دوره پانزدهم، شماره بهار و تابستان ۱۴۰۱

مجله فناوری اطلاعات در طراحی مهندسی

Information Technology in Engineering Design

http://ited.sinaweb.net

پیش‌بینی روند بازارهای مالی مبتنی بر مدلسازی مفاهیم نهفته‌ی اقتصادی در اسناد خبری

سعیده انبایی فریمانی[۱] مجید وفایی جهان*[۲] امین میلانی فرد[۳] سیدرضا کامل طباخ[۴]

(۱) گروه مهندسی کامپیوتر، واحد مشهد، دانشگاه آزاد اسلامی، مشهد، ایران

(۲) گروه مهندسی کامپیوتر، واحد مشهد، دانشگاه آزاد اسلامی، مشهد، ایران*

(۳) گروه علوم کامپیوتر، دانشگاه نیویورک، ونکور، کانادا

(۴) گروه مهندسی کامپیوتر، واحد مشهد، دانشگاه آزاد اسلامی، مشهد، ایران

(تاریخ دریافت :۱۴۰۱/۰۷/۲۷ تاریخ پذیرش: ۱۴۰۱/۱۰/۰۷)

چکیده

انتشار اخبار سیاسی، فرهنگی و اجتماعی، باعث تلاطم و بعضا ناهنجاری‌هایی در بازارهای مالی می‌شود، لذا پیش‌بینی روند بازار مبتنی بر اخبار اهمیت زیادی دارد. روند صعودی یا نزول بازار با انتشار اخبار مهم دارای موضوع‌های مختلف سیاسی، اجتماعی و فرهنگی تغییر می‌کند. شیوه‌ی بازنمایی یک سند خبری چالشی اساسی در سازماندهی به داده‌های متنی در راستای پی‌بردن به ارتباط موضوعی میان اسناد خبری است. تا کنون اغلب روش‌های پیشگو، بدون در نظر گرفتن ارتباط موضوعی میان اسناد خبری به پیش‌بینی بازار بر اساس اخبار پرداخته‌اند و اغلب از روش تجمیعی از واژه‌ها برای برداری سازی اسناد استفاده کرده‌اند در حالیکه این روش ارتباط معنایی پنهان میان واژه‌ها را در نظر نمی‌گیرد. این تحقیق با در نظر گرفتن ارتباط موضوعی اسناد خبری مبتنی بر الگوریتم مدلسازی مفاهیم معنایی پنهان اقتصادی روشی برای پیش‌بینی روند بازار مبتنی بر اخبار ارائه داده است. برای بازنمایی موضوعی اسناد، ابتدا پیکره‌ای از اخبار اقتصادی مرتبط با بازار فارکس ساخته شد و سپس با استفاده از روش جاسازی واژه‌ها، برای هر واژه برداری که قادر به بازنمایی ارتباط معنایی و نحوی آن با سایر واژه‌هاست ساخته شد و پس از خوشه‌بندی بردارهای حاصل تعدادی مفهوم معنایی پنهان اقتصادی استخراج و در نهایت بازنمایی سند بر اساس توزیع واژه‌هایش روی این مفاهیم معنایی محاسبه گردید. بهترین زمان برای بررسی تاثیر خبر بر روند بازار آزموده شد و در نهایت از دسته‌بند XGBoost به عنوان روش پیشگو استفاده شد. نتایج آزمایش‌ها نشان می‌دهد که صحت پیش‌بینی روند برابر ۷۹٪ درصد بوده و نسبت به سایر روشهای پایه در مدلسازی عنوان به ترتیب در معیارهای AUC ، صحت و F۱ به میزان ۲۴٪، ۱۲٪ و ۳۶٪ بهبود داشته است، درحالی‌که از هیچ‌گونه ویژگی دیگری نظیر شاخص‌های مالی و تحلیل احساس، استفاده نشده است.

