فصلنامه علمی- پژوهشی
فناوری اطلاعات و ارتباطات ایران

سال یازدهم، شماره‌های ۳۹ و ۴۰، بهار و تابستان ۱۳۹۸
صص: ۱۹-۴۸

به‌کارگیری وب‌کاوی در پیش‌بینی جهت قیمت سهام گروه محصولات شیمیایی در بورس اوراق بهادار

[*]امیر دایی [**]امید مهدی عبادتی [***]کیوان برنا
[*] کارشناسی ارشد مدیریت فناوری اطلاعات، دانشکده مدیریت، دانشگاه خوارزمی، تهران
[**] استادیار گروه مدیریت عملیات و فناوری اطلاعات دانشگاه خوارزمی، تهران
[***] استادیار گروه علوم کامپیوتر دانشگاه خوارزمی، تهران

تاریخ دریافت: ۱۳۹۸/۱۰/۰۹ تاریخ پذیرش: ۱۳۹۹/۰۴/۱۸

چکیده

پیش‌بینی بازارها از جمله سهام به دلیل حجم بالای معاملات و نقدینگی برای محققان و سرمایه‌گذاران دارای جذابیت بوده است. توانایی پیش‌بینی جهت قیمت ما را قادر می‌سازد با کاهش ریسک و اجتناب از ضرر و زیان مالی، به بازده بالاتری دست‌یابیم. اخبار نقش مهمی در فرایند ارزیابی قیمت فعلی سهام دارد. توسعه روش‌های داده‌کاوی، هوش محاسباتی و الگوریتم‌های یادگیری ماشین سبب ایجاد مدل‌های جدیدی در پیش‌بینی شده‌اند. هدف از این پژوهش ذخیره سازی اخبار خبرگزارها و استفاده از روش‌های متن کاوی و الگوریتم ماشین بردار پشیبان به منظور پیش‌بینی جهت قیمت روز آینده سهم است. بدین منظور خبرها منتشر شده در ۱۷ خبرگزاری با استفاده از یک خزگشر موضوعی به زبان پی‌اچ‌پی[۱] ذخیره و دسته‌بندی شده است. سپس با استفاده از روش‌های متن‌کاوی و الگوریتم ماشین بردار پشتیبان و کرنل‌های مختلف به پیش‌بینی جهت قیمت سهام گروه محصولات شیمیایی در بورس اوراق بهادار پرداخته می‌شود. دراین مطالعه از ۳۰۰ هزار خبر در دسته‌های سیاسی و اقتصادی و قیمت‌های سهام ۲۵ شرکت منتخب در بازه زمانی آبان تا اسفند ۹۷ در ۱۲۲ روز معاملاتی استفاده شده است. نتایج نشان می‌دهد با مدل ماشین بردار پشتیبان با کرنل خطی می‌توان به صورت میانگین ۸۳ درصد جهت قیمت‌ها را پیش‌بینی کرد. با استفاده از کرنل‌های غیرخطی و معادله درجه ۲ ماشین بردار پشتیبان صحت پیش‌بینی به صورت میانگین تا ۸۵ درصد افزایش می‌یابد و سایر کرنل‌ها نتایج ضعیف‌تری از خود نشان می‌دهند.

کلید واژه: متن‌کاوی، کاوش محتوای وب، خزشگر وب، پیش‌بینی بورس اوراق بهادار، ماشین بردار پشتیبان

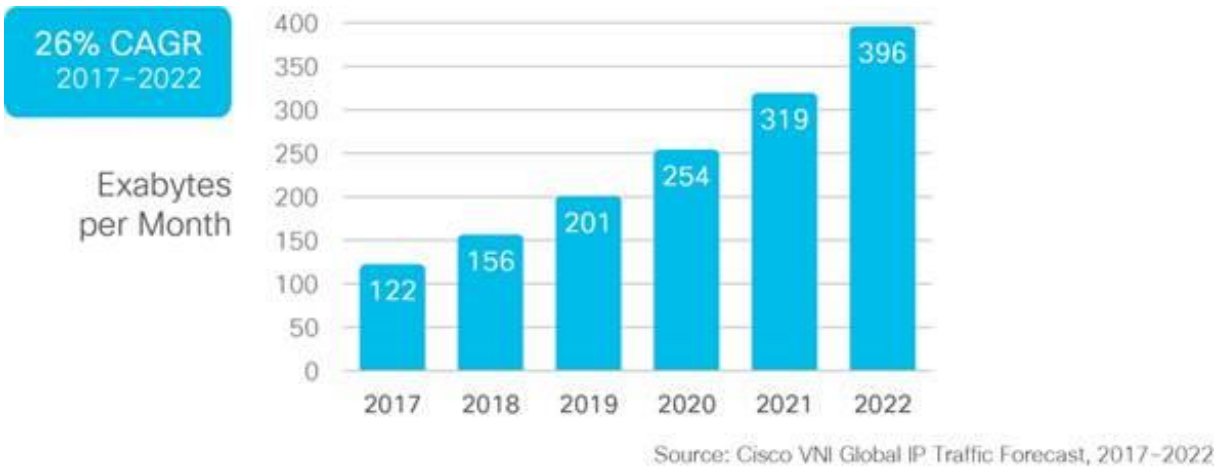
۱-مقدمه

با رشد روز افزون اینترنت و تولید محتوا در محیط وب، داده بسیار زیادی در بانک‌های اطلاعاتی ذخیره می‌شوند. این داده‌ها ممکن است از طریق صفحات وب به نمایش درآیند. اگرچه دسترسی به پایگاه داده وب سایت‌ها و گرفتن داده‌ها از مدیران سایت‌ها کار ساده‌ای نیست با این حال می‌توان با خزش وب سایت‌ها، این داده‌های ارزشمند را ذخیره کرد. این موضوع فرصتی را فراهم می‌آورد تا بتوان به مباحث داده‌کاوی و متن‌کاوی و استخراج دانش پرداخت. قبل از ظهور وب اکثر کاربران مجموعه عظیمی از اسناد نداشتند و بنابراین نیازی به سامانه‌ای پیچیده برای جستجو میان اسناد مختلف نبود. با ظهور وب، کاربران به داده زیاد در حال رشدی دسترسی پیدا کردند. مرتب‌سازی و جستجو در میان انبوهی از داده‌های وب مسائل جدیدی را ایجاد کرد که با عنوان بازیابی اطلاعات وب[۱] شناخته می‌شود [1]. بر اساس شاخص cisco VNI پیش‌بینی می‌شود ترافیک آی پی جهانی از سال ۲۰۱۷ تا ۲۰۲۲ سه برابر شود. پیش‌بینی می‌شود ترافیک آی پی در سال ۲۰۲۲ به صورت ماهانه به ۳۹۶EB برسد.

ترافیک کل برای اینترنت در دو دهه اخیر رشد فوق‌العاده‌ای را تجربه کرده است. بیش از ۲۰ سال قبل، در سال ۱۹۹۲، شبکه‌های اینترنت جهانی، حدود ۱۰۰ گیگ ترافیک را در هر روز منتقل می‌کردند. در سال ۲۰۱۷، ترافیک جهانی اینترنت به ۴۵۰۰۰ گیگ در هر ثانیه رسید. [۲].

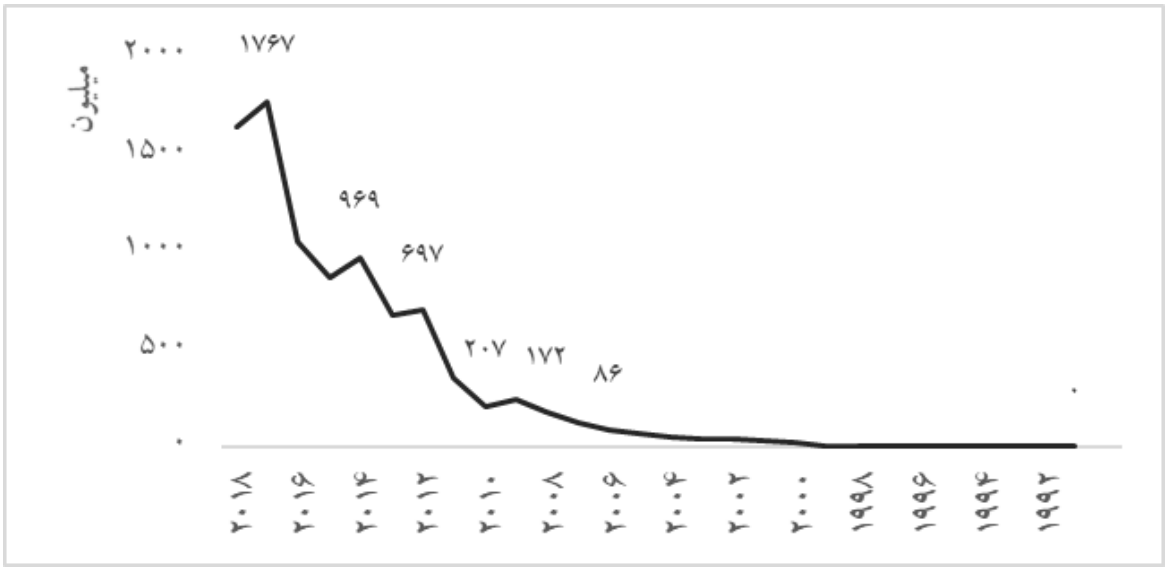
نویسنده عهده‌دار مکاتبات: امیدمهدی عبادتی ebadati@khu.ac.ir

[۱] PHP



شکل ۱-پیش‌بینی شرکت سیسکو از ترافیک آی‌پی [۲]

در حال حاضر بیش از یک و نیم میلیارد وب‌سایت در شبکه جهانی اینترنت وجود دارد. کمتر از ۲۰۰ میلیون از این وب‌سایت‌ها فعال هستند. این موضوع در تحقیقات web server اکتبر ۲۰۱۴ netcraft تائید شد و برای اولین بار توسط Internet live stats برآورد و اعلام شد. شکل ۲ تعداد کل وب‌سایت‌ها را در طول تاریخ از ۱ وب‌سایت در سال ۱۹۹۱ تا سال ۲۰۱۸ نمایش می‌دهد [۳]



شکل ۲- تعداد وب‌سایت‌ها از سال ۱۹۹۲ تا ۲۰۱۸

رشد اینترنت و گسترش وب‌سایت‌ها و تولید محتواهای غیر ساختاریافته مطالعه در این حوزه را روز به روز با اهمیت‌تر می‌کند.

وب کاوی در حوزه‌های مختلف مورد مطالعه قرارگرفته است: فن‌های آماده‌سازی و پالایش داده‌های صفحات وب[۲] استخراج و ذخیره‌سازی الگوی وب[۳] کاوش ساختار وب[۴] کاوش استفاده‌ی از وب[۵]و کاوش محتوای وب[۶]که هدف آن کشف دانش مفید از داده‌های موجود در اینترنت است [۱]. وب کاوی بسته به اهداف کاوش و داده‌های ذخیره‌شده به سه دسته تقسیم می‌شود: استخراج ساختار وب، استخراج کاربرد وب و استخراج محتوای وب [۴]. رشد روز افزون وب سایت‌ها و مطالب منتشرشده در محیط وب اهمیت استفاده از کاورش محتوای وب را دو چندان می‌کند.

[۲] Web Data Pre-processing and Cleaning
[۳] Web Pattern Extraction and Storage
[۴] Web structure mining
[۵] Web usage mining
[۶] Web content mining

خزش اولین قدم در وب کاوی یا ساخت یک موتور جستجو است. به صورت کلی خزشگرها به دو دسته تقسیم می‌شوند: خزشگر عمومی[۷] و خزشگر موضوعی.[۸] خزشگرهای عمومی همه صفحات را بدون در نظر گرفتن محتوای آنها دانلود می‌کند، اما در خزشگرهای موضوعی فقط صفحاتی با موضوعات خاص دانلود می‌شود [۵].

به‌طور کلی ابزارهای پیش‌بینی در بازار سرمایه به دو بخش تحلیل تکنیکال[۹] و بنیادی[۱۰] تقسیم می‌شوند. تفاوت این روش‌ها در داده‌های ورودی است؛ در روش تکنیکال از داده‌های تاریخی بازار استفاده می‌شود در صورتی که در تحلیل بنیادی از دیگر انواع اطلاعات یا اخبار درباره کشور، جامعه، شرکت و غیره استفاده می‌شود. بیشتر تحقیقات در گذشته بر روی رهیافت تکنیکال انجام‌شده است، که دلیل عمده آن در دسترس بودن داده‌های کمّی تاریخی در بازار و تمایل عمومی معامله‌گران برای استفاده از روش‌های کمّی تکنیکال است. داده‌های بنیادی در صورتی که بدون ساختار باشد دارای چالش‌های بیشتری برای استفاده به عنوان ورودی هستند. داده‌های بنیادی ممکن است از منابع ساختاریافته و عدد مانند داده‌های اقتصاد کلان یا گزارش صورت‌های مالی بانکی و دولتی گرفته شوند. [۶]

توسعه روش‌های داده‌کاوی، هوش محاسباتی و الگوریتم‌های یادگیری ماشین سبب ایجاد مدل‌های جدیدی در پیش‌بینی شده‌اند. به صورت کلی در مقالات متعدد نشریات مختلف دو دسته تحقیق در استفاده از روش‌های داده‌کاوی برای پیش‌بینی بازارهای مالی صورت گرفته است. روش اول استفاده از داده‌های دارای ساختار است که در بیشتر مطالعات از داده‌های دارای ساختار مانند قیمت گذشته، درآمد و سود تاریخی استفاده کرده‌اند. [۷]

همچنین در مقاله‌ای دیگر که توسط چن جانگ هانگ[۱۱] و همکاران منتشرشده است تمرکز سیستم‌های پیش‌بینی مالی را صرفاً بر روی داده‌های کمی مانند قیمت سهام و شاخص بازار می‌دانند. [۸]

اما بخش عمده‌ای از داده‌هایی که در محیط وب تولید می‌شوند بدون ساختار هستند و از این رو می‌توان از روش دوم یعنی استفاده از تکنیک‌های متن‌کاوی و داده‌کاوی بر روی داده‌های بدون ساختار استفاده کرد. متن‌کاوی شامل وظایف بسیاری مانند خوشه‌بندی اسناد، دسته‌بندی اسناد، خلاصه‌سازی متن، تحلیل احساسات، تحلیل شبکه‌های اجتماعی، تشخیص موضوع، دسته‌بندی صفحات وب، شناسایی نویسنده، تشخیص سرقت ادبی، تحلیل فیشینگ/هرزنامه و نرم‌افزارهای مخرب، تحلیل الگو، تصمیم‌گیری مالی و غیره است؛ اما اصلی‌ترین چالش در متن‌کاوی، داده‌های بدون ساختار است که قبل از شروع داده‌کاوی نیاز است آن‌ها را به فرمت ساختاریافته تبدیل کرد. [۹]

اخبار[۱۲] نقش مهمی در فرایند ارزیابی قیمت فعلی سهام، که توسط تحلیل‌گران، سرمایه‌گذاران و سرمایه‌گذاران نهادی[۱۳] صورت می‌گیرد، دارد. بر اساس یک دیدگاه تئوریک، ارزیابی کارآمد از یک شرکت باید بر ارزش فعلی و جریان وجوه نقد آینده شرکت تأثیر بگذارد. در اخبار نه تنها آمار و ارقام مالی بلکه اجزای متنی کیفی نیز قیمت سهام را تحت تأثیر قرار می‌دهند. [۱۰]

در این پژوهش فرض شده است که با استفاده از اخبار منتشر شده می‌توان جهت قیمت سهم را در روز بعد پیش‌بینی کرد. بدین منظور اخبار منتشر شده در خبرگزاری‌ها به عنوان متغییر تحقیق در نظر گرفته شده‌اند و دیتای اخبار ذخیره می‌شود. سپس با استفاده از این اخبار و روش‌های متن‌کاوی و داده‌کاوی و استفاده از الگوریتم ماشین بردار پشتیبان با کرنل‌های مختلف به دسته‌بندی اخبار برای پیش‌بینی صعودی یا نزولی بودن قیمت در روز بعد پرداخته می‌شود.

۲- ادبیات موضوع

برای مطالعات وب کاوی می‌توان از داده‌هایی که توسط خزشگرها و موتورهای جستجو ذخیره‌شده است یا با نوشتن خزشگر و ذخیره‌سازی دیتا دسترسی داشت. اگر بخواهیم از داده‌های خزشگرها و موتورهای جستجو استفاده کنیم این دیتا در مدت زمان کوتاه و به راحتی ممکن است در اختیار قرار گیرد. بدین منظور می‌توانید از مجموعه داده‌های آماده استفاده کنیم، اما این دیتا بیشتر برای آزمایش یک الگوریتم مفید است. شاید در این زمینه بهترین مجموعه داده ترک (trec.org) باشد که شامل مجموعه از صفحات وب با ساختار html است [۱۱]. ممکن است داده‌ها به روز نباشد، داده‌هایی

---

[۷] Universal crawler

[۸] Topic crawlers

[۹] Technical

[۱۰] Fundamental

[۱۱] Chenn-Jung Huang

[۱۲] News

[۱۳] Institutional Traders

مورد نیاز ذخیره نشده باشد، دیتا ساختار مناسب نداشته باشد، دیتا بیش از نیازها باشد یا متحمل هزینه شود. برای دستیابی به اطلاعات دقیق، به روز و مورد اعتماد نیاز است که از خزشگر مورد نیاز نوشته شود [۱۱]. اما این روش چالش‌هایی مانند نیاز به دانش برنامه‌نویسی، زمان‌بر بودن نوشتن و آزمایش خزشگر و در نهایت فرایند ذخیره‌سازی داده دارد. در این پژوهش ابزار قابل توسعه‌ای طراحی و ارائه‌شده است تا بتوان محتوای مورد نیاز را از وب ذخیره کرد.

