش‌بینی‌ها را بهبود می‌بخشد. به عنوان مثال، استفاده از اطلاعات مربوط به هویت‌های تأثیرگذار در رسانه‌های اجتماعی یا شاخص‌های محبوبیت می‌تواند به مدل‌ها کمک کند تا احساسات و تغییرات مرتبط با آن‌ها را بهتر تحلیل کنند. چنین اطلاعات زمینه‌ای، شامل تحلیل احساسات و داده‌های شبکه‌های اجتماعی، به وضوح می‌تواند بر قیمت‌گذاری دارایی‌های دیجیتال مانند بیت‌کوین تأثیرگذار باشد و دقت پیش‌بینی‌ها را افزایش دهد.

افزایش دقت پیش‌بینی‌ها از طریق یکپارچه‌سازی داده‌های مختلف، به‌ویژه در حوزه‌های مالی پیچیده، یک ضرورت است. مطالعه *Critien et al. (۲۰۲۲)* و *Kraaijeveld and Smedt (۲۰۲۰)* نیز این امر را تأیید می‌کنند؛ آن‌ها نشان داده‌اند که با تحلیل هم‌زمان داده‌های شبکه‌های اجتماعی و حجم داده‌های موجود، می‌توان تغییرات قیمت ارزهای دیجیتال را با دقت بیشتری پیش‌بینی کرد. این نوع داده‌های زمینه‌ای، که شامل احساسات عمومی و نوسانات در شبکه‌های اجتماعی هستند، نه‌تنها به افزایش دقت پیش‌بینی‌ها کمک می‌کنند، بلکه ارتباط مستقیم بین اطلاعات متنی و روندهای قیمتی را نشان می‌دهند. [۱۵][۱۳]

در نهایت، این نوع مطالعات اهمیت استفاده از اطلاعات زمینه‌ای برای بهبود پیش‌بینی‌های مالی را به‌خوبی به تصویر می‌کشند. مدل‌هایی که به‌طور خاص بر تحلیل اطلاعات متنی و داده‌های مرتبط با شبکه‌های اجتماعی تمرکز دارند، می‌توانند تحلیل‌های دقیق‌تر و قابل اعتمادتری از نوسانات بازار ارائه دهند. ترکیب اطلاعات هویتی، محبوبیت‌ها، و داده‌های خبری با سایر داده‌های سنتی مالی می‌تواند دقت پیش‌بینی‌های مرتبط با بازارهای مالی را به‌طور چشم‌گیری افزایش دهد.

[۸] Embeddings

فصل ۳ : روش تحقیق

۳ – ۱ مقدمه

در این بخش به شرح کامل روش تحقیق پرداخته شده است. ابتدا به دلایل انتخاب روش‌های مورد استفاده پرداخته شده و سپس به توضیح جامع هر یک از مراحل پیاده‌سازی مدل‌های ارائه شده در این تحقیق می‌پردازیم.

در این روش تحقیق، ابتدا یک مدل خودرمزگذار برای استخراج ویژگی‌های نهفته از داده‌ها توسعه داده شده است که شامل یک LSTM (به عنوان رمزگذار)، یک لایه چگال[۹] (به عنوان برون‌یاب)، و سپس یک LSTM (به عنوان رمزگشا) می‌باشد. این مدل خودرمزگذار به عنوان بخش اولیه فرآیند استخراج ویژگی‌های اساسی از داده‌ها استفاده شده است.

پس از آن، یک مدل دو لایه چگال طراحی شده است که وظیفه آن تحلیل داده‌های نهفته استخراج شده و تعیین برچسب مناسب برای هر نمونه می‌باشد. این مدل به صورت مستقیم از خروجی رمزگذار خودرمزگذار استفاده می‌کند تا پیش‌بینی‌های دقیق‌تری انجام دهد.

در مرحله بعد، داده‌های تکنیکال و تحلیل‌های بدست آمده از بیت‌کوین و مدل‌های قبلی به این مدل داده می‌شود تا برچسب نهایی تعیین شود. همچنین در مدل تحلیل احساسات، یک LSTM برای تحلیل ویژگی‌های استخراج شده از متن استفاده شده است. این ویژگی‌ها شامل تحلیل احساسات، تاثیرات کوتاه‌مدت، برچسب قبلی، افزایش ارزش در طول زمان، ROC، و RSI می‌باشد.

در پایان، مدل نهایی فیوژن که شامل دو لایه چگال است، احتمال‌های محاسبه شده از مدل‌های مختلف را دریافت کرده و برچسب نهایی را تعیین می‌کند. این روش ترکیبی به منظور افزایش دقت و قابلیت تعمیم‌پذیری مدل‌های پیش‌بینی ایجاد شده و به بهبود عملکرد کلی سیستم منجر شده است.

۳ – ۲ علت انتخاب روش

انتخاب روش‌های مورد استفاده در این پژوهش بر اساس نیاز به بهبود دقت و کارایی مدل‌های پیش‌بینی در حوزه تحلیل تکنیکال و تحلیل احساسات صورت گرفته است. با توجه به پیچیدگی داده‌های بیت‌کوین و

---

[۹] Dense

همچنین تأثیرات مختلف عوامل خارجی مانند احساسات بازار، استفاده از یک مدل ترکیبی که بتواند اطلاعات مختلف را به صورت همزمان پردازش کند، ضروری به نظر می‌رسید.

مدل خودرمزگذار (Autoencoder) انتخاب شد تا ویژگی‌های نهفته و الگوهای پنهان در داده‌های تکنیکال و تاریخی را استخراج کند، چرا که این ویژگی‌ها می‌توانند اطلاعات ارزشمندی را در مورد رفتار بازار فراهم کنند. این امر به بهبود دقت مدل‌های بعدی در تحلیل داده‌ها کمک می‌کند.

در کنار این، مدل تحلیل احساسات برای بررسی و تجزیه و تحلیل داده‌های متنی و تأثیرات کوتاه‌مدت آن‌ها بر روی بازار بیت‌کوین استفاده شد. به دلیل اهمیت روزافزون احساسات عمومی و تأثیر آن‌ها بر تصمیم‌گیری‌های سرمایه‌گذاری، این بخش از تحقیق نیز ضروری به نظر می‌رسید.

در نهایت، به منظور ترکیب نتایج مدل‌های مختلف و دستیابی به یک برچسب نهایی با دقت بالا، از یک مدل فیوژن استفاده شد که قادر به پردازش همزمان اطلاعات مختلف و ارائه یک پیش‌بینی جامع است. این رویکرد ترکیبی به دلیل قابلیت تطبیق و انعطاف‌پذیری بالای آن در مواجهه با داده‌های پیچیده و نامتجانس انتخاب شده است.

۳ – ۳ تشریح کامل روش تحقیق

همانطور که گفته شد، روش تحقیق در این پروژه شامل چهار مرحله اصلی توسعه مدل‌ها، ترکیب نتایج، ارزیابی عملکرد نهایی، و اجرای استراتژی معاملاتی با استفاده از بک‌تریدر [۱۰] است.

در مرحله اول، به توسعه مدل‌های خودرمزگذار و تحلیل احساسات پرداخته شده است. مدل خودرمزگذار برای استخراج ویژگی‌های نهفته از داده‌های تکنیکال و تاریخی بیت‌کوین طراحی شده است. این مدل شامل یک LSTMبه عنوان رمزگذار، یک لایه چگال برای برون‌یابی، و یک LSTM به عنوان رمزگشا می‌باشد. هدف از این مرحله، استخراج ویژگی‌هایی است که بتواند رفتار بازار را با دقت بیشتری مدل‌سازی کند. [۲۹]

در مرحله دوم، یک مدل تحلیل احساسات بر اساس داده‌های متنی طراحی شده است. این مدل با استفاده از LSTM، ویژگی‌های مختلفی مانند تحلیل احساسات، تأثیرات کوتاه‌مدت، و شاخص‌های تکنیکال مانند ROC و RSIرا تجزیه و تحلیل کرده و به پیش‌بینی برچسب نهایی کمک می‌کند. هدف از این مرحله، بررسی تأثیرات احساسات عمومی و داده‌های متنی بر بازار بیت‌کوین است.

[۱۰] Backtrader

در مرحله سوم، نتایج به دست آمده از دو مدل مذکور با استفاده از یک مدل فیوژن ترکیب شده است. این مدل فیوژن شامل یک ساختار دو لایه چگال می‌باشد که وظیفه آن ترکیب احتمالات محاسبه شده از مدل‌های قبلی و ارائه برچسب نهایی است. هدف از این مرحله، دستیابی به یک پیش‌بینی جامع و دقیق از رفتار بازار با توجه به تمامی اطلاعات موجود است.

در نهایت، در مرحله چهارم، از استراتژی معاملاتی با استفاده از پلتفرم بک‌تریدر برای ارزیابی و اجرای عملی نتایج به دست آمده استفاده شده است. این استراتژی به منظور بررسی سودآوری و کارایی پیش‌بینی‌های مدل‌ها در شرایط واقعی بازار طراحی و اجرا شده است. هدف از این مرحله، اعتبارسنجی نتایج مدل‌ها و بررسی امکان پیاده‌سازی عملی آن‌ها در معاملات واقعی است.

در ادامه، به جزئیات پیاده‌سازی هر یک از این مراحل و نتایج حاصل از آن‌ها پرداخته خواهد شد.

۳ – ۴ پیش پردازش داده های عددی

در این بخش، به تشریح کامل فرآیند پیش‌پردازش داده‌های عددی که برای مدل‌سازی استفاده شده‌اند، پرداخته می‌شود. این فرآیند شامل جمع‌آوری داده‌ها از منابع مختلف، آماده‌سازی و ترکیب داده‌ها، محاسبه اندیکاتورهای تکنیکال، و نرمال‌سازی ویژگی‌ها برای بهبود عملکرد مدل‌های پیش‌بینی است.

۳ – ۴ – ۱ جمع‌آوری داده‌ها

داده‌های مورد استفاده در این تحقیق شامل داده‌های تاریخی بیت‌کوین(BTC) ، اتریوم(ETH) ، و طلا (Gold) است که از منابع مختلف جمع‌آوری شده‌اند:

- **داده‌های بیت‌کوین** از وب‌سایت CryptoCompare با استفاده از API این سرویس جمع‌آوری شده است. داده‌های OHLCV (قیمت آغازین، بالاترین قیمت، پایین‌ترین قیمت، قیمت بسته شدن و حجم) روزانه برای بازه زمانی از اول ژانویه ۲۰۱۵ تا ۳۱ مه ۲۰۲۱ استخراج شده است.
- **داده‌های اتریوم** نیز از همان منبع و برای همان بازه زمانی دریافت شده است. تنها قیمت بسته شدن روزانه (Close) مورد استفاده قرار گرفته است.
- **داده‌های طلا** از سرویس Yahoo Finance با استفاده از کتابخانه yfinance استخراج شده است. در اینجا نیز فقط قیمت بسته شدن روزانه طلا مورد استفاده قرار گرفته است.

۳ – ۴ – ۲ آماده‌سازی و ترکیب داده‌ها

پس از جمع‌آوری داده‌ها، لازم بود که داده‌های اتریوم و طلا با داده‌های بیت‌کوین هماهنگ شوند. این کار با استفاده از روش پر کردن رو به جلو [۱۱]انجام شده است تا مقادیر گمشده در تاریخ‌های خاص با آخرین مقدار موجود جایگزین شوند. پس از هماهنگ‌سازی، داده‌های اتریوم و طلا به عنوان ستون‌های جدید به نگاشت داده ها[۱۲] بیت‌کوین اضافه شدند.

سپس، حجم معاملات بیت‌کوین به دلار (USD) محاسبه و به نگاشت داده ها اضافه شد. در نهایت، ستون‌های غیرضروری مانند volumefrom، conversionType، conversionSymbolو volumetoحذف شدند تا فقط داده‌های مورد نیاز برای مدل‌سازی باقی بمانند.

۳ – ۴ – ۳ محاسبه اندیکاتورهای تکنیکال

محاسبه اندیکاتورهای تکنیکال

اندیکاتورهای تکنیکال ابزارهای قدرتمندی هستند که در تحلیل داده‌های مالی و پیش‌بینی رفتار بازار استفاده می‌شوند. این اندیکاتورها به منظور شناسایی روندها، الگوها و سیگنال‌های معاملاتی مورد استفاده قرار می‌گیرند. در این بخش، چندین اندیکاتور تکنیکال مهم محاسبه و به داده‌های بیت‌کوین اضافه شده‌اند تا مدل‌ها بتوانند با دقت بیشتری به تحلیل داده‌ها بپردازند. در ادامه، توضیح هر یک از این اندیکاتورها و نحوه محاسبه آنها آمده است:

۱ .میانگین متحرک نمایی(EMA)[۱۳]

میانگین متحرک نمایی (EMA) نوعی میانگین متحرک است که به داده‌های جدید وزن بیشتری می‌دهد و در نتیجه به تغییرات قیمت حساس‌تر است. در این تحقیق، سه نوع میانگین متحرک نمایی محاسبه شده است:

EMA۰.۶۷: این EMA با بازه زمانی ۰.۶۷ روزه محاسبه شده است و برای شناسایی تغییرات لحظه‌ای در بازار به کار می‌رود.

EMA۱۲: این میانگین متحرک نمایی ۱۲ روزه است و برای تحلیل روندهای کوتاه‌مدت مورد استفاده قرار می‌گیرد.

---

[۱۱] forward fill

[۱۲] DataFrame

[۱۳] Exponential Moving Average

EMA۲۶:این میانگین متحرک نمایی ۲۶ روزه است که برای تحلیل روندهای بلندمدت‌تر نسبت به EMA۱۲ به کار گرفته می‌شود.

این میانگین‌ها به گونه‌ای محاسبه می‌شوند که به قیمت‌های اخیر وزن بیشتری می‌دهند، بنابراین به تغییرات سریع قیمت واکنش بیشتری نشان می‌دهند.

۲ .اندیکاتور میانگین متحرک همگرا[۱۴](MACD)

MACDیکی از پرکاربردترین اندیکاتورهای تکنیکال است که تفاوت بین دو میانگین متحرک نمایی با بازه‌های مختلف را محاسبه می‌کند. در این تحقیق، MACDبه صورت تفاضل بین EMA۱۲ و EMA۲۶ محاسبه شده است.

MACDبه عنوان یک اندیکاتور حرکت[۱۵] استفاده می‌شود و به شناسایی نقاط خرید و فروش بالقوه کمک می‌کند. زمانی که MACD مثبت باشد، نشان‌دهنده فشار خرید بیشتر است و زمانی که منفی باشد، فشار فروش را نشان می‌دهد.

۳ .انحراف استاندارد ۲۰ روزه(۲۰dSTD)

انحراف استاندارد یکی از معیارهای پراکندگی داده‌ها است که نوسانات قیمت را اندازه‌گیری می‌کند. در این پژوهش، انحراف استاندارد قیمت بسته شدن بیت‌کوین در طول ۲۰ روز گذشته محاسبه شده است.

این اندیکاتور به منظور تعیین باندهای بولینگر و شناسایی دوره‌های با نوسانات بالا یا پایین استفاده می‌شود. افزایش انحراف استاندارد نشان‌دهنده افزایش نوسانات بازار است و کاهش آن نشان‌دهنده کاهش نوسانات.

۴ .باندهای بولینگر[۱۶]

باندهای بولینگر یکی دیگر از ابزارهای تحلیل تکنیکال است که از دو باند تشکیل شده است :باند بالایی و باند پایینی .این باندها با استفاده از میانگین متحرک ساده ۲۱ روزه (MA۲۱) و انحراف استاندارد ۲۰ روزه (۲۰dSTD)محاسبه می‌شوند.

باند بالایی :دو برابر انحراف استاندارد ۲۰ روزه بالاتر از میانگین متحرک ۲۱ روزه قرار دارد.

---

[۱۴] Moving Average Convergence Divergence

[۱۵] Momentum

[۱۶] Bollinger Bands

باند پایینی :دو برابر انحراف استاندارد ۲۰ روزه پایین‌تر از میانگین متحرک ۲۱ روزه قرار دارد.

باندهای بولینگر به عنوان سطوح حمایتی و مقاومتی عمل می‌کنند و معمولاً برای شناسایی نقاط اشباع خرید یا فروش و همچنین نوسانات غیرعادی استفاده می‌شوند.

۵ .اختلاف قیمت بالا و پایین روزانه[۱۷]

اختلاف قیمت بالا و پایین روزانه تفاوت بین بالاترین قیمت (High) و پایین‌ترین قیمت (Low) در طول روز است.

این اندیکاتور میزان نوسانات در طول روز را نشان می‌دهد و می‌تواند به شناسایی روزهای با نوسانات بالا یا پایین کمک کند.

۶ .شاخص MA

این شاخص با مقایسه میانگین متحرک ۷ روزه (MA۷) و میانگین متحرک ۲۱ روزه (MA۲۱) محاسبه شده است. اگر میانگین ۷ روزه بالاتر از میانگین ۲۱ روزه باشد، نشان‌دهنده روند مثبت و اگر پایین‌تر باشد، نشان‌دهنده روند منفی است.

این شاخص به طور ساده نشان می‌دهد که آیا روند کوتاه‌مدت (MA۷) بالاتر از روند میان‌مدت (MA۲۱) قرار دارد یا خیر. اگر MA۷ بالاتر باشد، نشان‌دهنده یک روند صعودی است و در غیر این صورت، روند نزولی محسوب می‌شود.

