

به کارگیری 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 پذیرش: ۱۳۹۹/۰۴/۱۸

تاریخ دریافت: ۱۳۹۸/۱۰/۰۹

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 خبرگزاری‌ها و استفاده از روش‌های متن‌کاوی و الگوریتم ماشین بردار پشتیبان به منظور پیش‌بینی جهت قیمت روز آینده سهام است. بدین منظور خبرها منتشر شده در ۱۷ خبرگزاری با استفاده از یک خزگشر موضوعی به زبان پی‌اچ‌پی 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. در این 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، خزگشر وب، پیش‌بینی بورس اوراق بهادار، ماشین بردار پشتیبان

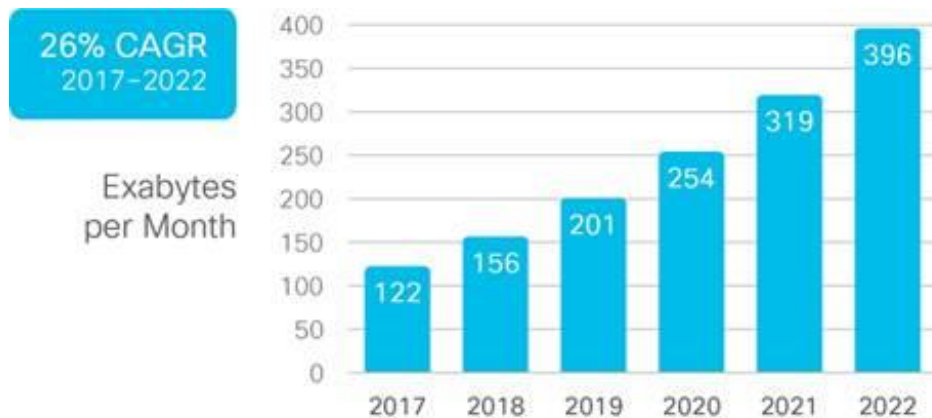
۱-مقدمه

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].

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،

نویسنده عهده‌دار مکاتبات: امید مهدی عبادتی ebadati@khu.ac.ir

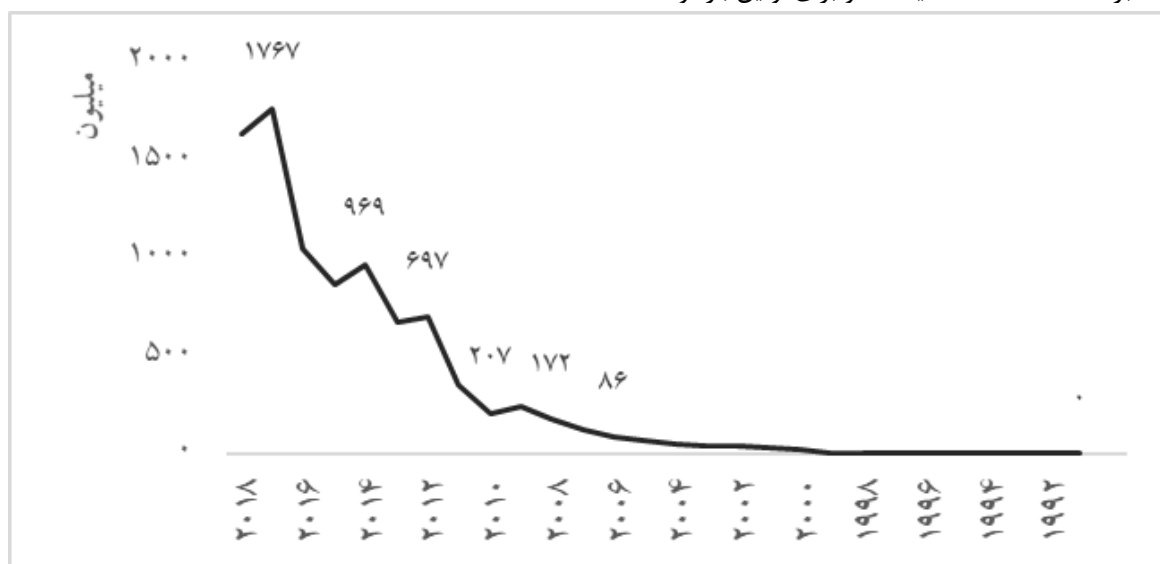


Source: Cisco VNI Global IP Traffic Forecast, 2017-2022

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 تأیید شد و برای اولین بار توسط

شکل ۲- تعداد وب‌سایت‌ها از سال ۱۹۹۲ تا ۲۰۱۸



شکل ۲- تعداد وب‌سایت‌ها از سال ۱۹۹۲ تا ۲۰۱۸

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 کاوش محتوای وب را دو چندان می‌کند.

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 قرار گرفته است: فن‌های آماده‌سازی و پالایش داده‌های صفحات وب؛ استخراج و ذخیره‌سازی الگوی وب؛ کاوش ساختار وب؛ کاوش

^۵Web usage mining

^۶Web content mining

^۱Web Data Pre-processing and Cleaning

^۲Web Pattern Extraction and Storage

^۳Web structure mining

خزش اولین قدم در وب‌کاوی 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 می‌کنند، 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 انجام شده است، که دلیل عمده آن در دسترس بودن داده‌های کمی تاریخی در بازار و تمایل عمومی معامله‌گران برای استفاده از روش‌های کمی 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 منتشر شده 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

^۷Universal crawler

^۸Topic crawlers

^۹Technical

^{۱۰}Fundamental

^{۱۱}Chenn-Jung Huang

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 سهام را در روز بعد پیش‌بینی کرد. بدین منظور اخبار منتشر شده در خبرگزاری‌ها به عنوان متغیر 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 وب‌کاوی می‌توان از داده‌هایی که توسط خزشگرها و موتورهای جستجو ذخیره شده 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

^{۱۲}News

^{۱۳}Institutional Traders

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]

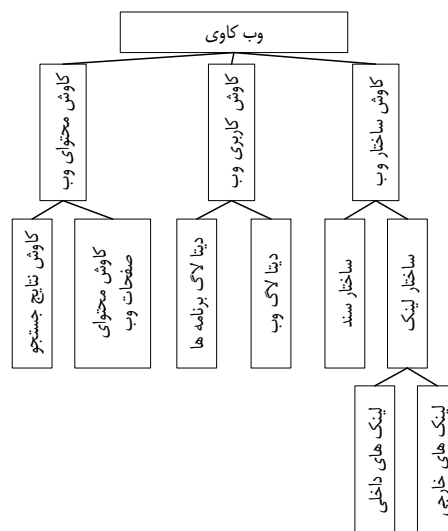
۲-۲- روش‌های وب کاوی

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]

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 ارائه شده است تا بتوان محتوای مورد نیاز را از وب ذخیره کرد.

۲-۱- وب کاوی

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]



شکل ۳- روش‌های وب کاوی [۱۵]

۲-۲-۳- 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 برجستگی از قبل برای آموزش الگوریتم وجود ندارد [۱۶]. [۱۹]

۲-۳- انواع روش‌های متن کاوی

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 بیان شده است [۱۶]:

۱-۲-۳- کلاس‌بندی یا دسته‌بندی^{۱۴}

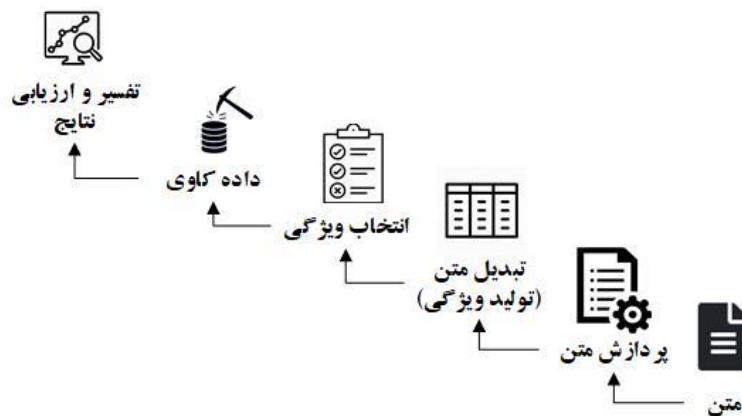
کلاس‌بندی یا دسته‌بندی یکی از روش‌هایی است که برای گروه‌بندی متن‌ها در متن‌کاوی مورد استفاده قرار می‌گیرد. هدف از کلاس‌بندی متون، نسبت دادن کلاس‌های از پیش تعریف‌شده به اسناد متنی موجود، مانند گروه اخبار، گروه کالاها، جناح سیاسی است. [۱۷، ۱۸].