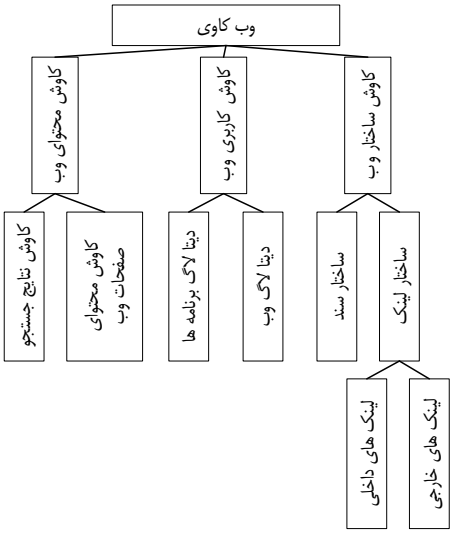
۲-۱-وب‌کاوی

در طی سال‌ها شبکه وب از طریق انتقال سنتی و به اشتراک گذاشتن رایانه‌ها و اسناد به عنوان «وب داده»، به اتصال فعلی مردم به عنوان «وب افراد» و به اتصال در حال ظهور میلیاردها اشیاء به عنوان «وب اشیاء» تغییر کرده است. [۱۲]

وب کاوی به معنی استفاده از فن داده‌کاوی به‌منظور خودکار کردن جستجو و استخراج اطلاعات از اسناد و خدمات وب است. با توجه به حجم عظیمی اطلاعاتی که در محیط اینترنت، شبکه جهانی وب حوزه بکری برای مطالعات داده‌کاوی است. این داده‌های عظیم باعث ایجاد مسائلی از قبیل پیدا کردن اطلاعات مرتبط، ایجاد دانش جدید از اطلاعات در محیط وب، شخصی‌سازی اطلاعات و شناخت رفتار مصرف کنند و کاربران شده است [۱۳]

۲-۲-روش‌های وب کاوی

وب کاوی بر اساس نوع دیتا به سه دسته کاوش کاربری وب، کاوش ساختار وب و کاوش محتوای وب تقسیم می‌شود. هر سه دیدگاه بر روی کشف دانش ضمنی، اطلاعات ناشناخته و بالقوه تمرکز دارند. [۱۴]



شکل ۳-روش‌های وب کاوی [۱۵]

۲-۳-انواع روش‌های متن‌کاوی

الگوریتم‌های متن‌کاوی به‌طور کلی به دو دسته الگوریتم یادگیری با نظارت و الگوریتم یادگیری بدون نظارت تقسیم می‌شوند. برای متن‌کاوی از ابزارها و روش‌های مختلفی استفاده می‌شود که به سه گروه اصلی تقسیم می‌شوند. این ابزارها و روش‌ها در زیر بیان‌شده است [۱۶]:

۲-۳-۱-کلاس‌بندی یا دسته‌بندی[۱۴]

کلاس‌بندی یا دسته‌بندی یکی از روش‌هایی است که برای گروه‌بندی متن‌ها در متن‌کاوی مورد استفده قرار می‌گیرد . هدف از کلاس‌بندی متون، نسبت دادن کلاس‌های از پیش تعریف‌شده به اسناد متنی موجود، مانند گروه اخبار، گروه کالاها، جناح سیاسی است. [۱۷, ۱۸].

۲-۳-۲-خوشه‌بندی[۱۵]

خوشه‌بندی یکی از روش‌های متن‌کاوی است که برای شناسایی گروه‌های داده‌ها بر اساس مشخصه‌ها یا ویژگی‌های آن‌ها به کار برده می‌شود که در آن هیچ گروه از پیش تعیین شده‌ای وجود ندارد [۱۷].

۳-۲-۳- کاوش قوانین وابستگی

کاوش قوانین وابستگی روشی است که با کشف و یافتن قوانینی در رخدادهای دیگر وقوع یک مورد را در آینده پیش‌بینی می‌کند. کاوش قوانین وابستگی در دسته الگوریتم‌های یادگیری بدون نظارت جای دارد زیرا هیچ برچسبی از قبل برای آموزش الگوریتم وجود ندارد [۱۶, ۱۹].

[۱۴] Classification

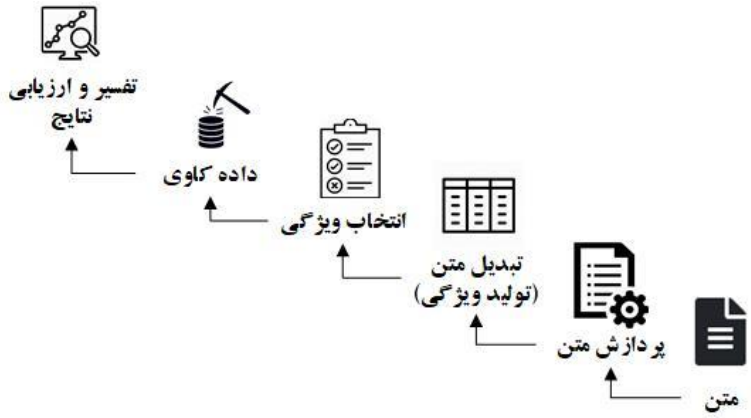
[۱۵] Clustering

۲-۴-مراحل متن‌کاوی

فرایند متن‌کاوی مراحل مختلفی دارد که در پژوهش هاشیمی[۶] و همکاران (۲۰۱۵)، پیش‌پردازش، استخراج ویژگی و انتخاب ویژگی از مراحل اصلی این فرایند عنوان‌شده است. مراحل کلی این فرایند در شکل ۴ نشان داده‌شده است. [۲۰] کومر و راوی[۷] به‌طور کلی فرایند متن‌کاوی شامل دو فاز اساسی می‌باشد: پیش‌پردازش متن، استخراج دانش [۱۶].

[۶] Hashimi

[۷] Kumar & Ravi



شکل ۴-مراحل متن‌کاوی

۱-۲-۴-پیش‌پردازش

برای کشف دانش از حجم قابل توجهی از اسناد، لازم است که بر روی اسناد پیش‌پردازش انجام شود؛ یعنی اطلاعات در یک ساختار داده‌ای مناسب برای پردازش‌های بعدی آماده و ذخیره‌سازی شود. در این مرحله داده‌های ورودی که در دسترس است باید برای ورود به الگوریتم یادگیری ماشین آماده شود، یعنی از حالت غیرساخت‌یافته به فرمت ساختاریافته و قابل تشخیص برای ماشین تبدیل شود. [۱۸]

۲-۲-۴-تولید و استخراج ویژگی

اگرچه برنامه‌های کاربردی زیادی درزمینه بازیابی اطلاعات مانند پالایش و جست‌وجوی اطلاعات مرتبط می‌توانند از تحقیقات درزمینه رده‌بندی متن سود ببرند، مشکل اصلی رده‌بندی متن، ابعاد بالای فضای ویژگی با توجه به تعداد زیاد لغات است. راه‌حل این مشکل استفاده از روش‌های استخراج و انتخاب ویژگی است [۲۰].

۳-۲-۴-انتخاب ویژگی

انتخاب ویژگی فرایندی است که زیرمجموعه‌ای از ویژگی‌های اصلی را با توجه به برخی از معیارها و یا اهمیت ویژگی‌ها انتخاب می‌کند. الگوریتم‌های انتخاب ویژگی به دو دسته زیر تقسیم می‌شوند:

۱- الگوریتم‌های رتبه‌بندی ویژگی

۲- الگوریتم‌های انتخاب زیر مجموعه ویژگی

۴-۲-۴-کشف دانش

گام بعدی استخراج و کشف دانش از فرم‌های میانی است که بر اساس نحوه نمایش هر سند می‌تواند متفاوت باشد. نمایش مبتنی بر سند برای گروه‌بندی، طبقه‌بندی و تجسم‌سازی استفاده می‌شود، در حالی‌که نمایش مبتنی بر مفهوم برای یافتن روابط میان مفاهیم، ساخت اتوماتیک آنتولوژی و غیره به‌کار می‌رود [۲۱].

۵-۲-۴-تفسیر و ارزیابی

در این مرحله، خروجی مراحل قبل مورد ارزیابی قرار می‌گیرد تا مشخص شود که دانشی کشف‌شده است و اینکه دانش کشف‌شده اهمیت دارد یا خیر. با اجرای الگوریتم‌ها، داده/ متن استخراج‌شده به فنون مختلفی تحویل داده می‌شود که امکان استفاده مستقیم از اطلاعات استخراج‌شده را از طریق ابزار کشف پیوند یا مصورسازی فراهم می‌کنند [۲۱].

۲-۵-یادگیری ماشین

یادگیری ماشین ارتباط نزدیکی با آمار محاسباتی دارد و اغلب با آن همپوشانی دارد. تمرکز این شاخه پیش‌بینی کردن با رایانه است و پیوند محکمی با بهینه‌سازی ریاضی دارد. مدل‌های تحلیلی به محققان، پژوهشگران علم داده و تحلیلگران اجازه می‌دهد تصمیمات و نتایج قابل اطمینان و تکرارپذیر به دست آورند و با یادگیری از روابط و روندهای مربوط به گذشته، از الگوهای پنهان پرده‌برداری کنند [۲۲].

۱-۲-۵-یادگیری بدون نظارت: خوشه‌بندی

الگوریتم خوشه‌بندی کا-مینز[۱۸] و کا-مد[۱۹]

الگوریتم کا مینز که در سال ۱۹۶۷ توسط مک کویین[۲۰] مطرح شد، یکی از محبوب‌ترین الگوریتم‌های خوشه‌بندی است که در زمینه‌های مختلف مورد استفاده قرار می‌گیرد. هدف الگوریتم کا-مینز، بهینه‌سازی تابع هدف[۲۱] می‌باشد که پاسخ‌های حاصل از خوشه‌بندی به کمینه‌سازی یا

۱۸K-means

۱۹K-mode

۲۰Mac Queen

۲۱Object Function

بیشینه‌سازی تابع هدف منجر می‌شود. در این الگوریتم باید تعداد خوشه‌ها از قبل مشخص‌شده باشد. این الگوریتم بر روی داده‌های پیوسته تعریف می‌شود. [۲۳].

۱-۱-۲-۵-الگوریتم DBSCAN

یکی از پرکاربردترین الگوریتم‌های خوشه‌بندی، تکنیک DBSCAN است که خوشه‌ها را بر اساس تراکم و غلظت آن‌ها تعیین کند. در این الگوریتم نیازی به مشخص کردن تعداد خوشه‌ها توسط کاربر نیست. این الگوریتم قادر است با اثربخشی بالایی خوشه‌هایی به شکل‌های دلخواه ایجاد کند. برای خوشه‌بندی تعدادی از نقاط توسط این الگوریتم، از دو پارامتر شعاع همسایگی[۳۲] و حداقل تعداد نقاط موجود[۳۳] در همسایگی، استفاده می‌شود و مجموعه نقاط را به نقاط مرکزی[۳۴]، نقاط مرزی[۳۵] و داده‌های پرت[۳۶] تقسیم می‌کند. [۲۴].

الگوریتم سلسه‌مراتبی[۳۷]

این الگوریتم به دو دسته از بالا به پایین[۳۸] و از پایین به بالا[۳۹] تقسیم می‌شود. در روش از بالا به پایین، تمامی داده‌ها با هم به‌عنوان یک خوشه بزرگ در نظر گرفته می‌شوند و در مرحله بعد به خوشه‌های کوچک‌تر تقسیم‌شده تا جایی که هر داده به‌عنوان یک خوشه در نظر گرفته شود. اما در روش پایین به بالا هر داده در ابتدا یک خوشه است، به ترتیب در هر مرحله داده‌هایی که بیشترین شباهت (کمترین فاصله) را به هم دارند به خوشه می‌پیوندند تا خوشه بزرگ‌تر شود و در نهایت همه داده‌ها با هم یک خوشه شوند. برای اندازه‌گیری فاصله بین دو خوشه از روش‌های پیوند[۴۰] استفاده می‌شود. انواع روش‌های پیوند عبارت‌اند از: نزدیک‌ترین همسایه[۴۱]، پیوند تکی[۴۲]، دورترین همسایه[۴۳]، پیوند کامل[۴۴] و یا پیوند میانگین[۴۵] [۲۵].

مدل ترکیبی گاوسی[۴۶]

منظور از نمودار گوسی این است که یک مقدار مشخص وجود دارد که حداکثر نمونه‌ها در آن قرار دارند و هرچه از این مقدار دورتر می‌شویم تعداد داده‌ها کمتر می‌شود. به این نوع پخش‌شدگی توزیع گوسی می‌گویند که بیشتر داده‌های جهان از این نوع توزیع پیروی می‌کنند. شکل توزیع گوسی همیشه متوازن نیست و ممکن است به سمت چپ یا راست چولگی[۴۷] داشته باشد. این الگوریتم از روش بیشینه‌سازی انتظار[۴۸] استفاده می‌کند.

۲-۲-۵-الگوریتم یادگیری بانظارت: کلاس‌بندی

الگوریتم درخت تصمیم[۴۹]

درخت تصمیم راه‌حلی سریع و مفید برای کلاس‌بندی مجموعه داده‌های بزرگ با تعداد زیادی از متغیرها را فراهم می‌کند. این الگوریتم که متغیرهای کمی و کیفی را پیش‌بینی می‌کند، اولین بار توسط برمان[۵۰] مطرح شد. نتیجه این الگوریتم مجموعه‌ای از شرط‌های منطقی با ساختار درختی است که برای پیش‌بینی یک ویژگی به‌کار می‌رود. الگوریتم درخت تصمیم به‌گونه‌ای عمل می‌کند که گوناگونی یا تنوع در گره‌ها را به حداقل برساند. ۴ نوع الگوریتم درخت تصمیم CART، QUEST، CHAID و C۰.۵ وجود دارد که تفاوت آن‌ها در معیار اندازه‌گیری عدم خلوص، شیوه شاخه‌بندی و هرس‌کردن گره‌های درخت می‌باشد [۲۶].

الگوریتم جنگل تصادفی[۵۱]

الگوریتم جنگل تصادفی یکی از الگوریتم‌های یادگیری ماشین و ابزاری برای کلاس‌بندی و رگرسیون است [۲۷]. جنگل تصادفی الگوریتمی ترکیبی است که بر اساس مدل درخت تصمیم شکل‌گرفته است. در این الگوریتم ابتدا با استفاده از روش نمونه‌گیری بوت استرپ[۵۲] تعداد K زیرمجموعه آموزش[۵۳] را از بین مجموعه داده‌های اصلی استخراج می‌کند و سپس با

---

[۳۲] Epsilon
[2] MinPoints
[۳۴] Core points
[۳۵] Border point
[۳۶] Outlier
[۳۷] Hierarchical
[۳۸] Partitioning
[۳۹] Agglomerative
[۴۰] Linkage
[۴۱] Nearest Neighbor
[۴۲] Single Linkage
[۴۳] Furthest Neighbor
[۴۴] Complete Linkage
[۴۵] Average Linkage
[۴۶] Gaussian Mixture Model (GMM)
[۴۷] Skew
[۴۸] Expectation Maximization (EM)
[۴۹] Decision Tree
[۵۰] Breman
[۵۱] Random Forest
[۵۲] Bootstrap
[۵۳] Train

آزمایش کردن این زیرمجموعه‌ها تعداد K درخت تصمیم ایجاد می‌شود. در نهایت یک جنگل تصادفی از این درختان تصمیم ایجاد می‌شود. [۲۸].

الگوریتم شبکه عصبی[۴۴]

شبکه عصبی ابزار قدرتمندی برای حل مشکلات پیچیده است که با پردازش داده‌ها، دانش پنهان آن‌ها را به ساختار شبکه منتقل می‌کند. نورون یک تابع غیرخطی، پارامتری و محدود است که برای راحتی، به این تابع نورون گفته می‌شود. به متغیرهای نورون ورودی نورون گفته می‌شود و مقدار آن در خروجی آن نشان داده می‌شود. نورون‌ها می‌توانند به‌راحتی به‌صورت گرافیکی نمایش داده شوند. ترکیب توابع غیرخطی دو یا چند نورون، شبکه‌ای از نورون‌ها را تشکیل می‌دهد. به‌طور کلی شبکه‌های عصبی دو نوع هستند: شبکه پیش‌خور[۴۵] و شبکه بازخور[۴۶] [۲۹].

ماشین بردار پشتیبان[۴۷]

مسأله طبقه بندی یکی از مسائل اصلی مطر شده در یادگیری ماشین است. بسیاری از مسائل را می توان بصورت یک مسأله کلاسه بندی مطرح کرده و حل نمود. روش های طبقه بندی خطی، سعی دارند که با ساختن یک ابر سطح (که عبارت است از یک معادله خط)، داده ها را از هم تفکیک کنند. چندین تفکیک کننده خطی می تواند داده دو کلاس را از هم جدا کند. یکی از روش هایی که بصورت گسترده برای اینگونه مسائل استفاده می شود، ماشین بردار پشتیبان است. ماشین بردار پشتیبان در سال ۱۹۹۳ توسط ولادیمیر واپنیک پیشنهاد شد. SVM بهترین ابر سطحی را پیدا می کند که با حداکثر فاصله، داده های مربوط به دو طبقه را از هم تفکیک کند. [۳۰]

۲-۶-ارزیابی مدل در یادگیری با ناظر

در یادگیری با ناظر چندین راه برای ارزیابی عملکرد مدل یادگیری و نتایج دسته‌بندی وجود دارد. شاخص‌های ارزیابی کیفیت کلاس‌بندی بر اساس ماتریس درهم ریختگی تعیین می‌شوند. در ماتریس درهم ریختگی تعیین می‌شود چه تعداد از پیش‌بینی‌های مدل صحیح و غلط بوده‌اند. جدول ۱ ماتریس درهم ریختگی برای یک مدل دوتایی را نمایش می‌دهد.

جدول ۱- ماتریس در هم ریختگی

کلاس / تشخیص	تشخیص مثبت	تشخیص منفی
مثبت	Tp	Fn
منفی	Fp	Tn

Tp (true positive): کلاس مثبتی که به درستی مثبت تشخیص داده‌شده است.