کلمات کلیدی: مدلسازی مفاهیم معنایی، مفاهیم پنهان اقتصادی، اخبار، پیش‌بینی بازارهای مالی، بازنمایی اسناد خبری

* عهده‌دار مکاتبات:

مجید وفایی‌جهان

نشانی: گروه مهندسی کامپیوتر، واحد مشهد، دانشگاه آزاد اسلامی، مشهد، ایران

پست الکترونیکی: *vafaeijahan@mshdiau.ac.ir*

۱-مقدمه

گسترش استفاده از اینترنت و رشد شبکه‌های اجتماعی و همچنین اهمیت ویژه‌ی اقتصاد بر رشدیافتگی ملت‌ها، مطالعه‌ی بر هم‌کنش آن‌ها همواره دارای اهمیت می‌باشد. تا قبل از سال ۱۹۶۵ تصور می‌شد سری‌های زمانی بورس رفتاری تصادفی و غیرقابل پیش‌بینی دارند، تا این که در این سال فاما[۱] نظریه‌ی بازار موثر[۲] را ارائه داد [۱]. او در این نظریه، تصادفی بودن رفتار سری‌های زمانی مالی[۳] را رد می‌کند و آن را تحت تاثیر اطلاعاتی می‌داند که در دسترس سرمایه‌گذاران قرار می‌گیرد. اقتصاد رفتاری[۴] به مطالعه‌ی رفتار سرمایه‌گذاران در پی وقوع رویدادهایی با موضوع‌های ملی، سیاسی، فرهنگی برای توجیه ناهنجاری-های بازار می‌پردازد [۲, ۳]. مطالعه‌ی رفتار سرمایه‌گذاران می‌تواند در راستای تدوین سیستم‌های پیشگو و پشتیبان تصمیم مطلوب باشد.

از این رو تاکنون پژوهش‌های زیادی با استفاده از روش‌های متن‌کاوی[۵] سعی بر سازماندهی متون بدون ساختار وبی و تحلیل احساس اسناد خبری برای ارائه‌ی سیستم‌های پیشگوی مالی داشته‌اند [۴-۶] ، در حالیکه به اطلاعات پنهان میان اخبار با موضوع‌های مشابه توجهی نداشته‌اند. اخبار با موضوع‌های مختلف، میزان تاثیر متفاوتی بر سرمایه‌گذاران دارند. به عنوان مثال، یک خبر با موضوع سیاسی راجع به جنگ میزان تاثیر متفاوتی با یک گزارش اقتصادی بانک جهانی دارد، لذا باید بازنمایی اسناد به نحوی انجام شود که موضوع در شیوه‌ی بازنمایی منعکس گردد و در نهایت پیش‌بینی بازار بر اساس میزان تشابه موضوعی که اسناد با یکدیگر دارند، انجام گیرد. واژه‌ها در کنار هم قرار می‌گیرند و متون با موضوع‌های متفاوت را می‌سازند و آن‌چه به یک متن مفهوم می‌بخشد ارتباط معنایی و نحوی است که این واژه‌ها در کنار هم تولید می‌کنند. بسیاری از روش‌های پیشگو در بازارهای مالی با ارائه‌ی یک روش انتخاب ویژگی از روش متداول تجمیعی از واژه‌ها۶[۷] برای بازنمایی اسناد استفاده کرده‌اند[۸-۱۲]، در حالیکه این روش ارتباط معنایی که میان واژه‌ها وجود دارد را در نظر نمی‌گیرد و با چالش ابعاد بالا و فضای خلوت ویژگی‌ها مواجه است. دسته‌ای دیگر از روش‌ها با استفاده از هستی‌شناسی‌ها۷ و ابزارهای تعیین نقش نحوی واژه‌ها، سعی در مدلسازی ارتباط معنایی میان واژه‌ها و انتخاب ویژگی بر این اساس داشته‌اند[۶, ۱۳] در حالیکه با افزایش حجم داده دو چالش ابعاد بالای ویژگی‌ها و همچنین وجود کلماتی با معانی مختلف در حوزه‌های مختلف در هستی‌شناسی‌ها وجود دارد [۱۴]. یکی از راه حل‌های غلبه بر این چالش، مدلسازی مفاهیم معنایی مبتنی بر ارتباط معنایی و نحوی[۸] میان واژه‌هاست[۱۵]. در روش‌های متداول مدلسازی عنوان نظیر LDA[16] و LSA [۱۷] توزیع واژه‌های هر سند حول تعدادی مفهوم پنهان به عنوان بازنمایی سند محاسبه می‌شود، در حالیکه روشLDA دارای پیچیدگی زمانی زیادی است.