۳ – ۴ – ۴ نرمال‌سازی داده‌ها

به منظور بهبود عملکرد مدل‌ها، ویژگی‌های محاسبه شده نرمال‌سازی شدند. نرمال‌سازی به این صورت انجام شده که هر ویژگی به نسبت به قیمت بسته شدن روز قبل (Close) مقیاس‌بندی شده است. این کار به کاهش مقادیر بسیار بزرگ یا کوچک و ایجاد تعادل بین ویژگی‌ها کمک می‌کند. به عنوان مثال:

- MACD به نسبت قیمت بسته شدن روز قبل نرمال شده است.
- انحراف استاندارد ۲۰ روزه نیز به همین روش نرمال شده است.
- حجم معاملات (volume) به صورت تغییرات نسبی نسبت به روز قبل نرمال‌سازی شده است.

[۱۷] High-Low Spread

این روش نرمال‌سازی در تمامی ویژگی‌ها اعمال شده تا مقادیر خروجی نهایی برای مدل‌های یادگیری ماشینی آماده شوند.

۳ – ۴ – ۵ پاکسازی داده‌ها

پس از نرمال‌سازی، برخی از مقادیر ممکن است به دلیل محاسبات نرمال‌سازی یا وجود داده‌های گمشده، به مقادیر نامعتبر (مانند NaN یا بی‌نهایت) تبدیل شده باشند. برای پاکسازی داده‌ها، ابتدا تمامی مقادیر نامعتبر با استفاده از روش پرگردن رو به جلو با مقادیر روز قبل جایگزین شدند. اگر در روز قبل نیز داده‌ای موجود نبود، مقدار ۰جایگزین آن شده است. این کار باعث می‌شود که هیچ داده گمشده یا نامعتبر در داده‌های نهایی وجود نداشته باشد.

۳ – ۵ تحلیل بر اساس احساسات و ویژگی های برآمده از متن

در دنیای ارزهای دیجیتال، تحلیل احساسات به یکی از ابزارهای کلیدی برای پیش‌بینی حرکت‌های بازار تبدیل شده است. با افزایش حجم اطلاعات متنی در رسانه‌های اجتماعی و انجمن‌های مرتبط با ارزهای دیجیتال، نیاز به ابزارهای پیشرفته برای تحلیل این داده‌ها بیش از پیش احساس می‌شود. این مقاله با عنوان "تحلیل بر اساس احساسات و ویژگی‌های برآمده از متن" به بررسی مدل CryptoBERT می‌پردازد. این مدل بر پایه معماری BERTتوسعه یافته و به منظور تحلیل احساسات در پست‌های مرتبط با ارزهای دیجیتال طراحی شده است. در ادامه، ساختار، فرآیند آموزش و قابلیت‌های این مدل به تفصیل مورد بررسی قرار می‌گیرد.

۳ – ۵ – ۱ روش SHAP

۳ – ۵ – ۱ – ۱ توضیح روشSHAP

یکی از ابزارهای قدرتمند برای تفسیر مدل‌های یادگیری ماشین، **SHAP** است[۱۸]. این روش که بر اساس تئوری بازی‌های تعاونی شاپلی[۱۹] بنا شده است، به عنوان یک روش جامع و قابل اطمینان برای تبیین چگونگی و چرایی تصمیمات مدل‌های پیچیده، به خصوص مدل‌های مبتنی بر شبکه‌های عصبی و جنگل‌های تصادفی، به کار می‌رود.

روش SHAP به ما کمک می‌کند تا سهم هر ویژگی ورودی را در تصمیم‌گیری نهایی مدل به صورت دقیق اندازه‌گیری کنیم. با استفاده از این روش، می‌توان توضیح داد که چگونه هر ویژگی در تغییر پیش‌بینی نهایی مدل تأثیر دارد و به این ترتیب، قابلیت تفسیرپذیری مدل به میزان قابل توجهی افزایش می‌یابد.

---

[۱۸] SHapley Additive exPlanations

[۱۹] Shapley Value

۳ – ۵ – ۱ – ۲ اهمیت استفاده از SHAP

استفاده از SHAP برای تحلیل مدل‌های یادگیری ماشین، به خصوص در پیش‌بینی‌های مالی و تحلیل داده‌های پیچیده‌ای مانند متون مالی و اجتماعی، اهمیت ویژه‌ای دارد. این ابزار به محققان و تحلیل‌گران امکان می‌دهد تا تاثیر هر یک از ویژگی‌های ورودی را در پیش‌بینی‌های بازار به دقت بسنجند. از آنجا که در مدل‌های مالی و اجتماعی، اغلب با حجم زیادی از داده‌ها و ویژگی‌های گوناگون مواجه هستیم، استفاده از SHAP می‌تواند به ما کمک کند تا مشخص کنیم کدام یک از این ویژگی‌ها بیشترین تأثیر را در حرکت‌های بازار داشته‌اند.

۳ – ۵ – ۱ – ۳ روش کار SHAP

SHAPارزش هر ویژگی ورودی را بر اساس یک چارچوب همکاری محاسبه می‌کند. به این معنا که برای هر ویژگی، مقدار تغییرات در پیش‌بینی نهایی مدل در نظر گرفته می‌شود. این تغییرات به صورت میانگین وزنی توزیع می‌شود و به ما نشان می‌دهد که چگونه ویژگی‌ها به صورت تعاملی با یکدیگر کار می‌کنند و تاثیر هر یک بر نتیجه نهایی چیست.

با این روش، نه تنها می‌توان به صورت کلی تأثیرات هر ویژگی را بر پیش‌بینی مدل مشاهده کرد، بلکه می‌توان بررسی کرد که چگونه ویژگی‌ها در کنار یکدیگر عمل می‌کنند و تعاملات بین ویژگی‌ها را در نظر گرفت. [۱۶]

۳ – ۶ برچسب زنی داده ها

برچسب‌گذاری بر اساس پنجره ثابت[۲۰] یکی از تکنیک‌های متداول در تحلیل داده‌های سری زمانی، به‌ویژه در زمینه‌های مالی و پیش‌بینی بازار است. در این روش، داده‌های تاریخی به صورت دوره‌های زمانی ثابت یا پنجره‌های مشخص تقسیم می‌شوند. این پنجره‌های زمانی می‌توانند بسته به نیاز و نوع تحلیل به صورت روزانه، هفتگی، ماهانه یا حتی سالانه تعریف شوند. هر پنجره زمانی به یک برچسب مشخص اختصاص داده می‌شود که می‌تواند نشان‌دهنده روند بازار یا رفتار قیمت دارایی در آن بازه باشد. این برچسب‌ها معمولاً به صورت "صعودی" یا "نزولی" تعریف می‌شوند، اما در برخی موارد ممکن است از برچسب‌های دیگری مانند "ثابت" یا "بی‌تغییر" نیز استفاده شود.

برای اختصاص برچسب به هر پنجره، ابتدا رفتار کلی قیمت یا شاخص موردنظر در طول بازه زمانی مربوطه بررسی می‌شود. اگر در طول آن بازه زمانی قیمت دارایی یا شاخص افزایش یابد، به آن پنجره برچسب "صعودی" تخصیص داده می‌شود. برعکس، اگر قیمت کاهش یابد، برچسب "نزولی" به آن تعلق می‌گیرد. این

[۲۰] Fixed Window Labeling

روش به طور موثری روندهای کلی بازار را شناسایی می‌کند و به عنوان یک ابزار پایه در بسیاری از مدل‌های پیش‌بینی مالی، از جمله پیش‌بینی قیمت دارایی‌ها، شاخص‌ها و سایر متغیرهای اقتصادی، مورد استفاده قرار می‌گیرد.

یکی از مزایای اصلی این روش، سادگی آن است. تقسیم داده‌ها به بازه‌های زمانی ثابت و اختصاص برچسب به هر بازه، فرآیندی ساده و مستقیم است که می‌تواند درک خوبی از روندهای عمومی بازار فراهم کند. به ویژه برای تحلیل‌های بلندمدت که در آن تغییرات جزئی و نوسانات لحظه‌ای بازار اهمیت کمتری دارند، این روش می‌تواند بسیار مفید باشد. به عنوان مثال، در تحلیل‌های مربوط به روندهای سالانه قیمت سهام یا شاخص‌های اقتصادی، استفاده از پنجره‌های زمانی ماهانه یا سه‌ماهه می‌تواند نتایج خوبی به همراه داشته باشد.

با این حال، یکی از محدودیت‌های بزرگ روش پنجره ثابت این است که نوسانات لحظه‌ای یا تغییرات ناگهانی بازار را به درستی منعکس نمی‌کند. از آنجا که در این روش تمام تحرکات و تغییرات بازار در طول یک پنجره زمانی به یک برچسب واحد خلاصه می‌شود، جزئیات مهم و تغییرات ناگهانی بازار که ممکن است در طول پنجره رخ دهند، نادیده گرفته می‌شوند. به عنوان مثال، ممکن است در طول یک بازه زمانی یک‌ماهه، قیمت دارایی ابتدا افزایش یابد و سپس کاهش شدیدی را تجربه کند. با این حال، برچسب اختصاص داده شده به آن بازه تنها نشان‌دهنده نتیجه نهایی خواهد بود و جزئیات مربوط به تغییرات بینابینی در نظر گرفته نمی‌شود. این موضوع می‌تواند دقت پیش‌بینی‌های کوتاه‌مدت را به شدت کاهش دهد و برای بازارهایی که نوسانات زیادی دارند، ناکارآمد باشد.

علاوه بر این، اندازه پنجره‌های زمانی نیز می‌تواند تاثیر قابل توجهی بر نتایج داشته باشد. انتخاب یک پنجره زمانی کوتاه می‌تواند منجر به برچسب‌گذاری مکرر و نادیده گرفتن روندهای بلندمدت شود، در حالی که استفاده از پنجره‌های طولانی‌تر ممکن است باعث شود که تغییرات کوتاه‌مدت و نوسانات مهم بازار از دید پنهان بمانند. بنابراین، انتخاب مناسب اندازه پنجره‌های زمانی یکی از چالش‌های اصلی این روش است و بسته به نوع داده‌ها و اهداف پیش‌بینی، نیاز به تنظیم دقیق دارد.

با توجه به این محدودیت‌ها، برخی از تحلیل‌گران ممکن است از روش‌های پیشرفته‌تر و ترکیبی برای افزایش دقت پیش‌بینی استفاده کنند. به عنوان مثال، می‌توان از تکنیک‌هایی مانند "پنجره‌های متغیر [۲۱]"استفاده کرد که در آن طول پنجره‌ها بسته به شرایط بازار تغییر می‌کند. همچنین می‌توان از الگوریتم‌های یادگیری ماشین برای بهبود برچسب‌گذاری و کاهش تاثیر نوسانات لحظه‌ای بهره برد. با این وجود، روش پنجره ثابت همچنان به

[۲۱] Variable Windowing

دلیل سادگی و کارآمدی در بسیاری از تحلیل‌ها مورد استفاده قرار می‌گیرد و می‌تواند به عنوان یک نقطه شروع مناسب برای مدل‌های پیش‌بینی مالی عمل کند .

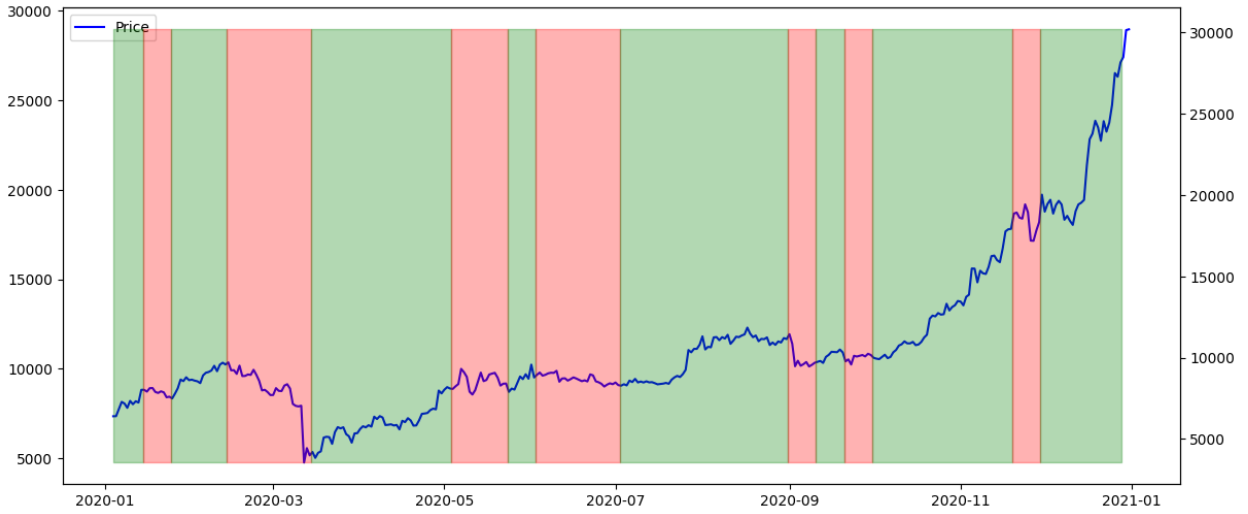

Figure ۱نمونه حالت پنجره ثابت برای برچسب زنی

۳ – ۶ – ۱ روش‌های پیشرفته‌تر در برچسب‌گذاری داده‌ها

روش‌های پیشرفته‌تر در برچسب‌گذاری داده‌ها به‌طور خاص به منظور حل مشکلات روش‌های ساده‌تر مانند پنجره ثابت و افزایش دقت پیش‌بینی‌ها توسعه یافته‌اند. یکی از این روش‌های پیشرفته که به‌طور گسترده مورد استفاده قرار گرفته است، برچسب‌گذاری سه‌گانه است. این روش در مقایسه با روش‌های قدیمی‌تر که از پنجره‌های زمانی ثابت استفاده می‌کنند، انعطاف‌پذیری بیشتری دارد و از معیارهای چندگانه برای تعیین برچسب‌ها بهره می‌برد. به جای تکیه بر یک بازه زمانی ثابت برای تعیین صعود یا نزول قیمت، این روش از سه مانع یا حد استفاده می‌کند: یک حد بالا برای تعیین نقطه صعود، یک حد پایین برای تعیین نقطه نزول، و یک محدودیت زمانی که مشخص می‌کند تا چه زمانی باید منتظر باشیم تا یکی از این دو حد اتفاق بیافتد.

این روش به‌طور خاص برای شناسایی دقیق‌تر نقاط ورود و خروج از بازار و ارزیابی بهتر رفتار قیمت‌ها در شرایط پیچیده طراحی شده است. در مقایسه با روش پنجره ثابت که به طور عمده به روند کلی قیمت در یک بازه زمانی مشخص توجه می‌کند، روش برچسب‌گذاری سه‌گانه با ایجاد معیارهای متنوعی برای شناسایی تغییرات قیمت، دقت بسیار بالاتری در پیش‌بینی ارائه می‌دهد. برای مثال، اگر قیمت دارایی در یک بازه زمانی مشخص به حد بالایی برسد، سیستم می‌تواند برچسب "صعودی" اختصاص دهد، حتی اگر تا پایان آن بازه زمانی قیمت دوباره کاهش یابد. به همین ترتیب، اگر قیمت به حد پایینی برسد، برچسب "نزولی" تخصیص می‌یابد، بدون توجه به تغییرات موقت قیمت در طول آن دوره. این انعطاف‌پذیری باعث می‌شود که این روش به‌ویژه برای بازارهایی که دارای نوسانات لحظه‌ای و سریع هستند، کارآمدتر باشد.

علاوه بر این، برچسب‌گذاری سه‌گانه توانایی ارزیابی و تحلیل عمیق‌تری از روندهای بازار ارائه می‌دهد. این روش، نه‌تنها تغییرات ساده قیمتی را در نظر می‌گیرد، بلکه با اضافه کردن محدودیت‌های زمانی و تحلیل دقیق شرایط بازار، می‌تواند به شناسایی روندهای پیچیده‌تر و حرکات غیرمنتظره قیمت کمک کند. از این رو، در بسیاری از مدل‌های پیش‌بینی مالی پیشرفته، برچسب‌گذاری سه‌گانه به عنوان یکی از ابزارهای کلیدی به کار می‌رود. این روش با کمک به تحلیل‌گران و سرمایه‌گذاران در شناسایی دقیق‌تر نقاط ورود و خروج از بازار، می‌تواند سودآوری استراتژی‌های معاملاتی را بهبود بخشد و در عین حال ریسک‌های ناشی از تغییرات ناگهانی بازار را کاهش دهد.

یکی از مفاهیم کلیدی در برچسب‌گذاری سه‌گانه، نحوه تعریف و تنظیم این سه حد (حد بالا، حد پایین و محدودیت زمانی) است. برای تعیین این حدود، معمولاً از داده‌های تاریخی قیمت‌ها و تحلیل‌های آماری استفاده می‌شود. برای مثال، تحلیل‌گران ممکن است با بررسی تغییرات تاریخی قیمت در بازار موردنظر، حد بالایی و پایینی را برای یک دارایی مشخص کنند که نشان‌دهنده سطحی است که اگر قیمت به آن برسد، احتمالاً روند قیمتی معکوس خواهد شد. به همین ترتیب، محدودیت زمانی ممکن است بر اساس تجربیات گذشته و تحلیل‌های الگوریتمی تنظیم شود، به طوری که نشان‌دهنده زمانی است که باید منتظر بمانیم تا یکی از این دو حد اتفاق بیفتد. در این صورت، سیستم می‌تواند برچسب مناسبی را برای پیش‌بینی قیمت‌ها تخصیص دهد.