Fn (false negative): کلاس مثبتی که به اشتباه منفی تشخیص داده‌شده است.

Fp (false positive): کلاس منفی که به اشتباه مثبت تشخیص داده‌شده است.

Tn (true negative): کلاس منفی که به درستی مثبت تشخیص داده‌شده است.

بر اساس این ماتریس شاخص‌های دقت، صحت و امتیاز-F محاسبه می‌شود.

۱-۲-۶-دقت[۴۸]

precision درستی پیش‌بینی‌ها را نسبت به کل موارد کلاس نشان می‌دهد. فرمول محاسبه به صورت زیر است:

$$\text{precision} = \frac{Tp}{Tp + Fp}$$

۲-۲-۶-صحت[۴۹]

در شاخص صحت نسبت مقدار موارد صحیح طبقه‌بندی‌شده توسط الگوریتم از یک کلاس به کل اعضای پیش‌بینی شده در آن گروه را محاسبه می‌کنیم. فرمول محاسبه به صورت زیر است:

$$\text{Recall} = \frac{Tp}{Tp + Fn}$$

۳-۲-۶-امتیاز-F1[۵۰]

این معیار دقت و صحت را با هم در نظر می‌گیرد. معیار F1 در بهترین حالت، یک و در بدترین حالت صفر است. این معیار توصیف‌کننده میانگین وزن‌دار مابین دو کمیت دقت و صحت است. این معیار میزان دقت و صحت یک مدل را به صورت هم‌زمان بررسی می‌کند و میزان کیفیت کلاس‌بندی را تعیین می‌کند. برای محاسبه امتیاز F1 از فرمول زیر استفاده می‌شود. برای محاسبه این شاخص از فرمول زیر استفاده می‌شود:

---

[۴۴] Neural Network

[۴۵] Feedforward

[۴۶] Feedback

[۴۷] Support vector machine

[۴۸] Precision

[۴۹] Recall

[۵۰] F1 Score

$$\text{F1 Score} = \frac{\text{precision} * \text{Recall}}{\text{precision} + \text{Recall}}$$

با استفاده از این شاخص‌های می‌توان مدل‌های یادگیری دوتایی با ناظر را ارزیابی کرد. [۳۱]

۲-۷-پیش‌بینی بازار بر اساس اخبار منتشرشده

مدیریت ریسک مالی یکی از کارهای بسیار پر چالش در واحدهای مالی است. در دو دهه گذشته روش‌ها و مدل‌های کمی[۵۱]بسیاری جهت مشخص کردن تأثیر بازارهای بی‌ثبات مالی بر تجارت، توسعه و گسترش یافته‌اند. اکثر این مطالعات بر روی داده‌های ساختاریافته مانند سری قیمت‌های تاریخی انجام‌شده است و توجه اندکی به داده‌های غیر ساختار یافته (متنی)[۵۲]شده است، در حالی که بیشتر حجم منابع اطلاعاتی موجود از این نوع محتوا هستند. تحقیقات تجربی که در گذشته صورت گرفته است نشان می‌دهد که اخبار خاص، مانند افشای شرکت‌های بزرگ می‌توانند سبب تغییر رفتار غیرطبیعی قیمت پس از انتشار شود. [۳۲]

فرضیه بازار کارا و نظریه گام تصادفی دو نظریه هستند که تأثیر قابل توجهی بر پیش‌بینی بازار داشته است. در فرضیه بازار کارا، قیمت سهام بازتابی از اطلاعات بازار کامل است و زمانی که اطلاعات جدیدی ارائه شود، بلافاصله در قیمت سهام منعکس می‌شود. در نظریه گام تصادفی اعتقاد بر این است که بازارها کارآمد هستند و اصلاح قیمت‌ها بلافاصله رخ می‌دهد و پیش‌بینی قیمت از داده بازار غیرممکن است. این نظریه‌ها بیان می‌کنند که قیمت‌ها به اطلاعات پیوند خورده‌اند و از اطلاعات امروز نمی‌توان برای پیش‌بینی قیمت‌ها در آینده استفاده کرد. [۳۳]

همچنین در پیش‌بینی قیمت سهام دو فلسفه کاملاً مخالف وجود دارد. فن‌های تجزیه و تحلیل بنیادی و تکنیکال. در حالی که تحلیلگران بنیادی به دنبال استفاده از داده‌های نسبی، نسبت‌ها و درآمد نسبی سهام هستند، تحلیلگران تکنیکال از نمودارها، فن‌های مدل‌سازی بر اساس حجم معاملات تاریخی و قیمت‌ها برای تحلیل خود استفاده می‌کنند. [۳۳]

در حالی که رفتار بازار تحت تأثیر اتفاقات محلی و جهانی است، جزئیات این اتفاقات در داده‌های ساختاریافته مشاهده نمی‌شود. انتظار می‌رود که داده‌کاوی نقش مهمی در طراحی راهبردهای پیش‌بینی رفتار بازار داشته باشد، زیرا می‌تواند با موفقیت برای تجزیه و تحلیل مقالات اخبار مالی و گزارش‌های همراه داده‌های سری زمانی بازار مورد استفاده قرار گیرد. متن‌کاوی گزارش‌های خبری مالی می‌تواند برای استخراج اطلاعات مهم در مورد وقایع مختلف سیاسی و اقتصادی که به طور کلی بازار مالی یک منطقه را تعیین می‌کند و نیز علل عملکرد ضعیف یا افزایش ناگهانی در بازار را توضیح دهد. امروزه با توجه به حجم اخبار موجود در اینترنت، نیاز روزافزونی به فناوری قابل اطمینانی وجود دارد که بتواند برای تجزیه و تحلیل خودکار گزارش‌های خبری و کشف اطلاعات کلیدی از طرف تحلیلگران و سرمایه‌گذاران استفاده شود. [۳۴]

اخبار[۵۳]نقش مهمی در فرایند ارزیابی قیمت فعلی سهام، که توسط تحلیل‌گران، سرمایه‌گذاران و سرمایه‌گذاران نهادی[۵۴] صورت می‌گیرد، دارد. بر اساس یک دیدگاه تئوریک، ارزیابی کارآمد از یک شرکت باید بر ارزش فعلی و جریان وجوه نقد آینده شرکت تأثیر بگذارد. در اخبار نه تنها آمار و ارقام مالی بلکه اجزای متنی کیفی نیز قیمت سهام را تحت تأثیر قرار می‌دهند [۱۰]

در سال‌های اخیر روزنامه‌های سنتی به دلیل افزایش فراگیر شدن شبکه جهانی وب، به گسترش سرویس‌های خبری بر خط را در محیط وب پرداخته‌اند. برای سرمایه‌گذاران، خبرهای بلادرنگ مالی[۵۵]در تصمیم‌گیری‌ها سرمایه‌گذاری بسیار مهم است، زیرا در محیط وب خبرها به طور مکرر در حال به روزرسانی هستند. اطلاعات بیش از حد یک مشکل قابل توجه است. برای سرمایه‌گذاران غیرممکن است که همه آنها را یکی یکی بخوانند. اگر چه تحقیقات رو به افزایشی در متن‌کاوی اسناد در حال انجام است، اما دقت کم و پایین بودن نرخ بازگشت سبب می‌شود سرمایه‌گذاران زمان زیادی را برای به دست آوردن اطلاعات معنی‌دار اندکی، در گشت و گذار وب از دست بدهند. [۸]

برای بررسی تأثیر اخبار بر روی قیمت سهام مطالعات مختلفی صورت گرفته است. مهاجان و همکاران به بررسی اخبار برای شناسایی رویدادهای مهم و تأثیر آن بر بازار سهام هند پرداختند. [۳۵] هوانگ و همکاران [۸] تأثیر تیترهای خبری منتشرشده توسط روزنامه‌های الکترونیکی بر روی شاخص

---

[۵۱] Quantitative

[۵۲] Nextual

[۵۳] News

[۵۴] Institutional traders

[۵۵] Real-time financial news

قیمت مالی بورس تایوان را مورد بررسی قرار دادند. هاگونا[۵۶] و همکاران سعی کرده‌اند با استفاده از روش‌های متن‌کاوی و انتخاب ویژگی‌ها از طریق بازخورد بازار به بهبود پیش‌بینی قیمت سهام بپردازند [۱۰]. مدل دیگری که بر روی پیش‌بینی قیمت سهام بر اساس اخبار منتشرشده می‌پردازد سیستم متن مالی آریزونا[۵۷] است که توسط شوماکر[۵۸] و همکاران طراحی‌شده است [۳۳]. در شرکت مشاوره‌ای تاتا در هند صورت گرفته است به دنبال ایجاد سامانه‌ای جهت تحلیل بازار سهام با استفاده از بررسی اخبار مالی منتشرشده به‌منظور شناسایی و تشخیص اتفاقات مهمی که بر روی بازار تأثیر می‌گذارند هستند [۳۴]. در تحقیقی که در دانشگاه میسوری[۵۹] در سال ۲۰۱۷ انجام شد، فرایند جامع داده‌کاوی را برای پیش‌بینی جهت روزانه شاخص اس‌اندپی ۵۰۰ بر اساس ۶۰ ویژگی مالی و اقتصادی ارائه گردید [۳۶]. در مدل دیگری که توسط آیمن و همکاران ارائه شد، به دنبال پیش‌بینی مبتنی بر تحلیل احساسات از اخبار مالی و قیمت‌های بازار سهام بودند [۳۷].

از آنجایی که سهامی که در یک گروه بورسی قرار می‌گیرند داری رفتارهای نسبتاً مشابهی هستند از این رو تحقیقات نشان می‌دهد برای کاهش ریسک سبد سهام، تمام سهم‌ها از یک گروه انتخاب نشود و با خرید متنوع از سهم در گروه‌های مختلف ریسک خود را کاهش دهند. [۳۸] از این رو بهتر است برای تحلیل، سهم‌هایی انتخاب شود که در یک گروه قرار دارند. برای مثال به نظر می‌رسد عوامل موثر بر قیمت سهم‌های گروه شیمیایی متفاوت از عوامل گروه غذایی باشد. گروه محصولات شیمیایی بزرگترین گروه سهام از نظر ارزش سهام در بورس اوراق بهادار است، این گروه به عنوان نمونه برای مطالعه انتخاب شده است.

۲-۸-یادگیری ماشین و پیش‌بینی قیمت‌ها

مطالعات انجام شده خارجی در این حوزه بر خلاف منابع فارسی بسیار غنی است. از جمله دلایل عدم تمایل محققین ایرانی در این حوزه پیچیده و دشوار بودن فرایند انجام مراحل متن‌کاوی و عدم دسترسی مناسب به منابع سخت‌افزاری است. از انجایی که اجرای الگوریتم‌ها بر روی داده‌های ساختار یافته تاریخی بسیار ساده است، اکثر مطالعات داخلی بر روی این حوزه انجام شده است.

در مطالعه‌ای که توسط تای وو[۶۰] و همکاران در سال ۲۰۲۰ با عنوان «یک مدل شبکه عصبی حلقه‌ای جدید برای پیش بینی قیمت سهام» بر روی سهام بورس تایوان انجام شده است، توضیح داده می‌شود که استفاده از شبکه عصبی در یادگیری عمیق بر اساس ویژگی‌ها بسیار موثر است. همچنین از چارچوب شبکه عصبی حلقوی می‌توان برای انتخاب ویژگی و پیش‌بینی قیمت سهام با استفاده از دیتای تاریخی استفاده کرد. در این مقاله از ترکیب این دو روش برای پیش‌بینی قیمت سهام استفاده شده است. [۳۹]

در تحقیق دیگری با عنوان «روش یادگیری کارآمد ماشین هیبریدی برای پیش بینی بازار سهام سری زمانی» توسط عبادتی و مرتضوی، با استفاده از یک روش ترکیبی از الگوریتم ژنتیک و تکنیک شبکه عصبی مصنوعی برای تهیه روشی برای پیش بینی قیمت سهام و سری های زمانی استفاده نموده اند. در این روش مقادیر خروجی الگوریتم ژنتیک وارد الگوریتم توسعه یافته تکنیک شبکه عصبی مصنوعی می شوند تا خطاها را در نقطه دقیق برطرف کنند. تجزیه و تحلیل نشان می دهد که روش ترکیبی الگوریتم ژنتیک و تکنیک شبکه عصبی مصنوعی می توانند دقت را در تکرارهای کمتری افزایش دهند. این تجزیه و تحلیل بر روی شاخص اصلی ۲۰۰ روزه و همچنین بر روی پنج شرکت ذکر شده در NASDAQ انجام شده است. [۴۰]

در مطالعه‌ای دیگر توسط رانجا سناپاتی[۶۱] و همکاران در سال ۲۰۱۸ با عنوان «یک مدل جدید برای پیش بینی قیمت سهام با استفاده از شبکه عصبی ترکیبی» با استفاده از دیتای تاریخی قیمت سهم بازار بومبای و با استفاده از شبکه عصبی مصنوعی به پیش‌بینی قیمت پرداخته شده است. همچنین از الگوریتم بهینه‌سازی ازدحام ذرات برای بهینه کردن وزن ورودی استفاده شده است. در نهایت نتایج بدست آمده با نتایج مدل‌های اندازه‌گیری فاصله و شبکه عصبی مصنوعی بیزین مقایسه شد و نتایج مطلوب‌تری از این روش بدست آمد. [۴۱]

در مقاله‌ای با عنوان «یک روش جامع خوشه‌بندی و طبقه‌بندی برای پیش‌بینی بازده روزانه بازار سهام» توسط

[۵۶] Hagenau

[۵۷] Arizona Financial Text

[۵۸] Robert P. Schumaker

[۵۹] Missouri University of Science and Technology

[۶۰] Jimmy Ming-Tai Wu

[۶۱] Manas Ranjan Senapati

ژونگ[۶۲] انکه[۶۳] در سال ۲۰۱۷ از روش fuzzy-means برای پاک‌سازی و PCA برای کاهش ابعاد داده‌ها استفاده شده است. خوشه‌بندی و دسته‌بندی با استفاده از روش‌های شبکه عصبی و رگرسیون لجستیک انجام و نشان داده شد شبکه عصبی نتایج مطلوب‌تری به همراه دارد. [۴۲]

در مقاله‌ای با عنوان «پیش‌بینی رفتار بازار سهام با استفاده از تکنیک داده‌کاوی و تجزیه و تحلیل احساسات اخبار» توسط آیمن خدر و همکاران در سال ۲۰۱۷ بر روی پیش‌بینی قیمت سهام انجام شد با استفاده از الگوریتم ناوی بایاس جهت‌گیری (مثبت یا منفی) اخبار را مشخص کردن بو با ترکیب جهت‌گیری‌ها با قیمت تاریخی و سپس استفاده از الگوریتم کاامین نزدیکترین همسایه جهت قیمت سهام (مثبت یا منفی) را مشخص کردن. [۴۳]

در تحقیقی دیگر توسط شوماکر[۶۴] و همکاران در سال ۲۰۱۲ با عنوان «ارزیابی احساسات در مقالات اخبار مالی» پیش‌بینی بازار سهام با استفاده از ابزار تحلیل احساسات بر اساس اخبار منتشرشده و روش استفاده از روش رگرسیون بردار پشتیبان پرداخته شد. [۳۳]

در مقاله‌ای که توسط هانگ[۶۵] و همکاران در سال ۲۰۱۰ با عنوان «درک عامل انتشار اخبار بر اساس قوانین انجمنی و تکنیک‌های داده‌کاوی» منتشر شد، با استفاده از سیستم پردازش اطلاعات و دانش چینی برای جداسازی کلمات و استفاده از الگوریتم قوانین انجمنی وزن دهی برای تشخیص دو یا چند عبارت تأثیرگذار در تیتر خبرها بر روی شاخص بورس تایوان استفاده شد. [۱۱]

در تحقیقی دیگر توسط گروس[۶۶] و مونترمن[۶۷] در سال ۲۰۱۱ با عنوان «یک رویکرد مدیریت ریسک در بازار روزانه بر مبنای تحلیل متنی» قبل از شروع فرایند داده‌کاوی اسناد از طریق سه قدم، شناسایی ویژگی‌ها، انتخاب ویژگی‌ها و نمایندگی ویژگی به داده‌های ساختاریافته و عددی تبدیل می‌شوند. سپس از طریق چهار روش ناوی بایاس، کاامین نزدیکترین همسایه، شبکه عصبی و ماشین بردار به دسته‌بندی داده‌ها می‌پردازد. در این مطالعه استفاده از روش ماشین بردار پشتیبان به نتایج بهتری منجر شده است. [۳۲]

در مطالعه‌ای دیگر توسط هانگنوا[۶۸] و همکاران با عنوان «خواندن خودکار اخبار: پیش‌بینی قیمت سهام بر اساس اخبار مالی با استفاده از ویژگی‌ها» در سال ۲۰۱۳ در مدلی چهار مرحله‌ای به استخراج ویژگی‌ها، انتخاب ویژگی‌ها و نمایندگی ویژگی‌ها با استفاده از بازخورد بازار و سپس طبقه‌بندی با استفاده از روش ماشین بردار پشتیبان پرداختند. [۴۴]

در مطالعه ای دیگر با عنوان «کاوش اخبار مالی برای وقایع مهم و و تأثیر آنها بر بازار» توسط ماهان جان[۶۹] و همکاران در سال ۲۰۰۸ روش تخصیص درکله پنهان برای شناسایی موضوعات و کلمات مرتبط مورد استفاده قرار گرفت. سپس با استفاده از روش ترکیبی شامل درخت تصمیم و ماشین بردار پشتیبان با هسته سیگموئید برای دسته‌بندی تأثیرگذاری اخبار بر بازار استفاده شد. [۴۱]

در مقاله‌ای دیگر با عنوان «خوشه‌بندی اسناد برای شناسایی رویدادها و تحلیل روند در اخبار بازار» توسط دی[۷۰] و همکاران در سال ۲۰۰۸ با استفاده از روش تخصیص دریکله پنهان برای استخراج اتفاقات مهم و استفاده از کرنل کاامین میانه برای خوشه‌بندی موضوعات استفاده شده است. [۴۱]

در این پژوهش به دلیل موجود بودن دیتای آموزش و تست می‌توان از الگوریتم‌های یادگیری با ناظر استفاده کرد. از آنجایی که می‌توان تغییر قیمت سهام را به صورت بزرگتر مساوی صفر و کوچکتر از صفر دسته بندی کرد (دسته‌بندی باینری) می‌توان از مدل ماشین بردار پشیبان سود برد. در سایر مطالعات انجام شده نیز این الگوریتم به عنوان یک الگوریتم بهینه با نتایج قابل قبول معرفی شده است. از مزایای الگوریتم ماشین بردار پشتیبان به دلیل سادگی در محاسبات نیاز کمتری به منابع سخت‌افزاری دارد.[45] به دلیل حجم بالای دیتای پردازش شده در این مطالعه و محدودیت‌های سخت‌افزاری این الگوریتم به عنوان مناسب‌ترین گزینه انتخاب شد.