^{۱۵}Clustering

^{۱۴}Classification

۲-۴- مراحل متن‌کاوی

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



شکل ۴- مراحل متن کاوی

یافتن روابط میان مفاهیم، ساخت اتوماتیک آنتولوژی و غیره به کار می‌رود [۲۱].

۴-۲-۵- تفسیر و ارزیابی

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

۴-۲-۵- 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].

۴-۲-۵- یادگیری بدون نظارت: خوشه‌بندی

الگوریتم خوشه‌بندی کا-مینز^۱ و کا-مد^۲

الگوریتم کا-مینز که در سال ۱۹۶۷ توسط مک کوپین^۳ مطرح شد، یکی از محبوب‌ترین الگوریتم‌های خوشه‌بندی است که در زمینه‌های مختلف مورد استفاده قرار می‌گیرد. هدف الگوریتم کا-مینز، بهینه‌سازی تابع هدف^۴ می‌باشد که پاسخ‌های حاصل از خوشه‌بندی به کمینه‌سازی یا

۴-۲-۱- پیش پردازش

برای کشف دانش از حجم قابل توجهی از اسناد، لازم است که بر روی اسناد پیش پردازش 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 غیرساخت یافته به فرمت ساختاریافته و قابل تشخیص برای ماشین تبدیل شود. [۱۸]

۴-۲-۲- تولید و استخراج ویژگی

اگرچه برنامه‌های کاربردی زیادی در زمینه بازیابی اطلاعات مانند پالایش و جست‌وجوی اطلاعات مرتبط می‌توانند از تحقیقات در زمینه 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].

۴-۲-۳- 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 حالی که نمایش مبتنی بر مفهوم برای

^۱K-means

^۲K-mode

^۳Mac Queen

^۴Object Function

مدل ترکیبی گاوسی^{۳۶}

منظور از نمودار گاوسی 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 گاوسی می‌گویند که بیشتر داده‌های جهان از این نوع توزیع پیروی می‌کنند. شکل توزیع گاوسی همیشه متوازن نیست و ممکن است به سمت چپ یا راست چولگی^{۳۷} داشته باشد. این الگوریتم از روش بیشینه‌سازی انتظار^{۳۸} استفاده می‌کند.

۲-۲-۵- الگوریتم یادگیری بانظارت: کلاس‌بندی

الگوریتم درخت تصمیم^{۳۹}

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 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 ۰.۵C وجود دارد که تفاوت آن‌ها در معیار اندازه‌گیری عدم خلوص، شیوه شاخه‌بندی و هرس کردن گره‌های درخت می‌باشد [۲۶].

الگوریتم جنگل تصادفی^{۴۱}

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 شکل گرفته است. در این الگوریتم ابتدا با استفاده از روش نمونه‌گیری بوت استرپ^{۴۲} تعداد K زیرمجموعه آموزش^{۴۳} را از بین مجموعه داده‌های اصلی استخراج می‌کند و سپس با

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

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^{۴۴} و حداقل تعداد نقاط موجود^{۴۵} در همسایگی، استفاده می‌شود و مجموعه نقاط را به نقاط مرکزی^{۴۶} تقاطع مرزی^{۴۷} و داده‌های پرت^{۴۸} تقسیم می‌کند. [۲۴].

الگوریتم سلسه‌مراتبی^{۴۹}

این الگوریتم به دو دسته از بالا به پایین^{۴۸} و از پایین به بالا^{۴۹} 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^{۵۰} استفاده می‌شود. انواع روش‌های پیوند عبارت‌اند از: نزدیک‌ترین همسایه^{۵۱}، پیوند تکی^{۵۲}، دورترین همسایه^{۵۳}، پیوند کامل^{۵۴} و یا پیوند میانگین^{۵۵} [۲۵].

^{۳۶}Epsilon

^{۳۷}MinPoints

^{۳۸}Core points

^{۳۹}Border point

^{۴۰}Outlier

^{۴۱}Hierarchical

^{۴۲}Partitioning

^{۴۳}Agglomerative

^{۴۴}Linkage

^{۴۵}Nearest Neighbor

^{۴۶}Single Linkage

^{۴۷}Furthest Neighbor

^{۴۸}Complete Linkage

^{۳۶}Average Linkage

^{۳۷}Gaussian Mixture Model (GMM)

^{۳۸}Skew

^{۳۹}Expectation Maximization (EM)

^{۴۰}Decision Tree

^{۴۱}Breman

^{۴۲}Random Forest

^{۴۳}Bootstrap

^{۴۴}Train

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 داده شده است.

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

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

Tn (true negative): کلاس منفی که به درستی مثبت تشخیص داده شده 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 طبقه‌بندی شده 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:

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 به راحتی به صورت 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}

مسئله طبقه بندی یکی از مسائل اصلی مطرح شده در یادگیری ماشین است. بسیاری از مسائل را می‌توان بصورت یک مسئله کلاسه بندی مطرح کرده و حل نمود. روش‌های طبقه بندی خطی، سعی دارند که با ساختن یک ابر سطح (که عبارت است از یک معادله خط)، داده‌ها را از هم تفکیک کنند. چندین تفکیک کننده خطی می‌تواند داده دو کلاس را از هم جدا کند. یکی از روش‌هایی که بصورت گسترده برای اینگونه مسائل استفاده می‌شود، ماشین بردار پشتیبان است. ماشین بردار پشتیبان در سال ۱۹۹۳ توسط ولادیمیر واپنیک پیشنهاد شد. SVM بهترین ابر سطحی را پیدا می‌کند که با حداکثر فاصله، داده‌های مربوط به دو طبقه را از هم تفکیک کند. [۳۰]

۲-۶-۲-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.

^{۴۴}Neural Network

^{۴۵}Feedforward

^{۴۶}Feedback

^{۴۷}Support vector machine

^{۴۸}Precision

^{۴۹}Recall

^{۵۰}F1 Score

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

با استفاده از این شاخص های می توان مدل های یادگیری دوتایی با ناظر را ارزیابی کرد. [۳۱]

۲-۷- پیش بینی بازار بر اساس اخبار منتشر شده

مدیریت ریسک مالی یکی از کارهای بسیار پر چالش در واحدهای مالی است. در دو دهه گذشته روش ها و مدل های کمی بسیاری جهت مشخص کردن تأثیر بازارهای بی ثبات مالی بر تجارت، توسعه و گسترش یافته اند. اکثر این مطالعات بر روی داده های ساختاریافته مانند سری قیمت های تاریخی انجام شده است و توجه اندکی به داده های غیر ساختار یافته (متنی) شده است، در حالی که بیشتر حجم منابع اطلاعاتی موجود از این نوع محتوا هستند. تحقیقات تجربی که در گذشته صورت گرفته است نشان می دهد که اخبار خاص، مانند افشای شرکت های بزرگ می توانند 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 پیش بینی 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 می شود. در نظریه گام تصادفی اعتقاد بر این است که بازارها کارآمد هستند و اصلاح قیمت ها بلافاصله رخ می دهد و پیش بینی قیمت از داده بازار غیرممکن است. این نظریه ها بیان می کنند که قیمت ها به اطلاعات پیوند خورده اند و از اطلاعات امروز نمی توان برای پیش بینی قیمت ها در آینده استفاده کرد. [۳۳]