یکی از مزایای دیگر این روش این است که می‌تواند تغییرات متنی و محیطی را نیز به تحلیل‌ها اضافه کند. به عنوان مثال، در یک محیط مالی که اطلاعات زیادی از طریق شبکه‌های اجتماعی و اخبار منتشر می‌شود، برچسب‌گذاری سه‌گانه می‌تواند تأثیرات این اطلاعات را به‌طور مستقیم بر روند قیمت‌ها اندازه‌گیری کند. این قابلیت به‌ویژه در بازاری که احساسات عمومی و اخبار می‌توانند به سرعت بر قیمت‌ها تأثیر بگذارند، اهمیت زیادی دارد.

این روش که برای اولین بار توسط مارکوس لوپز دو پرادو در کتاب خود تحت عنوان **پیشرفت‌های یادگیری ماشینی مالی** معرفی شد، انقلابی در نحوه تحلیل داده‌های مالی ایجاد کرد. دو پرادو این روش را برای تحلیل داده‌های مالی پیچیده و پیش‌بینی رفتار بازارها توسعه داد و نشان داد که برچسب‌گذاری سه‌گانه با ارائه تحلیل‌های عمیق‌تر از روندهای قیمتی، می‌تواند دقت پیش‌بینی‌ها را بهبود بخشد. این رویکرد به ما این امکان را می‌دهد که تأثیرات واقعی رویدادهای متنی و خارجی را بر بازار مالی بررسی کنیم، به جای اینکه تنها به ارزیابی سطحی از احساسات یا داده‌های تاریخی اکتفا کنیم.

در نهایت، برچسب‌گذاری سه‌گانه یکی از روش‌های نوین و قدرتمند در زمینه یادگیری ماشینی مالی است که توانسته است مشکلات روش‌های سنتی‌تر مانند پنجره ثابت را بهبود بخشد و دقت بیشتری در پیش‌بینی بازارها و تحلیل رفتارهای قیمتی فراهم کند. این روش با ترکیب معیارهای مختلف و ایجاد یک ساختار منعطف‌تر برای

تحلیل داده‌ها، به تحلیل‌گران و سرمایه‌گذاران کمک می‌کند تا تصمیمات بهتری در مورد استراتژی‌های معاملاتی خود بگیرند و در مواجهه با نوسانات بازار، بهتر عمل کنند.

۳ – ۶ – ۲ اهمیت برچسب‌گذاری سه‌گانه

این روش با فراهم آوردن یک چارچوب جامع، توییت‌ها و محتوای متنی مرتبط با بازار را بر اساس حرکت‌های واقعی بازار ارزیابی می‌کند. هدف از این رویکرد، این است که به جای تمرکز بر نیت یا محتوای احساسی، تأثیر واقعی آن محتوا را بر پویایی‌های بازار اندازه‌گیری کند. به عبارت دیگر، این روش یک سنجش واقعی از تأثیر توییت بر نوسانات بازار فراهم می‌آورد.

توضیح فنی

۱. فیلتر CUSUM : یکی از اجزای مهم این روش استفاده از فیلتر CUSUM برای تعیین مرزهای بالا و پایین در ابتدای هر پنجره زمانی است. این فیلتر تغییرات قیمتی قابل توجه را تشخیص می‌دهد و در نتیجه، مرزهایی برای تحلیل حرکات قیمتی در چارچوب زمانی مشخص ایجاد می‌کند.
۲. بهینه‌سازی پارامترها :بهینه‌سازی پارامترها هر شش ماه یکبار و بر اساس شاخص نسبت شارپ [۲۲]در استراتژی خرید و نگه‌داری انجام می‌شود. این روند به ما کمک می‌کند تا بهترین پارامترها را برای هر دوره زمانی انتخاب کنیم. در این فرآیند، بازده‌های روزانه و نوسانات قیمت محاسبه شده و سپس مرزهای بالا و پایین تعیین می‌شوند. برچسب‌گذاری مشاهدات بر اساس این است که آیا قیمت در طول پنجره زمانی مشخص به مرزهای بالا یا پایین برخورد کرده است یا خیر.

این روش جامع تضمین می‌کند که مدل‌های ما به داده‌های دقیق و معناداری دسترسی دارند که به پیش‌بینی حرکات واقعی بازار کمک می‌کند. با استفاده از برچسب‌گذاری سه‌گانه، ما قادر خواهیم بود تا تحلیل‌هایی با دقت بیشتر و وابسته به داده‌های واقعی و منسجم انجام دهیم، که این امر به بهبود عملکرد مدل‌های پیش‌بینی بازار منجر می‌شود.

[۲۲] Sharpe Ratio

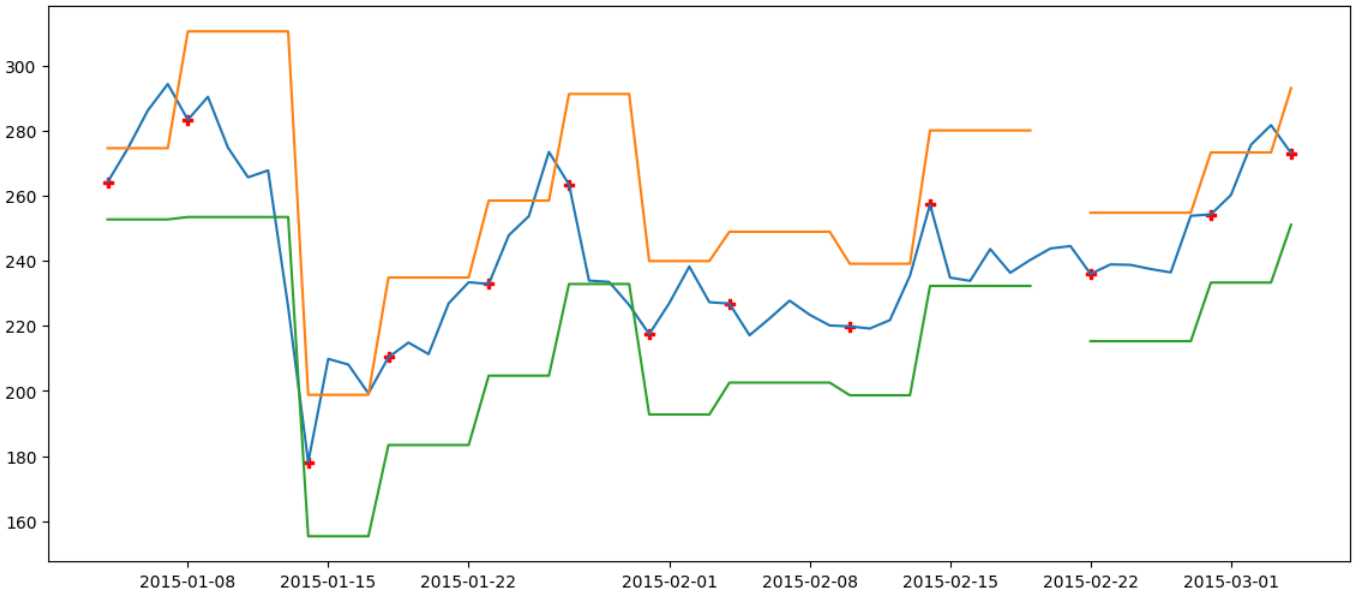

Figure ۲ برچسب زنی سه ماهه به عنوان نمونه جهت بررسی

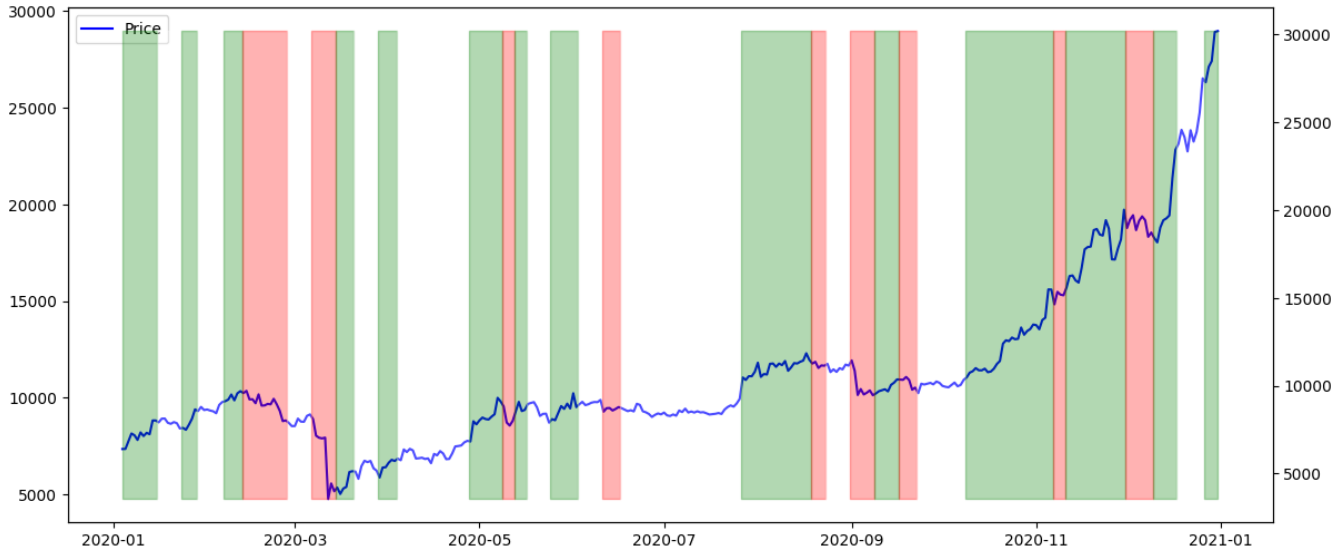

Figure ۳نمونه برچسب زنی TBL

۳ – ۶ – ۳ مدل Cryptobert

۳ – ۶ – ۳ – ۱ توضیح جامع ساختار مدل CryptoBERT

CryptoBERT یک مدل زبان مبتنی بر معماری BERT است که برای تحلیل احساسات در پست‌های مرتبط با ارزهای دیجیتال در رسانه‌های اجتماعی توسعه یافته است. این مدل با هدف طبقه‌بندی احساسات پست‌ها به

سه دسته صعودی(bullish) ، نزولی (bearish) و خنثی طراحی شده است. در ادامه به تشریح دقیق ساختار، مراحل آموزش و جزئیات فنی این مدل می‌پردازیم.

۱ .جمع‌آوری و پیش‌پردازش داده‌ها

مجموعه داده‌ها:

- حجم کل داده‌ها ۳.۲۰۷ :میلیون پست از منابع مختلف.
- منابع داده‌ها:
  - توییتر: ۴۹۶,۰۰۰ پست (از ۱۱ تا ۲۴ ژوئیه ۲۰۱۸).
  - ردیت: ۱۷۲,۰۰۰ پست (از ۱ مه ۲۰۲۱ تا ۳۰ آوریل ۲۰۲۲).
  - تلگرام: ۶۶۴,۰۰۰ پست (از ۱۶ نوامبر ۲۰۲۰ تا ۳۰ ژانویه ۲۰۲۱).

برچسب‌گذاری داده‌ها:

- تنها پست‌های StockTwits دارای برچسب‌های صعودی[۲۳] یا نزولی [۲۴] هستند که توسط نویسندگان آنها تعیین شده‌اند. اگر برچسبی تعیین نشده باشد، احساس خنثی در نظر گرفته می‌شود.
- داده‌های سایر منابع بدون برچسب هستند و فقط برای آموزش بدون نظارت[۲۵] مدل استفاده می‌شوند.

پیش‌پردازش داده‌ها:

- حذف کاراکترهای غیرضروری :حروف چینی، ژاپنی و کره‌ای، آدرس‌های کیف پول کریپتو، لینک‌ها، نمادهای خاص مانند (cashtags) $، (#هشتگ‌ها)، @(یوزرنیم‌ها) و RT(ریتوییت‌ها) حذف می‌شوند.
- اصلاح خطاهای رمزگذاری :خطاهای شناخته‌شده در کاراکترهای خاص، چندین نقطه و فاصله‌ها اصلاح می‌شوند.
- تبدیل به حروف کوچک :تمام کاراکترها به حروف کوچک تبدیل می‌شوند.
- حذف پست‌های تکراری و کوتاه :پست‌های تکراری و پست‌هایی که کمتر از چهار کلمه دارند حذف می‌شوند.

---

[۲۳] bullish

[۲۴] bearish

[۲۵] unsupervised

۳ – ۶ – ۳ – ۲ آموزش مدل پایه(Post-Training)

انتخاب مدل پایه:

- مدل BERTweet به عنوان نقطه شروع انتخاب شده است. این مدل بر اساس معماری BERT بوده و برای داده‌های توییتر بهینه شده است.

روش آموزش بدون نظارت:

- مدل زبانی ماسک شده[۲۶] (MLM) : این روش شامل ماسک کردن تقریباً ۱۵٪ از توکن‌های ورودی و پیش‌بینی آنها بر اساس زمینه است.
- استفاده از روش :RoBERTa برای بهبود فرآیند آموزش، از روش بهینه‌سازی‌شده RoBERTa استفاده می‌شود که تغییراتی در روش اصلی BERT ایجاد کرده است.
- توکنایزر تعبیه جفت بیت [۲۷]: هم BERTweet و هم RoBERTa از این نوع توکنایزر استفاده می‌کنند که امکان کار با کاراکترهای خاص و زبان‌های مختلف را فراهم می‌کند.

جزئیات آموزش:

- مرحله اول آموزش:
  - طول توالی کوتاه :ابتدا مدل با طول توالی ۳۲ توکن آموزش می‌بیند.
  - تعداد اپک‌ها ۱۲۰ :اپک با ۱۰ ماسک مختلف (۱۲ اپک برای هر ماسک).
- مرحله دوم آموزش:
  - طول توالی بلندتر :سپس طول توالی به ۱۲۸ توکن افزایش می‌یابد.
  - تعداد اپک‌ها ۱۲ :اپک اضافی برای طول توالی ۱۲۸.
- :Multiple Maskingبا الهام از کار Liu و همکاران، از ماسک‌گذاری چندگانه در طول آموزش استفاده می‌شود.
- بهینه‌سازی وزن‌ها:
  - بهینه‌ساز :Adam برای بهینه‌سازی پارامترهای مدل از بهینه‌ساز Adam استفاده می‌شود.
  - نرخ یادگیری و هایپرپارامترها :تنظیمات بهینه برای نرخ یادگیری و سایر هایپرپارامترها اعمال می‌شوند.

---

[۲۶] Masked Language Modeling

[۲۷] Byte-Level Encoding

نتیجه آموزش:

- مدل حاصل از این مرحله CryptoBERTنامیده می‌شود که با داده‌های مرتبط با ارزهای دیجیتال آموزش دیده و برای تحلیل‌های بعدی آماده است.

۳ – ۶ – ۳ – ۳ Fine-Tuning برای طبقه‌بندی احساسات

مجموعه داده‌های مورد استفاده:

- مجموعه آموزشی StockTwits :
  - شامل پست‌های مرتبط با سه ارز دیجیتال پرمناقشه: بیت‌کوین(BTC.X) ، اتریوم (ETH.X) و شیبا اینو.(SHIB.X)
  - بازه زمانی: از ۱ نوامبر ۲۰۲۱ تا ۱۵ ژوئن ۲۰۲۲.
  - تعداد پست‌ها: ۱.۳۳۲ میلیون پست.

مشکل عدم توازن کلاس‌ها:

- کلاس‌های احساسات دارای توزیع نامتوازن هستند:
  - صعودی: بزرگترین کلاس با بیشترین تعداد پست‌ها.
  - نزولی: کوچکترین کلاس با کمترین تعداد پست‌ها.
  - خنثی[۲۸] :بین دو کلاس دیگر.

روش‌های متوازن‌سازی داده‌ها:

- نمونه برداری کند [۲۹]:
  - کاهش تعداد نمونه‌های کلاس‌های بزرگتر به اندازه کلاس کوچکتر
  - تعداد نهایی پست‌ها برای هر کلاس: ۱۲۴,۴۵۱ پست.
  - مجموعه داده متوازن با ۳۷۳,۳۵۳ پست برای آموزش.
- نمونه برداری تند[۳۰]:

[۲۸] neutral
[۲۹] Undersampling
[۳۰] Oversampling

- افزایش تعداد نمونه‌های کلاس‌های کوچکتر به اندازه کلاس بزرگتر (bullish) با نمونه‌برداری با جایگذاری.
- تعداد نهایی پست‌ها برای هر کلاس: ۶۷۶,۷۰۱ پست.
- مجموعه داده بزرگ با ۲.۰۳ میلیون پست برای آموزش مدل‌های بزرگتر) برچسب.(XL

فرآیند میزان سازی دقیق[۳۱]:

- مدل CryptoBERT با استفاده از مجموعه داده‌های متوازن شده آموزش می‌بینند.
- تنظیمات آموزش:
  - نسبت تقسیم داده‌ها ٪۱۰ :از داده‌های آموزشی برای اعتبارسنجی کنار گذاشته می‌شود.
  - هایپرپارامترها :تنظیمات بهینه برای تعداد اپک‌ها، نرخ یادگیری و سایر پارامترها اعمال می‌شود.

