سال دوازدهم، شماره‌های ۴۳ و ۴۴، بهار و تابستان ۱۳۹۹
صفحات: ۱۱۹_۱۳۲

دو فصلنامه علمی
فناوری اطلاعات و ارتباطات ایران

تحلیل متنی خبرهای بانک مرکزی در پیش‌بینی بلندمدت شاخص بورس اوراق بهادار تهران

میثم هاشمی* مهران رضایی** مرجان کائدی***
*دانش آموخته کارشناسی ارشد مهندسی کامپیوتر، دانشگاه اصفهان
**استادیار دانشکده مهندسی کامپیوتر، دانشگاه اصفهان
***دانشیار دانشکده مهندسی کامپیوتر، دانشگاه اصفهان
تاریخ دریافت: ۱۳۹۹/۰۳/۱۶ تاریخ پذیرش: ۱۳۹۹/۰۸/۵
نوع مقاله: پژوهشی

چکیده

بازارهای مالی همواره تحت تاثیر انتشـارات رسـانه‌های خبری بوده‌اند. به همین دلیل تحلیل اسـناد خبری به عنوان یک رهیافت برای پیش‌بینی بورس اوراق بهادار به کار رفته است. در تحقیقات پیشین در این زمینه، تحلیل اسناد متنی با استفاده از روش‌های رایج در بازیابی اطلاعات انجام گرفته است. مبنای آماری این روش‌های رایج بر این است که کلماتی که در مجموعه اسناد کم‌تکرار هستند ولی در یک سند پرتکرار هستند، نسبت به کلمات پرتکرار مجموعه و سند، وزن بالاتری بگیرند. ولی مشکل این است که برخلاف آنچه در تحقیقات قبلی در نظر گرفته شده است، در اسناد خبری، کلمات پرتکرار نشان‌دهنده خبرهای مهم و تاثیرگذار هستند. در این تحقیق برای رفع این مشکل، یک روش جدید برای وزن‌دهی کلمات اسناد خبری ارائه شده است. روش پیشنهادی روی داده‌های شاخص کل بورس اوراق بهادار تهران و اسناد خبری بانک مرکزی ایران در بازه زمانی ۱۳۸۴ تا ۱۳۹۹ ارزیابی شده است. نتایج حاکی از ۶۴ در صد صعودی و ۴۱ در صد نزولی دقت پیش‌بینی نوسانات شاخص کل و کاهش ۱۰ در صد میانگین درصد خطای مطلق نسبت به بهترین روش رایج می‌باشد. همچنین نتایج نشان می‌دهد که اگرچه تغییرات در نسبت بین تعداد کلمات مثبت و منفی شواهد پیش گویانه ای ارائه نمیکند اما بین خبرهای منتشر شده از سوی بانک مرکزی و نوسانات شاخص کل بورس تهران ارتباط وجود دارد.

واژگان کلیدی: شاخص کل بورس تهران، پیش‌بینی بلندمدت، تحلیل متنی، اخبار مالی، وزن‌دهی DF

نویسنده مسئول: مهران رضایی m.rezaei@eng.ui.ac.ir

۱- مقدمه

اگر در زمان حال ادعایی درباره رویدادهای آینده مطرح شود یک پیش‌بینی انجام شده‌است. تحقیقاتی که در زمینه پیش‌بینی‌های بورس اوراق بهادار انجام شده‌است به دو دسته پیش‌بینی قیمت و گرایش آن تقسیم می‌شود که هر کدام بنا به زمان مورد پیش‌بینی به دو گروه بلندمدت و کوتاه‌مدت تقسیم می‌شوند. برای انجام یک پیش‌بینی، به داده‌های مرتبط با رویداد مورد پیش‌بینی نیاز است. داده‌های مورد استفاده در پیش‌بینی می‌توانند کمّی و یا کیفی باشند. داده‌های کمّی از ساختار بهره می‌برند و داده‌های کیفی فاقد ساختار هستند. روش‌های سری زمانی که از رایج‌ترین روش‌های پیش‌بینی به شمار می‌روند، از نوع داده کمّی برای مدل سازی و پیش‌بینی استفاده می‌کنند. داده‌های کمّی، داده تاریخچه‌ای هستند. در مقابل، داده‌های کیفی طیف وسیعی از داده‌ها را در بر می‌گیرند، از اطلاعات مربوط به عزل و نصب‌های داخل شرکت تا مدیران اقتصادی و بازارهای مالی و از جریانات و رویدادی داخلی تا رویدادهای کلان یک کشور و حتی اطلاعات داخلی رقبا همگی از نوع داده‌های کیفی محسوب می‌شوند. به طور کلی می‌توان گفت انواع داده در مورد رویدادهای خصوصی و همچنین رویدادهای عمومی جزء داده کیفی به شمار می‌آیند [۵-۱].

پیش‌بینی سری زمانی به عنوان روشی که از داده‌های کمی استفاده می‌کند سعی دارد براساس داده‌های گذشته، یک مدل به‌دست آورد. سپس با استفاده از مدل، مقادیر آینده را پیش‌بینی کند. پیش‌بینی سری زمانی یکی از مسائل مهم در سرمایه‌گذاری و تصمیم‌گیری به شمار می‌رود. با این وجود، در زمینه پیش‌بینی سری‌های زمانی، مشکلاتی نظیر نویز، پویایی و درهم برهم بودن داده کمّی مطرح است. عدم استفاده از داده کیفی نیز به عنوان یک مساله در زمینه سری‌های زمانی شناخته شده است که برای رفع آن، تحقیقات بسیاری انجام گرفته است [۱۱-۵].

اخبار، یکی از منابع اطلاعاتی موثر بر قیمت سهام است و بازار سهام نسبت به انتشار اخبار با محتوای مرتبط واکنش نشان می‌دهد [۱، ۴ و ۱۱]. در این زمینه، رابرتسون و همکاران بررسی کرده‌اند که آیا محتوای اخبار عمومی می‌تواند رفتار نابهنجار بازار سهام را پیش‌بینی کند و در بررسی خود با تحلیل خروجی دسته‌بندها، ثابت کرده‌اند که بازار به اخبار عمومی واکنش نشان می‌دهد [۱۲]. کلوپچنکو و همکاران با ترکیب روش‌های داده و متن‌کاوی در تحلیل داده‌های کمّی و کیفی گزاراشات مالی، ثابت کرده‌اند که تحلیل متنی گزارشات مالی شرکت‌ها، حاوی نمایه‌هایی راجع به عملکرد مالی شرکت در آینده می‌باشد [۲]. بیکر و همکاران در پیش‌بینی نوسانات کوتاه مدت قیمت سهام ثابت کرده‌اند که اخبار مالی بر روی قیمت سهام تاثیرگذار می‌باشد و افراد حرفه‌ای در بازار سهام به رسانه‌های خبری توجه ویژه‌ای دارند [۱۳]. آسائه با بررسی ارتباط بین قیمت سهام و اسناد خبری، ثابت کرده است که اخبار نقش بسزایی در نوسانات بازار سهام دارد [۸]. گوپتا و بانرجی [۱۴] نشان داده‌اند که حس مثبت یا منفی نهفته در اخبار سازمان OPEC بر روی بازده شرکت‌های فعال در بخش انرژی اثرگذار بوده است. به عنوان مثال، اخبار منفی OPEC باعث افزایش بازده شرکت‌های آمریکایی شده است. وو و همکاران [۱۵] و وِی و همکاران [۱۶] نیز در پژوهش‌هایی جداگانه نشان دادند که بر اساس اخبار اقتصادی می‌توان بازده بازار سهام در کشور تایوان را پیش‌بینی کرد.

بنابراین نتایج مطالعات یادشده در مجموع نشان می‌دهند که محتوای اخبار منتشر شده بر گرایشات بازار مالی تاثیرگذار بوده و این دو مرتبط هستند.

برای اینکه عوامل کیفی از متن اخبار استخراج شوند، از روش‌های متن‌کاوی استفاده می‌شود. رویکردهای تحلیل متنی ارائه شده در تحقیقات قبلی مبتنی بر روش‌های بازیابی اطلاعات و پردازش زبانی طبیعی می‌باشند [۱، ۸ و ۱۰]. در روش‌های یادشده، کلمات کم تکرار مجموعه نسبت به کلمات پرتکرار مجموعه وزن‌های بالاتری می‌گیرند و به ویژگی‌های خاص اسناد خبری توجه نشده است، حال آنکه در اسناد خبری کلمات پرتکرار نشان‌دهنده خبرهای مهم می‌باشند. این نکته در پژوهش پیش رو مدنظر قرار خواهد گرفت و براساس آن، روش برای وزن‌دهی به کلمات در متن‌های خبری ارائه خواهد شد. در ادامه، در بخش ۲ پژوهش‌های پیشین مرور می‌شوند و در بخش ۳، مساله پژوهشی شرح داده می‌شود. سپس در بخش‌های ۴ و ۵ به ترتیب به شرح رویکرد پیشنهادی و ارزیابی آن پرداخته می‌شود. در بخش ۶، نتیجه‌گیری و کارهای آینده بررسی خواهند شد.