همچنین در پیش بینی قیمت سهام دو فلسفه کاملاً مخالف وجود دارد. فن های تجزیه و تحلیل بنیادی و تکنیکال. در حالی که تحلیلگران بنیادی به دنبال استفاده از داده های نسبی، نسبت ها و درآمد نسبی سهام هستند، تحلیلگران تکنیکال از نمودارها، فن های مدل سازی بر اساس حجم معاملات تاریخی و قیمت ها برای تحلیل خود استفاده می کنند. [۳۳]

در حالی که رفتار بازار تحت تأثیر اتفاقات محلی و جهانی است، جزئیات این اتفاقات در داده های ساختاریافته مشاهده نمی شود. انتظار می رود که داده کاوی نقش مهمی در طراحی راهبردهای پیش بینی رفتار بازار داشته باشد، زیرا می تواند با

موفقیت برای تجزیه و تحلیل مقالات اخبار مالی و گزارش های همراه داده های سری زمانی بازار مورد استفاده قرار گیرد. متن کاوی گزارش های خبری مالی می تواند برای استخراج اطلاعات مهم در مورد وقایع مختلف سیاسی و اقتصادی که به طور کلی بازار مالی یک منطقه را تعیین می کند 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 گزارش های خبری و کشف اطلاعات کلیدی از طرف تحلیلگران و سرمایه گذاران استفاده شود. [۳۴]

اخبار نقش مهمی در فرایند ارزیابی قیمت فعلی سهام، که توسط تحلیلگران، سرمایه گذاران و سرمایه گذاران نهادی^{۵۴} صورت می گیرد، 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 می دهند [۱۰]

در سال های اخیر روزنامه های سنتی به دلیل افزایش فراگیر شدن شبکه جهانی وب، به گسترش سرویس های خبری بر خط را در محیط وب پرداخته اند. برای سرمایه گذاران، خبرهای بلادرنگ مالی^{۵۵} در تصمیم گیری ها سرمایه گذاری 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 سرمایه گذاران غیرممکن است که همه آنها را یکی یکی بخوانند. اگر چه تحقیقات رو به افزایشی در متن کاوی اسناد در حال انجام است، اما دقت کم و پایین بودن نرخ بازگشت سبب می شود سرمایه گذاران زمان زیادی را برای به دست آوردن اطلاعات معنی دار 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 تیتراهای خبری منتشر شده توسط روزنامه های الکترونیکی بر روی شاخص

^{۵۴}News

^{۵۵}Institutional traders

^{۵۶}Real-time financial news

^{۵۷}Quantitative

^{۵۸}Nextual

تاریخی بسیار ساده است، اکثر مطالعات داخلی بر روی این حوزه انجام شده است.

در مطالعه‌ای که توسط تای وو 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 سری‌های زمانی استفاده نموده‌اند. در این روش مقادیر خروجی الگوریتم ژنتیک وارد الگوریتم توسعه یافته تکنیک شبکه عصبی مصنوعی می‌شوند تا خطاها را در نقطه دقیق برطرف کنند. تجزیه و تحلیل نشان می‌دهد که روش ترکیبی الگوریتم ژنتیک و تکنیک شبکه عصبی مصنوعی می‌توانند 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 سنپاتی 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 سهام 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

قیمت مالی بورس تایوان را مورد بررسی قرار دادند. هاگونا 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^۴ است که توسط شوماکر^۵ و همکاران طراحی شده 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^۶ 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 سهام‌ها از یک گروه انتخاب نشود و با خرید متنوع از سهام در گروه‌های مختلف ریسک خود را کاهش دهند. [۳۸] از این رو بهتر است برای تحلیل، سهام‌هایی انتخاب شود که در یک گروه قرار دارند. برای مثال به نظر می‌رسد عوامل موثر بر قیمت سهام‌های 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 آنجایی که اجرای الگوریتم‌ها بر روی داده‌های ساختار یافته

^۴Hagenau

^۵Arizona Financial Text

^۶Robert P. Schumaker

^۷Missouri University of Science and Technology

^۸Jimmy Ming-Tai Wu

^۹Manas Ranjan Senapati

می‌پردازد. در این مطالعه استفاده از روش ماشین بردار پشتیبان به نتایج بهتری منجر شده است. [۳۲]

در مطالعه‌ای دیگر توسط هانگنوا^۶ 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 مطالعه‌ای دیگر با عنوان «کاوش اخبار مالی برای وقایع مهم و و تأثیر آنها بر بازار» توسط ماهان جان^۷ 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^۸ و همکاران در سال ۲۰۰۸ با استفاده از روش تخصیص دریکله پنهان برای استخراج اتفاقات مهم و استفاده از کرنل کامین 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 پشتیبان 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

ژونگ^۹ و آنکه^{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 کامین نزدیکترین همسایه جهت قیمت سهام (مثبت یا منفی) را مشخص کردن. [۴۳]

در تحقیقی دیگر توسط شوماکر^۹ 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^{۱۰} 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 تیتیر خبرها بر روی شاخص بورس تایوان استفاده شد. [۱۱]

در تحقیقی دیگر توسط گروس^{۱۱} و مونترمن^{۱۲} 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، کامین نزدیکترین همسایه، شبکه عصبی و ماشین بردار به دسته‌بندی داده‌ها

^۶Xiao Zhong

^۷David Enke

^۸Robert P. Schumaker

^۹Chenn-Jung Huang

^{۱۰}Sven S. Groth

^{۱۱}Jan Muntermann

^{۱۲}Michael Hagenau

^{۱۳}Anuj Mahajan

^{۱۴}Lipika Dey

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.

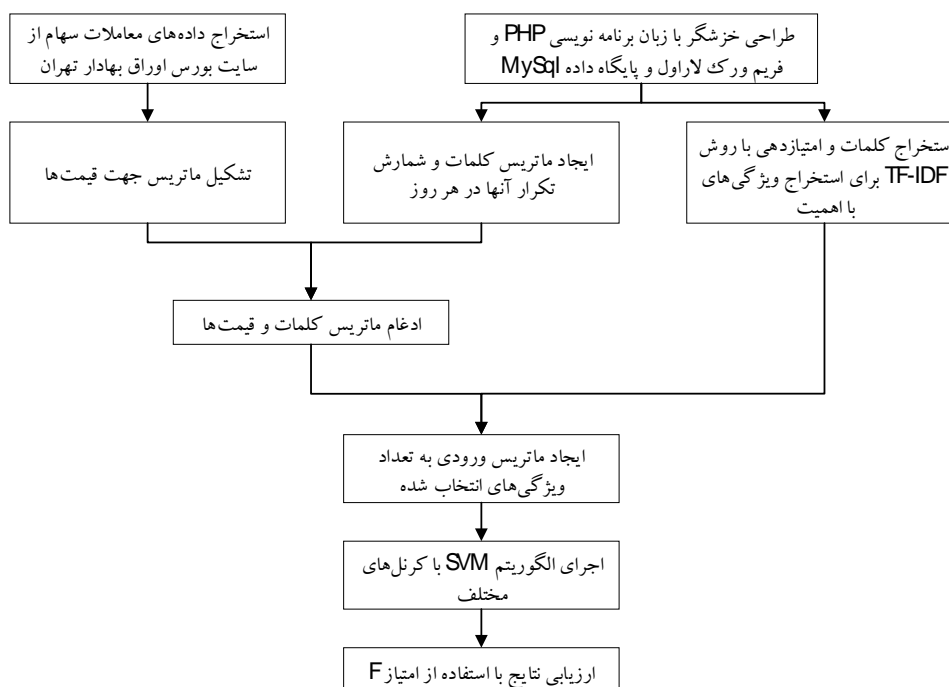
در قدم پنجم با استفاده از الگوریتم SVM مدل آموزش داده می‌شود و نتایج حاصل از پیش‌بینی مدل مورد ارزیابی قرار می‌گیرد. به‌غیر از مرحله خزش تمام مراحل با استفاده از زبان برنامه‌نویسی Python و محیط توسعه PyCharm انجام می‌شود.