۴ .ساختار داخلی مدل

معماری مدل:

- تنها رمزگشا : مانند BERT اصلی، CryptoBERTاز بخش انکودر ترانسفورمر استفاده می‌کند که شامل چندین لایه توجه چندسر[۳۲] و شبکه‌های عصبی پیش‌خور [۳۳]است.

ویژگی‌های ورودی:

- توکن‌های ورودی :پس از پیش‌پردازش و توکنایز شدن، پست‌ها به توکن‌های عددی تبدیل می‌شوند.
- جاسازی‌های توکنی :هر توکن با یک بردار جاسازی نشان داده می‌شود.
- جاسازی‌های مکانی :برای حفظ ترتیب توکن‌ها، جاسازی‌های مکانی اضافه می‌شوند.
- جاسازی‌های سگمنت :اگر نیاز به تفکیک بخش‌های مختلف ورودی باشد، از این جاسازی‌ها استفاده می‌شود.

مکانیزم توجه:[۳۴]

---

[۳۱] Fine-Tuning

[۳۲] Multi-Head Attention

[۳۳] Feedforward Neural Network

- توجه به خود[۳۵]: مدل می‌تواند وابستگی‌های بین توکن‌های ورودی را بدون توجه به فاصله آنها تشخیص دهد.
- لایه توجه چندسر :امکان یادگیری روابط مختلف در فضاهای برداری متنوع را فراهم می‌کند.

لایه‌های نرمال‌سازی و اتصال باقیمانده:

- نرمال سازی لایه : به پایداری آموزش کمک می‌کند.
- رابطه های اضافه :جریان گرادیان‌ها را بهبود می‌بخشد و از مشکل ناپدید شدن گرادیان جلوگیری می‌کند.

خروجی مدل:

- بردارهای نهفته :پس از عبور از لایه‌های انکودر، هر توکن به یک بردار نهفته تبدیل می‌شود که نمایانگر معنای آن در متن است.
- توکن :[CLS] برای وظایف طبقه‌بندی، بردار مربوط به این توکن به عنوان نماینده کل جمله یا پست استفاده می‌شود.

۳ – ۶ – ۳ – ۴ استفاده از ویژگی‌های تکنیکال و احساسی

ویژگی‌های اضافی:

- تحلیل احساسات : :مدل قادر است احساسات موجود در متن را تشخیص دهد.
- تأثیرات کوتاه‌مدت[۳۶]: بررسی تأثیرات لحظه‌ای اخبار یا پست‌ها بر بازار.
- برچسب قبلی: استفاده از اطلاعات مربوط به برچسب‌های قبلی برای بهبود پیش‌بینی.
- شاخص‌های تکنیکال:
  - RSI شاخص قدرت نسبی :اندازه‌گیری قدرت حرکات قیمت.
  - ROC نرخ تغییرات :(اندازه‌گیری سرعت تغییرات قیمت.

---

[۳۴] Attention Mechanism

[۳۵] Self-Attention

[۳۶] Short-Term Impact

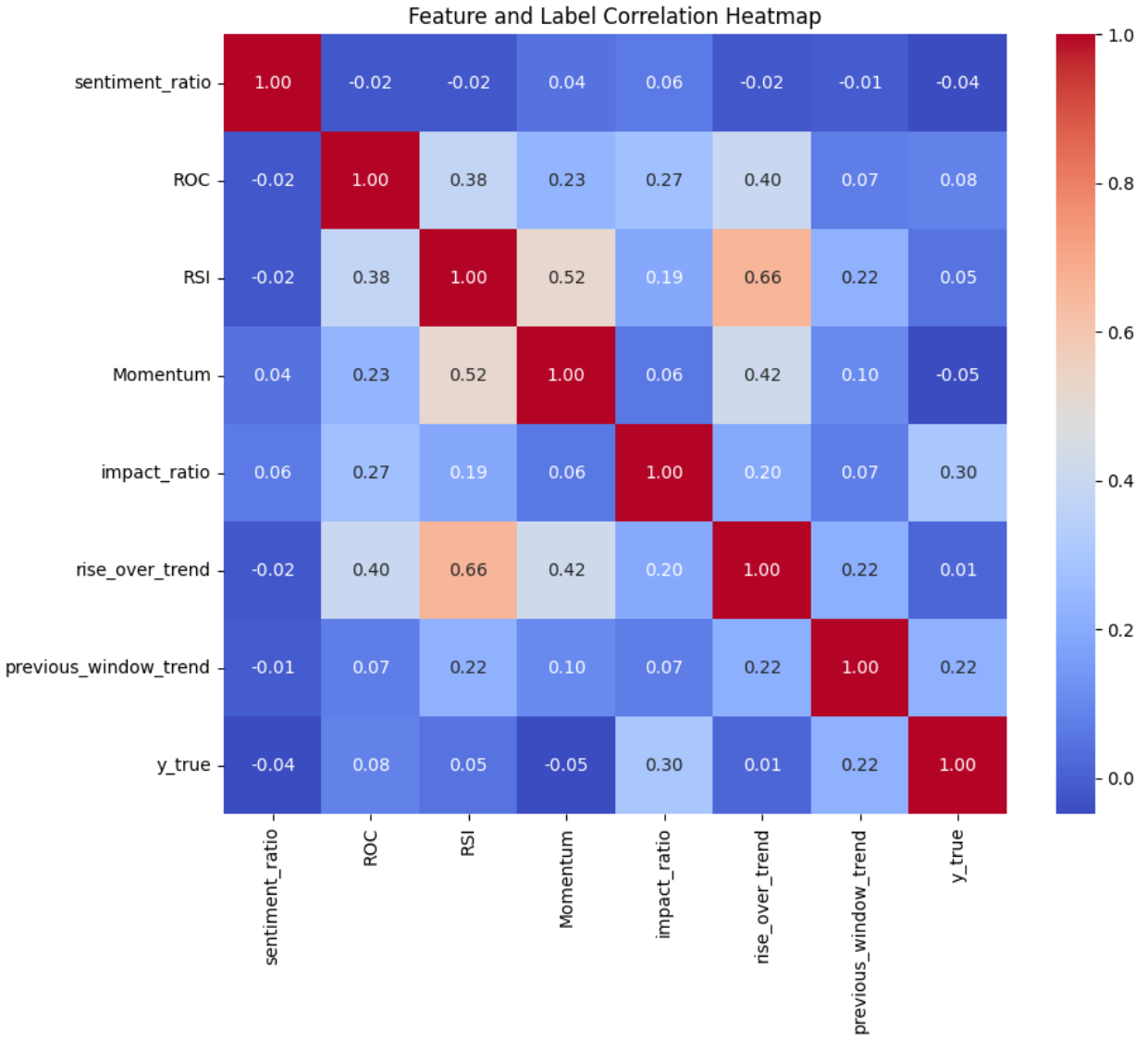

Figure ۴ اثرگذاری ویژگی ها بر هم

ادغام ویژگی‌ها:

- Fusion Layer : ممکن است لایه‌ای برای ترکیب ویژگی‌های متنی و تکنیکال وجود داشته باشد تا مدل بتواند تصمیم‌گیری بهتری انجام دهد.
- نمایش چندوجهی :ترکیب داده‌های متنی و عددی برای بهبود عملکرد مدل.

۳ – ۶ – ۳ – ۵ ارزیابی و مقایسه مدل

مقایسه با مدل‌های دیگر:

- BERTعمومی :مدل اصلی BERT که برای زبان عمومی آموزش دیده است.
- :FinBERTمدل BERT آموزش‌دیده بر داده‌های مالی.
- :BERTweetمدل BERT آموزش‌دیده بر داده‌های توییتر.
- :CryptoBERT XL نسخه بزرگتر مدل.

معیارهای ارزیابی:

- دقت :(Accuracy) نسبت پیش‌بینی‌های درست به کل پیش‌بینی‌ها.
- Precision, Recall, F۱-Score : معیارهای دقیق‌تری برای ارزیابی عملکرد در کلاس‌های نامتوازن.

نتایج:

- CryptoBERT عملکرد بهتری نسبت به مدل‌های دیگر در طبقه‌بندی احساسات پست‌های مرتبط با ارزهای دیجیتال نشان داده است.
- بهبود دقت :استفاده از داده‌های تخصصی و آموزش مجدد مدل باعث افزایش دقت پیش‌بینی‌ها شده است.

منابع و ابزارهای مرتبط

- کدهای منبع :در GitHub به آدرس https://github.com/mikik۱۲۳۴/CryptoBERT-LUKE در دسترس است.
- مدل آموزش‌دیده :CryptoBERT از طریق Huggingface به آدرس https://huggingface.co/ElKulako/cryptobert قابل دانلود است.
- مجموعه داده‌ها:
  - :StockTwits Dataset https://huggingface.co/datasets/ElKulako/stocktwits-crypto
  - :CryptoBERT Post-Training Corpus https://huggingface.co/datasets/ElKulako/cryptobert-posttrain

- StockTwits Emoji Dataset:
https://huggingface.co/datasets/ElKulako/stocktwits-emoji

نتیجه‌گیری

مدل CryptoBERT با بهره‌گیری از معماری قدرتمند BERT و آموزش مجدد بر روی داده‌های مرتبط با ارزهای دیجیتال، توانسته است عملکرد بالایی در طبقه‌بندی احساسات پست‌های رسانه‌های اجتماعی نشان دهد. استفاده از تکنیک‌های پیشرفته مانند مدل سازی ماسک شده زبان، تنظیمات دقیق هایپرپارامترها، در کنار پیش‌پردازش دقیق داده‌ها، باعث شده است که این مدل به عنوان یک ابزار مؤثر در تحلیل احساسات بازار ارزهای دیجیتال مورد استفاده قرار گیرد.

۳ – ۷ تحلیل تکنیکال داده ها

تحلیل تکنیکال به معنای بررسی و تحلیل رفتار قیمت‌ها و حجم معاملات در بازارهای مالی است تا از طریق الگوها و شاخص‌های آماری مختلف، روندهای آینده پیش‌بینی شود. در این بخش، ما از داده‌های تاریخی و اندیکاتورهای مختلفی مانند میانگین متحرک(MA) ، شاخص قدرت نسبی(RSI) ، و نرخ تغییر (ROC) استفاده می‌کنیم تا ویژگی‌های مهمی از داده‌ها استخراج کنیم. این ویژگی‌ها به مدل یادگیری ماشین ورودی داده می‌شوند تا بتواند به پیش‌بینی دقیق‌تر روندهای آینده کمک کند.

۳ – ۷ – ۱ مدل خود رمزگذار[۳۷]

خودرمزگذارها نوعی شبکه عصبی مصنوعی هستند که برای یادگیری فشرده‌سازی داده‌ها و بازسازی آن‌ها از روی نسخه فشرده‌شده استفاده می‌شوند. این مدل‌ها به‌طور خاص در کاربردهایی مانند کاهش ابعاد، تشخیص ناهنجاری‌ها، و استخراج ویژگی‌های مهم داده‌ها استفاده می‌شوند. در این بخش، مدل خودرمزگذار طراحی شده در این تحقیق را بررسی خواهیم کرد.

ساختار مدل

این مدل خودرمزگذار شامل سه بخش اصلی است :**رمزگذار**، **پیش‌بینی‌کننده** (Extrapolator)و **رمزگشا** .
در ادامه به توضیح هر یک از این بخش‌ها می‌پردازیم:

[۳۷] autoencoder

۱. رمزگذار[۳۸] :

- بخش رمزگذار وظیفه دارد تا داده‌های ورودی را به یک نمایش نهان (Latent Representation) فشرده کند. این نمایش نهان یک فضای با ابعاد کمتر از فضای اصلی داده‌هاست که ویژگی‌های اصلی داده‌ها را حفظ می‌کند.
- در کد ارائه شده، این بخش شامل یک شبکه عصبی بازگشتی (RNN) با لایه‌های متعدد است که داده‌ها را به یک فضای پنهان (Hidden Space) تبدیل می‌کند. سپس، این ویژگی‌های فشرده توسط یک لایه‌ی کاملاً متصل (Fully Connected Layer) پردازش می‌شوند.

پیش‌بینی‌کننده[۳۹] :

- این بخش به‌عنوان یک لایه اضافی بین رمزگذار و رمزگشا عمل می‌کند که ویژگی‌های فشرده‌شده را پردازش و اصلاح می‌کند تا مدل بتواند عملکرد بهتری داشته باشد. این بخش از یک لایه کاملاً متصل برای پردازش داده‌های پنهان استفاده می‌کند.

رمزگشا :

- بخش رمزگشا وظیفه دارد که ویژگی‌های فشرده‌شده را به فضای اصلی داده‌ها بازگرداند. در این مدل، رمزگشا نیز شامل یک شبکه عصبی بازگشتی (RNN) و یک لایه‌ی کاملاً متصل است که داده‌ها را به

[۳۸] Encoder

[۳۹] Extrapolator

شکل اصلی خود بازسازی می‌کند

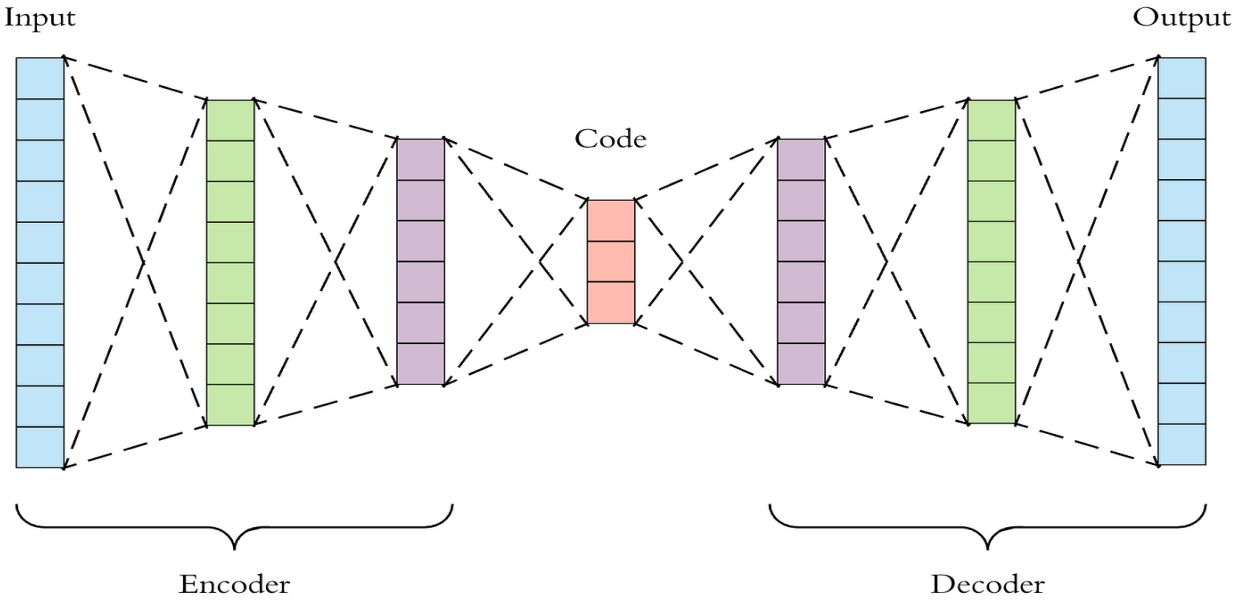

Figure ۵ ساختار اتوانکودر

مدل پیشنهادی شامل سه بخش اصلی است :**انکودر، اکستراپولاتور** و **دیکودر،** که هر کدام وظیفه‌ای مشخص در پردازش داده‌های ورودی بر عهده دارند.

انکودر:

انکودر مسئول پردازش دنباله داده‌های ورودی و تبدیل آن به یک نمایش مخفی (نهان) است. این بخش شامل یک لایه شبکه عصبی بازگشتی[۴۰] (RNN) و یک لایه کاملاً متصل (Fully Connected) است. ابتدا، لایه RNNداده‌های ورودی را با بعد مشخص پردازش کرده و آن را از طریق چندین لایه مخفی عبور می‌دهد. خروجی حاصل از RNN سپس از طریق لایه کاملاً متصل عبور داده می‌شود که نمایشی دقیق‌تر از دنباله ورودی ارائه می‌کند. استفاده از RNN به مدل کمک می‌کند تا وابستگی‌های زمانی موجود در داده‌های دنباله‌دار را کشف کند و مدل را برای پردازش داده‌های ترتیبی مؤثرتر سازد.

اکستراپولاتور:

[۴۰] Recurrent neural network

اکستراپولاتور بخشی ساده‌تر از مدل است که حالت مخفی تولید شده توسط انکودر را دریافت کرده و آن را از طریق یک لایه کاملاً متصل پردازش می‌کند. این بخش به بهبود و پیش‌بینی ویژگی‌های رمزگذاری شده کمک می‌کند و اطمینان حاصل می‌کند که دیکودر ورودی به خوبی ساختار یافته‌ای دریافت می‌کند. وظیفه اصلی اکستراپولاتور تقویت ویژگی‌های رمزگذاری شده به‌گونه‌ای است که برای بخش‌های بعدی مدل آماده باشد.

دیکودر:

دیکودر، مشابه انکودر، از یک لایه RNN و یک لایه کاملاً متصل تشکیل شده است. حالت مخفی که توسط اکستراپولاتور پردازش شده است به لایه‌های RNN دیکودر تغذیه می‌شود. وظیفه دیکودر تولید دنباله خروجی بر اساس نمایش‌های مخفی تولید شده توسط انکودر است. پس از عبور از لایه‌هایRNN ، خروجی از طریق لایه کاملاً متصل به بعد مورد نظر پروژه شده و به شکل خروجی نهایی درمی‌آید. ساختار دیکودر به شکلی متقارن با انکودر است تا روابط و وابستگی‌های زمانی یاد گرفته شده در ورودی، در خروجی نیز منعکس شوند.