۳-فرآیند انجام پژوهش

به‌منظور ورود داده‌ها به مدل، در قدم اول سیستمی با زبان برنامه‌نویسی PHP و محیط توسعه PHPStorm طراحی شده

[۶۲] Xiao Zhong
[۶۳] David Enke
[۶۴] Robert P. Schumaker
[۶۵] Chenn-Jung Huang
[۶۶] Sven S. Groth
[۶۷] Jan Muntermann
[۶۸] Michael Hagenau
[۶۹] Anuj Mahajan
[۷۰] Lipika Dey

است که با خزش وب‌سایت‌های خبری پربازدید لینک آخرین خبرهای منتشر شده به‌صورت ۱۵ دقیقه یک‌بار در دیتابیس Mysql به‌عنوان وظیفه‌ای که در آینده باید انجام شود، ذخیره می‌شود. سپس لینک‌های ذخیره‌شده خوانده می‌شود و اطلاعاتی مانند عنوان خبر، متن خبر و دسته خبر در دیتابیس ذخیره می‌شود. پس از پیش‌پردازش متن‌ها، داده‌ها آماده ورود به الگوریتم یادگیری ماشین شود.

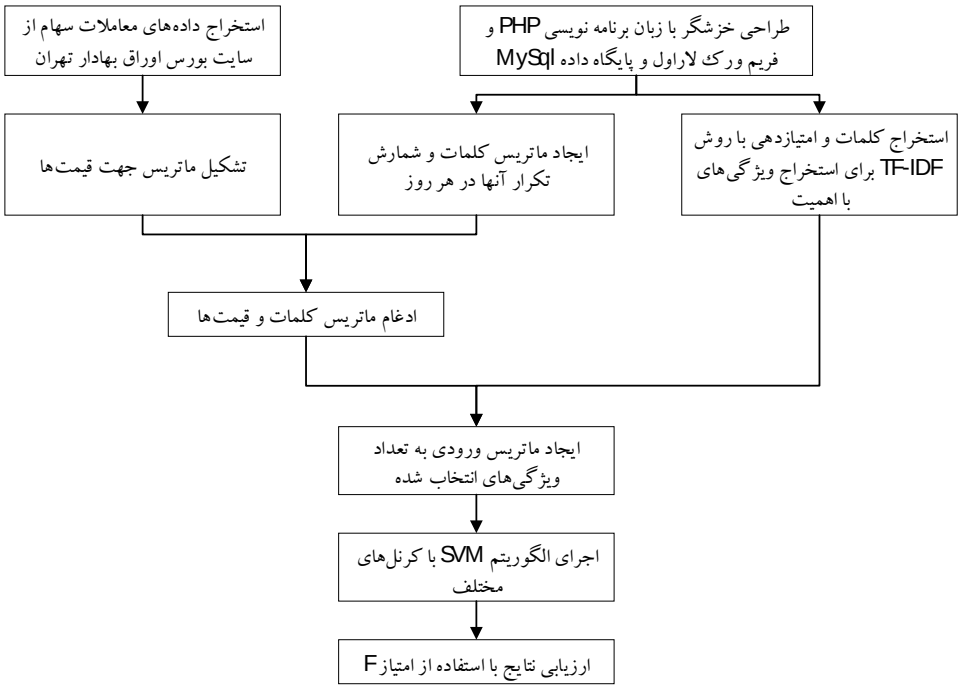
در قدم دوم داده‌های مربوط به معامله سهم از سایت بورس اوراق بهادار استخراج می‌شود. برای ذخیره‌سازی این داده‌ها ابتدا پیش‌پردازش بر روی دیتا انجام می‌شود و داده‌های سهم در دیتابیس ذخیره می‌شود.

در مرحله سوم سه ماتریس ایجاد می‌شود. ماتریس کلمات خبری که نشان دهنده تکرار کلمات در هرروز است. ماتریس TF-IDF که امتیاز هر کلمه را برای مرحله انتخاب ویژگی مشخص می‌کند. ماتریس جهت قیمت که نشان دهنده جهت قیمت در هرروز است.

در مرحله چهارم پس از انتخاب کلماتی که بیشترین امتیاز را دارند ویژگی‌ها مشخص می‌شوند و از ترکیب ماتریس کلمات و جهت قیمت سهام، ماتریس ورودی مدل تشکیل می‌شود.

در قدم پنجم با استفاده از الگوریتم SVM مدل آموزش داده می‌شود و نتایج حاصل از پیش‌بینی مدل مورد ارزیابی قرار می‌گیرد. به‌غیراز مرحله خزش تمام مراحل با استفاده از زبان برنامه‌نویسی Python و محیط توسعه PyCharm انجام می‌شود.

برای انتخاب الگوریتم مناسب برای پروژه و تحقیق داده‌کاوی باید به ۵ عامل، دقت، مدت زمان آموزش الگوریتم، خطی یا غیرخطی بودن مدل، تعداد متغیرهای مسئله و تعداد ویژگی‌های انتخابی دقت کنید.



شکل ۵- روش جمع‌آوری و تجزیه و تحلیل داده‌ها

۳-۱-سیستم خزش و ذخیره‌سازی اخبار روزانه

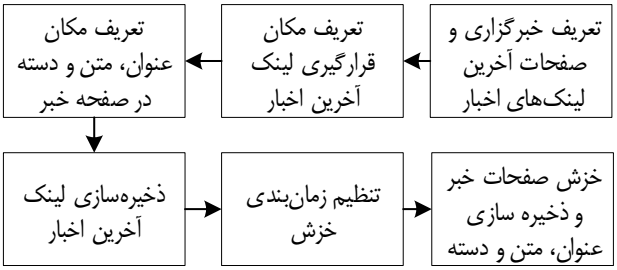
۱-۳-۱-ذخیره‌سازی اخبار خبرگزاری‌ها

یک برنامه خزشگر برای جمع‌آوری دیتای مورد نیاز از سطح وب یکی از ضروری‌ترین بخش مطالعات وب‌کاوی است. در اینجا به بررسی یک خزشگر که در این مطالعه با زبان پی‌اچ‌پی نوشته شده است معرفی می‌شود. در شکل (۴) فرایند کار این برنامه نمایش داده‌شده است.

این سیستم با استفاده از زبان برنامه‌نویسی پی‌اچ‌پی و چهارچوب[۱]لاراول[۲]نوشته شد. برای دانلود صفحات خبر از پکیج گازل و برای ذخیره‌سازی عنوان، متن و دسته خبر بر اساس آدرس مکان سی‌اس‌اس[۳]از پکیج دام کراولر استفاده می‌شود. در فاز اول ۱۷ سایت خبرگزاری پربازدید ایران برای

[۱] Framework
[۲] Laravel
[۳] Cascading Style Sheets

ذخیره‌سازی تعیین شده‌اند. همچنین ۲ دسته اصلی اخبار سیاسی و اخبار اقتصادی و ۱۶۰ زیر دسته برای آن‌ها تعریف‌شده است.



شکل ۶- نحوه کار پی‌اچ‌پی کراولر

جدول ۲- سایت‌های خبرگزاری خزش‌شده

نام خبر گذاری	نام خبر گذاری
ایسنا	شبکه خبر
عصر ایران	اقتصاد آنلاین
باشگاه خبرنگاران	خبر آنلاین
تسنیم	مشرق
خبرگزاری فارس	مهر
افکار نیوز	تابناک
موج	دنیای اقتصاد
فردا	برتر ین‌ها
اقتصاد نیوز	

الگوریتم کار این خزشگر در نمودار زیر ارائه شده است. در قدم اول داده‌های اولیه یک وب‌سایت جهت خزش ذخیره می‌شود. این داده‌ها شامل عنوان خبرگزاری، لینک اصلی خبرگزاری، صفحه‌ای که آخرین خبرها در آن منتشر می‌شود، مکان درج لینک آخرین خبرها براساس انتخابگر سی‌اس‌اس، مکان درج عنوان خبر بر اساس انتخابگر سی‌اس‌اس و مکان درج دسته خبر بر اساس انتخابگر سی‌اس‌اس ذخیره می‌شود.

id	task_url	rand_number	complete	error	created_at
1	https://www.isna.ir/news/98060402201/ترامپ-از-فران...	45	2	NULL	2019-08-26 22:24:15
2	https://www.isna.ir/news/98060402200/سیاوشی-اداره-...	152	2	NULL	2019-08-26 22:24:15
3	https://www.isna.ir/news/98060402199/فولادگر-درجه...	26	2	NULL	2019-08-26 22:24:15
4	https://www.isna.ir/news/98060402198/سند-نهایی-گرو...	42	2	NULL	2019-08-26 22:24:15

شکل ۷- نمونه دیتای جدول crawl_tasks و لینک‌های ذخیره‌شده جهت خزش

در قدم دوم لینک آخرین خبرها با استفاده از کران جاب‌ها هر ۱۰ دقیقه به عنوان یک وظیفه ذخیره می‌شوند. این امر به این دلیل است که در طول روز بار سرور زیاد است و کار ذخیره‌سازی اخبار در انتهای روز و قبل از صبح انجام می‌شود. شبه‌کد این مرحله در زیر آورده شده است:

```
Select all agents
For each agent in agents do
   Download page for last news links
   Remove JavaScript code
   For each link in last news links do
      If link has base_agent_url
         task_link = link
      Else
         task_link = base_agent_url  + link
      If task_link not in tasks then
         Create new task
         Add random number to task
         Add task_link to task
         Insert task to database
```

در قدم سوم بررسی می‌شود آیا لینکی برای ذخیره‌سازی خبر وجود دارد یا خیر. اگر وظیفه وجود داشته باشد، لینک‌هایی که به عنوان یک وظیفه ذخیره شده بودند به صورت تصادفی انتخاب می‌شوند و اخبار آنها ذخیره می‌شود. این فرایند به

خاطر جلوگیری از ارسال درخواست‌های زیاد پشت سر هم به یه خبرگزاری است.

شبه‌کد فرایند ذخیره‌سازی اخبار به صورت زیر است:

```
Select 12 tasks randomly that not crawled
Change tasks status to is crawling
If number of tasks is 0 then
   Select 3 tasks randomly that had error in crawl
For each task in task do
   Download task page
   Remove JavaScript code
   Remove tab and newline character from text
   Create new news
   Select and add title to news
   Select and add content to news
   Select news category
   Insert news to database
If news page has category then
   If news category is defined then
      Add category to news
   Else
      If category not defined in suggested category then
         Add category to suggestion
```

id	crawl_task_id	title	content
1	839	اولین جایزه ادبی شهید اندرزگو برگزیدگانش را شناخت	به گزارش مشرق، نخستین دوره جایزه ادبی شهید سیدعلی ...
2	193	روزنامه پیروزی - ۵ شهریور	
3	53	راز بزرگ دوچرخه‌سوار مشهور ایتالیایی چه بود؟	انتخاب: دوچرخه سوار مشهور ایتالیایی، جینو بارتالی ...
4	58	نخست‌وزیر انگلیس: ایران نباید به سلاح هسته‌ای دست ...	نخست‌وزیر انگلیس در کفرانسی خبری اعلام کرد که اعضا...
5	581	ماجرای اصرار هاشمی بر حذف شعار مرگ بر آمریکا	خبرگزاری تسنیم: آیت‌الله اکبر هاشمی‌رفسنجانی از آ...

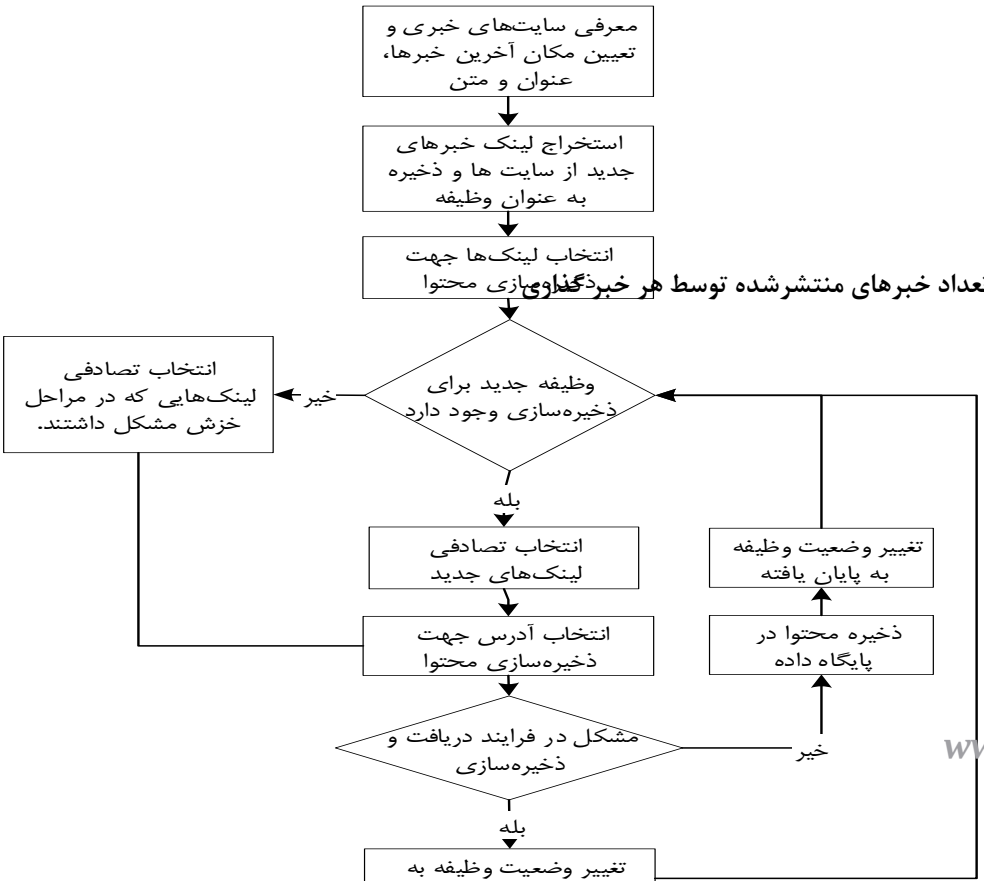
شکل ۸- نمونه دیتای جدول news (خبرهای منتشرشده)

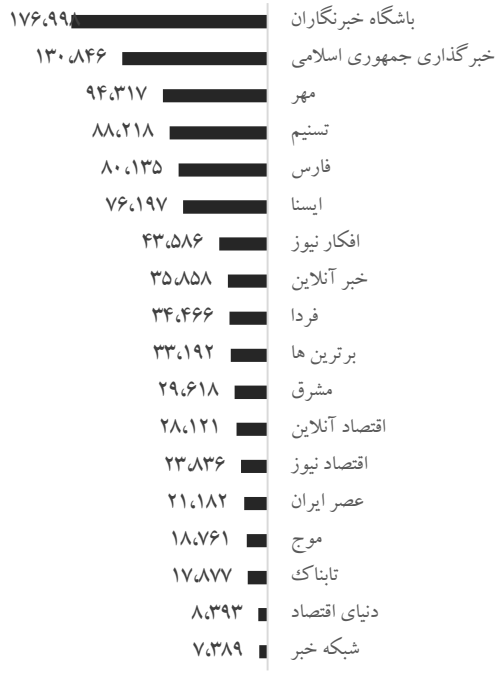
در قدم چهارم عنوان خبر و دسته خبر ذخیره می‌شود و اگر این کار با موفقیت انجام شود وضعیت ذخیره‌سازی به پایان یافته تغییر می‌کند. اگر سرور خبرگزاری پاسخگو نباشد، وضعیت به دارای خطا تغییر می‌کند. اگر لینکی دارای خطا باشد یکبار دیگر پس از ذخیره همه خبرها خوانده می‌شود.

وجود کدهای جاوا اسکریپت به دلیل داشتن برخی از تگ‌های اچ‌تی‌ام‌ال در صفحات وب باعث ایجاد خطا در پکیج دام کراولر در تشخیص ساختار صفحه و در نتیجه ذخیره‌سازی می‌شود. به این دلیل پس از دانلود محتوای صفحه وب تگ‌های <script> و محتوای داخل آن از صفحه حذف می‌شود.

همچنین وجود کاراکترهای خط بعد (\n) و تب (\t) در صفحات دانلود شده، کاراکترهای زائدی هستند که برای کاهش حجم دیتابیس و مرتب شدن نمایش محتوای ذخیره‌شده، حذف می‌شوند.