۲- مرور تحقیقات پیشین

در این بخش، تحقیقات انجام شده در زمینه استفاده از داده‌های متنی برای پیش‌بینی مالی مرور خواهند شد. فوریگل و گوردون از داده‌های متنی خبری در بهینه‌سازی پیش‌بینی شاخص‌های مالی استفاده کرده‌اند. در سیستم پشتیبان

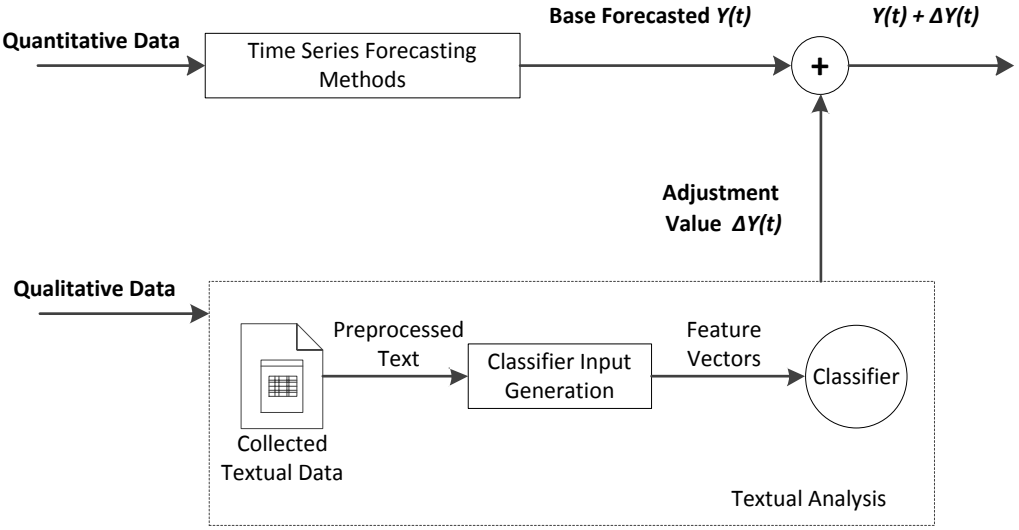
تصمیم که این پژوهشگران پیاده‌سازی کرده‌اند از روش TF-IDF برای وزن‌دهی استفاده شده است. روش ترکیبی ایشان با کاهش خطای RMSE به مقدار ۱۹/۵، ۱۹/۴ و ۳۵/۶ درصد به ترتیب در پیش‌بینی شاخص‌های DAX, CDAX و STOXX Europe 600 همراه بوده است [۱۷]. رن و همکاران با این باور که داده‌های متنی تولیدشده توسط کاربران می‌تواند در پیش‌بینی بازار سهام مورد استفاده قرار بگیرد رویکردی جدید مرکب از روش‌های یادگیری ماشین و بازیابی اطلاعات ارائه کرده‌اند. در این رویکرد به تحلیل احساسات بر روی داده‌های متنی پرداخته شده و در پیش‌بینی شاخص SSE 50 مقدار ۸۹/۹۳ درصد صحت جهت‌گیری به دست آمده است [۱۸]. یکی دیگر از مطالعات انجام شده با استفاده از تحلیل احساسات به عنوان رویکرد متن کاوی در پیش‌بینی مالی توسط رحمان و همکاران با هدف پیش‌بینی تغییرات قیمت سهام در بورس مالزی انجام شده است. در رویکرد ترکیبی ایشان از ماشین بردار پشتیبان و TF-IDF استفاده شده و روند تغییرات قیمت در بازار سهام با نرخ ۵۶ درصد صحت انجام شده است [۱۹]. نادری و همکاران با استفاده از تحلیل متنی رویدادهای خبری به پیش‌بینی روند تغییرات قیمت در بازار ارز پرداخته‌اند. ایشان با استفاده از روش‌های یادگیری ماشین و صرفا روش وزن‌دهی TF بجای TF-IDF به ۶۶/۳ درصد صحت پیش‌بینی دست یافته‌اند [۲۰]. لین یو و همکاران با بکارگیری تئوری مجموعه‌های ناهموار بر روی داده‌های به‌دست آمده از تحلیل متنی، نوسانات قیمت بازار نفت خام را پیش‌بینی کردند. روش آنها نسبت به شبکه‌های عصبی ۱۱ درصد و نسبت به رگرسیون خطی ۳۰ درصد بهبود داشته است [۲۱]. لاس و اسپرمانت برای پیش‌بینی تغییرات نابهنجار قیمت در بازار سهام از تحلیل متنی، استفاده کرده‌اند که حداکثر دقت دسته بندی اسناد خبری در پژوهش ایشان ۷۱ درصد گزارش شده است [۲۲]. پینتو و اسنانی با استفاده از اخبار و گفتگوهای معاملاتی، گرایش قیمت سهام را پیش‌بینی کرده‌اند که دقت ۵۰ درصد برای پیش‌بینی شاخص را گزارش کرده‌اند [۲۳]. ناصری و احمدی برای پیش‌بینی قیمت نفت در بلندمدت یک رویکرد ترکیبی مبتنی بر خوشه بندی نزدیک ترین همسایه، ژنتیک و شبکه عصبی مصنوعی بر روی داده‌های کمی و کیفی ارائه داده‌اند که میانگین دقت جهت گیری قیمت نفت را در بلندمدت به میزان ۷۸ درصد به دست آورده‌اند [۲۴]. آنتویلر و فرانک با استفاده از روش‌های پردازش زبان طبیعی، پیام‌های کاربران یاهو درباره شرکت‌های تجاری را برای به دست آوردن توصیه‌ای جهت خرید، فروش و نگهداری یک سهام خاص، مورد تحلیل قرار داده‌اند [۲۵]. گواردیا سباوم و همکاران برای پیش‌بینی فروش هفتگی فیلم، از روش‌های نظرکاوی بر روی داده‌های توئیتر استفاده کرده‌اند که خطای میانه در بین سه معیار خطای متفاوت، ۵۶ درصد بوده است [۲۶]. جیوجیانگ و همکاران با استفاده از روش‌های متن کاوی و شبکه عصبی مصنوعی به پیش‌بینی قیمت طلا پرداخته‌اند که مقدار میانگین مربع خطا را ۲۷۱۴ به دست آورده‌اند [۲۷]. در زمینه پیش‌بینی فروش هفتگی فیلم، آسور و همکاران از داده‌های توئیتر استفاده کرده‌اند که میانگین درصد خطای مطلق را برابر با ۰/۵۶ گزارش کرده‌اند [۲۸]. فیوریگل و گوردون برای پیش‌بینی شاخص‌های اقتصاد کلان، روش‌های مختلف یادگیری ماشین را بر روی فراوانی کلمات مختلف در اخبار اعمال کردند و سپس برای مواجهه با مشکل بیش-برازش که ناشی از ابعاد زیاد داده بود، مهندسی ویژگی‌ها انجام دادند. مثلا کلماتی که از دسته‌های معنایی یکسانی هستند را به ساختارهای پنهانی نگاشت کردند تا بدین‌ترتیب، ابعاد فضای ویژگی به میزان قابل توجهی کاهش یابد[۲۹]. چن و همکاران [۳۰] حس عمومی جامعه را از طریق تحلیل اخبار موجود در شبکه‌های اجتماعی استخراج کرده‌اند و از آن برای پیش‌بینی نواسانات بازار سهام در کشور چین استفاده کرده‌اند. نَم و سئونگ [۳۱] روابط سببی بین شرکت‌هایی که با هم ارتباط دارند، را برای پیش‌بینی قیمت سهام یک شرکت درنظر گرفته‌اند و اخباری که در مورد شرکت‌های مرتبط با شرکت موردنظر منتشر می‌شود را برای پیش‌بینی نوسانات قیمت سهام شرکت موردنظر استفاده می‌کنند. دایی و همکاران با استناد به خبرهای سیاسی و اقتصادی ۱۷ خبرگزاری در طول ۱۲۲ روز آخر سال ۱۳۹۷ شمسی، به پیش‌بینی جهت قیمت سهام در بورس اوراق بهادار مبادرت ورزیدند [۴۳]. آنها در پژوهش خود از روش متن‌کاوی استفاده کرده و برای انتخاب و وزن‌دهی ویژگی‌ها فرمول TF-IDF (فرمول ۱) را بکار گرفته‌اند. سپس روش خود را با الگوریتم بردار ماشین پشتیبان برای پیش‌بینی جهت قیمت ترکیب کرده و سهام گروه محصولات شیمیایی را مورد مطالعه قرار داده‌اند.