روند آموزش مدل

۱. بهینه‌سازی :برای آموزش مدل، از ترکیب سه بخش رمزگذار، پیش‌بینی‌کننده و رمزگشا استفاده می‌شود. هدف از آموزش مدل کاهش خطای بازسازی داده‌ها است؛ یعنی مدلی آموزش داده می‌شود که خروجی نهایی آن کمترین تفاوت ممکن را با ورودی اصلی داشته باشد. این هدف با استفاده از الگوریتم بهینه‌سازی Adamو تابع خطای MSEانجام می‌شود.
۲. مراحل آموزش :داده‌های ورودی به رمزگذار داده می‌شوند، که آنها را به یک نمایش نهان تبدیل می‌کند. سپس، این نمایش نهان توسط پیش‌بینی‌کننده پردازش شده و در نهایت به رمزگشا داده می‌شود تا بازسازی داده‌ها انجام شود. با مقایسه خروجی بازسازی‌شده با داده‌های ورودی اصلی، خطا محاسبه و مدل به‌روزرسانی می‌شود.

برای بهینه‌سازی مدل پیشنهادی، از یک تابع هزینه و بهینه‌ساز مناسب استفاده می‌شود. در این بخش، از تابع زیان میانگین مربعات خطا (MSE) به عنوان معیار اندازه‌گیری خطا بین خروجی پیش‌بینی شده و مقادیر واقعی استفاده می‌شود. این تابع زیان میزان اختلاف را به شکل مربعی محاسبه می‌کند که باعث می‌شود خطاهای بزرگ‌تر تأثیر بیشتری در بهینه‌سازی مدل داشته باشند. این تابع برای مسائل رگرسیون بسیار مناسب است و به خوبی می‌تواند انحراف پیش‌بینی‌ها از مقادیر واقعی را مشخص کند.

به منظور بهینه‌سازی پارامترهای مدل، از الگوریتم Adam استفاده می‌شود. این الگوریتم به‌عنوان یکی از الگوریتم‌های پیشرفته بهینه‌سازی شناخته می‌شود که ترکیبی از مزایای روش‌های مومنتوم و رسمی‌سازی RMSprop را به همراه دارد. Adam با تنظیم خودکار نرخ یادگیری و استفاده از میانگین موزون لحظه‌ای گرادیان‌ها و مربعات آنها، به فرآیند بهینه‌سازی سرعت می‌بخشد و از نوسانات زیاد در به‌روزرسانی پارامترها جلوگیری می‌کند.

در این روش، پارامترهای مربوط به انکودر، اکستراپولاتور و دیکودر به صورت یکجا ترکیب شده و به الگوریتم Adam برای به‌روزرسانی و بهینه‌سازی ارائه می‌شوند. نرخ یادگیری این بهینه‌ساز نیز برابر با مقدار ۰.۰۰۱ تنظیم شده است که به آرامی و با دقت به سمت کمینه‌سازی تابع زیان حرکت می‌کند.

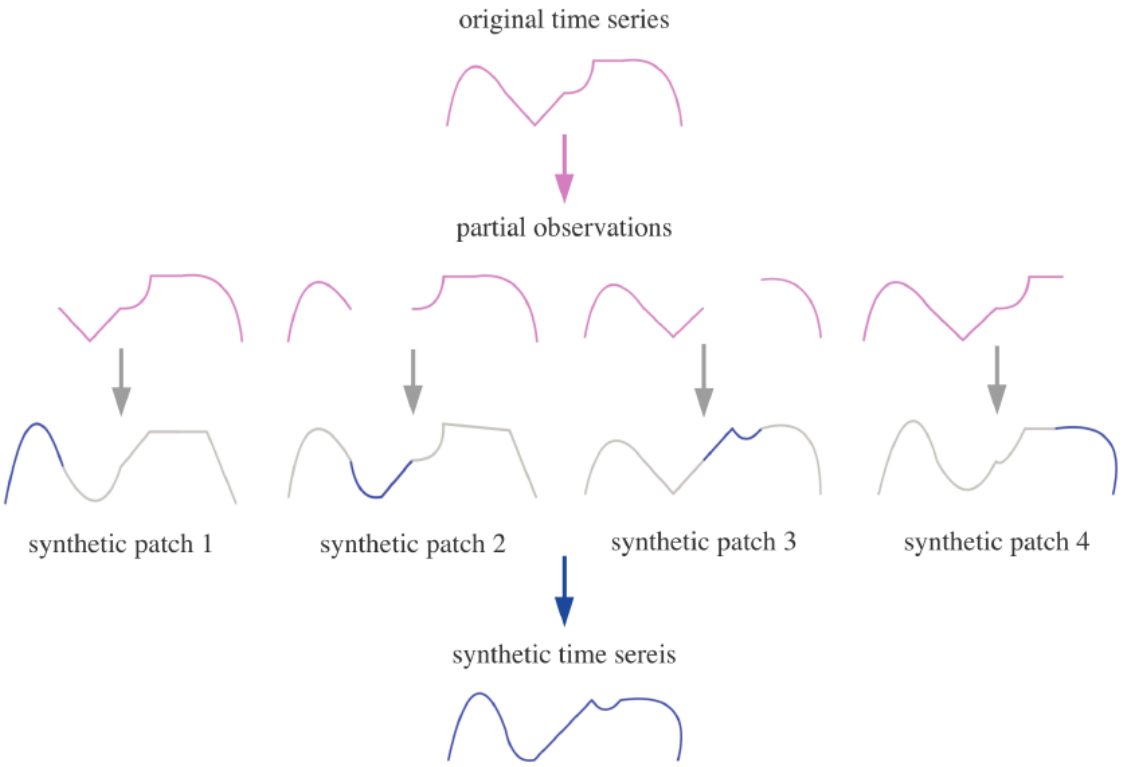

Figure ۶ نحوه یادگیری مدل انکودر [۳۶]

نتیجه‌گیری

خودرمزگذارهای ارائه شده در این تحقیق به‌عنوان ابزاری قدرتمند برای فشرده‌سازی داده‌ها و استخراج ویژگی‌های مهم از آنها استفاده می‌شوند. این مدل‌ها امکان کاهش ابعاد داده‌ها و در عین حال حفظ ویژگی‌های

کلیدی را فراهم می‌کنند که می‌تواند در مراحل بعدی برای تحلیل‌های دقیق‌تر و پیش‌بینی‌های دقیق‌تر استفاده شود.

۳ – ۷ – ۲ مدل تصمیم گیر(decider)

مدل تصمیم‌گیر در این تحقیق، وظیفه دسته‌بندی داده‌های ورودی را پس از استخراج ویژگی‌های نهفته از مدل رمزگذار (Autoencoder) بر عهده دارد. برای این منظور، ابتدا ویژگی‌های نهفته با استفاده از مدل رمزگذار به دست می‌آیند. از این ویژگی‌ها برای ورودی مدل تصمیم‌گیر استفاده می‌شود. در اینجا از یک طبقه‌بند چندلایه (MLP Classifier)استفاده شده است که شامل چندین لایه کاملاً متصل (Fully Connected) است.

استخراج ویژگی‌های نهفته

ویژگی‌های نهفته از داده‌های سری زمانی توسط رمزگذار استخراج می‌شوند. با تقسیم داده‌ها به تکه‌های کوچک‌تر و اعمال مدل رمزگذار، ویژگی‌های مربوطه استخراج و برای هر پچ زمانی به دست می‌آید. این ویژگی‌ها در نهایت با استفاده از تابعی که در زیر آمده است، در قالب یک بردار نهفته یکپارچه می‌شوند:

برای استخراج الگوهای پنهان از داده‌های سری زمانی، از روشی مبتنی بر شبکه عصبی انکودر استفاده می‌شود. این فرآیند شامل تقسیم داده‌های ورودی به بخش‌های کوچکتر به نام پچ و سپس اعمال انکودر بر روی این بخش‌ها است. هر پچ از داده‌ها، بسته به طول آن (که با اندازه پچ مشخص می‌شود)، به‌صورت ماسک شده به انکودر داده می‌شود. انکودر مسئول استخراج ویژگی‌های پنهان از سری داده ماسک شده است.

برای انجام این کار، مراحل زیر انجام می‌شود:

داده‌های سری زمانی ورودی به چندین پچ با اندازه مشخص تقسیم می‌شوند.

هر پچ به‌صورت ماسک شده (که به این معناست که بخشی از داده‌ها پنهان یا حذف می‌شوند) به انکودر داده می‌شود.

انکودر پس از پردازش این داده‌ها، ویژگی‌های پنهانی را استخراج کرده و به عنوان خروجی ارائه می‌دهد.

ویژگی‌های پنهان استخراج شده از هر پچ به صورت یک ماتریس به هم پیوسته و نهایی ترکیب می‌شوند که نماینده کلی الگوهای پنهان سری زمانی است.

این ویژگی‌های پنهان می‌توانند در مدل‌های بعدی برای پیش‌بینی یا تحلیل‌های بیشتر مورد استفاده قرار گیرند و از اهمیت بالایی در تحلیل سری‌های زمانی برخوردارند.در اینجا، مدل رمزگذار با استفاده از توالی‌های داده‌های سری زمانی، ویژگی‌های نهفته را استخراج کرده و در نهایت یک بردار نهفته ایجاد می‌کند.

طبقه‌بندی و آموزش

برای طبقه‌بندی، از مدل طبقه‌بند MLP استفاده می‌شود. این مدل دارای سه لایه کاملاً متصل با تابع فعال‌سازی ReLU است که به دنبال شناسایی کلاس‌های مختلف خروجی می‌باشد. فرمول ساختار طبقه‌بند MLPبه شکل زیر است:

برای ایجاد یک مدل تصمیم‌گیر، از یک شبکه عصبی چندلایه (MLP) استفاده شده است. این مدل شامل لایه‌های متعددی است که وظیفه دسته‌بندی داده‌های ورودی را بر عهده دارند. ساختار مدل به شرح زیر است:

۱. **لایه ورودی** :اولین لایه مدل، داده‌های ورودی با تعداد ویژگی‌های مشخص را دریافت می‌کند و آن‌ها را به فضای ویژگی‌های پنهان منتقل می‌کند. این لایه یک لایه تمام‌متصل [۴۱] است که ابعاد ورودی را به تعداد نورون‌های لایه پنهان تبدیل می‌کند.

۲. **لایه فعال‌ساز** :پس از هر لایه تمام‌متصل، از تابع فعال‌سازی ReLU استفاده می‌شود که به‌عنوان یک تابع غیرخطی به مدل کمک می‌کند تا الگوهای پیچیده‌تر را شناسایی کند.

۳. **لایه پنهان** :یک لایه دیگر به عنوان لایه پنهان وجود دارد که داده‌ها را پس از عبور از لایه ورودی پردازش می‌کند. این لایه نیز از همان معماری تمام‌متصل به همراه تابع ReLU بهره می‌برد.

۴. **لایه خروجی** :در نهایت، لایه خروجی داده‌های پردازش شده را به تعداد دسته‌های نهایی (کلاس‌ها) تبدیل می‌کند. این لایه مسئول ارائه پیش‌بینی نهایی است.

مدل پس از طراحی با استفاده از داده‌های آموزشی بهینه‌سازی و آموزش داده می‌شود تا بتواند تصمیم‌گیری‌های دقیقی را در مسائل دسته‌بندی انجام دهد.

در این مدل، داده‌های آموزشی پس از پردازش توسط رمزگذار و طبقه‌بند برای یادگیری به کار می‌رود و در طی چند دوره آموزش، مدل بهینه‌سازی می‌شود.

[۴۱] Fully Connected

۳ – ۸ مدل فیوژن

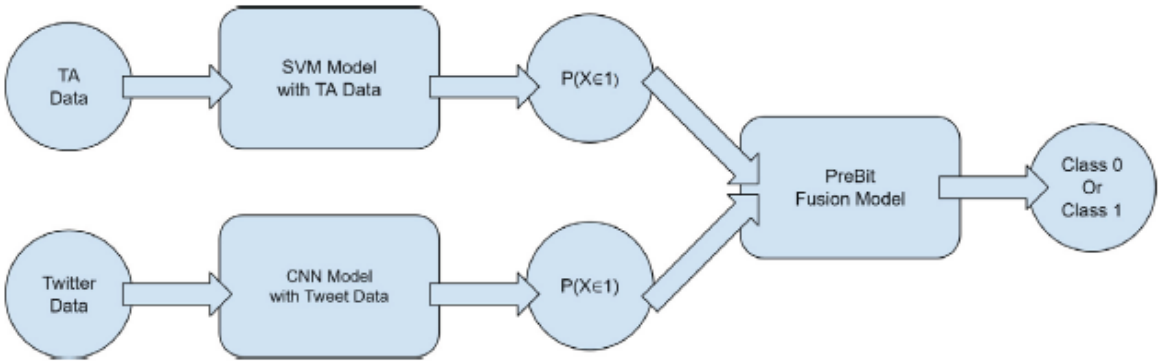

Figure ۷ شکل نهایی مدل های بررسی شده [۱۸]

مدل فیوژن در این پژوهش به منظور ترکیب احتمالات به دست آمده از دو منبع متفاوت، یعنی **CryptoBERT**و **SVM**، و در نهایت پیش‌بینی برچسب نهایی داده‌ها استفاده شده است. این مدل فیوژن یک شبکه عصبی ساده است که با استفاده از ویژگی‌های ورودی، خروجی نهایی را پیش‌بینی می‌کند. در اینجا از یک لایه فیوژن شامل لایه‌های کاملاً متصل استفاده شده که در نهایت با یک تابع **Softmax**، پیش‌بینی دسته‌بندی نهایی انجام می‌شود.

ساختار مدل به صورت زیر پیاده‌سازی شده است:

در این مدل، از یک معماری ساده اما قدرتمند برای ادغام (فیوژن) اطلاعات مختلف استفاده شده است. ساختار مدل فیوژن به‌گونه‌ای طراحی شده است که داده‌های ورودی را به‌صورت بهینه پردازش کرده و نتیجه نهایی را ارائه دهد. مراحل کار مدل فیوژن به شرح زیر است:

۱. **لایه ورودی** :داده‌های ورودی به مدل از طریق یک لایه تمام‌متصل به لایه پنهان منتقل می‌شوند. این لایه وظیفه دارد تا داده‌های ورودی را با تعداد نورون‌های مشخص پردازش کند.

۲. **تابع فعال‌سازی (ReLU)** : پس از انتقال داده‌ها به لایه پنهان، از تابع فعال‌سازی ReLU استفاده می‌شود. این تابع به مدل کمک می‌کند تا روابط غیرخطی در داده‌ها را شناسایی کرده و الگوهای پیچیده را بهتر استخراج کند.

۳. **لایه پنهان** :داده‌های پردازش‌شده در لایه پنهان به لایه خروجی منتقل می‌شوند. این لایه نیز یک لایه تمام‌متصل است که خروجی نهایی مدل را محاسبه می‌کند.

۴. **تابع :Softmax** در انتهای مدل، از تابع Softmax برای دسته‌بندی استفاده می‌شود. این تابع به هر دسته احتمالی تخصیص می‌دهد که مجموع این احتمالات برابر ۱ است. این مرحله کمک می‌کند تا مدل پیش‌بینی دقیقی ارائه دهد و مشخص کند کدام کلاس (برچسب) بیشترین احتمال را دارد.

این مدل به‌عنوان یک ابزار کارآمد برای ادغام داده‌ها از چند منبع استفاده می‌شود و به ما امکان می‌دهد تا نتیجه‌ای دقیق و شفاف از داده‌های چندوجهی به‌دست آوریم.

در این مدل، ابتدا ویژگی‌های ورودی از لایه اول شبکه عبور کرده و به وسیله تابع فعال‌سازی **ReLU** پردازش می‌شوند. سپس، خروجی به لایه دوم ارسال شده و در نهایت با استفاده از تابع **Softmax**، احتمالات دسته‌بندی نهایی به دست می‌آید. این مدل از خروجی‌های مدل‌های قبلی برای نهایی‌سازی تصمیمات دسته‌بندی بهره می‌برد.

۳ – ۹ استراتژی معامله

در بخش **استراتژی معامله**، از کتابخانه **Backtrader** برای توسعه، آزمایش و اجرای استراتژی‌های معاملاتی استفاده شده است. این کتابخانه که یکی از محبوب‌ترین ابزارها در زمینه معامله‌گری الگوریتمی است، بستری انعطاف‌پذیر برای شبیه‌سازی و بررسی استراتژی‌های معاملاتی با استفاده از داده‌های تاریخی فراهم می‌کند . Backtraderبه کاربران این امکان را می‌دهد تا با داده‌های واقعی و گذشته‌نگر، عملکرد استراتژی‌های خود را ارزیابی کرده و آن‌ها را قبل از پیاده‌سازی در معاملات زنده بهینه‌سازی کنند.