شکل ۱۰- تعداد خبرهای منتشرشده توسط هر خبرگزاری





شکل ۱۱- تعداد خبرهای منتشرشده توسط هر خبر گذاری

پس از مراجعه به لینک آرشیو معاملات نقد در سایت بورس اوراق بهادار تهران به نشانی tse.ir/archive.html پس از انتخاب تاریخ موردنظر می‌توانید داده‌های سهام معامله‌شده در آن تاریخ را ذخیره کنید. لینک دانلود به‌صورت زیر است.

http://tse.ir/archive/Trade/Cash/TradeOneDay/TradeOneDay_1398_5_14.xls

حال با داشتن تاریخ می‌توان قسمت آخر لینک موردنظر را تولید و تمام فایل‌های معاملات را دانلود کرد. برای گرفتن تاریخ‌های معاملاتی کافی است در قسمت آرشیو سایت بورس اوراق بهادار فایل خلاصه معاملات برای سال ۹۷ را دانلود کرد و سپس از ستون تاریخ برای دانلود فایل ریز معاملات استفاده کرد. بدین منظور ستون تاریخ در یک فایل csv ذخیره‌شده است و در کد از سطرهای این ستون برای ساخت لینک استفاده می‌شود.

ازآنجایی‌که فایل‌های دانلود جدول‌های HTML هستند، اگر با فرمت html به‌جای xls ذخیره شوند، می‌توان توسط کتابخانه Pandas خوانده و یکپارچه شوند و سپس در دیتابیس ذخیره شوند.

۳-۲-ذخیره‌سازی داده‌های سهام

id	stock_namad	stock_name	stock_volume	stock_va	stock_transa
1	اخابر1	مخابرات ایران	142594	3179589	40
2	وخارزم1	سرمایه گذاری خوارزمی	508973	4336016	47
3	شپدا1	پالایش نفت اصفهان	2105211	9110653	231
4	حسابا1	سایپا	5666758	4785844	439
5	وبوعلی1	سرمایه‌گذاری بوعلی	187658	2169823	38
6	ملی1	ملی صنایع مس ایران	3746245	9874156	177
7	[illegible]	[illegible]	6686	2496984	10

شکل ۱۲- نمونه دیتای جدول transaction (تراکنش‌های سهام)

در مرحله بعد داده مربوط به شرکت‌های بورس اوراق بهادار را ذخیره می‌شود تا بتوانیم بر اساس دسته تعیین‌شده برای هر سهم، سهام مربوط به شرکت‌های گروه پتروشیمی را تعیین و الگوریتم را بر روی آن‌ها اجرا کنیم. برای این کار از لینک زیر استفاده می‌کنیم و داده‌ها را آماده ورود به سیستم می‌کنیم.

http://www.tsetmc.com/Loader.aspx?ParTree=111C1417

در این لینک داده‌های مربوط به هر سهم موجود است و می‌توان از داده‌های ذخیره شده در مرحله قبل، سهم های مربوط به گروه محصولات شیمیایی را ذخیره کرد.

stock_code	stock_group	stock_industry
IRB5IKCO8751	N2	خودرو و ساخت قطعات
IRO1APPE0001	N2	رایانه و فعالیت‌های وابسته به آن
IRO1ASIA0001	N1	ه وصندوق بازنشستگی به جزتامین اجتماعی
IRO1CONT0001	N1	ابزاریزشکی، اپتیکی و اندازگیری
IRR1CONT0101	N1	ابزاریزشکی، اپتیکی و اندازگیری
[illegible]	N1	[illegible]

شکل ۱۳- نمونه دیتا جدول company (داده‌های مربوط به شرکت‌ها)

۳-۳-پیش‌پردازش متن

پیش‌پردازش متن از مراحل بسیار مهم در متن‌کاوی است. ازاین‌رو در این پژوهش چندین مرحله جهت پیش‌پردازش و حذف داده‌های اضافه و یکسان‌سازی استفاده شده است. به دلیل حجم بالای خبرها، قبل از پیش‌پردازش خبرهایی انتخاب شدند که فقط در دو دسته سیاسی و اقتصادی دسته‌بندی‌شده بودند. دیتای جدید در جدول news_back ذخیره شد. در این مرحله ۸۰۶۱۸ خبر انتخاب شد.

۱-۳-۳-نرمال‌سازی متن

در این مرحله با استفاده از کتابخانه هضم عنوان و متن خبرها نرمال‌سازی شد. به دلیل حجم داده‌ها برای استفاده از حداکثر ظرفیت سیستم، از روش multi thread در کدهای پایتون استفاده شد. یعنی شناسه خبرها در دسته‌های ۱۰۰هزارتایی در یک حلقه قرار می‌گرفتند و کد نرمال‌سازی به‌صورت موازی برای خبرها اجرا شد.

در حالت ساده عملیات نرمال‌سازی متن با مراحل زیر انجام‌پذیر است:

اصلاح انواع حرف «ک» به معادل فارسی آنان.

اصلاح انواع حرف «ی» به معادل فارسی آنان.

بررسی همزه و انواع مختلف املاهای موجود و اصلاح هرکدام (به‌عنوان‌مثال تبدیل ؤ به و ، ئ به ی ، أ به ا ، إ به ا و...)

حذف شناسه‌ی همزه از انتهای واژه‌هایی مثل شهداء

حذف شناسه «آ» به «ا» مانند: آب به اب

اصلاح نویسه‌ی «طور» در واژه‌هایی مانند به‌طور، آن‌طور، این‌طور و ...

بررسی وجود حرف «ی» در انتهای لغاتی مانند خانه‌ی ما و اصلاح آنان

حذف تشدید از واژه‌ها

تبدیل ارقام عربی و انگلیسی به معادل فارسی.

اصلاح نویسه‌ی نیم‌فاصله

اصلاح اعراب و حذف فتحه، کسره و ضمه و همچنین تنوین‌ها

حذف نیم‌فاصله‌های تکراری

حذف نویسه‌ی «ـ» که برای کشش نویسه‌های چسبان مورداستفاده قرار می‌گیرد. مانند تبدیل«بـــــــر» و «بـــر» به «بر»

چسباندن پسوندهای «تر»، «ترین» و ... به آخر واژه‌ها

اصلاح فاصله‌گذاری «ها» در انتهای واژه‌ها و همچنین پسوندهای «های»، «هایی»، «هایم»، «هایت»، «هایش» و ...

اصلاح فاصله‌گذاری «می»، «نمی»، «درمی»، «برمی»، «بی» در ابتدای واژه‌ها

تبدیل «ه» به «ه‌ی»

تبدیل «ب» متصل به ابتدای واژه‌ها به «به»

اصلاح فاصله‌گذاری پسوندها

حذف فاصله‌ها و نیم‌فاصله‌های اضافه بکار رفته در متن

تصحیح فاصله‌گذاری در مورد علائم سجاوندی بدین‌صورت که علائم سجاوندی به لغات قبل از خود می‌چسبند و با لغت بعد از خود فاصله خواهند داشت. [۴۶]

۲-۳-۳-حذف کلمات توقف

در مرحله دوم پیش‌پردازش متن تمامی کلمات توقف از متن حذف شدند. برای جلوگیری افزایش حجم جدول‌ها داده‌های هر مرحله در جدول جدیدی ذخیره شدند. در این مرحله ابتدا داده‌های مرحله قبل در جدول news_stop ذخیره و سپس مرحله حذف کاراکترهای اضافه و کلمات توقف بر روی دیتا اجرا شد و ستون جدید content_nomalize_stop به دیتا اضافه شد.

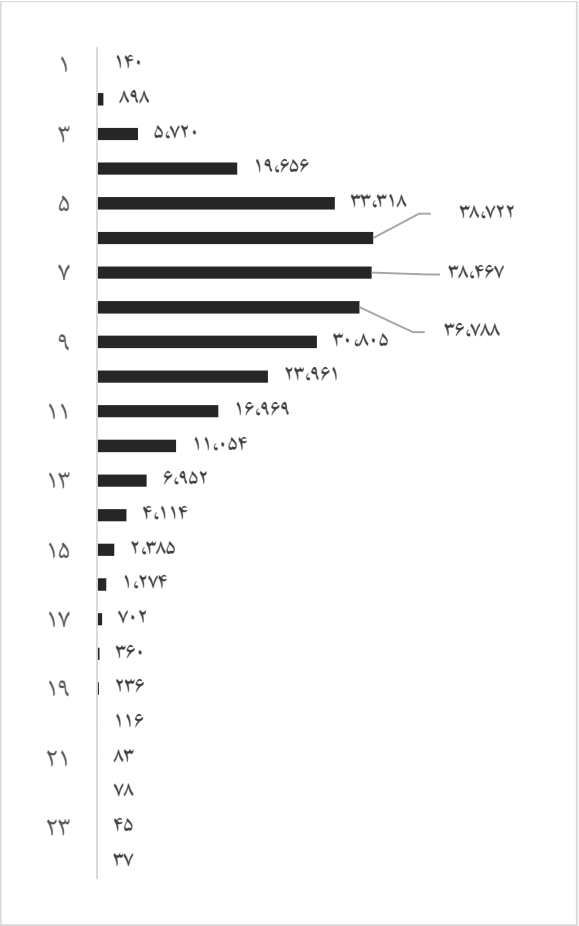
۳-۳-۳-ذخیره تمامی کلمات و حذف کلمات و کاراکترهای زائد

برای تشخیص کاراکترها و کلمات زائد ابتدا تمام کلمات در جدول جدید words ذخیره شدند. در این مرحله ابتدا با متد tokenize در کتابخانه هضم خبرها به کلمات تجزیه شدند، سپس تمامی کاراکترهای اضافه که لیست آن در زیر آمده است و همچنین تمامی کاراکترهای انگلیسی از کلمات حذف شدند.

{}⁎$#@!s-\،«»<>:؛/()

_ 《》 €﴾﴿±@؟?&×+۴۳۵۲۷;=

بعدازاین مرحله اگر کلمه موردنظر به‌صورت کامل حذف نشد و در جدول کلمات موجود نباشد در جدول ذخیره می‌شود. درمجموع ۲۷۲۸۸۰ کلمه یکتا ذخیره شد.



شکل ۱۴-تعداد کلمات بر اساس تعداد کاراکترها

از این تعداد کلمات ۱۰۹۷ کلمه به‌عنوان کلمات زائد شناسایی و به‌عنوان کلماتی که نباید در محاسبات شمارش شوند علامت‌گذاری شدند. که عمدتاً کلمات یک و دو کاراکتری هستند که شامل علامت‌های خاص و نگارشی، حروف، حرف‌های اضافه و غیره هستند. در این مرحله اگر لغات نامه استانداردی در اختیار باشد که بتوان کلمات استخراج‌شده را با آن مقایسه کرد و کلماتی که در لغت‌نامه نیستند را مشخص کرد می‌توان ابعاد مسئله را کاهش داد. همچنین اگر برنامه‌ای باشد تا غلط‌های املایی را اصلاح کرد می‌تواند به کاهش ابعاد کمک کند.

۳-۳-۴-ایجاد جدول کلمات و شمارش کلمات در هرروز

پس از مرحله پیش‌پردازش متن، ماتریس کلمات ورودی به مدل ایجاد شد. در این بخش مراحل زیر بر روی دیتا انجام شد. در قدم اول تعداد تکرارهای هر کلمه در روز موردنظر شمارش شد و شناسه کلمه به همراه تعداد تکرار در جدول extract_word ذخیره شد. ازآنجایی‌که ساعت معاملات در بورس اوراق بهادار تهران از ساعت ۸:۳۰ تا ۱۲:۳۰ دقیقه است، خبرهایی که در ساعت بعد از ۱۲:۳۰ منتشر می‌شوند جزء اخبار روز بعد تلقی می‌شوند و بر روی قیمت‌های روز آینده تأثیرگذار هستند. بدین‌جهت کلمات آن‌ها در کلمات روز آینده شمارش می‌شوند. تعداد کلمات به تفکیک روزهای انتشار خبر در این مرحله ۳۴۳۳۰۸۳۸ است.

۳-۳-۵-انتخاب ویژگی‌ها با استفاده از روش TF-IDF

روش tf × idf (تکرار کلمه × تکرار سند معکوس)، یک روش نمایندگی است که معمولاً در روش‌های استخراج ویژگی‌ها از متن استفاده می‌شود. به‌صورت پیش‌فرض، تکرار کلمات[۷۴](tf) در سند نشان دهنده میزان اهمیت اصطلاح در این سند است و تکرار سند در این اصطلاح[۷۵](df)، درصد اسناد حاوی این اصطلاح) نشان دهنده میزان اهمیت این اصطلاح در کل متن‌ها است. مقدار df پایین نشان می‌دهد که این اصطلاح در بسیاری از اسناد ظاهر نمی‌شود و این نشان دهنده منحصربه‌فرد بودن این اصطلاح در اسناد است. بنابراین، به‌جای استفاده از df، idf (معکوس df) را به‌عنوان برنامه وزن دهی انتخاب می‌شود. بنابراین وزن بالا در روش tf × idf نشان دهنده تکرار زیاد یک کلمه در یک سند و اسناد کم حاوی این متن است. [۴۷]

در این بخش با استفاده از Pandas ماتریس کلمات جهت محاسبه TF-IDF تشکیل شد. تمام خبرهایی که در یک روز منتشرشده‌اند به‌عنوان یک سند در نظر گرفته می‌شوند. پس باید برای محاسبه TF تعداد تکرار کلمات در هرروز را به دست آورد. سپس با تقسیم تکرار هر کلمه در هر روز TF محاسبه می‌شود. برای محاسبه IDF از فرمول $IDF = log(\frac{c_d}{i_d})$ استفاده شده است که در آن c_d نشان‌دهنده کل سندها است. که در پژوهش ما برابر تعداد روزهای تولید خبر و برابر ۱۳۶ است. i_d نیز نشان‌دهنده تعداد اسناد است که ما هرروز را برابر یک سند گرفتیم. سپس با ضرب TF در IDF شاخص TF-IDF برای هر کلمه محاسبه می‌شود. سپس ماتریس کلمات تشکیل‌شده و براساس تعداد ویژگی‌های موردنظر فیلتر می‌شود. مثلاً ابتدا ۱۰۰۰ ویژگی که بیشترین امتیاز TF-IDF را گرفته‌اند انتخاب می‌شود و از بین ویژگی‌ها فقط ۱۰۰۰ ستون انتخاب می‌شود.

۳-۴-آماده‌سازی دیتای بورس

ازآنجایی‌که نوع مدل انتخابی یادگیری با ناظر بر اساس جهت قیمت سهام است، باید دیتای سهم موردنظر را بر اساس اینکه

[۷۴] Term frequency

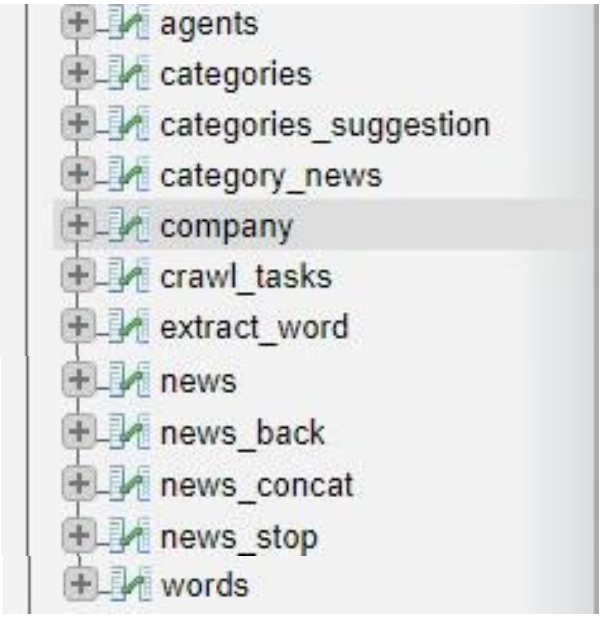
[۷۵] Document frequency

کاهشی یا افزایشی بوده است برچسب‌گذاری کنیم. سپس ماتریس کلمات و جهت قیمت را ادغام می‌کنیم تا دیتا جهت ورود به الگوریتم آماده شود. قیمت منفی و صفر نسبت به‌روز قبل را با برچسب صفر و قیمت مثبت را با برچسب یک برای یک سهم نمایش داده می‌شود.

created_at	change_end_price
2018-03-25	1
2018-03-26	0
2018-03-27	0
2018-03-28	0
2018-04-03	0
2018-04-04	0
2018-04-07	0
2018-04-08	1
2018-04-09	1

شکل ۱۵- برچسب‌گذاری قیمت‌های سهم در هرروز

از ادغام ماتریس کلمات و جهت قیمت یک سهم خاص در روزهای معاملاتی، ماتریس ورودی به الگوریتم ماشین بردار پشتیبان تشکیل می‌شود. به دلیل حجم بالای دیتا همان‌طور که در شکل می‌بینید انجام این فرایند حداقل به ۲۴ گیگ حافظه موقت جهت ایجاد ماتریس کلمات نیاز است.