تحلیل متن در مطالعات مرور شده، با استفاده از روش‌های رایج در زمینه‌های بازیابی اطلاعات و پردازش زبان طبیعی انجام شده است. اما با توجه به عدم کارایی مناسب اینگونه روش‌ها در زمینه تحلیل متن مالی، در این مطالعه یک روش جدید برای تحلیل متن مالی ارائه خواهد شد که به طور خاص منطبق با اسناد خبری عمل خواهد کرد.

۳- شرح مساله

در شکل ۱، معماری عمومی پژوهش‌های انجام شده در زمینه به‌کارگیری تکنیک‌های متن‌کاوی در پیش‌بینی‌های مالی، نشان‌داده شده‌است.

اسناد متنی جمع‌آوری شده پس از عملیات پیش‌پردازش متن، در مرحله تولید ورودی دسته‌بند، به بردارهای ویژگی تبدیل می‌شوند. همانطور که در معماری عمومی (شکل ۱) مشاهده می‌شود، در موازات تحلیل داده‌های کمّی که منجر به تولید یک مقدار پایه برای پیش‌بینی می‌شوند، داده‌های کیفی نیز مورد تحلیل قرار گرفته و منجر به تولید مقدار تعدیلی می‌شوند.



شکل ۱. معماری عمومی یک سیستم پیش‌بینی مالی با استفاده از داده‌های متنی (با الهام از ون-بین یو [۲۸])

در برخی پژوهش‌های نیز از داده‌های کمّی استفاده نشده است و پیش‌بینی مالی صرفا با استفاده از داده‌های کیفی انجام گرفته است. در این گونه پژوهش‌ها در بخش تحلیل متنی، قیمت با استفاده از روش‌های یادگیری ماشین قیمت پیش‌بینی می‌شود. این رویکردی است که در پژوهش پیش رو نیز مد نظر است.

بخش تحلیل متنی شامل گام‌های جمع‌آوری اسناد متنی، پیش‌پردازش، تبدیل سند به بردار ویژگی، وزن‌دهی ویژگی، کاهش ویژگی و انتخاب ویژگی می‌باشد. روش رایجی که به طور معمول مورد استفاده قرار می‌گیرد پشته کلمات است که براساس وزن‌دهی TF-IDF بنانهاده شده‌است و روش محاسبه آن برای کلمه W_i در سند d_j در رابطه (۱) نشان‌داده شده‌است [۳۳]. در این رابطه، مقدار N تعداد اسناد مجموعه، n_i تعداد اسناد حاوی کلمه W_i و $f_{i.j}$ نرخ تکرار کلمه W_i در سند d_j می‌باشد. رابطه (۲) به نام احتمالی شناخته شده و مبتنی بر احتمالات و قانون بیز بوده که یکی از مدل‌های وزن‌دهی پایه می‌باشد [۳۳].

$$w_{i,j} = \left(1 + Log\ f_{i,j}\right) Log \frac{N}{n_i} \qquad (۱)$$

$$w_{i,j} = \frac{P_i R}{1 - P_i R} \qquad (۲)$$

$$w_{i,j} = \begin{cases} 1 & if\ \exists c(q) \mid c(q) = c(d_j) \\ 0 & otherwise \end{cases} \qquad (۳)$$

در این رابطه P_i احتمال وجود کلمه i در مجموعه اسناد R می‌باشد. رابطه (۳) به نام بولی شناخته می‌شود و مقادیر بدست آمده از آن در بازه بسته اعداد طبیعی [۱ ۰] قرار میگیرند [۳۳]. در این رابطه $c(q)$ ویژگی درخواست شده است و بقیه متغیرها مانند رابطه (۱) می‌باشند.

براساس رابطه (۱)، به عنوان روش اصلی وزن‌دهی در متن کاوی مالی، کلمه‌ای که در اسناد کمتری تکرار شده باشد وزن بالاتری خواهد داشت. چرا که می‌تواند در زمان بازیابی اطلاعات برای یک کوئری، شانس بیشتری برای بازیابی سندی فراهم کند که حاوی کلمه‌ای است که در تعداد اسناد کمتری تکرار شده‌است. اما در اسناد خبری وضعیت متفاوتی وجود دارد. همه خبرها به نوبه خود می‌توانند بر بازارهای مالی تاثیرگذار باشند اما جریان اصلی تاثیرگذار بر نوسانات بورس اوراق بهادار حاوی خبرهای مهم است. از طرفی خبرهای مهم

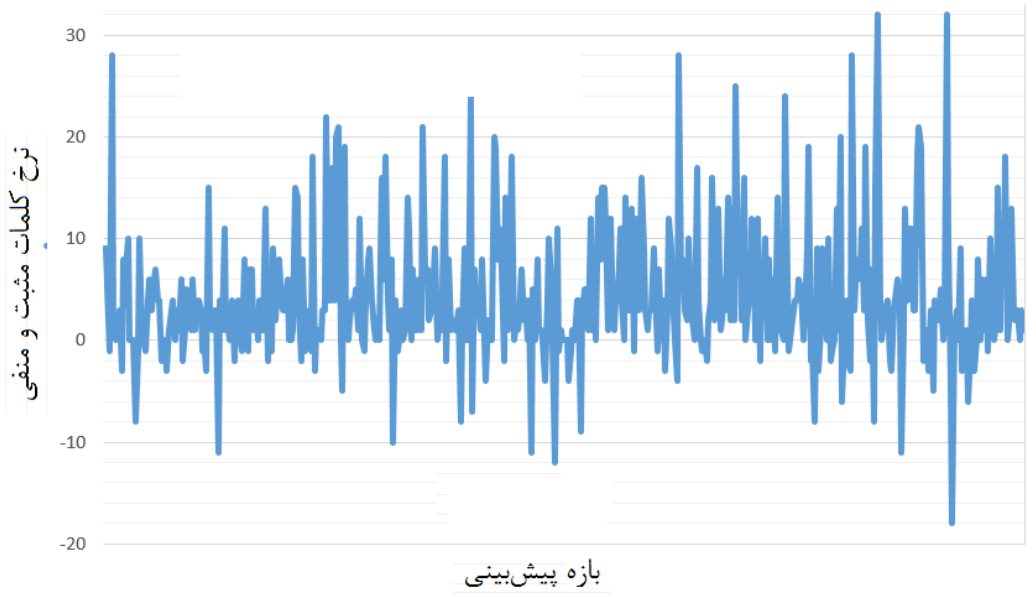
صرفا توسط یک خبرگزاری منتشر نخواهند شد و بالعکس خبرهای مهم در یک روز بارها توسط خبرگزاری‌های متفاوت منتشر خواهند شد و بنابراین در میان کلیه اسناد خبری، خبرهای مهم نرخ انتشار بالاتری خواهند داشت. چنانچه از رابطه (۱) برای وزن‌دهی کلمات خبرها استفاده کنیم خبرهای با نرخ انتشار بالا وزن‌های کمتری به‌دست می‌آورند. نتیجه آن خواهد شد که به خبرهای مهم و به عبارتی جریان اصلی خبری ارزش کمتری داده شود. این مساله‌ای است که در این پژوهش پیش رو، مدنظر است. در این مقاله روشی ارائه خواهد شد که برخلاف رابطه (۱)، کلمات کم تکرار و پرتکرار مجموعه به ترتیب وزن‌های کمتر و بیشتری دریافت کنند. پیش‌بینی مالی براساس این وزن‌دهی انجام می‌شود.

۴- رویکرد پیشنهادی

در این بخش یک روش جدید برای وزن‌دهی کلمات اسناد خبری ارائه و شرح داده می‌شود که در آن کلمات پرتکرار مجموعه اسناد خبری با مقادیر بزرگتر نسبت به کلمات کم تکرار مجموعه، وزن‌دهی می‌شوند.