---

۱ Fama

۲ Efficient Market Hypothesis

۳ Financial Market time series

۴ Behavioural Finance

۵ Text Mining

۶ یک روش متداول بازنمایی متن که در آن ماتریسی از اسناد-لغات تشکیل می‌شود و هر درایه در این ماتریس فرکانس تکرار آن واژه در سند را نشان می‌دهد.

۷ Ontology

۸ Semantic and Syntactic

در روش مدلسازی مفاهیم معنایی[۱] ارائه شده در ، ابتدا بازنمایی برداری واژه‌ها بر اساس ارتباط معنایی و مفهوم پنهانی که از کنار هم قرار گرفتن واژه‌ها ساخته شده است، توسط روش جاسازی واژه‌ها[۲][۱۸] تولید می‌شود. این روش، بازنمایی برداری واژه‌ها را بر اساس میزان بیشینه آنتروپی متقابل رخداد آن‌ها در یک پنجره‌ی همپوشان به طول d با استفاده از یک شبکه عصبی محاسبه می‌کند. هر چه مقدار d کوچکتر باشد بردارهای تولید شده مشابهت معنایی میان واژه‌ها را منعکس می‌کنند و هر چه d بزرگتر شود، بردارهای تولید شده روابط نحوی میان واژه‌ها را مدل می‌کنند. مزیت روش جاسازی واژه‌ها، در تولید بردارهایی مشابه برای واژه‌هایی است که اغلب در سند متنی در یک همسایگی از هم ظاهر می‌شوند. پس از ساخت فضای برداری جاسازی‌شده برای واژه‌ها، بردارهای تولید شده خوشه‌بندی[۳] می‌شوند و هر خوشه معادل یک مفهوم معنایی پنهان[۴] در نظر گرفته می‌شود. از این رو می‌توان بازنمایی سند متنی[۵] را مبتنی بر موضوعی که از کنار هم قرار گرفتن واژه‌ها در آن پدید آمده است، بر اساس فراوانی رخداد واژه‌های آن سند در هر کدام از مفاهیم پنهان ساخته شده، انجام داد. پژوهش پیش‌رو یک روش پیشگو در بازار مالی فارکس[۶] بر اساس اخبار مرتبط با جفت ارز مبدا ارائه داده است. در این روش با توجه به مسئله‌ی حجم زیاد داده‌های متنی خبری، ابتدا یک روش بازنمایی اسناد خبری مبتنی بر مدلسازی مفاهیم معنایی پنهان در اخبار اقتصادی استفاده شده و سپس بهترین زمان برای بررسی تاثیر خبر بر روند بازار بر اساس زمان انتشار و میزان تغییر قیمت بستن معاملات جفت ارز مبدا[۷]، آزموده شد. به عنوان یک دسته‌بند پایه[۸] از روش‌های ماشین بردار پشتیبان[۹]، XGBoost و جنگل تصادفی استفاده شد. نتایج ارزیابی‌ها نشان می‌دهد روش پیشنهادی بازنمایی سند خبری، نسبت به سایر روش‌ها، دقت پیش‌بینی را بهبود داده است. نوآوری‌های ارائه شده در این مقاله به شرح زیر است:

- بررسی بهترین زمان تاثیر خبر بر جفت‌ارز EUR/USD مورد معامله در بازار تبادل ارزهای خارجی.
- به‌کارگیری روش مدلسازی مفاهیم پنهان جهت بازنمایی موضوعی اخبار اقتصادی و بررسی تفسیر پذیری موضوعی اخبار مرتبط با بازار فارکس.
- انجام مطالعات فرسایشی در بررسی تاثیر عنوان و محتوای خبر بر صحت پیش‌بینی

بنابراین، ابتدا در بخش مرور ادبیات به مرور فنون بازنمایی اسناد خبری اقتصادی پرداخته شده است. سپس، در بخش ۳، روش پیشنهادی پژوهش تشریح گردید. در بخش ۴ و ۵ ارزیابی بیان شد. بخش ۶ به نتیجه‌گیری پرداخته است.

۱ Topic modeling

۲Word embedding

۳ Cluster

۴ Latent Semantic Concepts

۵ Document representation method

۶ بازار معاملات ارزهای خارجی(در این بازار معاملات بر اساس نسبت دو جفت ارز پایه به یکدیگر انجام می‌شود)

۷ Source currency pair

۸ Baseline Classifier

۹ Support Vector Machine(SVM)· یک طبقه بند که یک صفحه در فضا پیدا می‌کند بطوریکه فاصله‌ی نقاط هر دسته از این صفحه بیشینه باشد.

۲-پیشینه‌ی پژوهش

بر اساس نظریه‌ی بازار کارامد، اطلاعاتی که از طریق گروه‌های خبری[۱] و رسانه‌های اجتماعی[۲] در اختیار سرمایه‌گذاران قرار می‌گیرد، شامل گزارش‌های فنی بانک جهانی، اخبار، تحلیل‌های فنی بانک‌های کشورهای مختلف، توئیت‌هایی که روزانه در رسانه‌های اجتماعی نظیر توئیتر منتشر می‌شوند، بر رفتار سرمایه‌گذاران موثر می‌باشند. در طی سال‌های اخیر، روش‌های مختلفی با استفاده از فنونی نظیر واکشی اطلاعات[۳] و پردازش زبان طبیعی[۴] به بررسی تاثیر این اطلاعات بر سرمایه‌گذاران بورس و بازارهای مالی پرداخته‌اند [۱۹–۲۲]. فنونی مانند نظرکاوی و تحلیل احساس در راستای ارائه‌ی سیستم‌های پیشگو [۲۳–۲۵] روش‌های تحلیل رفتار سرمایه‌گذاران [۲۶–۲۸] همچنین سیستم‌های توصیه‌گر استراتژی‌های تجاری [۲۹–۳۲] در این حوزه به کار گرفته شده‌اند. اخبار، داده‌های شبکه‌های اجتماعی نظیر توییتر [۲۶]، آمار جستجوهای آنلاین در موتورهای جستجو [۳۳]، آمار مراجعه به صفحات ویکی پدیا [۳۴] و همچنین بوردهای تخصصی گفتگو در بورس[۵] ، انواع منابع داده‌ای هستند که از طریق اینترنت در اختیار سرمایه‌گذاران قرار می‌گیرند [۳۵]. در میان روش‌های بررسی شده بخش زیادی از مراجع با توجه ویژگی قابل اعتماد بودن اخبار نسبت به سایر منابع اطلاعاتی، تنها از تحلیل اخبار استفاده نموده‌اند [۳۶, ۳۷] و دسته‌ی کمی برای بهبود دقت پیش‌بینی از ترکیبی از چند منبع داده‌ای استفاده کرده‌اند [۳۳, ۳۸–۴۰] .