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.

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 سهام از سایت بورس اوراق بهادار استخراج می‌شود. برای ذخیره‌سازی این داده‌ها ابتدا پیش‌پردازش بر روی دیتا انجام می‌شود و داده‌های سهام 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



شکل ۵- روش جمع‌آوری و تجزیه و تحلیل داده‌ها

بین سیستم با استفاده از زبان برنامه‌نویسی پی‌اچ‌پی و چهارچوب لاراول^۳ نوشته شد. برای دانلود صفحات خبر از پکیج گازل و برای ذخیره‌سازی عنوان، متن و دسته خبر بر اساس آدرس مکان سی‌اس‌اس‌آز^۴ پکیج دام کراولر استفاده می‌شود. در فاز اول ۱۷ سایت خبرگزاری پربازدید ایران برای

۳-۱- سیستم خزش و ذخیره‌سازی اخبار روزانه

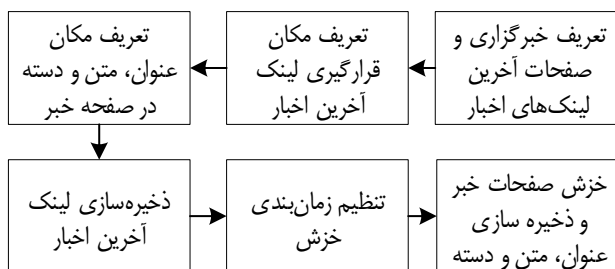
۳-۱-۱- ذخیره‌سازی اخبار خبرگزاری‌ها

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 داده شده است.

^۳Framework

^۴Laravel

^۵Cascading Style Sheets



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

شکل ۶- نحوه کار پی‌اچ‌پی کراولر

جدول ۲- سایت‌های خبرگزاری خزش شده

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

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.

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

id	task_url	rand_number	complete	error	created_at
1	https://www.isna.ir/news/98060402201/ ...ترامپ-از-قران...	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-در-جه...	26	2	NULL	2019-08-26 22:24:15
4	https://www.isna.ir/news/98060402198/ ...سند-تهایی-گرو...	42	2	NULL	2019-08-26 22:24:15

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

```

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

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
    
```

```

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
    
```

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

شبه‌کد فرایند ذخیره‌سازی اخبار به صورت زیر است:

```

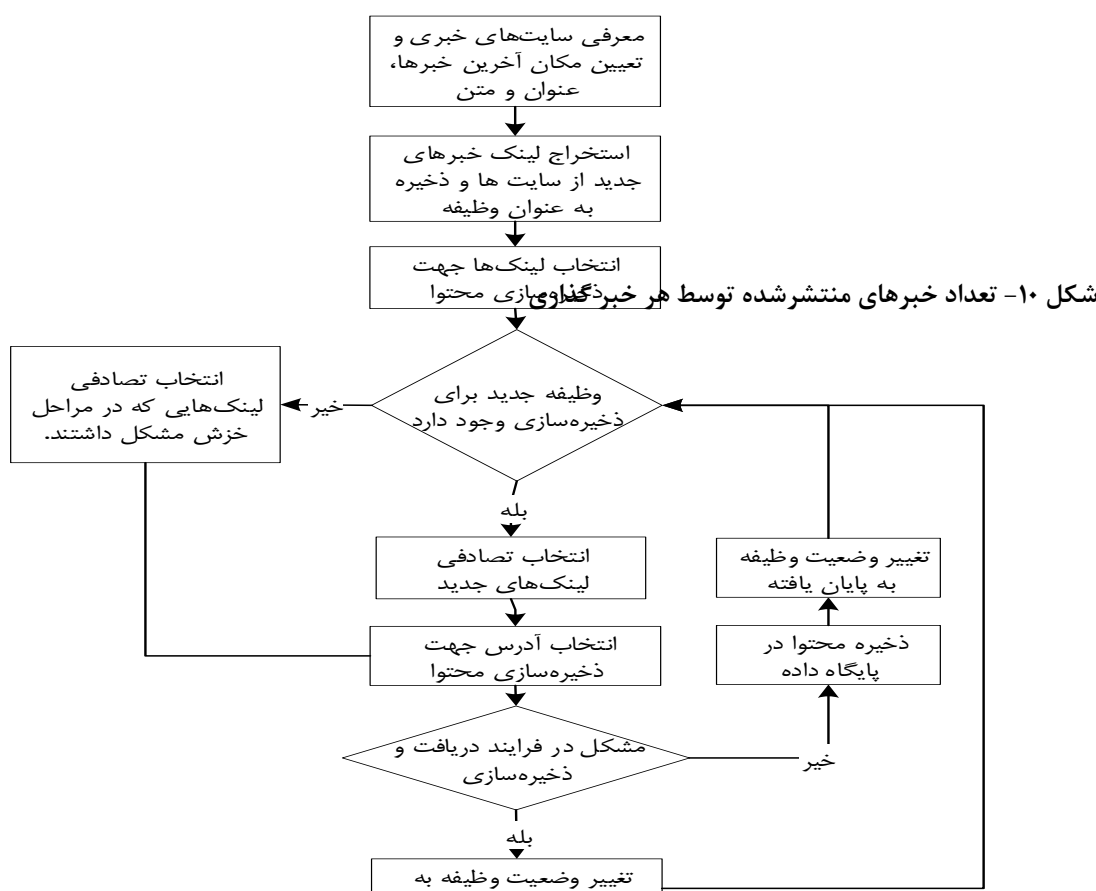
Select 12 tasks randomly that not crawled
Change tasks status to is crawling
If number of tasks is 0 then
  Select 3 tasks randomly that had error in crawl
For each task in task do
  Download task page
  Remove JavaScript code
  Remove tab and newline character from text
  Create new news
  
```

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 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.

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

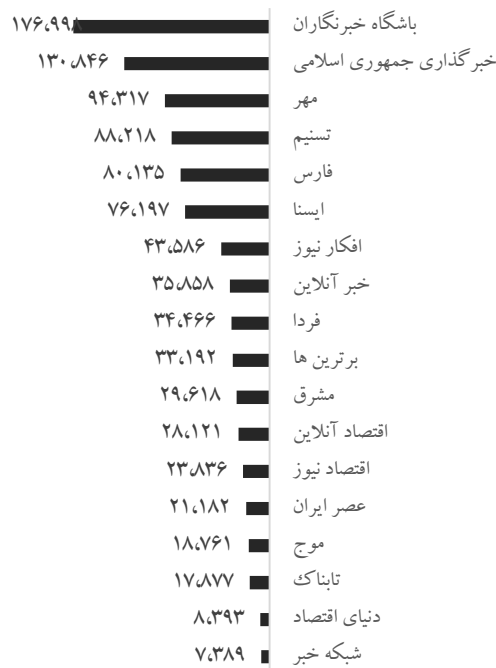


پس از مراجعه به لینک آرشیو معاملات نقد در سایت بورس اوراق بهادار تهران به نشانی tse.ir/archive.html پس از انتخاب تاریخ موردنظر می‌توانید داده‌های سهام معامله‌شده در آن تاریخ را ذخیره کنید. لینک دانلود به صورت زیر است.

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 ذخیره شده است و در کد از سطرهای این ستون برای ساخت لینک استفاده می‌شود.

از آنجایی که فایل‌های دانلود جدول‌های HTML هستند، اگر با فرمت html به جای xls ذخیره شوند، می‌توان توسط کتابخانه Pandas خوانده و یکپارچه شوند و سپس در دیتابیس ذخیره شوند.