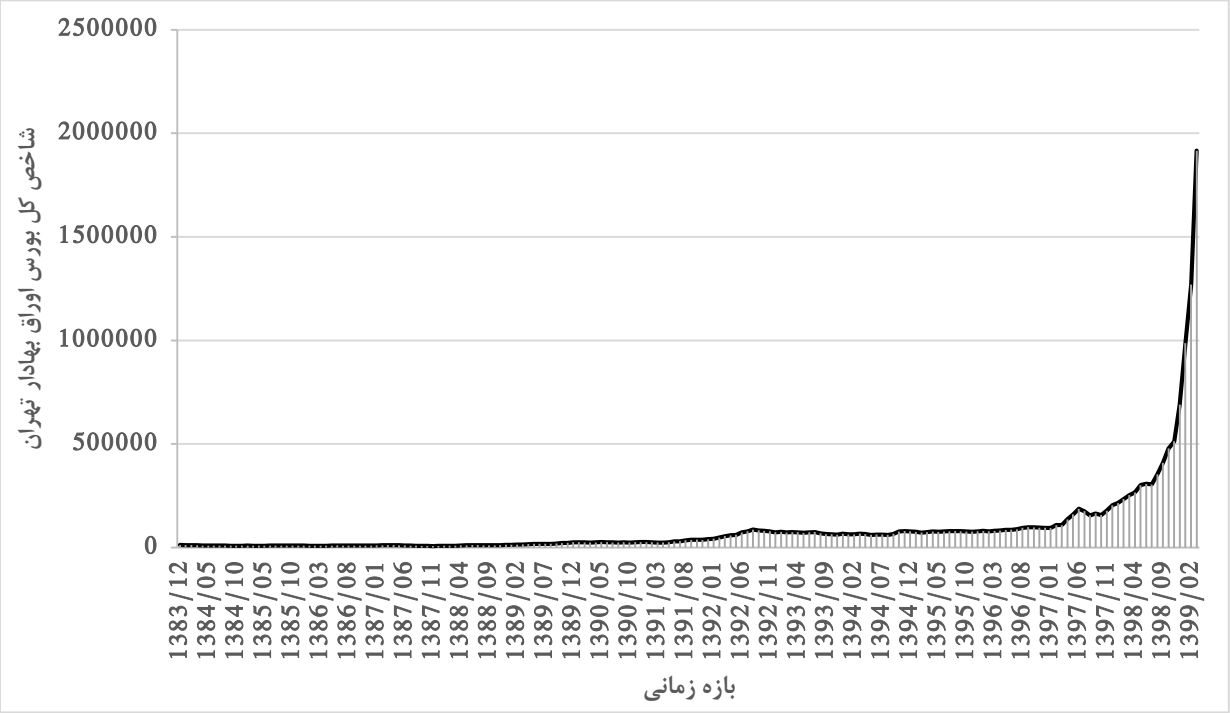
در شکل ۲، نسبت تکرار کلمات مثبت و منفی در خبرهای بانک مرکزی در بازه زمانی مورد تحقیق (۱۳۸۴ تا ۱۳۹۹) نشان داده شده است. در این شکل، محور افقی بازه زمانی و محور عمودی نرخ کلمات مثبت و منفی می‌باشد.

در شکل ۳ نیز شاخص کل بورس اوراق بهادار تهران مشاهده می‌شود.



شکل ۲. نرخ کلمات مثبت و منفی در خبرهای بانک مرکزی

در این شکل، محور افقی زمان و محور عمودی مقدار شاخص می‌باشد. همانطور که از مقایسه شکل ۲ و ۳ مشاهده می‌شود، نرخ تکرار کلمات مثبت و منفی در بازه زمانی زمانی ۹ ساله تقریبا مساوی است، حال آنکه روند تغییرات شاخص کل بورس اوراق بهادار تهران در همین بازه زمانی بیش از ۵ برابر رشد داشته است. بنابراین اگر کلمات مثبت و منفی در کل دوره در نظر گرفته شوند پیش‌بینی گرایش شاخص کل به درستی انجام نخواهد شد. زیرا مثلا چه زمانی که مقدار تقریبی شاخص کل ۱۰۰۰۰ و چه زمانی که مقدار این شاخص برابر با ۵۰۰۰۰ بوده، تعداد کلمات مثبت منفی در اسناد خبری بانک مرکزی تقریبا مساوی است. بنابراین اگر وزن کلمات در هر دو مقطعِ زمانی یادشده، یکسان محاسبه شود پیش‌بینی گرایش قیمت به درستی انجام نمی‌گیرد [۳۴-۳۶]. در ادامه روش پیشنهادی برای مواجهه با این موضوع شرح داده می‌شود.



شکل ۳. روند شاخص کل بورس اوراق بهادار تهران از ۱۳۸۴ تا ۱۳۹۹ [۳۶]

اسناد متنی تقریبا از هیچ گونه ساختاری بهره نمی‌برند. حتی قالب کدهای HTML یکسانی نیز ندارند. برای آنکه از یک سند HTML صرفا و دقیقا محتوای اصلی آن استخراج گردد نیاز به تفسیر کدهای آن سند می‌باشد. خوشبختانه اسنادی که از یک وب‌سایت دریافت می‌شوند قالب‌های شبیه به یکدیگر دارند. خبرهای منتشر شده در وب‌سایت بانک مرکزی نیز همین ویژگی را دارند [۷ و ۳۲].

پس از پیش‌پردازش اولیه و تبدیل اسناد درهم HTML به اسناد پالایش شده Text، گام تحلیل واژگان آغاز می‌شود. در این تحقیق گام تحلیل واژگان به شیوه استفاده شده در [۳۹] انجام می‌شود. در این گام یک رشته متنی به اجزائی مانند کلمات، نمادها، حروف ربط و ... به نام توکن تقسیم می‌شود. واژه‌های سراسری، کلماتی با نرخ تکرار تقریبا کامل در کل اسناد هستند و از نظر اطلاعاتی ارزش نداشته و قابل توجه نمی‌باشند. یکی از گام‌های پیش‌پردازش اسناد متنی، حذف این طیف از کلمات است. لیست‌های گوناگونی از واژه‌های سراسری وجود دارد که در این تحقیق از یک لیست کوتاه [۴۰] استفاده شده‌است.

مثال (۱):

ورودی: نرخ ارز غیر مرجع در مرکز مبادلات ارزی اعلام شد.

خروجی: [«نرخ»، «ارز»، «غیر»، «مرجع»، «در»، «مرکز»، «مبادلات»، «ارزی»، «اعلام»، «شد»، «.»]

در مرحله تحلیل واژگان، هر رشته کلمات به اجزاء آن تقسیم می‌شود و هر جزء بیانگر یک ویژگی است. از این‌رو تعداد ویژگی‌ها بسیار زیاد است و نمی‌توان برای دسته‌بندی از همه ویژگی‌ها استفاده کرد. یکی از مراحل بسیار حساس در فرآیند تحلیل متن، مرحله انتخاب ویژگی‌ها است. در این مرحله کلماتی از بین کل اسناد انتخاب می‌شوند. در برخی پژوهش‌های پیشین [۳۲ و ۳۷]، از روش دستی برای انتخاب ویژگی استفاده شده است. در پژوهش ما نیز از مجموعه کلمات مثبت و منفی استخراج شده توسط یو و همکاران [۳۲] استفاده شده است (پس از انجام بومی‌سازی). این کلمات در پیوست (۱) آورده شده‌اند.

برای آنکه محاسبات عددی بر روی متن قابل انجام باشد، هر سند متنی باید به شکلی بازنمایی شود. سند متنی در متن‌کاوی و بازیابی اطلاعات براساس یک مدل فضای برداری بازنمایی می‌شود [۳۸]. ابعاد در این مدل متناظر با ویژگی‌های مستخرج از متن می‌باشند و هر بردار توصیفی از یک سند متنی خواهد بود. بنابراین هر سند توسط یک بردار و هر ویژگی توسط یک بعد از بردار بازنمایی می‌شود. روش‌های متنوعی برای بازنمایی بوسیله فضای برداری قابل پیاده‌سازی

است. روش رایج مبنی بر این است که متناظر با تکرار هر ویژگی در هر سند، وزن ویژگی قرار می‌گیرد. در تحقیقات قبلی متداول بوده است که این وزن از رابطه (۱) محاسبه شود و سایر ویژگی‌ها که در یک سند تکرار نشده‌اند مقدار صفر بگیرند [۷، ۳۲ و ۴۱]. ولی در این مقاله روش جدیدی برای وزن‌دهی ویژگی‌ها ارائه شده که با مقدار DF معرفی می‌شود و مقدار آن از رابطه (۴) به دست می‌آید:

$$DF_i = \frac{F_i}{LogN} \quad (۴)$$

در این رابطه، مقادیر F_i و N به ترتیب مربوط به تعداد تکرار ویژگی i در مجموعه و تعداد ویژگی‌های مجموعه هستند.