اجزای کلیدی Backtrader

۱. **موتور** :**Cerebro** این موتور اصلی Backtrader است که وظیفه مدیریت داده‌ها، اجرای استراتژی‌ها و شبیه‌سازی معاملات را بر عهده دارد Cerebro .وظیفه اجرای سفارشات خرید و فروش، ثبت نتایج، و حتی ترسیم نمودارهای عملکرد استراتژی را انجام می‌دهد. به‌علاوه، کاربران می‌توانند با استفاده از این موتور، تنظیماتی همچون میزان سرمایه اولیه و کارمزدهای کارگزار را تعریف کنند.
۲. **کلاس استراتژی** :استراتژی‌ها در Backtrader به صورت کلاس‌هایی تعریف می‌شوند که در آن‌ها منطق معاملاتی پیاده‌سازی می‌شود. کاربران می‌توانند شرایط خرید و فروش، سیگنال‌ها و نحوه اجرای سفارشات را در این کلاس‌ها تعریف کنند. این استراتژی‌ها می‌توانند شامل شاخص‌های مختلفی مانند میانگین متحرک ساده(SMA) ، شاخص قدرت نسبی (RSI) و یا استراتژی‌های ترکیبی باشند.
۳. **داده‌های تاریخی** Backtrader :قابلیت دریافت داده‌های تاریخی از منابع مختلف همچون Yahoo Financeیا فایل‌های CSV را دارد. این داده‌ها به کاربران امکان می‌دهد که عملکرد استراتژی‌های

معاملاتی خود را با داده‌های واقعی گذشته آزمایش کنند و از آن‌ها برای پیش‌بینی بازارهای آینده استفاده نمایند.

استراتژی‌های معاملاتی معمول:

یکی از رایج‌ترین استراتژی‌هایی که می‌توان با استفاده از Backtrader پیاده‌سازی کرد، **تقاطع میانگین‌های متحرک** است. در این روش، دو میانگین متحرک (یکی سریع و یکی کند) برای پیش‌بینی نقاط خرید و فروش استفاده می‌شود. به عنوان مثال، زمانی که میانگین متحرک سریع از میانگین متحرک کند بالاتر می‌رود، سیگنال خرید صادر می‌شود، و بالعکس زمانی که میانگین متحرک سریع از میانگین کند پایین‌تر می‌آید، سیگنال فروش صادر می‌گردد. این استراتژی‌ها به دلیل سادگی و کارایی در تحلیل‌های تکنیکال بسیار محبوب هستند.

اجرا و بهینه‌سازی استراتژی:

یکی از مزایای استفاده از Backtrader این است که می‌توان استراتژی‌های مختلف را به‌صورت همزمان آزمایش کرد و نتایج آن‌ها را با هم مقایسه نمود. به عنوان مثال، پس از تعریف یک استراتژی معاملاتی، با استفاده از داده‌های تاریخی، می‌توان عملکرد آن را بررسی کرده و تغییرات مورد نیاز را برای بهبود آن اعمال کرد. این فرآیند آزمایش و بهینه‌سازی به معامله‌گران این امکان را می‌دهد که استراتژی‌هایی با ریسک کمتر و بازدهی بیشتر طراحی و پیاده‌سازی کنند.

۳ – ۱۰ بازآزمایی

بازآزمایی یکی از مفاهیم کلیدی است که به‌طور گسترده مورد استفاده قرار می‌گیرد. بازآزمایی فرایندی است که در آن یک استراتژی معاملاتی یا مدلی که برای پیش‌بینی قیمت‌ها یا تحرکات بازار طراحی شده، با استفاده از داده‌های تاریخی گذشته ارزیابی می‌شود تا عملکرد آن در شرایط واقعی سنجیده شود. این روش به پژوهشگران و معامله‌گران اجازه می‌دهد تا بدون ریسک‌کردن سرمایه واقعی، نحوه عملکرد استراتژی خود را بررسی کنند.

در بازآزمایی، استراتژی تعریف‌شده بر روی داده‌های تاریخی بازار پیاده‌سازی می‌شود و بر اساس نتایجی که حاصل می‌شود، معیارهای مختلفی مانند **بازده کل**[۴۲] ، **نسبت شارپ**[۴۳] ، **حداکثر کاهش سرمایه**[۴۴] و **تعداد معاملات بسته‌شده**[۴۵] اندازه‌گیری می‌گردد. این شاخص‌ها به تحلیل‌گران کمک می‌کنند تا بتوانند توازن میان بازدهی و ریسک را در استراتژی‌های معاملاتی خود ارزیابی کنند.

یکی از اهداف اصلی بازآزمایی، شناسایی مشکلات احتمالی در یک استراتژی و جلوگیری از وقوع زیان‌های بالقوه در معاملات واقعی است. همچنین این روش به معامله‌گران کمک می‌کند تا استراتژی‌های خود را بهبود بخشیده و با تنظیمات دقیق‌تری در شرایط واقعی بازار استفاده کنند. با این حال، باید توجه داشت که بازآزمایی تنها معیاری از عملکرد گذشته است و تضمینی برای عملکرد آینده نیست.

به‌طور خلاصه، بازآزمایی ابزاری قدرتمند برای ارزیابی کارایی استراتژی‌های معاملاتی است و به معامله‌گران کمک می‌کند تا با تحلیل دقیق داده‌های تاریخی، استراتژی‌های بهتر و کم‌ریسک‌تری برای معاملات خود اتخاذ کنند.

تحلیل بازآزمایی چهار استراتژی معاملاتی شامل **خرید و نگه‌داری(B&H)** ، **استراتژی ورود و خروج (سیگنال مثبت)**، **استراتژی ورود و خروج (سیگنال منفی)** و **برچسب‌گذاری سه‌گانه(TBL)** ، بین بازه زمانی **اول ژانویه ۲۰۱۵ تا اول فوریه ۲۰۲۱** نتایج ارزشمندی را درباره عملکرد آن‌ها به دست می‌دهد. این نتایج در جدول زیر خلاصه شده است.

استراتژی **خرید و نگه‌داری (B&H)** بیشترین بازده روزانه را به دست آورد، اما ریسک قابل‌توجهی نیز متحمل شد، که این امر با میزان حداکثر کاهش سرمایه (Max DD) بالای آن مشخص می‌شود. این نتایج نشان‌دهنده وجود تعادلی میان بازدهی و ریسک در این رویکرد منفعلانه سرمایه‌گذاری است.

در استراتژی‌های **ورود و خروج**، استراتژی سیگنال مثبت بازدهی بهتری نسبت به سیگنال منفی داشت و نسبت شارپ بالاتری را نشان داد. تعداد معاملات بسته شده نیز در این دو استراتژی متعادل بود. استراتژی سیگنال منفی اگرچه بازدهی کمتری داشت، اما مدیریت ریسک بهتری ارائه داد، که نشان‌دهنده کارآمدی استخراج سیگنال با استفاده از برچسب‌های میانگین است.

---

[۴۲] Total Return

[۴۳] Sharpe Ratio

[۴۴] Max Drawdown

[۴۵] Closed Trades

استراتژی **برچسب‌گذاری سه‌گانه (TBL)** از نظر بازدهی تعدیل شده به ریسک عملکرد بهتری نسبت به سایر استراتژی‌ها داشت و بالاترین نسبت شارپ را به دست آورد. این استراتژی با کاهش سرمایه متوسط و تعداد معاملات بسته‌شده بالا، مدیریت ریسک مؤثری را نشان داد و نشان‌دهنده ثبات و قابلیت اطمینان آن در پیش‌بینی حرکات بازار بود.

این نتایج اهمیت توجه به هر دو عامل بازدهی و ریسک در ارزیابی استراتژی‌های معاملاتی را برجسته می‌کند. عملکرد برتر استراتژی TBL از نظر بازدهی تعدیل‌شده به ریسک، آن را به‌عنوان یک رویکرد امیدبخش برای پیش‌بینی و معامله در بازارهای مالی مطرح می‌کند. همچنین این نتایج پتانسیل ادغام مدل‌های پیشرفته زبان و اطلاعات متنی را برای بهبود دقت پیش‌بینی در بازارهای مالی نشان می‌دهد.

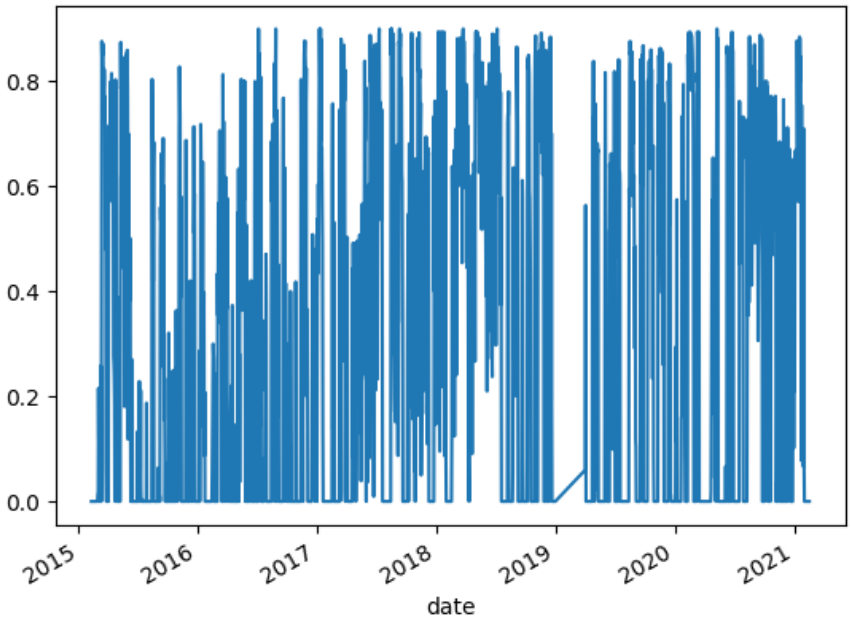

Figure ۸تعداد سیگنال ها جهت خرید و فروش در بازه 6 سال

شکل ۱۰ تعداد سیگنال‌های خرید و فروش را در یک بازه زمانی ۶ ساله از سال ۲۰۱۵ تا ۲۰۲۱ نشان می‌دهد. این نمودار نوسانات فرکانس سیگنال‌ها را در طول زمان به تصویر می‌کشد، به طوری که در برخی از دوره‌ها، به ویژه در سال‌های ۲۰۱۷ و ۲۰۱۸، تراکم بالای سیگنال‌ها دیده می‌شود. این تراکم نشان‌دهنده فعالیت شدید در بازار و تولید تعداد زیادی سیگنال در بازه زمانی کوتاه است. در مقابل، دوره‌هایی مانند سال ۲۰۱۹ و اوایل ۲۰۲۰ وجود دارند که در آن‌ها سیگنال‌های کمتری تولید شده است و به وضوح کاهش فعالیت بازار یا کاهش واکنش مدل به تغییرات بازار را نشان می‌دهند.

این نمودار به وضوح نشان‌دهنده الگوهای مختلف فعالیت در بازارهای مالی است، به طوری که در برخی از سال‌ها و دوره‌ها بازار نوسانات بیشتری را تجربه کرده و بنابراین تعداد بیشتری از سیگنال‌های خرید و فروش تولید شده است. همچنین وجود فاصله‌ها و نوسانات در این سیگنال‌ها ممکن است نمایانگر تغییرات در استراتژی‌های معاملاتی، تنظیمات الگوریتم یا تغییرات بازارهای مالی و میزان نوسانات آن‌ها باشد.

فصل ۴ : نتایج و تفسیر آنها

۴ – ۱ مقدمه

در این فصل به ارائه و تفسیر نتایج مدل‌های پیشنهادی پرداخته می‌شود. با توجه به اهداف پژوهش و اهمیت پیش‌بینی دقیق حرکات بازار مالی، نتایج به‌دست‌آمده از سه رویکرد مختلف تحلیل داده‌های متنی و قیمتی بررسی خواهد شد. نخست، نتایج دسته‌بندی میل بازار بر اساس متن ارائه می‌شود که شامل تحلیل احساسات و استخراج ویژگی‌های متنی است. سپس، دسته‌بندی میل بازار بر اساس داده‌های قیمتی بررسی خواهد شد. در ادامه، عملکرد مدل اتوانکودر (Autoencoder) به عنوان ابزاری برای استخراج ویژگی‌های مخفی از داده‌های سری‌زمانی و متنی مورد تحلیل قرار می‌گیرد.

در نهایت، نتایج استراتژی معاملاتی مورد استفاده در این پژوهش که بر مبنای داده‌های پیش‌بینی شده و رویکردهای تحلیل تکنیکال و داده‌های متنی است، ارائه و تحلیل خواهد شد. نتایج این فصل به تبیین کارایی روش‌های پیشنهادی در بهبود پیش‌بینی‌های کوتاه‌مدت و بهینه‌سازی تصمیمات معاملاتی پرداخته و اهمیت استفاده از مدل‌های پیشرفته مانند شبکه‌های عصبی و یادگیری عمیق در تحلیل مالی را برجسته خواهد کرد.

۴ – ۲ دسته بندی میل بازار بر اساس متن

در این بخش به تحلیل تأثیر کوتاه‌مدت توییت‌ها بر بازار با استفاده از روش "برچسب‌گذاری سه‌گانه "پرداخته می‌شود. هدف از این روش، ارتباط دادن توییت‌ها به حدود زمانی و سود و زیان است، به‌گونه‌ای که مفاهیم "نزولی"، "صعودی" و "خنثی" به‌طور دقیق تعریف شده و قابل استفاده مجدد باشند. فرضیه ما این است که این روش می‌تواند منجر به دسته‌بندی بهتری برای پیش‌بینی روندهای روز بعد نسبت به تحلیل احساسات سنتی شود که تعاریف مبهمی از احساسات صعودی، نزولی و خنثی دارد.

این روش با محدودیت‌هایی نیز همراه است. از جمله فرض این‌که هر توییت تأثیر کوتاه‌مدتی بر بازار دارد، که همیشه صحیح نیست. همچنین، این روش فرض می‌کند که حرکات بازار از نیت‌های خاصی ناشی می‌شوند، در حالی که ممکن است چنین نباشد، و یا توییت‌ها ممکن است از نظر زمانی به بازار مرتبط نباشند. با این وجود، هدف این است که مشخص شود آیا دسته‌بندی توییت‌ها بر اساس حرکات کوتاه‌مدت بازار می‌تواند به پیش‌بینی حرکات مشابه با داده‌های دیده‌نشده کمک کند یا خیر.

داده‌های مورد استفاده در این بخش شامل توییت‌های مربوط به سال ۲۰۲۰ است که از دیتاست **PreBit**توسط **Zou et al.**به‌دست آمده‌اند. این دیتاست شامل حدود ۶۰ هزار توییت است که به منظور تعادل داده‌ها به حدود ۴۰ هزار توییت کاهش یافته‌اند. برای تنظیم مدل، از روش اعتبارسنجی متقابل [۴۶]با ۵ بخش استفاده شده است که از روش "k-fold" طبقه‌بندی‌شده" بهره می‌برد. معیارهای ارزیابی شامل **دقت، دقت دسته‌بندی، یادآوری، امتیازF1 ، خسارت آنتروپی متقاطع، منحنی ROC** و **ماتریس اغتشاش** هستند. نتایج برای مدل پایه و مراحل آموزش و ارزیابی (برای همه دسته‌ها) و همچنین مجموعه داده خارج از نمونه (داده‌های مربوط به سال‌های ۲۰۱۵ تا ۲۰۲۱) ارائه می‌شود.

نتایج این تحلیل نشان می‌دهد که مدل پایه در ارزیابی با دقت ٪۳۳.۳ در دیتاست ۲۰۲۰ و ٪۳۳.۱ در دیتاست خارج از نمونه عملکرد ضعیفی داشته است. مدل ناآگاه به محیط[۴۷] (CUA)نیز در ارزیابی خارج از نمونه افت قابل‌توجهی نسبت به عملکرد آن در سال ۲۰۲۰ نشان داده و دقت آن در ارزیابی ٪۴۴.۱ برای دیتاست ۲۰۲۰ و ٪۳۵.۱ برای دیتاست خارج از نمونه بوده است. این کاهش شدید در عملکرد مدل CUA نشان‌دهنده وابستگی زمانی قوی آن است، به این معنا که این مدل به‌سختی قادر است داده‌های خارج از بازه زمانی آموزش دیده را تعمیم دهد. ولی به صورت کلی میتوان آن را بهتر از مدل ساده در نظر گرفت.