شکل ۱۶-جدول‌های ذخیره‌سازی دیتا در سیستم خزشگر

agent: جدول ذخیره‌سازی داده‌های خبرگزاری (عنوان، لینک صفحه اصلی و غیره)

Crawl_task: جدول ذخیره‌سازی لینک خبرها به عنوان وظیفه

news: اخبار ذخیره شده هر لینک

news_back: متن خبرهای پیش پردازش شده

news_stop: حذف کلمات توقف از متن خبرها

news_concat: تجمع متن خبرهای یک روز

categories: جدول ذخیره‌سازی دسته‌بندی‌ها

categories suggestion: جدول شناسایی سایر دسته‌ها به غیر از سیاسی و اقتصادی

category_news: جدول تعیین دسته خبرها

company: جدول ذخیره‌سازی داده‌های شرکت‌های بورسی

transaction: جدول ذخیره‌سازی داده‌های بورسی سهم

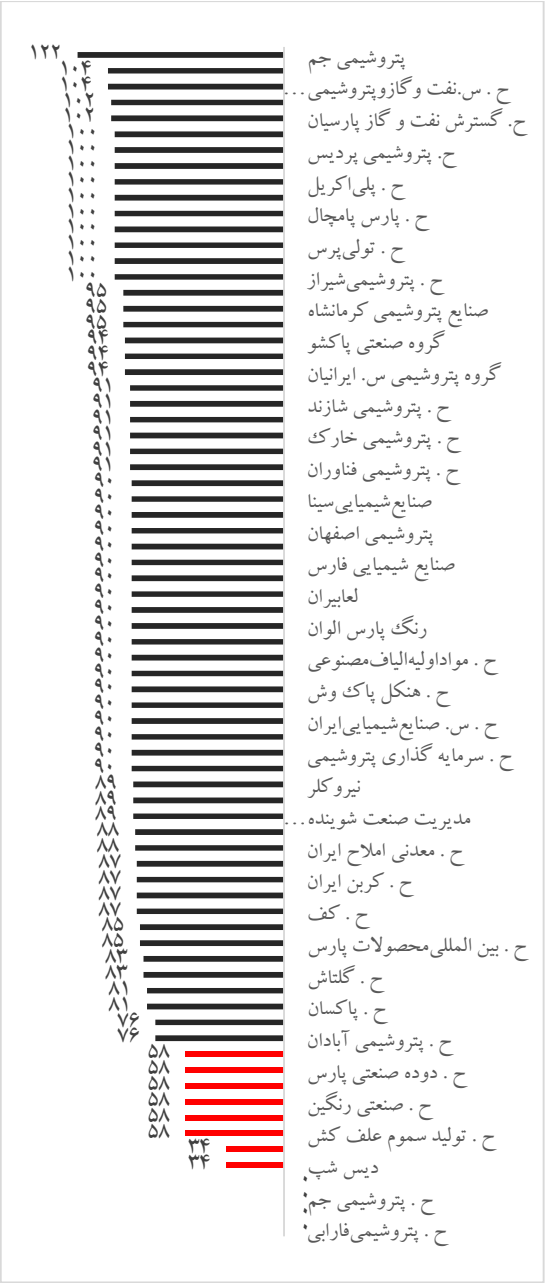
words: کلمات استخراج شده یکتا از خبرها بعد از پاکسازی

Extract_word: کلمات موجود در هر خبر بر اساس جدول words

۳-۵-اجرای مدل ماشین بردار پشتیبان با کرنل‌های خطی و غیرخطی

در این پژوهش از ۴ الگوریتم ماشین بردار پشتیبان با تنظیمات مختلف استفاده شده است که درنهایت به بررسی ۷ مدل خواهیم پرداخت. مدل‌های بررسی‌شده شامل: کرنل خطی، کرنل چندجمله‌ای درجه ۲ با گاما auto، کرنل چندجمله‌ای درجه ۲ با گاما scale، کرنل شعاعی با گاما auto ، کرنل شعاعی با گاما scale، کرنل سیگموئید با گاما auto، کرنل سیگموئید با گاما scale.

برای آموزش مدل از دو پارامتر متغیر برای انجام محاسبات مختلف استفاده شده است. پارامتر اول ویژگی‌های انتخابی و پارامتر دوم تعداد داده‌های آزمون است. برای هر سهم پارامتر ویژگی‌های انتخابی از ۱۰۰۰ تا ۱۹۰۰۰ ویژگی است. مدل برای هر ویژگی انتخابی با تعداد داده‌های آزمون ۱۰ درصد تا ۳۵ درصد اجرا می‌شود. سهم‌هایی انتخاب‌شده‌اند که بیش از ۷۰ روز در بازه زمانی موردنظر معامله‌شده‌اند.



شکل ۱۷- تعداد روزهای معاملاتی هر سهم

پس از آموزش مدل نتایج ارزیابی پیش‌بینی‌ها در شاخص‌های دقت، صحت و امتیاز و میانگین ساده و وزنی هر سهم برای هر پارامتر محاسبه می‌شود.

```
number of future:1000
[[9 0]
 [3 4]]
              precision    recall  f1-score   support

         0.0       0.75      1.00      0.86         9
         1.0       1.00      0.57      0.73         7

   micro avg       0.81      0.81      0.81        16
weighted avg       0.86      0.81      0.80        16
```

شکل ۱۸- نمونه‌ای از نتایج ارزیابی مدل

در شکل ۱۸ خط اول: نشان‌دهنده تعداد ویژگی‌های انتخابی
خط دوم و سوم: ماتریس در هم ریختگی
خط چهارم: شاخص‌های ارزیابی
خط ششم: نتایج ارزیابی پیش‌بینی جهت‌های منفی
خط هفتم: نتایج ارزیابی پیش‌بینی جهت‌های مثبت
خط هشتم: میانگین ساده هر شاخص
خط نهم: میانگین وزنی هر شاخص

ازآنجایی‌که امتیاز F معیار دقت و صحت را به صورت همزمان در نظر می‌گیرد شاخص مناسبی برای ارزیابی یک مدل دسته‌بندی است، برای تحلیل نتایج این پژوهش از این شاخص استفاده شده است.

نتیجه نهایی هر الگوریتم برای هر سهم در یک فایل CSV برای تحلیل در نرم‌افزار اکسل ذخیره می‌شود. نتایج حاصل از هر مدل در یک پوشه به اسم کرنل مورداستفاده شده ذخیره می‌شود. برای تهیه گزارش‌ها علاوه بر نتایج ارزیابی تعداد ویژگی‌های انتخاب‌شده و اندازه نمونه‌های آزمون نیست در فایل ذخیره می‌شود.

	A	B	C	D	E	F	G
1	index	f1-score	precision	recall	support	number_of	test_size
2	0	0.6	0.5	0.75	4	1000	0.1
3	1	0.5	0.666667	0.4	5	1000	0.1
4	micro avg	0.555556	0.555556	0.555556	9	1000	0.1
5	macro avg	0.55	0.583333	0.575	9	1000	0.1
6	weighted a	0.544444	0.592593	0.555556	9	1000	0.1
7	0	0.285714	0.25	0.333333	3	2000	0.1
8	1	0.545455	0.6	0.5	6	2000	0.1
9	micro avg	0.444444	0.444444	0.444444	9	2000	0.1
10	macro avg	0.415584	0.425	0.416667	9	2000	0.1
11	weighted a	0.458874	0.483333	0.444444	9	2000	0.1
12	0	0.222222	1	0.125	8	3000	0.1
13	1	0.222222	0.125	1	1	3000	0.1

شکل ۱۹- نمونه فایل ارزیابی پیش‌بینی مدل

۶-۳-نتایج ارزیابی مدل برای هر سهم

در این بخش به بررسی نتایج ارزیابی هر سهم خواهیم پرداخت و برای هر سهم بهترین کرنل با پارامترهای مختلف را انتخاب خواهیم کرد. سهم‌ها به ترتیب بیشترین روز معاملاتی موردبررسی قرار می‌گیرند. برای تحلیل نتایج از نرم‌افزار اکسل استفاده شده است. در قدم اول فایل‌های ارزیابی مدل توسط ابزار پاور کوئری بارگذاری و آماده شدند. سپس توسط ابزار پیوت تیبل خروجی‌ها و نمودارها ایجاد شد. ازآنجایی‌که برای ارزیابی از میانگین وزنی امتیاز F استفاده شد، تعداد سطرهای بارگذاری شده توسط اکسل ۲۰۰۶۴ سطر است.

تعداد روش‌ها: ۷

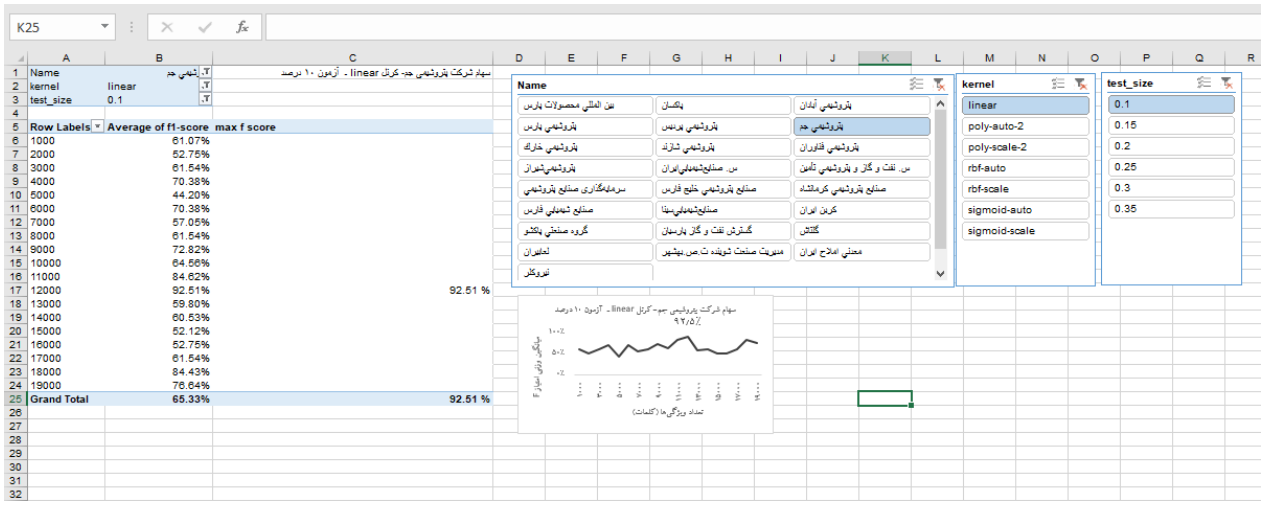
تعداد انتخاب شاخص‌ها: ۱۹

تعداد انتخاب نمونه آزمایش: ۶

تعداد سهم تحلیل‌شده: ۲۵

پس تعداد نتایج برابر است:

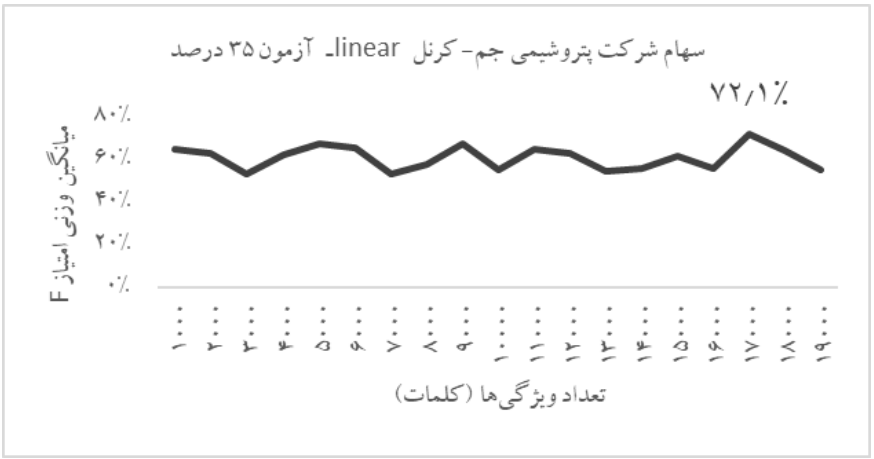
۷×۱۹×۶×۲۵=۲۰۰۶۴

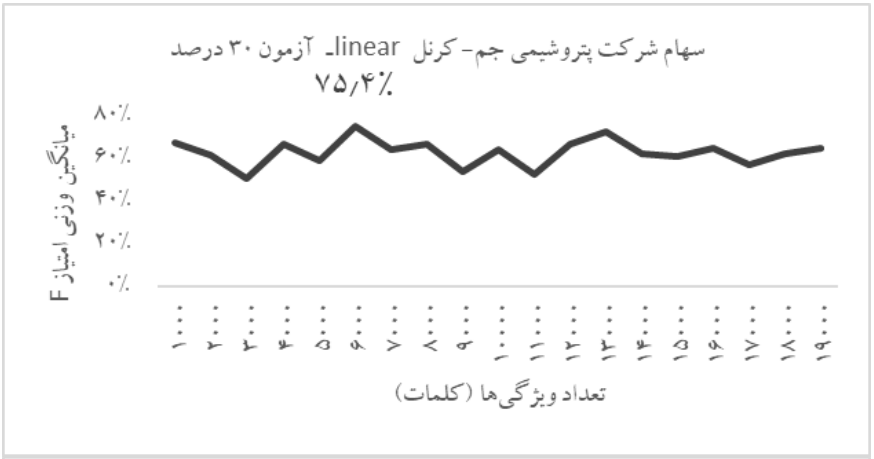


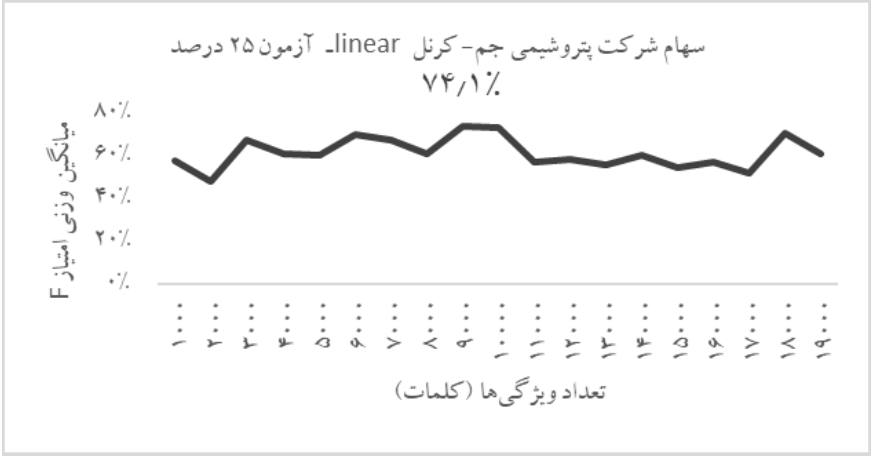
شکل ۲۰- تحلیل نتایج در اکسل

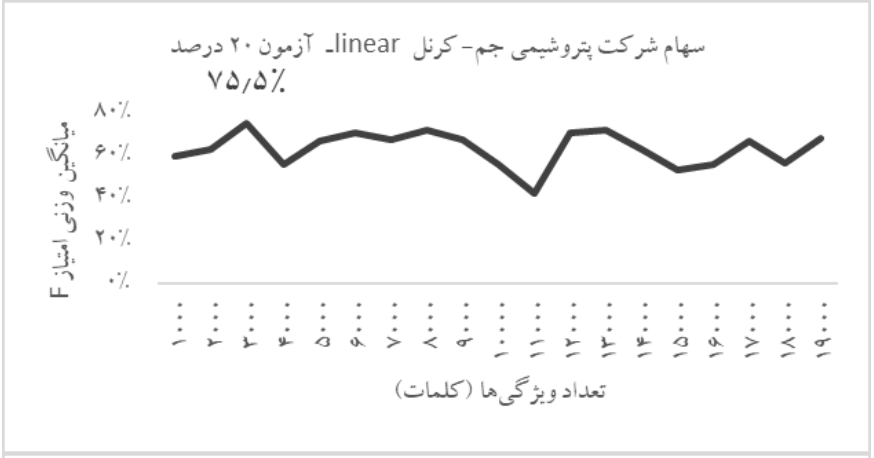
۶-۳-۱-نمونه ارزیابی نتایج برای سهم شرکت پتروشیمی جم

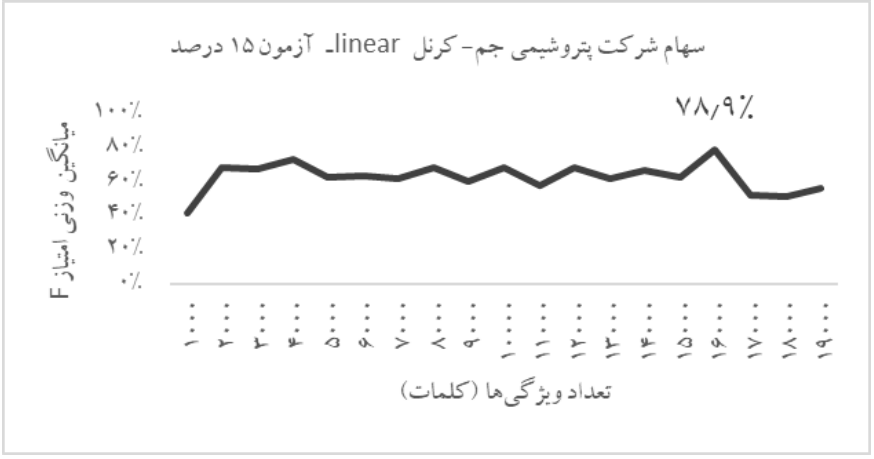
نتایج آزمون با کرنل خطی برای سهم پتروشیمی جم به‌صورت زیر است.

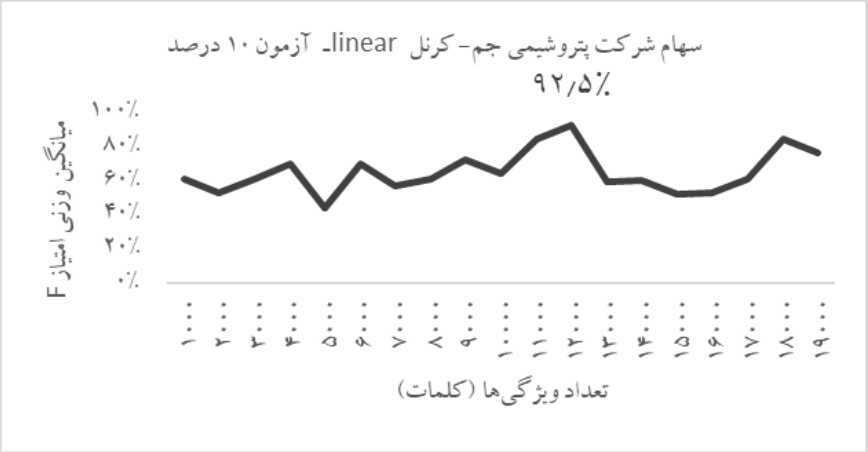












۱-۳-۶-ارزیابی کلی نتایج

بررسی ۲۰۰۶۴ حالتی که برای این ۲۵ سهم اتفاق می‌افتد به‌صورت تک‌به‌تک بسیار کار دشواری است. بدین منظور برای بررسی بهترین حالت از پیوت تیبل در اکسل استفاده شده است و برای هر کرنل بهترین حالت انتخاب می‌شود. برای کرنل خطی برای هر سهم تنظیمات زیر بهترین نتیجه را خواهد داشت. اطلاعات این جداول به شرح زیر است:

شرکت: نام شرکت در گروه محصولات شیمیایی

ویژگی‌ها: تعداد کلمات انتخاب شده بر اساس بیشترین امتیاز TF-IDF

درصد آزمایش: درصد روزهای معاملاتی که به عنوان بخش تست در نظر گرفته شده اند.