در مرحله اجرای محاسبات پیش‌بینی می‌توان از روش‌های یادگیری ماشین و یا آماری استفاده کرد. ما در این مطالعه از روش جستجوی محلی استفاده کرده‌ایم. با استفاده از جستجوی محلی تعداد رخدادهای ویژگی‌های منتخب در داده‌های متنی به دست می‌آید. سپس برای هر ویژگی رخداده (یافته شده) وزن متناظر با استفاده از رابطه ۴ محا سبه می شود. سرانجام، با ا ستفاده از مجموع مقادیر به دست آمده از وزن ویژگی‌ها به همراه مقدار شاخص کل در بازه زمانی فعلی پیش‌بینی شاخص کل در بازه زمانی آتی انجام می‌گیرد. توصیفی از جریان داده و مراحل کار در شکل ۴ آورده شده است. در این شکل کلیه پردازش‌هایی که برای پالایش یک سند HTML باید طی شود تا داده متنی خالص به دست آید در یک مرحله با عنوان پالایش متن نشان داده شده است. در این مرحله، فرآیندهای پاک سازی شامل حذف انواع تگ‌های HTML و داده‌های نامرتبط و فرآیندهای پیش پردازش متن شامل توکن‌سازی، حذف استاپ‌وردها و برچسب‌گذاری زمانی است. پس از انجام فرآیندهای پیش‌پردازش متن، مهم‌ترین مرحله یعنی وزن‌دهی کلمات آغاز می شود. خروجی این مرحله، بردارهای ویژگی است که در محاسبه مقدار تعدیلی مورد استفاده قرار می‌گیرند. حاصل‌ضرب مقدار تعدیلی و مقدار پایه شاخص کل، عددی است که نشان‌دهنده مقدار پیش‌بینی شده برای شاخص کل در پایان ماه می‌باشد.

۵- ارزیابی

معیارهای متنوعی برای ارزیابی رویکردهای پیش‌بینی مالی وجود دارد. با توجه به اینکه در معیارهای دقت جهت‌گیری (Directional Accuracy) و میانگین خطای مطلق در پیش‌بینی‌های بلندمدت برای ارزیابی نتایج انتخاب شده‌اند [۱، ۸ و ۳۶]، در این پژوهش نیز از این دو معیار برای ارزیابی روش پیشنهادی استفاده می‌شود. معیار دقت جهت‌گیری، نشان‌دهنده توانایی روش در پیش‌بینی روند تغییرات قیمت است. همچنین معیارهای درصد خطای مطلق و مجذور میانگین خطای مربع پیش‌بینی نیز برای ارزیابی نتایج به کار خواهند رفت. رابطه‌های (۵) تا (۸) به ترتیب روش محاسبه دقت جهت گیری، خطای مطلق، درصد خطای مطلق و مجذور میانگین خطای مربع پیش‌بینی را نشان می‌دهند.

$$DA = \frac{TR + TF}{TR + TF + FR + FF} \quad (۵)$$

$$MAE = \frac{1}{N}\sum_{t=1}^{N} |E_t| \quad (۶)$$

$$MAPE = \frac{100}{N}\sum_{t=1}^{N} \left|\frac{E_t}{Y_t}\right| \quad (۷)$$

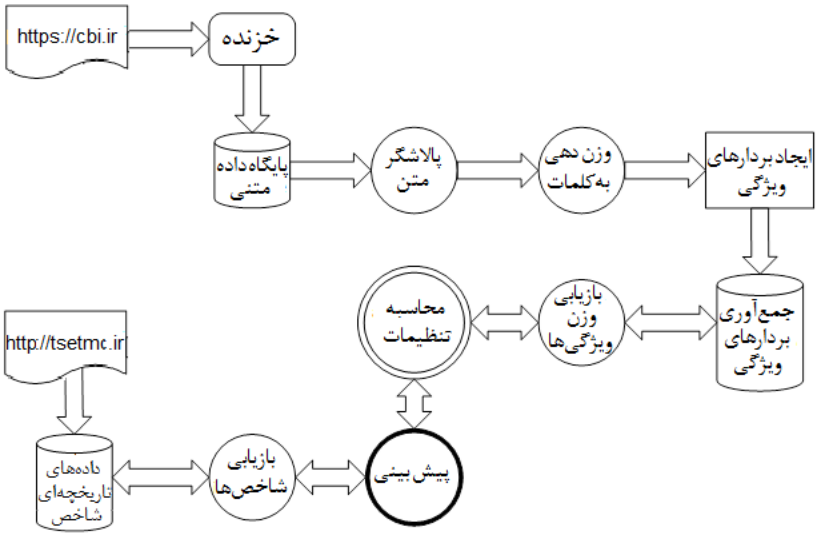
$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{N} E_t^2} \quad (۸)$$

در این روابط، تعداد آزمایش‌ها با N، خطای نقطه‌ای با E_t، جهت‌گیری‌های صحیح با $TR + TF$ و جهت‌گیری‌های غلط با $FR + FF$ نشان‌داده شده‌اند.

۵-۱ مجموعه داده

داده‌های مورد استفاده در تحقیق شامل دو نوع داده کمی (تاریخچه‌ای) و کیفی می‌باشند. مجموعه داده کیفی متشکل از اسناد خبری بانک مرکزی از آبان ماه سال ۱۳۸۴ تا تیرماه سال ۱۳۹۹ است و داده‌های کمی نیز شامل شاخص کل بورس اوراق بهادار مربوط به همین بازه زمانی می‌باشد. بخشی از داده‌های کمی (از ۱۳۸۷ تا ۱۳۹۹) از وب‌سایت رسمی بورس اوراق بهادار تهران (به آدرس http://tsetmc.ir) و بقیه داده‌ها (از ۱۳۸۳ تا ۱۳۸۷) از وب‌سایت رهاورد۳۶۵ (به آدرس https://rahavard365.com) گردآوری شده‌اند. داده‌های کیفی (یا به عبارتی اسناد متنی خبری) نیز به طور کامل از وب‌سایت رسمی بانک مرکزی ایران (به آدرس https://cbi.ir) قابل دریافت است. این اخبار به صورت سالانه تفکیک شده‌اند و دریافت آنها به صورت اسناد HTML امکان‌پذیر است. کل داده متنی بانک مرکزی شامل ۳۹۳۳ خبر بوده و شاخص بورس برای ۱۸۴ ماه پیش‌بینی شده است. از کل بازه زمانی مورد نظر این پژوهش، ماه‌های

مرداد و شهریور سال ۸۴ بدون خبر بوده و مورد پیش‌بینی قرار نگرفته است.



شکل ۴: توصیفی از جریان داده و مراحل کار

جدول ۱. میانگین دقت جهت‌گیری

DA	TF-IDF	Boolean	Probabilistic	DF
Observed Up	57.14	57.14	57.14	57.14
True Forecasted Up	81.73	**84.61**	83.65	64.42
Observed Down	42.85	42.85	42.85	42.85
True Forecasted Down	26.92	23.08	24.36	**41.02**
Observed Bias St.	14.29	14.29	14.29	14.29
Forecasted Bias St.	54.81	61.53	59.29	**23.4**

۵-۲ پیش‌بینی شاخص کل در بلند مدت

شکل ۵ پیش‌بینی گرایش قیمت را با استفاده از رویکرد پیشنهادی نشان می‌دهد. در بازه زمانی مربوط به اواخر سال ۱۳۸۴ تا اواسط سال ۱۳۸۸ نوسان شاخص کل بسیار کم بوده و مشاهده می‌شود که مقدار پیش‌بینی شده نیز به مقدار واقعی نسبتا نزدیک است. از سال ۱۳۸۸ تا ۱۳۹۲ و از سال ۱۳۹۷ تا ۱۳۹۹ شاخص کل با تغییرات زیادی همراه بوده که پیش‌بینی آن نیز با خطای بیشتری مواجه شده است. بیشترین خطای پیش‌بینی مربوط به فروردین ماه ۱۳۹۹ بوده است. در این مقطع زمانی مقدار واقعی ۶۹۰۰۳۷ و مقدار پیش‌بینی شده ۵۵۳۶۹۱ می‌باشد اما جهت‌گیری صحیح پیش شده است.

تغییرات شاخص کل با استفاده از تحلیل خبرهای بانک مرکزی و در بازه ماهانه، در وضعیت‌هایی که تغییرات شاخص به صورت صعودی و نزولی بوده است، پیش‌بینی شد. در پایان ماه، تغییرت شاخص برای روز پایانی ماه بعد پیش‌بینی شد. در این پیش‌بینی، برای معیار درصد بایاس جهت‌گیری، مقدار ۲۳/۴ درصد به دست آمده است. بنابراین، بین خبرهای منتشرشده از سوی بانک مرکزی و نوسانات شاخص کل ارتباط وجود داشته و کشف آن با استفاده از روش پیشنهادی میسر شده است.