شکل ۱۱- تعداد خبرهای منتشرشده توسط هر خبرگزاری

۳-۲- ذخیره‌سازی داده‌های سهام

id	stock_namad	stock_name	stock_volume	stock_v2	stock_transa
1	اخبار 1	مخابرات ایران	142594	3179589	40
2	وکارزار 1	سرمایه‌گذاری خوارزمی	508973	4336016	47
3	شید 1	پالایش نفت اصفهان	2105211	9110653	231
4	حساب 1	ساینا	5666758	4785844	439
5	ویروعلی 1	سرمایه‌گذاری ویروعلی	187658	2169823	38
6	مندی 1	مندی صنایع من ایران	3746245	9874156	177
7	داده 1	داده صنایع من ایران	6686	2496984	10

شکل ۱۲- نمونه دیتای جدول transaction (تراکنش‌های سهام)

<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، سهم‌های مربوط به گروه محصولات شیمیایی را ذخیره کرد.

در مرحله بعد داده مربوط به شرکت‌های بورس اوراق بهادار را ذخیره می‌شود تا بتوانیم بر اساس دسته تعیین شده برای هر سهم، سهم 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.

stock_code	stock_group	stock_industry
IRB5IKCO8751	N2	خودرو و ساخت قطعات
IRO1APPE0001	N2	رایانه و فعالیت‌های وابسته به آن
IRO1ASIA0001	N1	ه و صندوق بازنشستگی به جز تأمین اجتماعی
IRO1CONT0001	N1	ابزار پزشکی، ایتیکی و اندازنگیری
IRR1CONT0101	N1	ابزار پزشکی، ایتیکی و اندازنگیری
IRO1MKPT0001	N1	ابزار پزشکی، ایتیکی و اندازنگیری

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

۳-۳-۳- پیش پردازش متن

پیش پردازش متن از مراحل بسیار مهم در متن کاوی است. از این رو در این پژوهش چندین مرحله جهت پیش پردازش و حذف داده‌های اضافه و یکسان سازی استفاده شده است. به دلیل حجم بالای خبرها، قبل از پیش پردازش خبرهایی انتخاب شدند که فقط در دو دسته سیاسی و اقتصادی دسته بندی شده بودند. دیتای جدید در جدول news_back ذخیره شد. در این مرحله ۸۰۶۱۸ خبر انتخاب شد.

۳-۳-۱- نرمال سازی متن

در این مرحله با استفاده از کتابخانه هضم عنوان و متن خبرها نرمال سازی 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 ۱۰۰ هزار تایی در یک حلقه قرار می گرفتند و کد نرمال سازی به صورت موازی برای خبرها اجرا شد.

در حالت ساده عملیات نرمال سازی متن با مراحل زیر انجام پذیر است:

اصلاح انواع حرف «ک» به معادل فارسی آنان.

اصلاح انواع حرف «گی» به معادل فارسی آنان.

بررسی همزه و انواع مختلف املاهای موجود و اصلاح هر کدام (به عنوان مثال تبدیل و به و، ی به ی، ا به ا، ا 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 نیم فاصله

اصلاح اعراب و حذف فتحه، کسره و ضمه و همچنین تنوین‌ها حذف نیم فاصله‌های تکراری

حذف نویسه‌ی «ب» که برای کشش نویسه‌های چسبان مورد استفاده قرار می‌گیرد. مانند تبدیل «ببر» و «بر» به «بر»

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

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

اصلاح فاصله گذاری «می»، «نمی»، «درمی»، «برمی»، «بی»

در ابتدای واژه‌ها

تبدیل «ه» به «هی»

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

اصلاح فاصله گذاری پسوندها

حذف فاصله‌ها و نیم فاصله‌های اضافه بکار رفته در متن

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

۳-۳-۲- حذف کلمات توقف

در مرحله دوم پیش پردازش 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_normalize_stop به دیتا اضافه شد.

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

برای تشخیص کاراکترها و کلمات زائد 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 - \ , « » < > . : / ()

_ « » € « » ± @ ! / ? & x + ۴۳۵۲۷ ; =

بعد از این مرحله اگر کلمه مورد نظر به صورت کامل حذف نشد و در جدول کلمات موجود نباشد در جدول ذخیره می‌شود. در مجموع ۲۷۲۸۸۰ کلمه یکتا ذخیره شد.

جزء اخبار روز بعد تلقی می‌شوند و بر روی قیمت‌های روز آینده تأثیرگذار هستند. بدین جهت کلمات آن‌ها در کلمات روز آینده شمارش می‌شوند. تعداد کلمات به تفکیک روزهای انتشار خبر در این مرحله ۳۴۳۳۰۸۳۸ است.

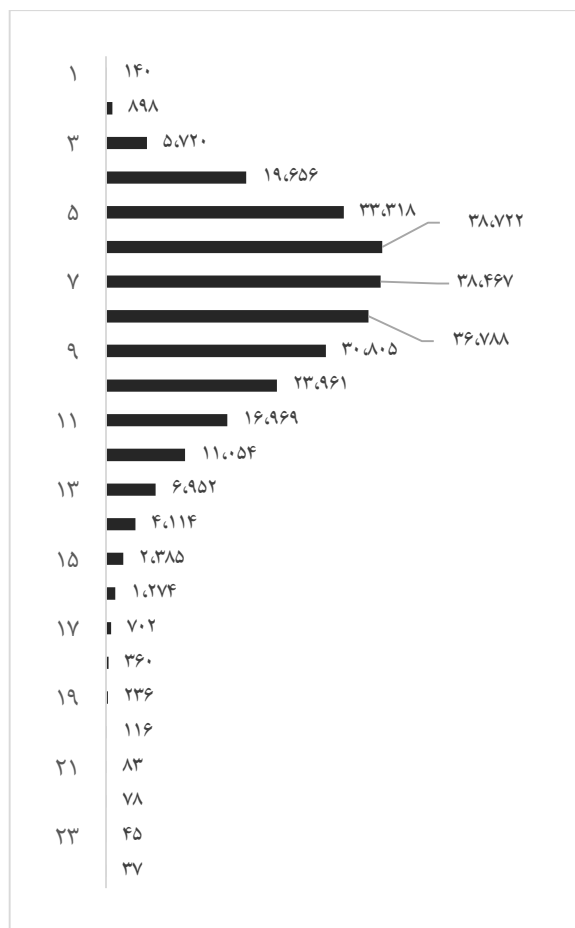
۳-۳-۵- انتخاب ویژگی‌ها با استفاده از روش TF-IDF

روش $tf \times idf$ (تکرار کلمه \times تکرار سند معکوس)، یک روش نمایندگی است که معمولاً در روش‌های استخراج ویژگی‌ها از متن استفاده می‌شود. به صورت پیش فرض، تکرار کلمات $(tf)^4$ در سند نشان دهنده میزان اهمیت اصطلاح در این سند است و تکرار سند در این اصطلاح 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 منحصر به فرد بودن این اصطلاح در اسناد است. بنابراین، به جای استفاده از df ، idf (معکوس df) را به عنوان برنامه وزن دهی انتخاب می‌شود. بنابراین وزن بالا در روش $tf \times idf$ 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 منتشر شده‌اند به عنوان یک سند در نظر گرفته می‌شوند. پس باید برای محاسبه TF تعداد تکرار کلمات در هر روز را به دست آورد. سپس با تقسیم تکرار هر کلمه در هر روز TF محاسبه می‌شود. برای محاسبه IDF از فرمول $IDF = \log\left(\frac{c_d}{i_d}\right)$ استفاده شده است که در آن c_d نشان دهنده کل سندها است. که در پژوهش ما برابر تعداد روزهای تولید خبر و برابر ۱۳۶ است. i_d نیز نشان دهنده تعداد اسناد است که ما هر روز را برابر یک سند گرفتیم. سپس با ضرب TF در IDF شاخص TF-IDF برای هر کلمه محاسبه می‌شود. سپس ماتریس کلمات تشکیل شده و بر اساس تعداد ویژگی‌های مورد نظر 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

از آنجایی که نوع مدل انتخابی یادگیری با ناظر بر اساس جهت قیمت سهام است، باید دیتای سهم مورد نظر را بر اساس اینکه



شکل ۱۴- تعداد کلمات بر اساس تعداد کاراکترها

از این تعداد کلمات ۱۰۹۷ کلمه به عنوان کلمات زائد شناسایی و به عنوان 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 استاندارد در اختیار باشد که بتوان کلمات استخراج شده 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 املائی را اصلاح کرد می‌تواند به کاهش ابعاد کمک کند.