کرنل: امتیاز F بدست آماده از این کرنل با توجه به تعداد ویژگی‌های انتخاب شده و درصد آزمایش

جدول ۳- ارزیابی نتایج کرنل خطی

شرکت	ویژگی‌ها	درصد آزمایش	کرنل خطی
بین‌المللی محصولات پارس	۹۰۰۰	۰.۲۵	۷۶.۹۱%
پاکسان	۱۱۰۰۰	۰.۱	۸۹.۵۷%
پتروشیمی آبادان	۱۰۰۰	۰.۳	۸۶.۹۶%
پتروشیمی پارس	۱۷۰۰۰	۰.۳	۸۱.۱۹%
پتروشیمی پردیس	۱۸۰۰۰	۰.۱۵	۷۳.۳۳%
پتروشیمی جم	۱۲۰۰۰	۰.۱	۹۲.۵۱%
پتروشیمی خارک	۵۰۰۰	۰.۱۵	۷۱.۴۳%
پتروشیمی شازند	۹۰۰۰	۰.۱۵	۷۱.۴۳%
پتروشیمی فناوران	۹۰۰۰	۰.۱	۱۰۰.۰۰%
پتروشیمی شیراز	۳۰۰۰	۰.۱۵	۸۶.۰۰%
س. صنایع‌شیمیایی‌ایران	۱۵۰۰۰	۰.۲۵	۷۳.۸۱%
س. نفت و گاز و پتروشیمی تأمین	۸۰۰۰	۰.۱	۱۰۰.۰۰%
سرمایه‌گذاری صنایع پتروشیمی	۱۷۰۰۰	۰.۱	۸۸.۸۹%
صنایع پتروشیمی خلیج فارس	۷۰۰۰	۰.۱	۹۰.۳۳%
صنایع پتروشیمی کرمانشاه	۷۰۰۰	۰.۱	۹۱.۳۷%
صنایع شیمیایی فارس	۶۰۰۰	۰.۱	۷۷.۷۸%
صنایع‌شیمیایی سینا	۱۱۰۰۰	۰.۱	۸۹.۱۸%
کربن ایران	۱۸۰۰۰	۰.۱	۸۱.۷۵%
گروه صنعتی پاکشو	۱۵۰۰۰	۰.۲	۸۴.۶۴%
گسترش نفت و گاز پارسیان	۶۰۰۰	۰.۱	۹۱.۰۶%
گلتاش	۳۰۰۰	۰.۱	۸۹.۵۷%
لعابیران	۱۷۰۰۰	۰.۱۵	۷۸.۴۶%
مدیریت صنعت شوینده ت.ص.بهشهر	۷۰۰۰	۰.۱	۸۸.۸۹%
معدنی املاح ایران	۷۰۰۰	۰.۱	۸۸.۲۱%
نیروکلر	۶۰۰۰	۰.۱	۸۹.۵۷%
میانگین			۸۵.۳۱%

با توجه به داده‌های بالا مدل خطی می‌تواند به‌صورت میانگین تا ۸۵,۳۱ درصد جهت قیمت‌ها را درست پیش‌بینی کند.

جدول ۴-ارزیابی نتایج کرنل چندجمله‌ای درجه ۲ با گاما auto

شرکت	تعداد ویژگی‌ها	درصد آزمایش	معادله درجه ۲ با گاما auto
بین‌المللی محصولات پارس	۱۳۰۰۰	۰.۲	۸۱.۵۸%
پاکسان	۵۰۰۰	۰.۱	۱۰۰.۰۰%
پتروشیمی آبادان	۱۳۰۰۰	۰.۱	۸۷.۷۰%
پتروشیمی پارس	۱۳۰۰۰	۰.۱	۸۰.۰۰%
پتروشیمی پردیس	۳۰۰۰	۰.۱	۸۹.۳۳%
پتروشیمی جم	۳۰۰۰	۰.۱	۸۴.۴۳%
پتروشیمی خارک	۶۰۰۰	۰.۱۵	۷۰.۸۳%
پتروشیمی شازند	۴۰۰۰	۰.۱	۷۰.۹۹%
پتروشیمی فناوران	۱۱۰۰۰	۰.۱	۱۰۰.۰۰%
پتروشیمی شیراز	۶۰۰۰	۰.۱	۱۰۰.۰۰%
س. صنایع‌شیمیایی‌ایران	۱۶۰۰۰	۰.۱	۸۷.۵۰%
س. نفت و گاز و پتروشیمی تأمین	۴۰۰۰	۰.۱	۱۰۰.۰۰%
سرمایه‌گذاری صنایع پتروشیمی	۱۴۰۰۰	۰.۱	۸۷.۴۱%
صنایع پتروشیمی خلیج فارس	۵۰۰۰	۰.۱	۸۰.۰۰%
صنایع پتروشیمی کرمانشاه	۵۰۰۰	۰.۱	۸۰.۰۰%
صنایع شیمیایی فارس	۱۶۰۰۰	۰.۱	۷۸.۳۳%
صنایع شیمیایی سینا	۱۸۰۰۰	۰.۱	۸۹.۱۸%
کربن ایران	۱۴۰۰۰	۰.۱	۹۰.۳۷%
گروه صنعتی پاکشو	۱۸۰۰۰	۰.۱	۹۰.۱۰%
گسترش نفت و گاز پارسیان	۱۱۰۰۰	۰.۱	۸۲.۱۲%
گلتاش	۱۲۰۰۰	۰.۱۵	۸۴.۶۲%
لعابیران	۱۵۰۰۰	۰.۱	۸۹.۱۸%
مدیریت صنعت شوینده ت.ص.بهشهر	۸۰۰۰	۰.۱	۸۹.۱۸%
معدنی املاح ایران	۹۰۰۰	۰.۱	۱۰۰.۰۰%
نیروکلر	۲۰۰۰	۰.۱	۸۸.۸۹%
میانگین			۸۷.۲۷%

با توجه به داده‌های بالا کرنل معادله درجه ۲ با گاما auto می‌تواند به‌صورت میانگین تا ۸۷,۲۷ درصد جهت قیمت‌ها را درست پیش‌بینی کند.

جدول ۵-ارزیابی نتایج کرنل چندجمله‌ای درجه ۲ با گاما scale

شرکت	تعداد ویژگی‌ها	تعداد آزمایش	معادل درجه ۲ با گاما scale
بین‌المللی محصولات پارس	۴۰۰۰	۰.۱۵	۸۴.۶۲%
پاکسان	۱۹۰۰۰	۰.۱۵	۷۷.۵۶%
پتروشیمی آبادان	۱۳۰۰۰	۰.۱	۷۵.۰۰%
پتروشیمی پارس	۱۹۰۰۰	۰.۱۵	۶۹.۱۴%
پتروشیمی پردیس	۱۰۰۰	۰.۲	۷۸.۱۱%
پتروشیمی جم	۴۰۰۰	۰.۱۵	۷۳.۲۵%
پتروشیمی خارک	۱۵۰۰۰	۰.۱۵	۶۲.۵۷%
پتروشیمی شازند	۴۰۰۰	۰.۱	۷۱.۱۱%
پتروشیمی فناوران	۳۰۰۰	۰.۱۵	۸۵.۷۱%
پتروشیمی شیراز	۱۲۰۰۰	۰.۱	۷۱.۱۱%
س. صنایع شیمیایی ایران	۱۵۰۰۰	۰.۱	۸۳.۶۶%
س. نفت و گاز و پتروشیمی تأمین	۷۰۰۰	۰.۱	۷۳.۶۴%
سرمایه‌گذاری صنایع پتروشیمی	۵۰۰۰	۰.۱	۸۷.۵۰%
صنایع پتروشیمی خلیج فارس	۲۰۰۰	۰.۱	۸۵.۲۶%
صنایع پتروشیمی کرمانشاه	۱۴۰۰۰	۰.۱	۱۰۰.۰۰%
صنایع شیمیایی فارس	۹۰۰۰	۰.۲۵	۶۰.۸۷%
صنایع‌شیمیایی‌سینا	۸۰۰۰	۰.۱۵	۷۲.۰۲%
کربن ایران	۱۶۰۰۰	۰.۱۵	۸۹.۴۲%
گروه صنعتی پاکشو	۱۴۰۰۰	۰.۲	۷۶.۹۹%
گسترش نفت و گاز پارسیان	۱۷۰۰۰	۰.۲۵	۶۵.۶۴%
گلتاش	۵۰۰۰	۰.۱	۶۸.۰۶%
لعابیران	۸۰۰۰	۰.۱	۸۸.۸۹%
مدیریت صنعت شوینده	۹۰۰۰	۰.۱	۸۸.۲۱%

ت.ص.بهشهر			
معدنی املاح ایران	۳۰۰۰	۰.۱۵	۷۲.۰۲%
نیروکلر	۱۰۰۰	۰.۱	۸۷.۴۱%
میانگین			۷۷.۹۱%

با توجه به داده‌های بالا کرنل معادله درجه ۲ با گاما scale می‌تواند به‌صورت میانگین تا ۷۷,۹۱ درصد جهت قیمت‌ها را درست پیش‌بینی کند.

جدول ۶-ارزیابی نتایج کرنل شعاعی با گاما auto

شرکت	تعداد ویژگی‌ها	تعداد آزمایش	کرنل شعاعی با گاما auto
بین‌المللی محصولات پارس	۱۲۰۰۰	۰.۱	۶۸.۰۶%
پاکسان	۱۰۰۰	۰.۱	۸۳.۶۶%
پتروشیمی آبادان	۹۰۰۰	۰.۱	۶۴.۲۹%
پتروشیمی پارس	۲۰۰۰	۰.۱	۶۱.۶۷%
پتروشیمی پردیس	۱۹۰۰۰	۰.۱	۱۰۰.۰۰%
پتروشیمی جم	۱۸۰۰۰	۰.۱	۹۱.۷۲%
پتروشیمی خارک	۱۶۰۰۰	۰.۳	۵۳.۳۹%
پتروشیمی شازند	۱۱۰۰۰	۰.۱	۷۱.۱۱%
پتروشیمی فناوران	۵۰۰۰	۰.۱	۷۱.۱۱%
پتروشیمی شیراز	۱۲۰۰۰	۰.۱۵	۷۸.۴۰%
س. صنایع‌شیمیایی‌ایران	۷۰۰۰	۰.۱	۸۷.۴۱%
س. نفت و گاز و پتروشیمی تأمین	۱۷۰۰۰	۰.۲۵	۸۲.۲۷%
سرمایه‌گذاری صنایع پتروشیمی	۵۰۰۰	۰.۱	۸۳.۶۶%
صنایع پتروشیمی خلیج فارس	۵۰۰۰	۰.۱	۸۸.۶۳%
صنایع پتروشیمی کرمانشاه	۱۰۰۰۰	۰.۱	۸۸.۶۳%
صنایع شیمیایی فارس	۱۷۰۰۰	۰.۲	۶۴.۰۷%
صنایع‌شیمیایی‌سینا	۳۰۰۰	۰.۱	۶۸.۰۶%
کربن ایران	۱۰۰۰۰	۰.۱	۸۷.۴۱%
گروه صنعتی پاکشو	۱۶۰۰۰	۰.۱	۸۰.۰۰%
گسترش نفت و گاز پارسیان	۱۱۰۰۰	۰.۱	۸۱.۸۲%
گلتاش	۸۰۰۰	۰.۱۵	۷۷.۵۶%
لعابیران	۸۰۰۰	۰.۱۵	۷۸.۲۳%
مدیریت صنعت	۱۹۰۰۰	۰.۲۵	۶۱.۰۲%
شوینده ت.ص.بهشهر			
معدنی املاح ایران	۱۶۰۰۰	۰.۲۵	۶۳.۳۳%
نیروکلر	۱۱۰۰۰	۰.۱	۸۳.۶۶%
میانگین			۷۶.۱۹%

با توجه به داده‌های بالا کرنل با گاما auto می‌تواند به‌صورت میانگین تا ۷۶,۱۹ درصد جهت قیمت‌ها را درست پیش‌بینی کند.

جدول ۷-ارزیابی نتایج کرنل شعاعی با گاما scale

شرکت	تعداد ویژگی‌ها	تعداد آزمایش	کرنل شعاعی با گاما scale
بین‌المللی محصولات پارس	۱۲۰۰۰	۰.۱	۷۷.۷۸%
پاکسان	۱۰۰۰۰	۰.۱	۸۳.۶۶%
پتروشیمی آبادان	۵۰۰۰	۰.۱۵	۷۶.۰۵%
پتروشیمی پارس	۸۰۰۰	۰.۱	۸۳.۶۶%
پتروشیمی پردیس	۱۱۰۰۰	۰.۱	۸۵.۲۶%
پتروشیمی جم	۱۸۰۰۰	۰.۱	۶۶.۸۹%
پتروشیمی خارک	۹۰۰۰	۰.۱	۶۹.۷۰%
پتروشیمی شازند	۱۸۰۰۰	۰.۱۵	۶۹.۱۴%
پتروشیمی فناوران	۱۰۰۰۰	۰.۱۵	۷۵.۵۱%
پتروشیمی شیراز	۵۰۰۰	۰.۱	۸۵.۲۶%
س. صنایع‌شیمیایی‌ایران	۱۰۰۰	۰.۱	۶۸.۰۶%
س. نفت و گاز و پتروشیمی تأمین	۹۰۰۰	۰.۱	۷۳.۶۴%
سرمایه‌گذاری صنایع پتروشیمی	۹۰۰۰	۰.۱	۷۳.۸۱%
صنایع پتروشیمی خلیج فارس	۴۰۰۰	۰.۱	۸۵.۲۶%
صنایع پتروشیمی کرمانشاه	۱۱۰۰۰	۰.۲	۷۴.۳۰%
صنایع شیمیایی فارس	۱۰۰۰	۰.۱	۵۵.۵۶%
صنایع‌شیمیایی‌سینا	۱۷۰۰۰	۰.۱	۶۸.۰۶%
کربن ایران	۱۳۰۰۰	۰.۱۵	۶۹.۱۴%
گروه صنعتی پاکشو	۱۰۰۰۰	۰.۱	۸۵.۲۶%
گسترش نفت و گاز پارسیان	۱۴۰۰۰	۰.۱	۷۳.۶۴%
گلتاش	۱۶۰۰۰	۰.۲	۷۴.۳۸%

لعابیران	۳۰۰۰	۰.۱۵	۵۵.۳۶%
مدیریت صنعت شوینده ت.ص.بهشهر	۱۰۰۰	۰.۱	۶۵.۸۰%
معدنی املاح ایران	۱۷۰۰۰	۰.۱	۷۷.۷۸%
نیروکلر	۴۰۰۰	۰.۲	۶۹.۴۴%
میانگین			۷۳.۷۰%

با توجه به داده‌های بالا کرنل شعاعی با گاما scale می‌تواند به‌صورت میانگین تا ۷۳.۷۰%درصد جهت قیمت‌ها را درست پیش‌بینی کند که نسبت به سایر کرنل‌ها نتیجه مناسبی محسوب نمی‌شود.

جدول ۸-ارزیابی نتایج کرنل سیگموئید با گاما auto

شرکت	تعداد ویژگی‌ها	تعداد آزمایش	کرنل سیگموئید با گاما auto
بین‌المللی محصولات پارس	۱۶۰۰۰	۰.۱	۸۳.۶۶%
پاکسان	۳۰۰۰	۰.۱۵	۸۸.۶۲%
پتروشیمی آبادان	۷۰۰۰	۰.۱	۶۴.۲۹%
پتروشیمی پارس	۷۰۰۰	۰.۱	۱۰۰.۰۰%
پتروشیمی پردیس	۱۷۰۰۰	۰.۱	۸۵.۲۶%
پتروشیمی جم	۱۳۰۰۰	۰.۱	۶۶.۸۹%
پتروشیمی خارک	۳۰۰۰	۰.۲	۳۶.۳۰%
پتروشیمی شازند	۶۰۰۰	۰.۱	۷۱.۱۱%
پتروشیمی فناوران	۸۰۰۰	۰.۱	۵۷.۶۵%
پتروشیمی شیراز	۲۰۰۰	۰.۱۵	۷۱.۱۱%
س. صنایع‌شیمیایی‌ایران	۲۰۰۰	۰.۱	۶۸.۰۶%
س. نفت و گاز و پتروشیمی تأمین	۱۱۰۰۰	۰.۱	۷۳.۶۴%
سرمایه‌گذاری صنایع پتروشیمی	۷۰۰۰	۰.۱۵	۷۹.۱۲%
صنایع پتروشیمی خلیج فارس	۱۶۰۰۰	۰.۱۵	۹۰.۱۱%
صنایع پتروشیمی کرمانشاه	۱۰۰۰	۰.۱۵	۷۱.۱۱%
صنایع شیمیایی فارس	۸۰۰۰	۰.۱	۳۹.۶۸%
صنایع‌شیمیایی‌سینا	۱۰۰۰	۰.۱	۶۸.۰۶%
کربن ایران	۶۰۰۰	۰.۱	۸۳.۶۶%
گروه صنعتی پاکشو	۵۰۰۰	۰.۱	۸۵.۲۶%
گلتاش	۴۰۰۰	۰.۱	۸۳.۶۶%
لعابیران	۳۰۰۰	۰.۱۵	۴۱.۵۶%
مدیریت صنعت شوینده ت.ص.بهشهر	۲۰۰۰	۰.۱۵	۴۱.۵۶%
معدنی املاح ایران	۱۰۰۰	۰.۲	۳۳.۳۳%
نیروکلر	۱۴۰۰۰	۰.۱	۸۳.۶۶%
میانگین			۶۹.۴۷%

با توجه به داده‌های بالا کرنل سیگموئید با گاما auto می‌تواند به‌صورت میانگین تا ۶۹,۴۷ درصد جهت قیمت‌ها را درست پیش‌بینی کند که نسبت به سایر الگوریتم‌ها نتایج نامناسب‌تری است.