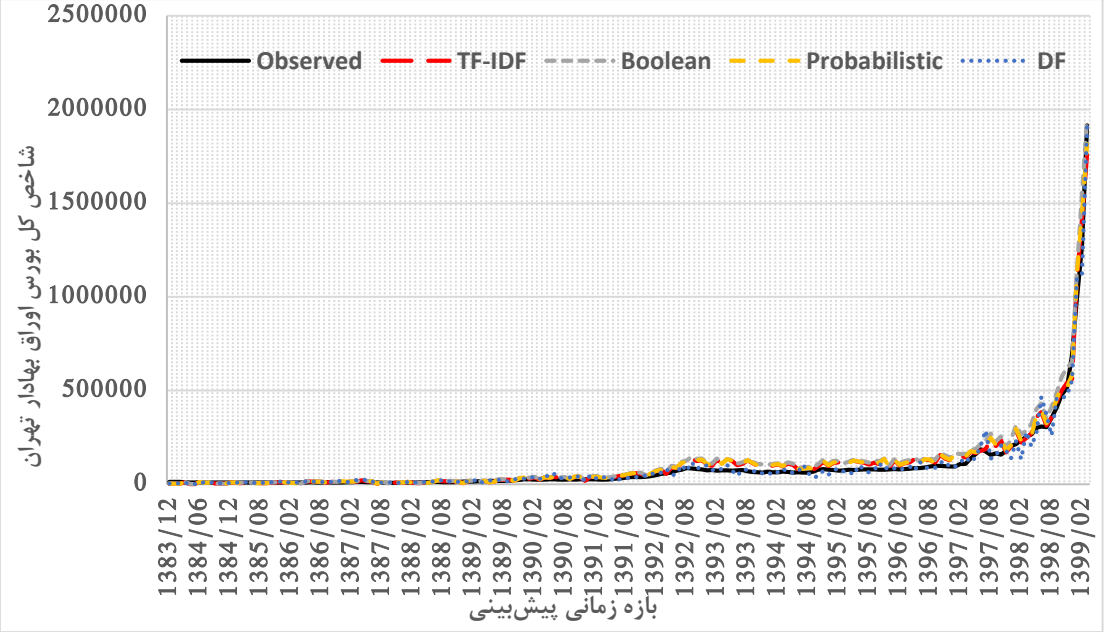
نتیجه پیش‌بینی روند تغییرات شاخص در شکل ۵ نشان داده شده است. در جدول‌های ۱ و ۲ روش پیشنهادی (DF) با روش‌های TF-IDF، احتمالی و بولی مقایسه شده است. جدول ۱ نیز میانگین دقت جهت‌گیری را با استفاده از رابطه

(۵) نشان می‌دهد. همچنین جدول ۲ میانگین خطا را با استفاده از روابط (۶) تا (۸) نشان می‌دهد.

میانگین خطای مطلق برای روش پیشنهادی برابر با ۱۵۱۶۹ است. اگر چه این خطا نسبتا زیاد به نظر می‌رسد، اما میانگین خطای مطلق برابر با تنها ۲۶ درصد است و باید در نظر داشت در پیش‌بینی گرایش بلندمدت به دلیل فاصله‌های یک ماهه بین هر دو مقطع زمانی مورد پیش‌بینی، مقدار ۱۵۱۶۹ برای میانگین خطای مطلق می‌تواند قابل قبول باشد. چرا که میزان تغییرات شاخص کل نسبت به تغییرات قیمت سهام بیشتر است، که این امر خود باعث می‌شود خطای پیش‌بینی شاخص کل در مقایسه با خطای پیش‌بینی قیمت سهام زیادتر باشد. حتی در برخی از بازه‌های مورد پیش‌بینی مقدار شاخص ۱۰۰ درصد افزایش یافته است که چنین نوسانی کار پیش‌بینی را بسیار مشکل می‌کند. مقدار بایاس روش پیشنهادی نسبت به روش‌های دیگر بسیار کمتر بوده و توانسته ۴۱ درصد روزهای نزولی را پیش‌بینی کند. مقدار بایاس رویت شده ۱۴ درصد و مقدار بایاس پیش‌بینی شده ۲۳ درصد است که نسبت به سایر روش‌ها بیش از ۵۰ درصد کمتر می‌باشد.

جدول ۲. میانگین خطا

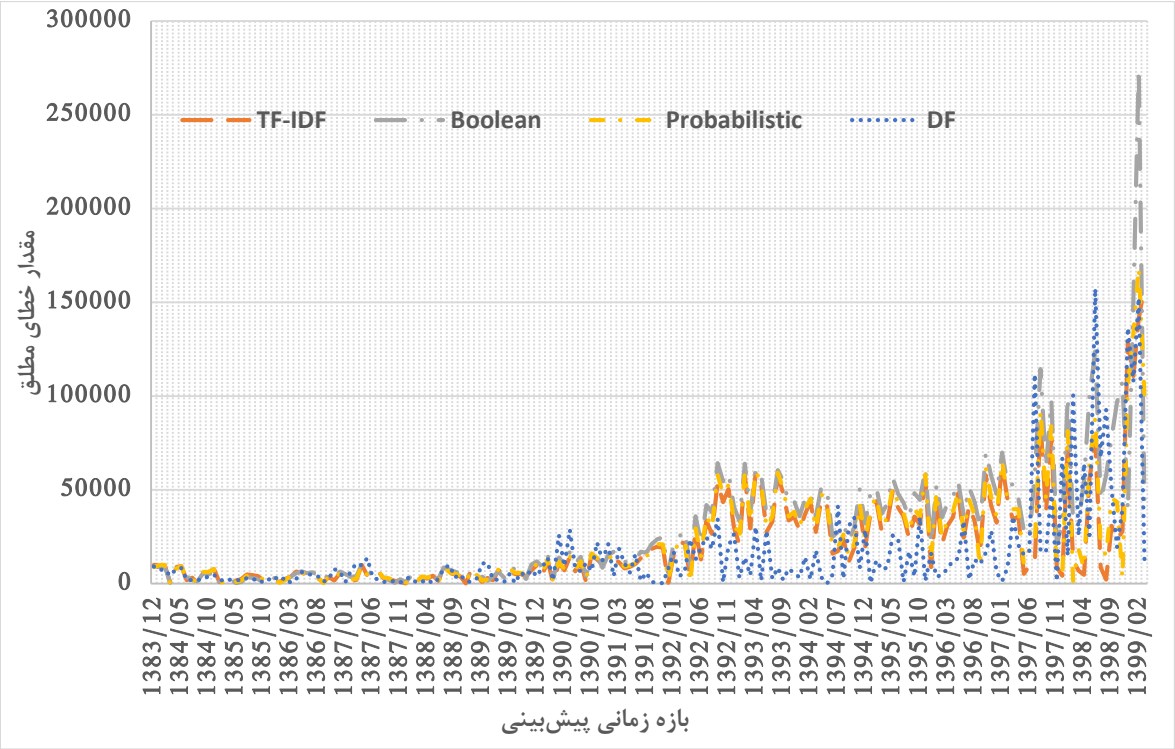
Error	TF-IDF	Boolean	Probabilistic	DF
MAE	21309	28532	22500	**15169**
MAPE	36.01	43.82	38.05	**26.99**
RMSE	32987	43535	34271	**27360**



شکل ۵. پیش‌بینی روند شاخص کل بورس اوراق بهادار تهران توسط روش‌های مورد مقایسه و روش پیشنهادی (DF)

به عنوان مثال سهام بانک ملت در ابتدای سال ۱۳۹۰ برابر با ۵۵۴ و در انتهای همان سال برابر با ۴۹۲ بوده است که بالاترین قیمت آن در همان سال به ۶۴۷ و کمترین قیمت آن به ۴۹۲ رسیده است. اما مقدار شاخص کل در ابتدای سال ۱۳۹۰ برابر با ۲۳۷۵۶ و در انتهای همان سال برابر با ۲۵۹۰۵ بوده است که بالاترین مقدار آن به ۲۷۰۹۸ و کمترین مقدار آن به ۲۳۷۵۶ رسیده است. تفاضل بین کمینه و بیشینه قیمت سهام بانک ملت برابر با ۱۵۵ بوده است که مقدار آن برای شاخص کل برابر با ۳۳۴۲ رقم خورده است، که حاکی از این است که نوسان شاخص کل ۲۱/۵ برابر نوسان سهام بانک ملت است [۳۶]. در کل بازه زمانی آزمایشی، تفاضل بین کمینه و بیشینه شاخص کل تقریبا ۲ میلیون واحد بوده که این مقدار ۱۳۱ برابر میانگین خطای مطلق روش پیشنهادی می‌باشد. برای دقت جهت‌گیری با استفاده از روش پیشنهادی مقدار ۶۴ و ۴۱ درصد به ترتیب صعودی و نزولی به دست آمده است. شوماخر و چن [۱] در پژوهش خود با استفاده از داده تاریخچه‌ای و متنی، مقدار میانگین ۵۸ درصد را برای دقت جهت‌گیری به دست آورده‌اند. روش پیشنهادی ما و

روش شوماخر و چن [۱] کمتر از ۲۰ درصد تفاوت دارند در حالی که در روش پیشنهادی ما صرفا از داده متنی برای انجام پیش‌بینی استفاده شده است ولی شوماخر و چن روش وزن‌دهی کلمات را با روش‌های یادگیری ماشین مانند بردار ماشین پشتیبان ترکیب کرده‌اند (برای ادامه این پژوهش، ترکیب روش پیشنهادی ما با روش‌های سری زمانی و یا هوش مصنوعی در دستور کار قرار گرفته است). در شکل ۶ خطای محاسبه شده برای رویکرد مبتنی بر مقدار DF و روش‌های مورد مقایسه نشان داده شده است. در این شکل، نقاطِ رویِ منحنی‌ها متناسب با میزان خطای متناظر، از محور افقی فاصله گرفته‌اند. کمترین مقدار خطای به دست آمده از رویکرد مبتنی بر مقدار DF مربوط به شهریور ماه سال ۱۳۸۸ و برابر با ۱۸,۳۵ می‌باشد.