۳-۳-۴- ایجاد جدول کلمات و شمارش کلمات در هر روز

پس از مرحله پیش پردازش 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 مورد نظر شمارش شد و شناسه کلمه به همراه تعداد تکرار در جدول `extract_word` ذخیره شد. از آنجایی که 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

^{۴۳}Term frequency

^{۴۴}Document frequency

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

۳-۵-۱ 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 در نهایت به بررسی ۷ مدل خواهیم پرداخت. مدل‌های بررسی شده 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 انتخاب شده‌اند که بیش از ۷۰ روز در بازه زمانی مورد نظر معامله شده‌اند.

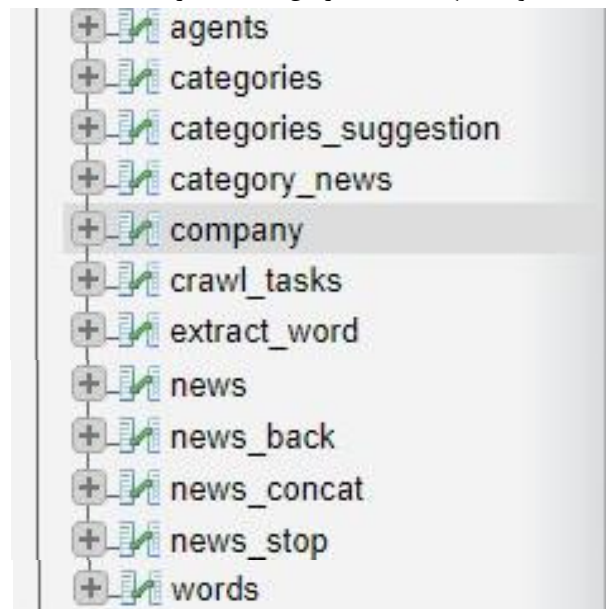
کاهش یا افزایشی بوده است برچسب‌گذاری کنیم. سپس ماتریس کلمات و جهت قیمت را ادغام می‌کنیم تا دیتا جهت ورود به الگوریتم آماده شود. قیمت منفی و صفر نسبت به روز قبل را با برچسب صفر و قیمت مثبت را با برچسب یک برای یک سهم نمایش داده می‌شود.

change_end_price

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 هر روز

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 خزگر

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

پس از آموزش مدل نتایج ارزیابی پیش بینی ها در شاخص های دقت، صحت و امتیاز و میانگین ساده و وزنی هر سهم برای هر پارامتر محاسبه می شود.

```
number of future:1000
[[9 0]
 [3 4]]
```

	precision	recall	f1-score	support
	0.0	0.75	1.00	9
	1.0	1.00	0.57	7
micro avg	0.81	0.81	0.81	16
weighted avg	0.86	0.81	0.80	16

شکل ۱۸- نمونه ای از نتایج ارزیابی مدل

در شکل ۱۸ خط اول: نشان دهنده تعداد ویژگی های انتخابی

خط دوم و سوم: ماتریس در هم ریختگی

خط چهارم: شاخص های ارزیابی

خط ششم: نتایج ارزیابی پیش بینی جهت های منفی

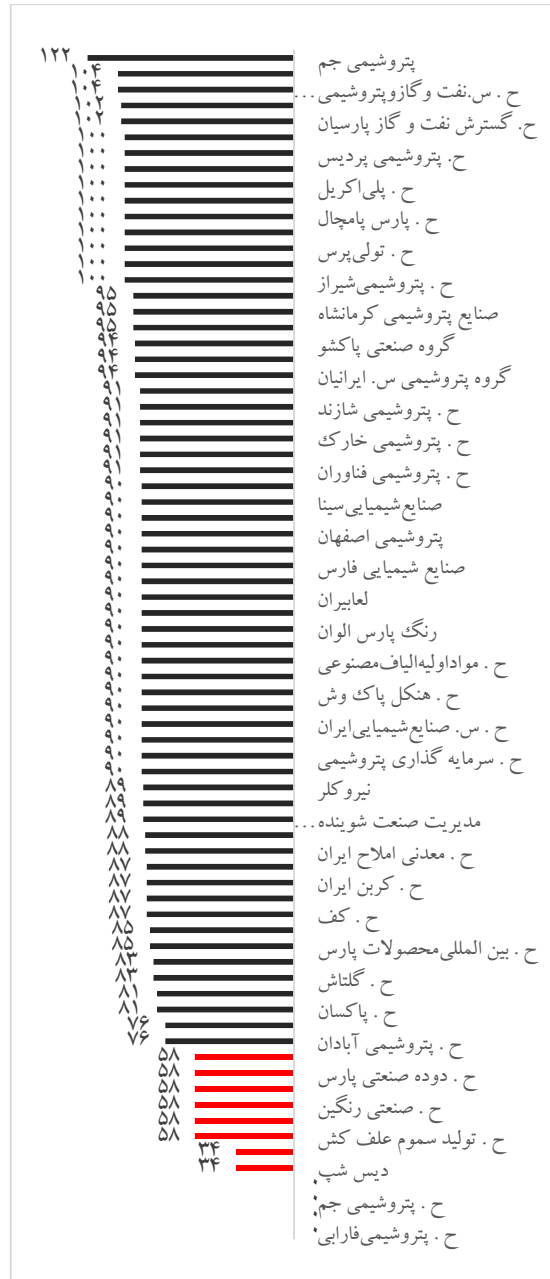
خط هفتم: نتایج ارزیابی پیش بینی جهت های مثبت

خط هشتم: میانگین ساده هر شاخص

خط نهم: میانگین وزنی هر شاخص

از آنجایی که امتیاز F معیار دقت و صحت را به صورت همزمان در نظر می گیرد شاخص مناسبی برای ارزیابی یک مدل دسته بندی 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 نرم افزار اکسل ذخیره می شود. نتایج حاصل از هر مدل در یک پوشه به اسم کرنل مورد استفاده شده ذخیره می شود. برای تهیه گزارش ها علاوه بر نتایج ارزیابی تعداد ویژگی های انتخاب شده و اندازه نمونه های آزمون نیست در فایل ذخیره می شود.



شکل ۱۷- تعداد روزهای معاملاتی هر سهم

	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 پیش بینی مدل

۳-۶- 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. از آنجایی که برای

ارزیابی از میانگین وزنی امتیاز F استفاده شد، تعداد سطرهای بارگذاری شده توسط اکسل ۲۰۰۶۴ سطر است.