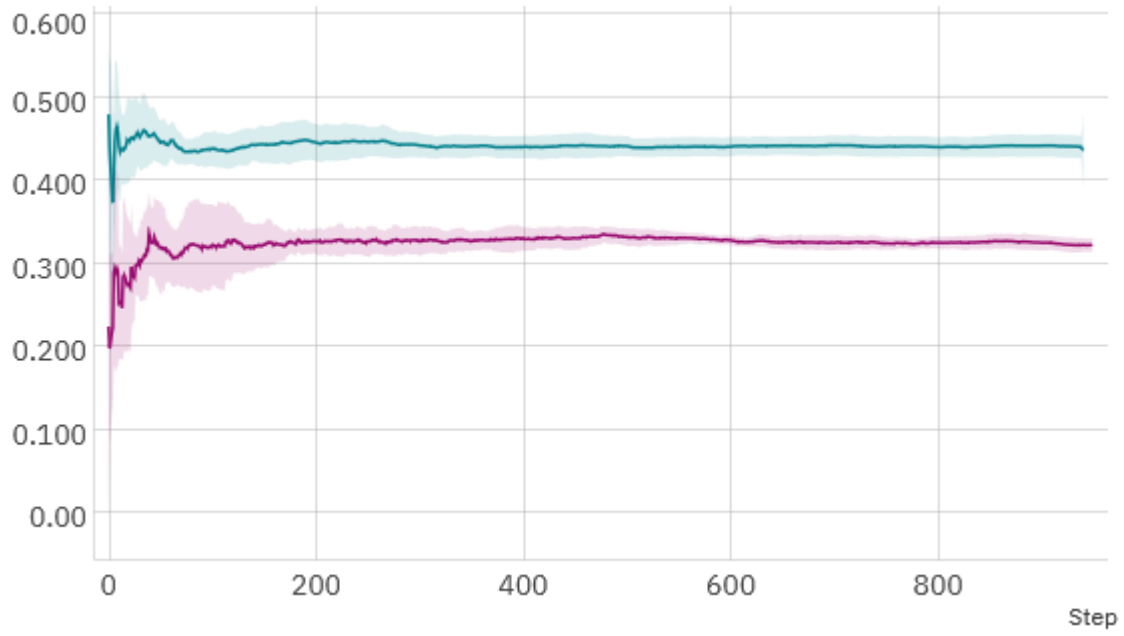

Figure ۹ بررسی precision مدل پایه(بنفش) و مدل CUA (سبز)

[۴۶] Cross Validation

[۴۷] Context Unaware

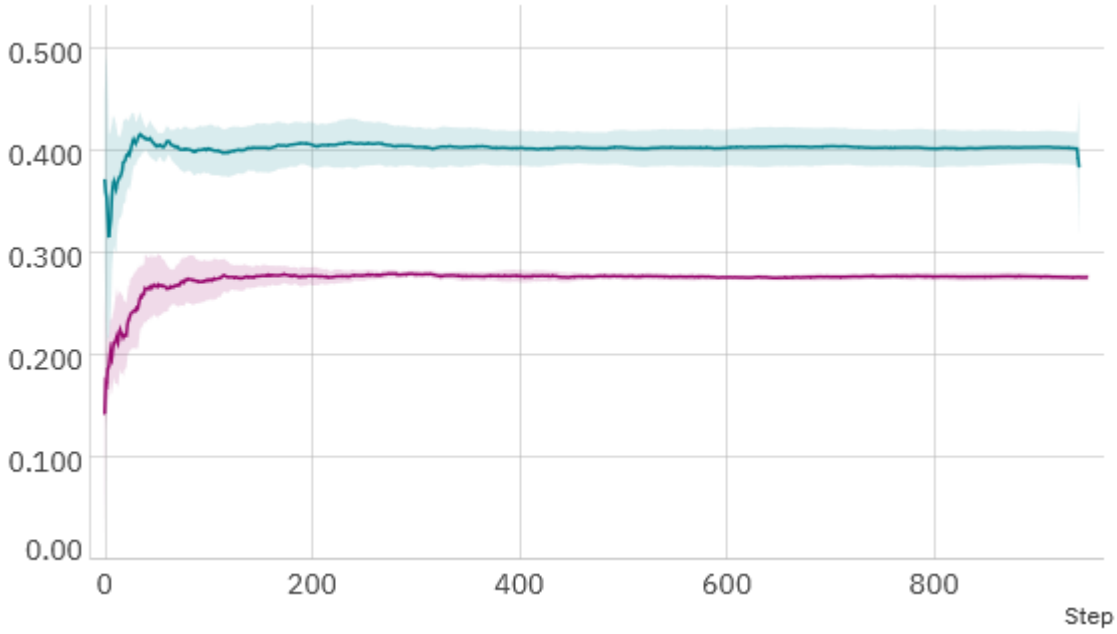

Figure ۱۰ بررسی f۱ هردو مدل

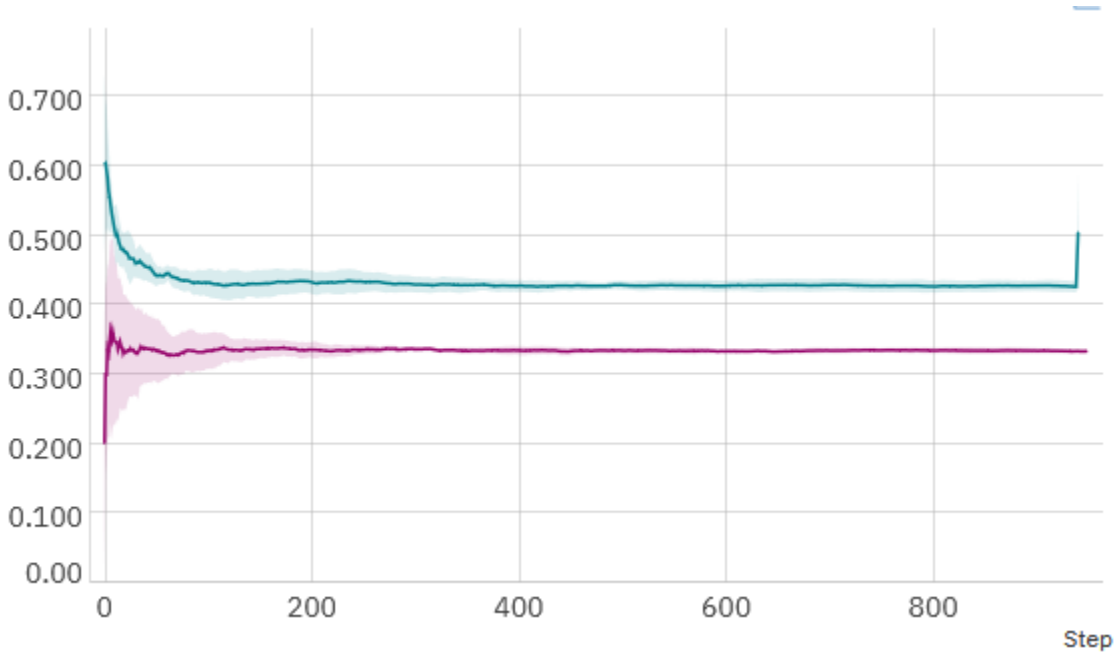

Figure ۱۱ بررسی دقت هردو مدل

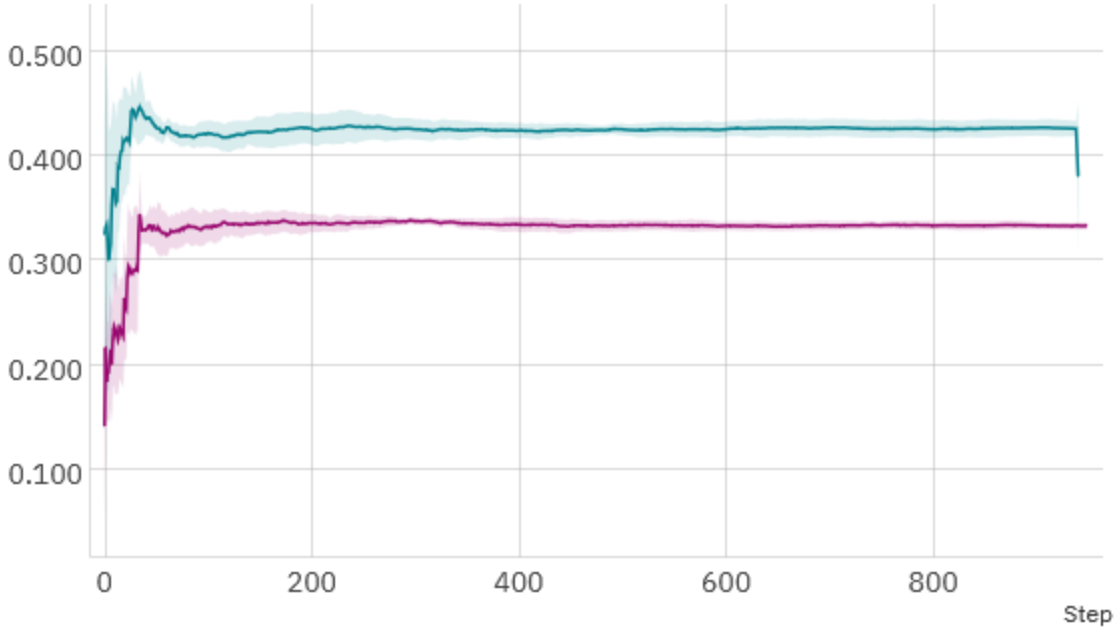

Figure ۱۲ بررسی بر اساس recall

۴ – ۳ دسته بندی میل بازار بر اساس قیمت

در این بخش، نتایج آزمایش‌های مختلف مدل‌های دسته‌بندی بازار بر اساس قیمت ارائه می‌شود. ابتدا مدل مبتنی بر مقاله Zou پیاده‌سازی شد که با استفاده از سیستم جدید برچسب‌گذاری توانست به دقت قابل قبولی دست یابد. سپس تلاش‌هایی برای بهبود این مدل انجام شد که منجر به بهبود نتایج شد. در نهایت، مدل جدیدی با استفاده از شبکه‌های عصبی LSTM به کار گرفته شد که نتایج بهتری نسبت به مدل SVM به دست آورد. همچنین یک مدل خودرمزگذار برای بهبود عملکرد به LSTM اضافه شد که نتایج نهایی را بهبود بخشید.

۴ – ۳ – ۱ نتایج مدل های مختلف

در ابتدا، مدل SVM پیاده‌سازی و آزمایش شد. نتایج این مدل به شرح زیر است:

۴ - ۱نتایج مدل SVM

معیار	مقدار
accuracy	۰.۴۳۶
precision	۰.۱۹۰
recall	۰.۴۳۶
F۱	۰.۲۶۴

همان‌طور که مشاهده می‌شود، مدل SVM در دسته‌بندی دقیق بازار بر اساس قیمت عملکرد مناسبی ندارد. هرچند دقت مدل بیش از ۴۰٪ است، اما معیار F۱ به وضوح نشان می‌دهد که این مدل نمی‌تواند به‌درستی دسته‌های مختلف بازار را از یکدیگر تفکیک کند.

پس از مشاهده نتایج ضعیف SVM، مدل شبکه عصبی LSTM پیاده‌سازی شد که در پردازش داده‌های سری زمانی کارایی بهتری دارد. نتایج مدل LSTM به شرح زیر است:

۴ - ۲نتایج مدل LSTM

معیار	مقدار
accuracy	۰.۵۲۸
precision	۰.۴۶۷
recall	۰.۵۲۸
F۱	۰.۴۹۲

در مقایسه با مدل SVM، مدل LSTM توانسته عملکرد بهتری داشته باشد. معیار دقت دسته‌بندی و یادآوری بهبود قابل توجهی پیدا کرده‌اند و معیار F۱ نیز به حدود ۰.۴۹۲ رسیده که نشان‌دهنده عملکرد بهتر این مدل در پیش‌بینی بازار است.

۴ – ۳ – ۲ اتوانکودر

در نهایت، مدل خودرمزگذار به عنوان یک مرحله پیش‌پردازش به مدل LSTM اضافه شد و نتایج نهایی آن آزمایش شد. نتایج مدل LSTM همراه با خودرمزگذار به شرح زیر است:

۴ - ۳نتایج مدل LSTM همراه با Autoencoder

معیار	مقدار
accuracy	۰.۵۴۲
precision	۰.۵۱۰
recall	۰.۵۴۲
F۱	۰.۵۰۰

۴ - ۴ مقایسه نتایج

مدل	accuracy	precision	recall	F۱
SVM	۰.۴۳۶	۰.۱۹۰	۰.۴۳۶	۰.۲۶۴
LSTM	۰.۵۲۸	۰.۴۶۷	۰.۵۲۸	۰.۴۹۲
LSTM + Autoencoder	۰.۵۴۲	۰.۵۱۰	۰.۵۴۲	۰.۵۰۰

از نتایج جدول ۴ - ۴ می‌توان نتیجه گرفت که مدل LSTM بهبود قابل توجهی نسبت به مدل SVM داشته و اضافه کردن مدل خودرمزگذار به این ترکیب منجر به نتایج بهتری شده است. دقت و معیار F۱ در مدل نهایی نشان می‌دهد که این مدل می‌تواند بهتر از مدل‌های قبلی دسته‌بندی‌های قیمت بازار را انجام دهد.

در نهایت، با توجه به پیشرفت‌هایی که با استفاده از مدل LSTM همراه با خودرمزگذار به دست آمده، می‌توان گفت که این روش برای دسته‌بندی قیمت‌ها در بازارهای مالی از دقت بالاتری برخوردار است و عملکرد بهتری نسبت به سایر روش‌ها دارد.

۴ – ۴ نتایج استراتژی

در این بخش، نتایج حاصل از تحلیل چهار استراتژی معاملاتی مختلف که شامل استراتژی خرید و نگه‌داری[۴۸] (B&H)، استراتژی ورود و خروج بر اساس سیگنال مثبت، استراتژی ورود و خروج بر اساس سیگنال منفی، و استراتژی برچسب‌گذاری سه‌گانه می‌باشد، بررسی می‌شود. این تحلیل‌ها بر روی داده‌های مربوط به بازه زمانی ۱ ژانویه ۲۰۱۵ تا ۱ فوریه ۲۰۲۱ انجام شده است. نتایج این بررسی‌ها در جدول زیر آورده شده است.

[۴۸] Buy and Hold

۴ - ۵نتایج استراتژی های مختلف در معامله

استراتژی	بازدهی روزانه٪	نسبت شارپ	تعداد معاملات بسته شده	بیشینه افت سرمایه٪
خرید و نگه‌داری(B&H)	۴.۷۰	۱.۴۰	۱	۸۳.۳۳
ورود و خروج +	۱.۸۴	۲.۲۱	۶۰	۳۲.۱۰
ورود و خروج -	۱.۶۶	۲.۰۹	۶۰	۱۶.۳۵
برچسب‌گذاری سه‌گانه (TBL)	۲.۶۸	۲.۹۴	۲۹۰	۲۰.۵۵

استراتژی خرید و نگه‌داری(B&H) ، که شامل خرید یک دارایی در ابتدای دوره و نگه‌داری آن تا انتهای دوره است، بالاترین بازده کلی (۱۰۴۶۲.۲۹٪) را به دست آورد. با این حال، این استراتژی بیشترین بیشینه افت سرمایه (۸۳.۳۳٪) را تجربه کرد که نشان‌دهنده ریسک بالای آن است. نسبت شارپ این استراتژی (۱.۴۰) نشان‌دهنده بازدهی نسبتاً متعادل نسبت به ریسک است، اما این استراتژی تنها یک معامله بسته داشت که نشان از ماهیت منفعل آن دارد.

استراتژی ورود و خروج (سیگنال مثبت) :این استراتژی با بازده کلی ۴۰۹۵.۵۴٪ و نسبت شارپ ۲.۲۱ عملکرد بهتری نسبت به استراتژی خرید و نگه‌داری داشته است. این استراتژی شامل ۶۰ معامله بسته بوده و بیشینه افت سرمایه آن ۳۲.۱۰٪ بوده است. این استراتژی که وارد موقعیت خرید در هنگام دریافت سیگنال مثبت شده و هنگام دریافت سیگنال منفی از موقعیت خارج می‌شود، تعادل مناسبی بین بازده و ریسک از خود نشان داده است.

استراتژی ورود و خروج (سیگنال منفی) :این استراتژی با بازده کلی ۳۷۱۳.۵۴٪ و نسبت شارپ ۲.۰۹ به دست آمده است. این استراتژی نیز ۶۰ معامله بسته داشته اما بیشینه افت سرمایه آن ۱۶.۳۵٪ بوده است که نسبت به استراتژی سیگنال مثبت کمتر است. استراتژی سیگنال منفی که هنگام دریافت سیگنال منفی وارد موقعیت فروش می‌شود و هنگام دریافت سیگنال مثبت از موقعیت خارج می‌شود، بازده کمتری داشته اما مدیریت ریسک بهتری داشته است.

استراتژی برچسب‌گذاری سه‌گانه ، این استراتژی نسبت به سایر استراتژی‌ها بهترین نسبت شارپ (۲.۹۴) را به دست آورده است که نشان‌دهنده بازدهی بهتر نسبت به ریسک است. این استراتژی با بازده کلی ۵۹۶۳.۳۲٪، ۲۹۰ معامله بسته و بیشینه افت سرمایه ۲۰.۵۵٪، عملکردی بهتر در مدیریت ریسک و بازده پایدار داشته است.

این استراتژی با استفاده از ترکیبی از موانع بالا، پایین و عمودی برای تعیین نقاط ورود و خروج، توانسته عملکرد مؤثرتری در مدیریت ریسک از خود نشان دهد.

نتایج این تحلیل‌ها نشان می‌دهد که استراتژی برچسب‌گذاری سه‌گانه بهترین تعادل بین بازده و ریسک را دارا بوده و نسبت شارپ بالای آن نشان‌دهنده بازدهی بهتر نسبت به ریسک است. استراتژی خرید و نگه‌داری اگرچه بازده کلی بالایی دارد، اما ریسک بالای آن در بیشینه افت سرمایه مشهود است. استراتژی‌های ورود و خروج نیز تعادلی بین این دو حالت ارائه می‌دهند که استراتژی سیگنال مثبت اندکی بازدهی بالاتر اما ریسک بیشتری نسبت به سیگنال منفی داشته است.

این نتایج نشان می‌دهند که انتخاب استراتژی مناسب در بازارهای مالی باید با توجه به تعادل میان بازده و ریسک صورت گیرد و استراتژی برچسب‌گذاری سه‌گانه می‌تواند به عنوان رویکردی مؤثر و پایدار در پیش‌بینی و معاملات مالی مورد استفاده قرار گیرد.