شکل ۶. خطای مطلق برای روش‌های مورد مقایسه و روش پیشنهادی (DF)

خطای حاصل از پیش‌بینی شاخص کل در بازه زمانی ۹۱ تا ۹۵ توسط راعی و همکاران [۴۲] که از تکنیک‌های داده کاوی و داده‌های سری زمانی بهره برده‌اند کمتر از ۱ درصد اعلام شده است. در اینجا فاکتورهایی تاثیرگذار وجود دارند مانند بازه زمانی مورد مطالعه که ما از سال ۸۴ تا سال ۹۹ را پوشش داده‌ایم ولی راعی و همکاران فقط از سال ۹۱ تا ۹۵ را مورد مطالعه قرار داده‌اند. لازم به ذکر است که شاخص کل در ابتدای سال ۹۱ مقدار ۲۶۲۸۰ واحد بوده و در شهریور ۹۵ به ۷۶۴۵۰ واحد رسیده که حدودا ۳ برابر شده است این در حالی است که شاخص کل در ابتدای سال ۸۴ مقدار ۱۲۷۰۲ واحد بوده و در تیرماه ۹۹ به مقدار ۱۹۱۶۱۹۴ واحد رسیده که حدودا ۱۵۰ برابر شده است. مسلما با چنین دامنه‌ای از تغییرات، پیش‌بینی شاخص کل در بازه ۸۴ تا ۹۹ مشکل‌تر و دارای خطای بسیار بیشتری نسبت به بازه ۹۱ تا ۹۵ خواهد بود.

ما با کاهش دادن بازه زمانی مورد مطالعه به محدوده سال ۹۱ تا ۹۵ (مشابه بازه زمانی که راعی و همکاران [۴۲] درنظر گرفته‌اند) به مقدار خطای ۹ درصد می‌رسیم که نسبت به مقدار خطا برای کل بازه مورد پیش‌بینی ۶۵ درصد کاهش دارد. به همین ترتیب، فاکتورهای موثر دیگری نیز همچون پنجره زمانی پیش‌بینی (که در پژوهش راعی و همکاران ۲۰ روزه و در روش ما ۳۰ روزه است) وجود دارند. اما مهمترین فاکتور که نقش کاملا تعیین کننده‌ای دارد استفاده از داده‌های سری زمانی و تکنیک‌های داده کاوی است که خطای پیش‌بینی را به شدت کاهش می‌دهند. اگر چه روش پیشنهادی ذاتا رویکردی برای بهبود روش وزن‌دهی در متن

کاوی مالی می‌باشد و در آن صرفا از داده‌های متنی برای ارزیابی رویکرد استفاده شده است، با این حال، نسبت به روش‌های داده کاوی خطای بسیار پایین و قابل قبولی دارد. از آنجا که روش‌های جدید در پیش‌بینی‌های مالی ترکیبی از تکنیک‌های متن کاوی و داده کاوی می‌باشند، برای بخش کاوش متن مالی رویکرد پیشنهادی می‌تواند روش کارآمدتری نسبت به روش‌های رایج به شمار آمده و در رویکردهای ترکیبی مورد استفاده قرار گیرد.

۶- نتیجه‌گیری

نرخ تکرار کلمات مثبت و منفی در اسناد خبری بانک مرکزی در بازه نه ساله اخیر تقریبا یکسان بوده و پیش‌بینی شاخص کل صرفا با استفاده از آن ممکن نیست. در این پژوهش، روشی برای وزن‌دهی کلمات ارائه شده و سپس بر اساس آن، پیش‌بینی گرایش شاخص کل در بلندمدت انجام شد. ارزیابی روش پیشنهادی نشان داد که مقدار میانگین درصد خطای مطلق و میانگین بایاس دقت جهت‌گیری نوسانات شاخص به ترتیب برابر با ۲۶/۹۹ و ۲۳/۴ درصد است. این نتایج نشان داد که خبرهای منتشر شده در وبسایت بانک مرکزی ایران با روند تغییرات شاخص کل بورس اوراق بهادار تهران مرتبط است و بنابراین بر اساس محتوای آنها می‌توان شاخص کل را در بلندمدت پیش‌بینی کرد.

در پایان، سه رویکرد پیشنهادی برای ادامه این تحقیق معرفی می‌گردند: پیشنهاد اول این است که برای کاهش مقدار خطای مطلق از روش‌های سری‌های زمانی در کنار رویکرد پیشنهادی استفاده شود. پیشنهاد دیگر این است که به غیر از خبرهای مالی، خبرهای سایر گروه‌ها (نظیر سیاسی و اجتماعی) نیز جمع‌آوری و مورد تحلیل قرار گیرند. در انتها نیز پیشنهاد می‌گردد که روش‌های هوشمند برای انتخاب بهترین ویژگی‌ها آزموده شوند و نتایج مورد تحلیل قرار گیرد.

مراجع

[1] R. P. Schumaker and H. Chen, "Textual analysis of stock market prediction using breaking financial news: The AZFin text system," ACM Transactions on Information Systems (TOIS), vol. 27, p. 12, 2009.

[2] A. Kloptchenko, T. Eklund, J. Karlsson, B. Back, H. Vanharanta, and A. Visa, "Combining data and text mining techniques for analysing financial reports," Intelligent systems in accounting, finance and management, vol. 12, pp. 29-41, 2004.

[3] C. Robertson, S. Geva, and R. C. Wolff, "Can the Content of Public News be used to Forecast Abnormal Stock Market Behaviour?," in Seventh IEEE International Conference on Data Mining, ICDM 2007, pp. 637-642.

[4] E. F. Fama, "The behavior of stock-market prices," The journal of Business, vol. 38, pp. 34-105, 1965.

[5] S. Chopra and P. Meindl, Supply Chain Managemnt Strategy, planning and operation. 3rd Edition, Pearson Prentice Hall, ISBN: 0-13-208608-5, 2007.

[6] P. Falinouss, "Stock trend prediction using news articles: a text mining approach," M.S. thesis, Lulea university of thechnology, Lulea, Sweden, 2007.

[7] K. G. Aase, "Text Mining of News Articles for Stock Price Predictions," M.S. thesis, Norwegian University of Science and Technology, Trondheim, Norway, 2011.

[8] M. Arias, A. Arratia, and R. Xuriguera, "Forecasting with Twitter Data," ACM Transactions on Intelligent Systems Technology, 5, 1, Article 8 (January 2014), 24 pages.

[9] M. A. Mittermayer, "Forecasting intraday stock price trends with text mining techniques," in Proceedings of the 37th Annual Hawaii International Conference on System Sciences, Big Island, 0-7695-2056-1, IEEE, 5-8 Jan, 2004.

[10] M. Butler and V. Kešelj, "Financial Forecasting Using Character N-Gram Analysis and Readability Scores of Annual Reports," In Advances in Artificial Intelligence (pp. 39-51). Springer Berlin Heidelberg, 2009.

[11] B. G. Malkiel, "A Random Walk Down Wall Street: The Time-Tested Strategy for Successful Investing," WW Norton & Company, New York, 1973.

[12] C. Robertson, S. Geva, and R. C. Wolff, "Can the Content of Public News be used to Forecast Abnormal Stock Market Behaviour?," in Seventh IEEE International Conference on Data Mining, ICDM 2007, pp. 637-642.

[13] S. Bacher and H. Stuckenschmidt, "Mining Unstructured Financial News to Forecast Intraday Stock Price Movements," M.S. thesis, University of Mannheim, Mannheim, Germany, Oct, 2012.