جدول ۹-ارزیابی نتایج کرنل سیگموئید با گاما scale

شرکت	تعداد ویژگی‌ها	تعداد آزمایش	کرنل سیگموئید با گاما scale
بین‌المللی محصولات پارس	۵۰۰۰	۰.۱	۶۸.۰۶%
پاکسان	۷۰۰۰	۰.۱	۱۰۰.۰۰%
پتروشیمی آبادان	۱۰۰۰	۰.۱	۶۴.۲۹%
پتروشیمی پارس	۱۷۰۰۰	۰.۱۵	۶۹.۱۰%
پتروشیمی پردیس	۱۸۰۰۰	۰.۱	۸۵.۲۶%
پتروشیمی جم	۱۰۰۰	۰.۱۵	۶۶.۴۱%
پتروشیمی خارک	۱۷۰۰۰	۰.۲۵	۵۰.۳۱%
پتروشیمی شازند	۱۵۰۰۰	۰.۱	۷۱.۱۱%
پتروشیمی فناوران	۳۰۰۰	۰.۱	۵۷.۶۵%
پتروشیمی شیراز	۱۵۰۰۰	۰.۱	۸۵.۲۶%
س. صنایع‌شیمیایی‌ایران	۴۰۰۰	۰.۱	۶۸.۰۶%
س. نفت و گاز و پتروشیمی تأمین	۸۰۰۰	۰.۱	۷۳.۶۴%
سرمایه‌گذاری صنایع پتروشیمی	۱۰۰۰	۰.۱	۶۸.۰۶%
صنایع پتروشیمی خلیج فارس	۱۰۰۰	۰.۱	۸۵.۲۶%
صنایع پتروشیمی کرمانشاه	۱۰۰۰	۰.۱	۷۱.۱۱%
صنایع شیمیایی فارس	۱۸۰۰۰	۰.۲	۷۷.۲۰%

صنایع‌شیمیایی‌سینا	۴۰۰۰	۰.۱	۶۸.۰۶%
کربن ایران	۲۰۰۰	۰.۱	۸۳.۶۶%
گروه صنعتی پاکشو	۱۱۰۰۰	۰.۱	۸۵.۲۶%
گسترش نفت و گاز پارسیان	۶۰۰۰	۰.۱۵	۷۷.۰۸%
گلتاش	۱۰۰۰	۰.۱	۶۸.۰۶%
لعابیران	۱۹۰۰۰	۰.۲	۵۴.۴۰%
مدیریت صنعت شوینده ت.ص.بهشهر	۱۲۰۰۰	۰.۲۵	۵۶.۶۹%
معدنی املاح ایران	۶۰۰۰	۰.۱۵	۴۷.۵۰%
نیروکلر	۱۰۰۰	۰.۱۵	۷۹.۱۲%
میانگین			۷۱.۱۴%

با توجه به داده‌های بالا کرنل سیگموئید با گاما scale می‌تواند به‌صورت میانگین تا ۷۱,۱۴ درصد جهت قیمت‌ها را درست پیش‌بینی کند.

۴-نتیجه‌گیری و پیشنهادات آتی

گسترش روز افزون محیط وب و رشد تولید محتوای غیرساختار یافته شامل متن، صوت و فیلم و عدم توانایی انسان در بررسی همه آنها و تصمیم‌گیری، روز به روز اهمیت متن‌کاوی افزایش می‌یابد. هر چند نتایج حاصل از داده‌کاوی با درصدی خطا همراه است، اما در تعداد تصمیم‌گیری زیاد، مانند تشخیص نامه‌های الکترونیک اسپم استفاده از این روش‌ها ضروری به نظر می‌رسد. هر چند استفاده از الگوریتم‌های یادگیری ماشین در حجم دیتای زیاد نیاز به قدرت پردازش بالایی است، اما با گسترش تکنولوژی و تولید کامپیوتری‌ها با قدرت پردازش بالا استفاده از روش‌های داده‌کاوی عمومیت بیشتری پیدا می‌کند.

یکی از کاربردهای متن‌کاوی بررسی تأثیر اخبار بر قیمت سهام در بورس است. از آنجایی که اخبار زیادی هر روز توسط خبرگذاری‌ها منتشر می‌شود بررسی همه خبرها توسط انسان کار دشواری به نظر می‌رسد. خرید و فروش سهام همواره با سود و ضرر همراه است و ریسک وجود دارد، توانایی پیش‌بینی قیمت‌ها می‌تواند تأثیر قابل توجهی در میزان سود ما داشته باشد. در این پژوهش سعی شده است با ذخیره‌سازی اخبار و استفاده از الگوریتم ماشین بردار پشتیبان با کرنل‌های مختلف میزان دقت پیش‌بینی سهم ارزیابی شود.

با توجه به الگوریتم‌ها، کرنل‌ها و پارامترهای مختلفی که وجود دارد، نمی‌توان یک الگوریتم خاص با تنظیمات خاص را برای همه سهم‌ها انتخاب کرد، بلکه باید با توجه به نتایج حاصل‌شده از هر مدل، بهترین مدل را برای سهم موردنظر انتخاب کرد. با گذشت زمان و تغییر خبرها و شرایط یک سهم، ممکن است پارامترها و کرنل‌های نتایج مختلفی را ایجاد کنند. بهتر است هرروز مدل‌ها آموزش داده شوند و بهترین مدل برای پیش‌بینی جهت قیمت فردا انتخاب شود. از آنجایی که تعداد حالت‌های زیادی که اتفاق می‌افتد، تحلیل نتایج به‌صورت دستی بسیار زمان‌بر خواهد بود و برای انجام این وظایف بهتر است از کدنویسی و اتوماتیک کردن فرایند استفاده کرد.

در این پژوهش اخبار ۱۷ ذخیره شد و داده‌های معاملات شرکت‌های بورس در گروه محصولات شیمیایی ذخیره شد و ۲۵ شرکت در این پژوهش مورد بررسی قرار گرفت که در بازه آبان تا اسفند ۹۷ بیش از ۷۰ روز معاملاتی داشتند.

شکل ۲۱ - میانگین نتایج کرنل‌ها

با توجه به نتایج حاصل از ارزیابی مدل‌ها به‌صورت شکل ۲۱ خواهد بود. بهترین میانگین مربوط به کرنل معادله درجه ۲ با گاما auto است که با دقت تقریبی ۸۷ درصد می‌توان جهت قیمت سهم را در شرکت‌های گروه محصولات شیمیایی پیش‌بینی کند. همچنین مشاهده می‌شود در این پژوهش کرنل خطی هرچند ساده‌تر است و زمان پردازش کمتری را به خود اختصاص می‌دهد و سرعت بالاتری دارد، نتایج قابل قبولی در پیش‌بینی دارد و می‌تواند به صورت میانگین ۸۵٪ پیش‌بینی درستی انجام دهد و کرنل‌های غیرخطی هر چند سرعت پایین‌تری دارند ولی نتایج مناسبی را ارائه نمی‌کنند.

به دلیل محدودیت‌های زمانی و مکانی موجود این تحقیقات می‌تواند با انجام سایر مطالعات تکمیل‌تر شود که در زیر به آنها اشاره می‌شود.

۱- استفاده از بازه زمانی بیشتر برای ذخیره‌سازی اخبار و بررسی روزهای معاملاتی بیشتر در الگوریتم ماشین بردار پشتیبان و بررسی نتایج حاصل از روزهای معاملاتی بیشتر می‌تواند مورد بررسی قرار گیرد.

۲- در تحلیل‌های تکنیکال اخبار جدیدتر می‌توانند تاثیرگذاری بیشتری بر روی قیمت داشته باشند در صورتی که در این تحقیق فقط از خبرهای منتشر شده در روز قبل برای پیش‌بینی جهت قیمت استفاده شده است، در مطالعات آتی می‌توان از اخبار چند روز قبل و وزن دهی به روزهای گذشته نیز استفاده کرد.

۳- استفاده از سایر الگوریتم‌های کلاس‌بندی مانند درخت تصمیم، جنگل تصادفی و بیزین برای دسته و نتایج حاصل با مدل ماشین بردار پشتیبان مقایسه گردد.

۴- اجرای مدل پژوهش برای سایر شرکت‌ها و نتایج حاصل در گروه‌های محصولات مختلف بررسی و مقایسه شود.

۵- در این مطالعه فقط به بررسی جهت قیمت (افزایش یا کاهش قیمت) پرداخت شده است. در معاملات مقدار افزایش قیمت نیز بسیار مهم است و معامله‌گران تمایل دارند سهمی را خریداری کنند که قیمت آن درصد بیشتری افزایش پیدا کند. می‌توان در مطالعات آتی بر روی میزان افزایش قیمت نیز تحقیقاتی صورت بگیرد و نتایج الگوریتم‌ها با هم مقایسه شود.

منابع

1.J. D. Velásquez, V. Palade, and L. C. Jain, *Advanced techniques in web intelligence*: Springer, 2013.

2. Cisco. (2019). *Cisco Visual Networking Index: Forecast and Trends, 2017–2022 White Paper*. Available: https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white-paper-c11-741490.html

3 .internetlivestats. (2019). *Total number of Websites*. Available: https://www.internetlivestats.com/total-number-of-websites/

4 .Z. Markov and D. T. Larose, *Data mining the Web :uncovering patterns in Web content, structure, and usage*: John Wiley & Sons, 2007.

5.B. Liu, *Web data mining: exploring hyperlinks, contents, and usage data*: Springer Science & Business Media, 2007.

6 A. Khadjeh Nassirtoussi, S. Aghabozorgi, T. Ying Wah, and D. C. L. Ngo, "Text mining for market prediction: A systematic review," *Expert Systems with Applications*, vol. 41, pp. 7653-7670, 11/15/ 2014.

7. M.-A. Mittermayer and G. Knolmayer, *Text mining systems for market response to news: A survey*: Institut für Wirtschaftsinformatik der Universität Bern, 2006.

8 .C.-J. Huang, J.-J. Liao, D.-X. Yang, T.-Y. Chang, and Y.-C. Luo, "Realization of a news dissemination agent based on weighted association rules and text mining techniques," *Expert Systems with Applications*, vol. 37, pp. 6409-6413, 2010.

9. B. S. Kumar and V. Ravi, "A survey of the applications of text mining in financial domain," *Knowledge-Based Systems*, vol. 114, pp. 128-147, 12/15/ 2016.

10. M. Hagenau, M. Liebmann, and D. Neumann, "Automated news reading: Stock price prediction based on financial news using context-capturing features," *Decision Support Systems*, vol. 55, pp. 685-697, 6// 2013.

11 .M. Thelwall, "Data cleansing and validation for multiple site link structure analysis," in *Web mining: Applications and techniques*, ed: IGI Global, 2005, pp. 208-227.

12 .M. Sheng, Y. Qin, L. Yao, and B. Benatallah, *Managing the web of things: linking the real world to the web*: Morgan Kaufmann, 2017.

13.R. Kosala and H. Blockeel, "Web mining research :A survey," *ACM Sigkdd Explorations Newsletter*, vol. 2, pp. 1-15, 2000.

14 .M. G. Da Costa and Z. Gong, "Web structure mining: an introduction," in *Information Acquisition, 2005 IEEE International Conference on*, 2005, p. 6 pp.

15 .F. Johnson and S. K. Gupta, "Web content mining techniques: a survey," *International Journal of Computer Applications*, vol. 47, 2012.

16 .Kumar and Ravi, "A survey of the applications of text mining in financial domain," vol. 114, pp. 128-147, 2016.

17.A. Hotho, A. Nürnberger ,and G. Paaß, "A brief survey of text mining," in *Ldv Forum*, 2005, pp. 19-62.

18.Gupta and Lehal, "A survey of text mining techniques and applications," vol. 1, pp. 60-76, 2009.
19.Y. Zhang, M. Chen, and L. Liu, "A review on text mining," in *Software Engineering and Service Science (ICSESS), 2015 6th IEEE International Conference on*, 2015, pp. 681-685.
20 .H. Hashimi, A. Hafez, and H. Mathkour, "Selection criteria for text mining approaches," *Computers in Human Behavior*, vol. 51, pp. 729-733, 2015.
21 .K .Javed, S. Maruf, and H. A. Babri, "A two-stage Markov blanket based feature selection algorithm for text classification," *Neurocomputing*, vol. 157, pp. 91-104, 2015.
22 .G. Hackeling, *Mastering Machine Learning with scikit-learn*: Packt Publishing Ltd, 2.۰۱۷
23 .Khan and A. Ahmad, "Cluster center initialization algorithm for K-means clustering," *Pattern recognition letters*, vol. 25, pp. 1293-1302, 2004.
24.J. Hou, H. Gao, and X. Li, "DSets-DBSCAN: a parameter-free clustering algorithm," *IEEE Transactions on Image Processing*, vol. 25, pp. 3182-3193, 2016.
25.Zhang and Z. Xu, "Hesitant fuzzy agglomerative hierarchical clustering algorithms," *International Journal of Systems Science*, vol. 46, pp. 562-576, 2015.
26 .D. M. Farid, L. Zhang, C. M. Rahman, M . A. Hossain, and R. Strachan, "Hybrid decision tree and naïve Bayes classifiers for multi-class classification tasks," *Expert Systems with Applications*, vol. 41, pp. 1937-1946, 2014.
27.V. Svetnik, A. Liaw, C. Tong, J. C. Culberson, R. P. Sheridan, and B .P. Feuston, "Random forest: a classification and regression tool for compound classification and QSAR modeling," *Journal of chemical information computer sciences*, vol. 43, pp. 1947-1958, 2003.
28 .Chen, K. Li, Z. Tang, K. Bilal, S. Yu, C. Weng, *et al*" ,.A parallel random forest algorithm for big data in a spark cloud computing environment," *IEEE Transactions on Parallel Distributed Systems*, pp. 1-1, 2017.
29.G. Dreyfus, *Neural networks: methodology and applications*: Springer Science & Business Media, 2.۰۰۵
30.C. J. Burges, "A tutorial on support vector machines for pattern recognition," *Data mining knowledge discovery*, vol. 2, pp. 121-167, 1998.
31 .M. Sokolova, N. Japkowicz, and S. Szpakowicz, "Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation," in *Australasian joint conference on artificial intelligence*, 2006, pp. 1015-1021.
32.S. S. Groth and J. Muntermann, "An intraday market risk management approach based on textual analysis," *Decision Support Systems* ,vol. 50, pp. 680-691, 2011.
33 .R. P. Schumaker, Y. Zhang, C.-N. Huang, and H. Chen, "Evaluating sentiment in financial news articles," *Decision Support Systems*, vol. 53, pp. 458-464, 6// 2012.
34 .L. Dey, A. Mahajan, and S. M. Haque, "Document clustering for event identification and trend analysis in market news," in *Advances in Pattern Recognition, 2009. ICAPR'09. Seventh International Conference on*, 2009, pp. 103-106.
35.A. Mahajan, L. Dey, and S. M. Haque, "Mining Financial News for Major Events and Their Impacts on the Market," in *2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, 2008, pp. 423-426.
36 .X. Zhong and D. Enke, "A comprehensive cluster and classification mining procedure for daily stock market return forecasting," *Neurocomputing*, vol. 267, pp. 152-168, 2017/12/06/ 2017.
37 .A. E. Khedr, S. Salama, and N. Yaseen, "Predicting Stock Market Behavior using Data Mining Technique and News Sentiment Analysis," *International Journal of Intelligent Systems and Applications (IJISA)*, vol. 9, pp. 22-30, 2017.
38 .H. Levy and M. Sarnat, "International Diversification of Investment Portfolios," *The American Economic Review*, vol. 60, pp. 668-675, 1970.
39 .J. M.-T. Wu, Z. Li, C.-W. Lin, and M. Pirouz, "A New Convolution Neural Network

Model for Stock Price Prediction," ed, 2020, pp. 581-585.
40 .O. M. Ebadati E and M. Mortazavi T, "An efficient hybrid machine learning method for time series stock market forecasting," *Neural Network World*, vol. 28, pp. 41.۲۰۱۸ ,۵۵-
41.A. Mahajan, L. Dey, and S. M. Haque, "Mining financial news for major events and their impacts on the market," in *Web Intelligence and Intelligent Agent Technology, 2008. WI-IAT'08. IEEE/WIC/ACM International Conference on*, 2008, pp. 423-42.۶
42.X. Zhong and D. J. N. Enke, "A comprehensive cluster and classification mining procedure for daily stock market return forecasting," vol. 267, pp. 152-168, 2017.
43.A. E. Khedr and N. Yaseen, "Predicting stock market behavior using data mining technique and news sentiment analysis," *International Journal of Intelligent Systems Applications*, vol. 9, p. 22, 2017.
44 .M. Hagenau, M. Liebmann, and D. J. D. S. S. Neumann, "Automated news reading: Stock price prediction based on financial news using context-capturing features," vol. 55, pp. 685-697, 2013.
45 .scikit-learn.org. *Choosing the right estimator*. Available: https://scikit-learn.org/stable/tutorial/machine_learning_map/index.html
۴۶. ع. نوریان. (۲۰۱۸). هضم برای پردازش زبان فارسی در پایتون. Available: http://www.sobhe.ir/hazm/
47.W.-H. Chen, S.-H. Hsu, and H.-P. Shen, "Application of SVM and ANN for intrusion detection," *Computers Operations Research*, vol. 32, pp. 2617-2634, 2005.