این نتایج نشان می‌دهد که مدل برچسب‌گذاری سه‌گانه در مقایسه با سایر استراتژی‌ها بهترین عملکرد را از نظر تعادل بازده و ریسک داشته است و توانسته با استفاده از مدیریت مؤثر ریسک، بازدهی مناسبی ارائه دهد. این استراتژی نه تنها بیشترین تعداد معاملات بسته شده را داشت، بلکه با نسبت شارپ بالای خود نشان داد که می‌تواند بازدهی بهتری نسبت به ریسک داشته باشد. در مقایسه با سایر استراتژی‌ها، استراتژی خرید و نگه‌داری B&H اگرچه بازدهی کلی بالایی داشت، اما با ریسک بالا همراه بود. در این میان، استراتژی‌های ورود و خروج (بر اساس سیگنال‌های مثبت و منفی) نیز عملکرد خوبی داشتند اما نتوانستند به همان سطح بازده و مدیریت ریسک استراتژی برچسب‌گذاری سه‌گانه برسند.

توضیح نتایج:

آزمایشاتی که انجام دادیم نشان‌دهنده چگونگی تحلیل و تفسیر مدل‌های ما از توییت‌ها برای پیش‌بینی بازارهای مالی است. یکی از نگرانی‌های اصلی در این فرآیند، احتمال بیش‌برازش (overfitting)مدل به داده‌های ورودی و وابستگی بیش از حد به برچسب‌های پیشین است. این موضوع می‌تواند باعث شود که مدل به دلیل همبستگی بالای برچسب‌های قبلی با برچسب‌های بعدی، تمرکز خود را بیشتر بر آن‌ها قرار دهد. برای مقابله با این چالش، از روش SHAP(توضیحات شاپلی افزایشی) استفاده شد تا نحوه پردازش اطلاعات توسط مدل را به طور دقیق تفسیر کنیم.

مقادیر SHAP به هر کلمه در توییت یک نمره اثرگذاری اختصاص می‌دهد که نشان می‌دهد چگونه آن کلمه بر پیش‌بینی نهایی مدل تأثیر گذاشته است. به عنوان مثال، در جمله‌ای مانند "بیت‌کوین سقوط خواهد کرد و این سقوط بسیار سریع خواهد بود"، مدل یک روند نزولی را پیش‌بینی می‌کند. این پیش‌بینی به دلیل تأثیر بالای کلمات "سقوط "و "سریع "است که مفهوم منفی‌ای را القا می‌کنند. مقادیر SHAP تأیید می‌کنند که مدل، علاوه بر برچسب‌های قبلی، محتوای واقعی توییت را تجزیه و تحلیل می‌کند و تنها به الگوهای گذشته وابسته نیست.

نتیجه‌گیری:

این روش باعث می‌شود مدل بتواند محتوای واقعی توییت‌ها را بهتر تفسیر کند و دقت بیشتری در پیش‌بینی‌های خود داشته باشد. توانایی مدل در تحلیل داده‌های شبکه‌های اجتماعی مانند توییتر، قدرت پیش‌بینی آن در بازارهای مالی را بهبود می‌بخشد و آن را به ابزاری کارآمد برای تحلیل سریع و دقیق تغییرات بازار تبدیل می‌کند.

.

فصل ۵ : مراجع

[۱] F. Z. Xing, E. Cambria, and R. E. Welsch, “Natural language based financial forecasting: a survey,” *Artificial Intelligence Review*, ۲۰۱۸.

[۲] J. Si, A. Mukherjee, B. Liu, Q. Li, H. Li, and X. Deng, “Exploiting topic based twitter sentiment for stock prediction,” ۲۰۱۳.

[۳] M. Kulakowski, F. Frasincar, and E. Cambria, “Sentiment classification of cryptocurrency-related social media posts,” *IEEE Intelligent Systems*, ۲۰۲۳

[۴] J. Bollen, H. Mao, and X.-J. Zeng, “Twitter mood predicts the stock market,” ۲۰۱۱

[۵] P. C. Tetlock, “Giving content to investor sentiment: The role of media in the stock market,” *Journal of Finance*, ۲۰۰۷

[۶] D. Zhou, L. Zheng, Y. Zhu, J. Li, and J. He, “Domain adaptive multi-modality neural attention network for financial forecasting,” *WWW*, ۲۰۲۰

[۷] S. Mohapatra, N. Ahmed, and P. Alencar, “KryptoOracle: A real-time cryptocurrency price prediction platform using twitter sentiments,” *arXiv: Computation and Language*, ۲۰۲۰

[۸] A. Gutiérrez-Fandiño, A. M. N, P. N. Kolm, and J. Armengol-Estap’e, “FinEAS: Financial embedding analysis of sentiment,” *SSRN Electronic Journal*, ۲۰۲۱

[۹] A. Huang, H. Wang, and Y. Yang, “FinBERT : A large language model for extracting information from financial text†,” *Contemporary Accounting Research*, ۲۰۲۲

[۱۰] X. Ding, Y. Zhang, T. Liu, and J. Duan, “Deep learning for event-driven stock prediction,” ۲۰۱۵.

[۱۱] Aysan, Ahmet Faruk, E. Muğaloğlu, A. Y. Polat, and H. Tekin, “Whether and when did bitcoin sentiment matter for investors? Before and during the COVID-۱۹ pandemic,” *Financial Innovation*, ۲۰۲۳

[۱۲] M. Z. Frank and W. Antweiler, “Is all that talk just noise? The information content of internet stock message boards,” ۲۰۰۱

[۱۳] J. V. Critien, A. Gatt, J. Ellul, J. V. Critien, A. Gatt, and J. Ellul, “Bitcoin price change and trend prediction through twitter sentiment and data volume,” *Financial Innovation*, ۲۰۲۲

[۱۴] X. Zhang, H. Fuehres, and P. A. Gloor, “Predicting stock market indicators through twitter ‘I hope it is not as bad as I fear,’” *Procedia - Social and Behavioral Sciences*, ۲۰۱۱

[۱۵] O. Kraaijeveld and Smedt, Johannes De, “The predictive power of public Twitter sentiment for forecasting cryptocurrency prices,” *Journal of international financial markets, institutions, and money*, ۲۰۲۰

[۱۶] L. Luo *et al*., “Beyond polarity: Interpretable financial sentiment analysis with hierarchical query-driven attention,” *IJCAI*, ۲۰۱۸

[۱۷] D. Valle-Cruz, V. Fernandez-Cortez, A. López-Chau, and R. Sandoval-Almazán, “Does twitter affect stock market decisions? Financial sentiment analysis during pandemics: A comparative study of the H۱N۱ and the COVID-۱۹ periods,” *Cognitive Computation*, ۲۰۲۱

[۱۸] Y. Zou and D. Herremans, “PreBit - A multimodal model with Twitter FinBERT embeddings for extreme price movement prediction of Bitcoin,” *Expert Systems With Applications*, ۲۰۲۳

[۱۹] C. Lamon, E. Nielsen, and E. Redondo, “Cryptocurrency price prediction using news and social media sentiment,” ۲۰۱۷.

[۲۰] Z. Ye, W. Liu, Q. Qu, Q. Jiang, and Y. Pan, “A cryptocurrency price prediction model based on twitter sentiment indicators,” ۲۰۲۲

[۲۱] S. Suardi, A. R. Rasel, and B. Liu, “On the predictive power of tweet sentiments and attention on bitcoin,” ۲۰۲۲

[۲۲] K. Wolk, “Advanced social media sentiment analysis for short-term cryptocurrency price prediction,” *Expert Syst. J. Knowl. Eng.*, ۲۰۲۰

[۲۳] G. Serafini *et al.*, “Sentiment-driven price prediction of the bitcoin based on statistical and deep learning approaches,” ۲۰۲۰

[۲۴] O. Sattarov, H. Jeon, R. Oh, and J. D. Lee, “Forecasting bitcoin price fluctuation by twitter sentiment analysis,” *۲۰۲۰ International Conference on Information Science and Communications Technologies (ICISCT)*, ۲۰۲۰

[۲۵] A. P. Ratto, S. Merello, Y. Ma, L. Oneto, and E. Cambria, “Technical analysis and sentiment embeddings for market trend prediction,” *Expert Syst. Appl.*, ۲۰۱۹

[۲۶] X. Guo and J. Li, “A novel twitter sentiment analysis model with baseline correlation for financial market prediction with improved efficiency,” *arXiv: Social and Information Networks*, ۲۰۲۰

[۲۷] D. Y. Aharon, E. Demir, C. Keung, and A. Zaremba, “Twitter-Based uncertainty and cryptocurrency returns,” ۲۰۲۲

[۲۸] S. Bounid, M. Oughanem, and S. Bourkadi, “Advanced financial data processing and labeling methods for machine learning,” *International Symposium on Computer Vision*, ۲۰۲۲

[۲۹] S. Usmani and J. Shamsi, “LSTM based stock prediction using weighted and categorized financial news,” *PLoS ONE*, ۲۰۲۳

[۳۰] A. M. E, M. Erkul, K. Kaempf, V. Vasiliauskaite, and N. Antulov-Fantulin, “Ask ‘Who’, not ‘What’: Bitcoin volatility forecasting with twitter data,” *Web Search and Data Mining*, ۲۰۲۱

[۳۱] M. Pouromid, A. Yekkehkhani, O. M. A, and A. Aminimehr, “ParsBERT post-training for sentiment analysis of tweets concerning stock market,” *International Computer Society of Iran Computer Conference*, ۲۰۲۱

[۳۲] Z. Hu, Z. Wang, S.-B. Ho, and A.-H. Tan, “Stock market trend forecasting based on multiple textual features: A deep learning method,” *IEEE International Conference on Tools with Artificial Intelligence*, ۲۰۲۱.

[۳۳] P. Eslamieh, M. Shajari, and A. Nickabadi, “User۲Vec: A novel representation for the information of the social networks for stock market prediction using convolutional and recurrent neural networks,” *Mathematics*, ۲۰۲۳

[۳۴] E. F. Fama, “Efficient Capital Markets: a Review of Theory and Empirical Work,” *The Journal of Finance*, vol. ۲۵, no. ۲, pp. ۳۸۳–۴۱۷,۱۹۷۰

[۳۵] A. W. Lo, “The Adaptive Markets Hypothesis,” *The Journal of Portfolio Management*, vol. ۳۰, no. ۵, pp. ۱۵–۲۹ ۲۰۰۴

[۳۶] “Time Series Generation with Masked Autoencoder,” *arXiv*, ۲۰۱۹.

واژه نامه انگلیسی به فارسی

Efficient Market Hypothesis ..... فرضیه بازار کارا

Adaptive Market Hypothesis ..... فرضیه بازار تطبیقی

Natural Language Programming ..... برنامه‌نویسی زبان طبیعی

Sentiment Analysis ..... تحلیل احساسات

Triple Barrier Labeling ..... برچسب‌گذاری سه‌گانه

Domain Adaptive Multi-Modality Neural Attention Network
..... شبکه عصبی توجهی چند وجهی تطبیق‌پذیر با دامنه

Autoencoders ..... خودرمزگذارها

Embeddings ..... جاسازی‌ها

Dense ..... چگال

Backtrader ..... بک‌تریدر (پلتفرم معاملاتی)

Forward Fill ..... پرکردن رو به جلو

DataFrame ..... دیتافریم (چارچوب داده‌ها)

Exponential Moving Average ..... میانگین متحرک نمایی

Moving Average Convergence Divergence ..... واگرایی همگرایی میانگین متحرک

Momentum ..... شتاب

Bollinger Bands ..... باندهای بولینگر

High-Low Spread ..... تفاوت قیمت بالا و پایین

SHapley Additive exPlanations (SHAP) ..... توضیحات جمعی شاپلی

Shapley Value ..... ارزش شاپلی

Fixed Window Labeling ..... برچسب‌گذاری با پنجره ثابت

Variable Windowing ..... پنجره‌بندی متغیر

Sharpe Ratio ..... نسبت شارپ

Bullish ..... صعودی

Bearish ..... نزولی

Unsupervised ..... بدون نظارت

Masked Language Modeling ..... مدل‌سازی زبان ماسک‌شده

Byte-Level Encoding ..... کدگذاری سطح بایت

Neutral ..... خنثی

Undersampling ..... نمونه‌برداری کمتر

Oversampling ..... نمونه‌برداری بیشتر

Fine-Tuning ..... تنظیم دقیق

Multi-Head Attention ..... توجه چند سری

Feedforward Neural Network ..... شبکه عصبی پیش‌خور

Attention Mechanism ..... مکانیزم توجه

Self-Attention ..... توجه به خود

Short-Term Impact ..... تاثیر کوتاه‌مدت

Autoencoder ..... خودرمزگذار

Encoder ..... رمزگذار

Extrapolator ..... برون‌یاب

Recurrent Neural Network..... شبکه عصبی بازگشتی

Fully Connected ..... کاملاً متصل

Total Return ..... بازده کل

Sharpe Ratio ..... نسبت شارپ

Max Drawdown ..... بیشترین کاهش

Closed Trades ..... معاملات بسته‌شده

Cross Validation ..... اعتبارسنجی متقابل

Context Unaware ..... بدون آگاهی از زمینه

Buy and Hold ..... خرید و نگه‌داری

واژه نامه فارسی به انگلیسی

فرضیه بازار کارا ..... Efficient Market Hypothesis

فرضیه بازار تطبیقی ..... Adaptive Market Hypothesis

برنامه‌نویسی زبان طبیعی ..... Natural Language Programming

تحلیل احساسات ..... Sentiment Analysis

برچسب‌گذاری سه‌گانه ..... Triple Barrier Labeling

شبکه عصبی توجهی چند وجهی تطبیق‌پذیر با دامنه .....
Domain Adaptive Multi-Modality Neural Attention Network

خودرمزگذارها ..... Autoencoders

جاسازی‌ها ..... Embeddings

چگال ..... Dense

بک‌تریدر (پلتفرم معاملاتی) ..... Backtrader

پرکردن رو به جلو ..... Forward Fill

دیتافریم (چارچوب داده‌ها) ..... DataFrame

میانگین متحرک نمایی ..... Exponential Moving Average

واگرایی همگرایی میانگین متحرک ..... Moving Average Convergence Divergence

شتاب ..... Momentum

باندهای بولینگر ..... Bollinger Bands

تفاوت قیمت بالا و پایین ..... High-Low Spread

توضیحات جمعی شاپلی ..... SHapley Additive exPlanations (SHAP)

ارزش شاپلی ..... Shapley Value

برچسب‌گذاری با پنجره ثابت ..... Fixed Window Labeling

پنجره‌بندی متغیر ..... Variable Windowing

نسبت شارپ ..... Sharpe Ratio

صعودی ..... Bullish

نزولی ..... Bearish

بدون نظارت ..... Unsupervised

مدل‌سازی زبان ماسک‌شده ..... Masked Language Modeling

کدگذاری سطح بایت ..... Byte-Level Encoding

خنثی ..... Neutral

نمونه‌برداری کمتر ..... Undersampling

نمونه‌برداری بیشتر ..... Oversampling

تنظیم دقیق ..... Fine-Tuning

توجه چند سری ..... Multi-Head Attention

شبکه عصبی پیش‌خور ..... Feedforward Neural Network

مکانیزم توجه ..... Attention Mechanism

توجه به خود ..... Self-Attention

تاثیر کوتاه‌مدت ..... Short-Term Impact

خودرمزگذار ..... Autoencoder

رمزگذار ..... Encoder

برون‌یاب ..... Extrapolator

شبکه عصبی بازگشتی ..... Recurrent Neural Network

واژه نامه فارسی به انگلیسی

کاملاً متصل ..... Fully Connected

بازده کل ..... Total Return

نسبت شارپ ..... Sharpe Ratio

بیشترین کاهش ..... Max Drawdown

معاملات بسته‌شده ..... Closed Trades

اعتبارسنجی متقابل ..... Cross Validation

بدون آگاهی از زمینه ..... Context Unaware

خرید و نگه‌داری ..... Buy and Hold

Abstract

In the era of rapid technological advancements, financial market prediction is evolving at an unprecedented pace. Traditional methods, which often rely on one-dimensional analyses such as sentiment analysis, are no longer sufficient to comprehend the complexities of modern markets. This research leverages advanced language models and textual information to enhance the accuracy of predictions.

In this study, sentiment analysis, particularly sentiments extracted from tweets, is used to forecast market trends. The sentiments present in tweets, which include positive, negative, and neutral emotions, are considered influential factors affecting market movements. Through this analysis, the system can detect short-term market trends based on real-time changes in user sentiments. For example, tweets with negative sentiment often lead to predictions of a downward market trend, while positive sentiments may signal an upward trend.

Additionally, advanced machine learning models such as LSTM neural networks and autoencoders have been employed. Autoencoders extract complex features from the data and feed them as input to the LSTM model for more accurate predictions. This combination of tweet sentiment analysis and the sequential processing of LSTM models has led to a significant improvement in prediction accuracy.

The results show that sentiment analysis and the use of tweets, alongside deep learning models, effectively enhance market movement predictions. This approach, which considers both market sentiments and more complex data features, represents an important step in applying machine learning and natural language processing in the financial domain.

Keywords: Natural Language Processing - Social Media Analysis - Financial Market Prediction - Sentiment Analysis

Iran University of Science and Technology
Computer engineering faculty

Analysis of Textual Features in the Domain of Sentiment and Their Impact on Financial Markets Using Natural Language Processing Methods

Bachelor thesis

By:
Ali Soltani

supervisor:
Dr. Reza Entezari Maleki

September 2024