[14] Kartick Gupta, Rajabrata Banerjee, "Does OPEC news sentiment influence stock returns of energy firms in the United States?" Energy Economics, vol. 77, pp. 34–45, 2019.

[15] George Guan-Ru Wu, Tony Chieh-Tse Hou, Jin-Lung Lin, "Can economic news predict Taiwan stock market returns?", Asia Pacific Management Review, vol. 24, pp. 54-59, 2019.

[16] Yu-Chen Wei, Yang-Cheng Lu, Jen-Nan Chen, Yen-Ju Hsu, "Informativeness of the market news sentiment in the Taiwan stock market", North American Journal of Economics and Finance, vol. 39, PP. 158-181, 2017.

[17] S. Feuerriegel, J. Gordon, "Long-term stock index forecasting based on text mining of regulatory disclosures," Decision Support Systems, vol. 112, pp. 88-97, 2018.
[18] R. Ren, D. D. Wu, and T. Liu, "Forecasting stock market movement direction using sentiment analysis and support vector machine," IEEE Systems Journal, vol. 13, no. 1, pp. 760-770, 2019.
[19] A.S. Ab. Rahman, S. Abdul-Rahman, and S. Mutalib, "Mining textual terms for stock market prediction analysis using financial news,"in Mohamed A., Berry M., Yap B. (eds) Soft Computing in Data Science. SCDS 2017. Communications in Computer and Information Science, vol. 788. Springer, Singapore, 2017.
[20] H. Naderi Semiromi, S. Lessmann, and Wiebke Peters, "News will tell: Forecasting foreign exchange rates based on news story events in the economy calendar," The North American Journal of Economics and Finance vol. 52, 101-181, 2020.
[21] L. Yu, S. Wang, and K. Lai, "A rough-set-refined text mining approach for crude oil market tendency forecasting," International Journal of Knowledge and Systems Sciences, vol. 2, pp. 33-46, 2005.
[22] R. Luss and A. d'Aspremont, "Predicting abnormal returns from news using text classification," Quantitative Finance, arXiv:0809.2792v3, 2009.
[23] M. V. Pinto and K. Asnani, "Stock Price Prediction Using Quotes and Financial News," International Journal of Soft Computing, vol. 1, Issue 5, November 2011.
[24] M. R. Amin-Naseri and E. A. Gharacheh, "A hybrid artificial intelligence approach to monthly forecasting of crude oil price time series," in The Proceedings of the 10th International Conference on Engineering Applications of Neural Networks, CEUR-WS284, 2007, pp. 160-167.
[25] W. Antweiler and M. Z. Frank, "Is all that talk just noise? The information content of internet stock message boards," The Journal of Finance, vol. 59, pp. 1259-1294, 2004.
[26] E. Guardia-Sebaoun, A. Rafrafi, V. Guigue, and P. Gallinari, "Cross-media sentiment classification and application to box-office forecasting," in Proceedings of the 10th Conference on Open Research Areas in Information Retrieval, 2013, pp. 201-208.
[27] X. Guo-Xiang, S. Ben-Chang, H. Yen-Bin, S. Po-Chih, and C. Kuo-Hao, "To Integrate Text Mining and Artificial Neural Network to Forecast Gold Futures Price," in International Conference on New Trends in Information and Service Science, NISS'09, 2009, pp. 1014-1020.
[28] S. Asur and B. A. Huberman, "Predicting the future with social media," In International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM (Vol. 1, pp. 492-499). IEEE.
[29] Stefan Feuerriegel, Julius Gordon, "News-based forecasts of macroeconomic indicators: A semantic path model for interpretable predictions", EuropeanJournal of Operational Research, vol. 272, pp. 162–175, 2019.
[30] Weiling Chen, Chai Kiat Yeo, Chiew Tong Lau, Bu Sung Lee, "Leveraging social media news to predict stock index movement using RNN-boost", Data & Knowledge Engineering, vol. 118, pp. 14-24, 2018.
[31] KiHwan Nam, NohY oon Seong, "Financial news-based stock movement prediction using causality analysis of influence in the Korean stock market", Decision Support Systems, Decision Support Systems, vol. 117, pp. 100-112, 2019.
[32] W. B. Yu, B. R. Lea, and B. Guruswamy, "A Theoretic Framework Integrating Text Mining and Energy Demand Forecasting," IJEBM, vol. 5, pp. 211-224, 2007.
[33] R. Baeza-Yates and B. Ribeiro-Neto, Modern information retrieval. Vol. 463. New York: ACM press, 1999.
[34] D. Thorleuchter and D. Van den Poel, "Predicting e-commerce company success by mining the text of its publicly-accessible website," Expert Systems with Applications, vol. 39, pp. 13026-13034, 2012.
[35] S. Mahfoud and G. Mani, Financial forecasting using genetic algorithms. Applied Artificial Intelligence, vol. 10, no. 6, 543-566, 1996.
[36] Tehran Stock Exchange, Available: http://www.tse.ir/market/Indices.aspx
[37] Cambridge University Press. Available: http://nlp.stanford.edu/IR-book/html/htmledition/tokenization-1.html
[38] Cambridge University Press. Available: http://nlp.stanford.edu/IR-book/html/htmledition/dropping-common-terms-stop-words-1.html
[39] X. Liang and R. C. Chen, "Mining Stock News in Cyberworid Based on Natural Language Processing and Neural Networks," in International Conference on Neural Networks and Brain, 2005, ICNN&B'05. Vol. 2, IEEE.
[40] G. Salton, A. Wong, and C. S. Yang, "A vector space model for automatic indexing," Communications of the ACM 18.11 (1975): 613-620.
[41] B. Drury, "A Text Mining System for Evaluating the Stock Market," Ph.D. Dissertation, Universities of Minho, Aveiro and Porto, 2009.
[۴۲] رضا راعی، علی نیک عهد قصیرائی و مصطفی حبیبی، «پیش بینی شاخص بورس اوراق بهادار تهران با ترکیب روشهای آنالیز مولفه‌های اصلی، رگرسیون بردارپشتیبان و حرکت تجمعی ذرات،» راهبرد مدیریت مالی، سال چهارم، ماه پانزدهم، ۱۳۹۵، صص: ۱-۲۳.
[۴۳] امیردایی، امیدعبادتی و کیوان برنا، «به‌کارگیری وب‌کاوی در پیش‌بینی جهت قیمت سهام گروه محصولات شیمیایی در

بورس اوراق بهادار»، فصل‌نامه علمی- پژوهشی فناوری اطلاعات و ارتباطات ایران، سال یازدهم، شماره‌های ۳۹ و ۴۰، بهار و تابستان ۱۳۹۸، صص: ۱۹ – ۴۸.

پیوست : لیست ویژگی‌ها (کلمات مورد بررسی) در اخبار مربوط به بورس

نقدینگی	اقتصادی	مسکن	پروژه	پایانی	عمومی	بانکداری
برنامه	صندوق	ارزی	پرداخت	فروش	خدمات	روابط
اوراق	تسهیلات	اعلام	شرکت	قانون	جلسه	موسسات
اجلاس	چین	نسبت	ارائه	صورت	توسعه	کار
نرخ	فعالیت	کالا	عامل	کمیسیون	ابلاغ	چک
سپرده	شاخص	لازم	معادل	استفاده	فرهنگی	نظارت
مردم	تعیین	مسافرتی	ضوابط	اسکناس	مصرفی	مناطق
شهری	قرض	مقررات	پولشویی	هیات	نظارتی	صادرات
اقتصاد	حضور	جهانی	بها	سیاستهای	مجموعه	دولتی
بانکی	اقدام	بخشنامه	شعب	نفت	جاری	درآمد

Textual analysis of central bank news in forecasting long-term trend of Tehran stock exchange index

Abstract

Financial markets have always been under influence of media news; therefore, text analysis of news is considered as an effective method of stock exchange forecasting. Research in this context has been conducted with the help of information retrieval techniques, in which high frequency words in a document that appeared sporadically in the whole corpus received higher weight than others. In contrast, the words which appeared in many news of a corpus, during a certain time, indicate the importance of an event. In our research, to address this contradiction, a new technique of assigning weight to influential words of news is presented. Financial news of Iran Central Bank (CBI) and actual data of Tehran Stock Exchange Index (TSEI) in the duration of 2005 to 2020 AD were utilized to evaluate the proposed method. The empirical results show 64% and 41% accuracy of trend prediction when TSEI moves upward and downward respectively and about 10% decreasing in Mean Absolute Error (MAE) to compare with prevalent techniques. While, the changes of the ratio between the number of positive and negative words in news does not offer predictive or analytical evidences, our results show that, there still exists a meaningful relationship between CBI news and TSEI fluctuations.

Keywords: Tehran Stock Exchange Index, long-term forecasting, textual analysis, word weighting.