تعداد روش‌ها: ۷

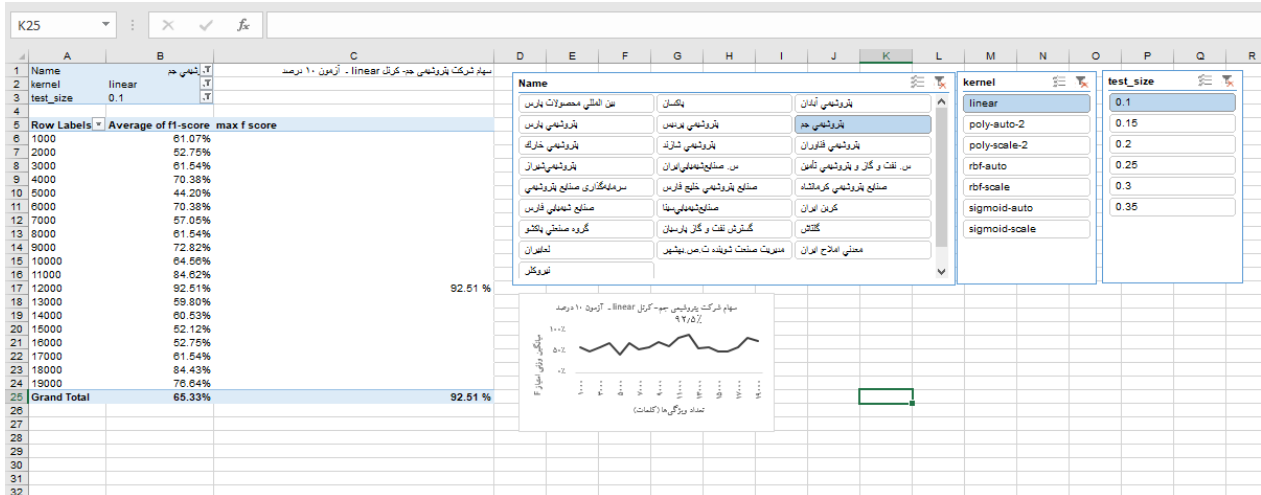
تعداد انتخاب شاخص‌ها: ۱۹

تعداد انتخاب نمونه آزمایش: ۶

تعداد سهم تحلیل‌شده: ۲۵

پس تعداد نتایج برابر است:

$$7 \times 19 \times 6 \times 25 = 20064$$

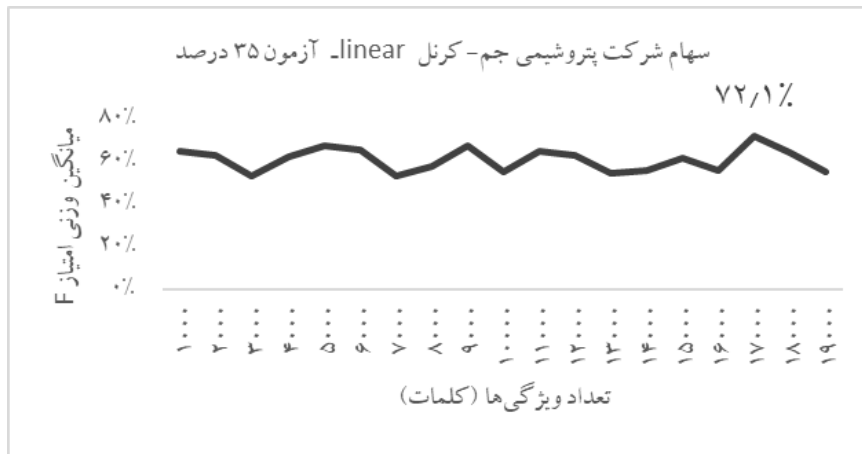


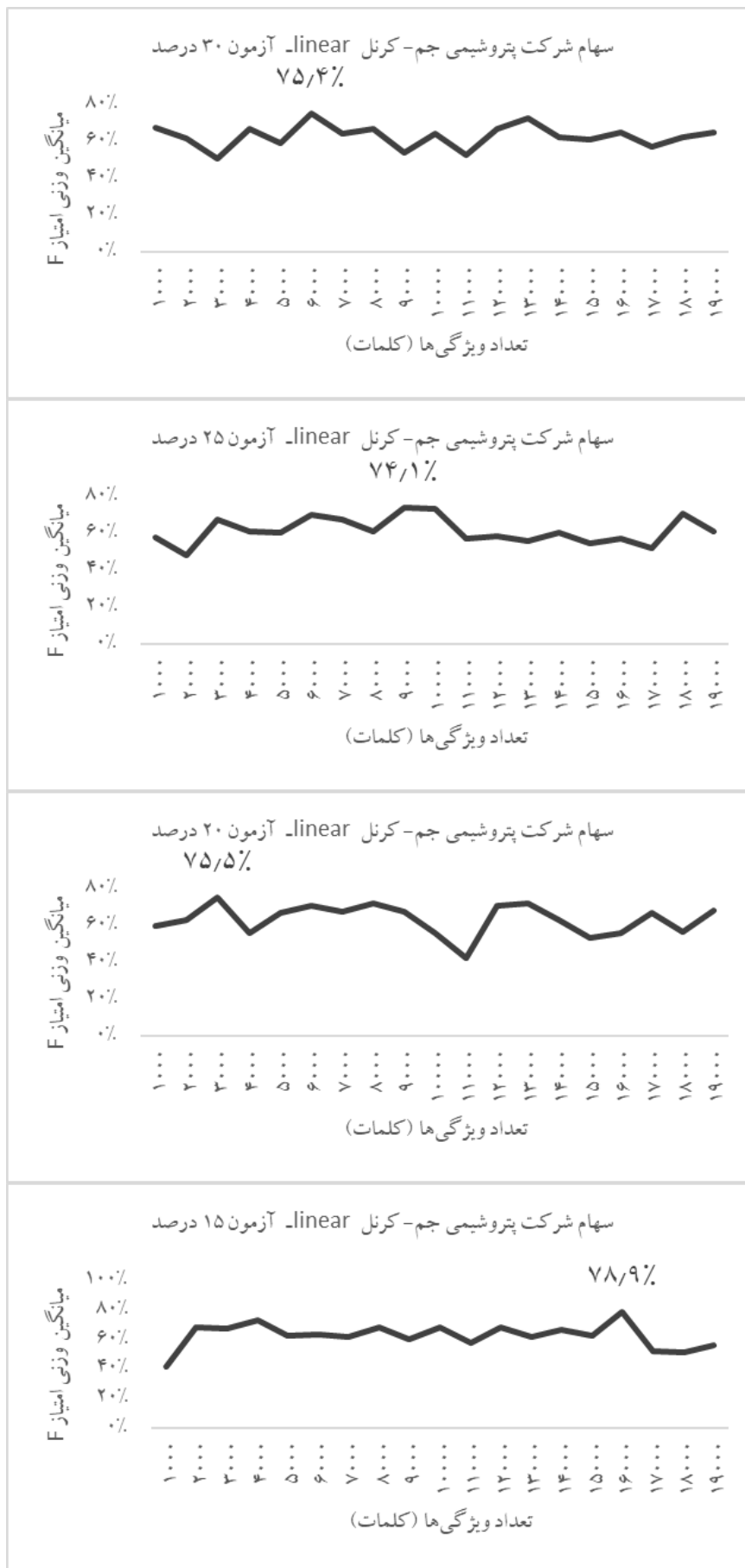
شکل ۲۰- تحلیل نتایج در اکسل

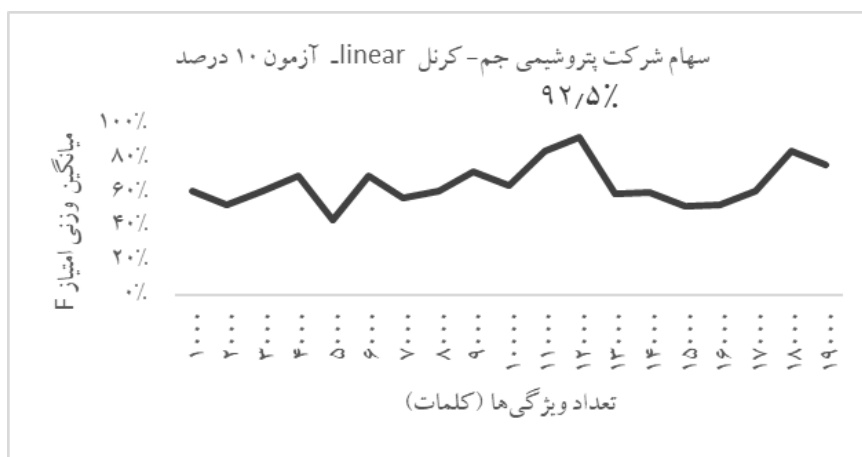
۱-۳-۶- نمونه ارزیابی نتایج برای سهم شرکت پتروشیمی

جم

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







۱-۳-۶-ارزیابی کلی نتایج

بررسی ۲۰۰۶۴ حالتی که برای این ۲۵ سهم اتفاق می‌افتد به صورت 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 آمده 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.			
صنایع شیمیایی ایران	۱۵۰۰۰	۰.۲۵	۷۳.۸۱%

س. نفت و گاز و پتروشیمی تأمین سرمایه‌گذاری	۸۰۰۰	۰.۱	۱۰۰.۰۰%
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%
صنایع شیمیایی 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 پارسین	۶۰۰۰	۰.۱	۹۱.۰۶%
گلتاش	۳۰۰۰	۰.۱	۸۹.۵۷%
لعابیران	۱۷۰۰۰	۰.۱۵	۷۸.۴۶%

مدیریت صنعت شوپنده

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 به صورت میانگین تا ۸۵.۳۱ درصد جهت قیمت‌ها را درست پیش‌بینی کند.

میانگین ۸۷.۲۷%

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

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

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

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

معادله درجه ۲ با گاما auto	درصد آزمایش	تعداد ویژگی‌ها	شرکت
			بین‌المللی
۸۱.۵۸%	۰.۲	۱۳۰۰۰	محصولات پارس
۱۰۰.۰۰%	۰.۱	۵۰۰۰	پاکسان
۸۷.۷۰%	۰.۱	۱۳۰۰۰	پتروشیمی آبادان
۸۰.۰۰%	۰.۱	۱۳۰۰۰	پتروشیمی پارس
۸۹.۳۳%	۰.۱	۳۰۰۰	پتروشیمی پردیس
۸۴.۴۳%	۰.۱	۳۰۰۰	پتروشیمی جم
۷۰.۸۳%	۰.۱۵	۶۰۰۰	پتروشیمی خارک
۷۰.۹۹%	۰.۱	۴۰۰۰	پتروشیمی شازند
۱۰۰.۰۰%	۰.۱	۱۱۰۰۰	پتروشیمی فناوریان
۱۰۰.۰۰%	۰.۱	۶۰۰۰	پتروشیمی شیراز
			س. صنایع شیمیایی ایران
۸۷.۵۰%	۰.۱	۱۶۰۰۰	س. نفت و گاز و پتروشیمی
۱۰۰.۰۰%	۰.۱	۴۰۰۰	پتروشیمی تأمین
			سرمایه‌گذاری
۸۷.۴۱%	۰.۱	۱۴۰۰۰	صنایع پتروشیمی
			صنایع پتروشیمی خلیج فارس
۸۰.۰۰%	۰.۱	۵۰۰۰	صنایع پتروشیمی کرمانشاه
۸۰.۰۰%	۰.۱	۵۰۰۰	صنایع شیمیایی فارس
۷۸.۳۳%	۰.۱	۱۶۰۰۰	صنایع شیمیایی سینا
۸۹.۱۸%	۰.۱	۱۸۰۰۰	کربن ایران
۹۰.۳۷%	۰.۱	۱۴۰۰۰	گروه صنعتی پاکشو
۹۰.۱۰%	۰.۱	۱۸۰۰۰	گسترش نفت و گاز
۸۲.۱۲%	۰.۱	۱۱۰۰۰	پارسیان
۸۴.۶۳%	۰.۱۵	۱۲۰۰۰	گلتاش
۸۹.۱۸%	۰.۱	۱۵۰۰۰	لعابیران
			مدیریت صنعت
			شوینده
۸۹.۱۸%	۰.۱	۸۰۰۰	ت.ص. بهشهر
۱۰۰.۰۰%	۰.۱	۹۰۰۰	مععدنی املاح ایران
۸۸.۸۹%	۰.۱	۲۰۰۰	نیروکلر

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 به صورت میانگین تا ۷۶,۱۹ درصد جهت قیمت‌ها را درست پیش‌بینی کند.

جدول ۷- ارزیابی نتایج کرنل شعاعی با گاما scale

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

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 به صورت میانگین تا ۷۷,۹۱ درصد جهت قیمت‌ها را درست پیش‌بینی کند.

جدول ۶- ارزیابی نتایج کرنل شعاعی با گاما auto

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

۸۵.۲۶%	۰.۱	۵۰۰۰	گروه صنعتی پاکشو
۸۳.۶۶%	۰.۱	۴۰۰۰	گلناتش
۴۱.۵۶%	۰.۱۵	۳۰۰۰	لعابیران
			مدیریت صنعت شونیده
۴۱.۵۶%	۰.۱۵	۲۰۰۰	ت.ص. بهشهر
۳۳.۳۳%	۰.۲	۱۰۰۰	معذنی املاح ایران
۸۳.۶۶%	۰.۱	۱۴۰۰۰	نیروکلر
۶۹.۴۷%			میانگین

با توجه به داده‌های بالا کرنل سیگموئید با گاما auto می‌تواند به صورت میانگین تا ۶۹,۴۷ درصد جهت قیمت‌ها را درست پیش بینی کند که نسبت به سایر الگوریتم‌ها نتایج نامناسب تری است.

جدول ۹- ارزیابی نتایج کرنل سیگموئید با گاما scale

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

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

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

جدول ۸- ارزیابی نتایج کرنل سیگموئید با گاما auto

کرنل سیگموئید با گاما auto	تعداد آزمایش	تعداد ویژگی‌ها	شرکت
			بین المللی محصولات پارس
۸۳.۶۶%	۰.۱	۱۶۰۰۰	پاکسان
۸۸.۶۲%	۰.۱۵	۳۰۰۰	پتروشیمی آبادان
۶۴.۲۹%	۰.۱	۷۰۰۰	پتروشیمی پارس
۱۰۰.۰۰%	۰.۱	۷۰۰۰	پتروشیمی پردیس
۸۵.۲۶%	۰.۱	۱۷۰۰۰	پتروشیمی جم
۶۶.۸۹%	۰.۱	۱۳۰۰۰	پتروشیمی خارک
۳۶.۳۰%	۰.۲	۳۰۰۰	پتروشیمی شازند
۷۱.۱۱%	۰.۱	۶۰۰۰	پتروشیمی فناوران
۵۷.۶۵%	۰.۱	۸۰۰۰	پتروشیمی شیراز
۷۱.۱۱%	۰.۱۵	۲۰۰۰	س. صنایع شیمیایی ایران
۶۸.۰۶%	۰.۱	۲۰۰۰	س. نفت و گاز و پتروشیمی تأمین سرمایه گذاری
۷۳.۶۴%	۰.۱	۱۱۰۰۰	صنایع پتروشیمی خلیج فارس
۷۹.۱۲%	۰.۱۵	۷۰۰۰	صنایع پتروشیمی کرمانشاه
۹۰.۱۱%	۰.۱۵	۱۶۰۰۰	صنایع پتروشیمی فارس
۷۱.۱۱%	۰.۱۵	۱۰۰۰	صنایع شیمیایی فارس
۳۹.۶۸%	۰.۱	۸۰۰۰	صنایع شیمیایی سینا
۶۸.۰۶%	۰.۱	۱۰۰۰	کربن ایران
۸۳.۶۶%	۰.۱	۶۰۰۰	

همه سهم‌ها انتخاب کرد، بلکه باید با توجه به نتایج حاصل شده 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.

صنایع شیمیایی سینا	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%
گلتنش	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 سیگنوید 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 متن کاوی افزایش می‌یابد. هر چند نتایج حاصل از داده کاوی 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 داده کاوی عمومیت بیشتری پیدا می‌کند.

یکی از کاربردهای متن کاوی بررسی تأثیر اخبار بر قیمت سهام در بورس است. از آنجایی که اخبار زیادی هر روز توسط خبرگزاری‌ها 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

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.

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.

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.

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

- 29.G. Dreyfus, *Neural networks: methodology and applications*: Springer Science & Business Media, 2۰۰۵
- 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
- 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, 2۰۱۷
- 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.

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.

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