هر سند خبری از دو بخش عنوان خبر و محتوای خبر سازمان‌یافته است. عنوان خبر به عنوان هسته‌ی موضوعی آن، حاوی واژه‌هایی است که بیشتر بیانگر موضوع هستند و محتوا به نوعی شرحی بر موضوع آن سند خبری است. در بسیاری از روش‌های ارائه شده در حوزه‌ی تحلیل اخبار، اغلب عنوان خبر بررسی شده است[۶] [۴۰, ۴۱] در حالیکه بخشی از اطلاعات ارزشمند راجع به موضوع سند در محتوای خبر وجود دارد. در روش [۴] با محاسبه‌ی میزان ارتباط معنایی واژه‌های عنوان خبر و محتوای خبر، با استفاده از یک شبکه‌ی یادگیری عمیق، با هدف ارزش بخشیدن به واژه‌های موجود در محتوا و مرتبط با عنوان خبر، به استخراج ویژگی‌هایی از اسناد خبری پرداخته است. در صورتیکه در روش مدلسازی مفاهیم پنهان، واژه‌های عنوان و محتوای خبر بر اساس مفهوم یکسانی که از کنار هم قرار گرفتن آن‌ها در یک سند خبری ساخته شده است، مفاهیم معنایی را می‌سازند. در روش پیشنهادی [۱۵]، فنی برای بازنمایی یک سند متنی ارائه شده، که با استفاده از توزیع واژه‌های آن سند روی تعدادی مفهوم معنایی پنهان بردار تولید می‌شود، در حالیکه در آن واژه‌ها در عنوان و محتوای سند متنی یکسان در نظر گرفته شده‌اند. برای مطالعه‌ی تاثیر واژه‌های عنوان و محتوا در بازنمایی سند متنی، روش پیشنهادی در [۴۲] سعی دارد با بسط واژه‌های عنوان خبر با استفاده از یافتن واژه‌های مشابه با آن از طریق نمایش برداری جاسازی شده، از واژه‌های عنوان و محتوا در کنار هم به گونه‌ای استفاده نماید که هم چالش ابعاد بالای ویژگی‌ها غلبه شود و هم موضوع سند در بازنمایی آن منعکس گردد. بر اساس

---

۱ News Group

۲ Social Network

۳ Information Diffusion

۴ Natural Language Processing

۵ Sina Weibo, Stock tweet

۶ بافزایش تعداد واژه‌ها در روش‌های متداولی مانند تجمیعی از واژه‌ها، تعداد ویژگی‌ها بسیار زیاد می‌شود و به همین دلیل در اغلب پژوهش‌ها تحلیل تنها بر واژه‌های عنوان خبر انجام گرفته است.

نتایج اعلام شده در [۴۲] در برتری روش ارائه شده مدلسازی مفاهیم پنهان اقتصادی بر سایر روش‌های مدلسازی عنوان، در این مقاله نیز از روش ارائه شده نویسنده در[۴۲] استفاده می‌شود.

تکنیک جاسازی واژه‌ها به دنبال بازنمایی برداری برای هر واژه است به طوریکه در بازنمایی برداری، نقش نحوی و معنایی آن واژه بر اساس احتمال رخداد متقابل آن در یک همسایگی از سایر واژه‌ها، منعکس شود. این روش در سال ۲۰۱۳ توسط میکولو با عنوان بردار برای هر واژه[۱] با پشتیبانی شرکت گوگل[۲] ارائه شد [۱۸]. در این روش برای هر واژه، بر اساس آنتروپی متقابل[۳] آن با سایر واژه‌های همسایه‌اش، با استفاده از یک شبکه عصبی، بازنمایی برداری جاسازی شده تولید می‌شود. هر چه مقدار آنتروپی متقابل دو واژه بیشتر باشد برداری با شباهت بیشتر برای آن دو واژه تولید می‌شود. بنابراین در زمینه‌ای که پیکره‌ی متنی از آن شکل یافته، این دو واژه بیشتر در کنار هم ظاهر شده‌اند و در کنار هم یک مفهوم یکسان را گزارش می‌دهند. نقطه‌ی قوت این روش، در تولید بردارهایی با میزان مشابهت زیاد برای واژه‌هایی است که اغلب در یک همسایگی از هم در پیکره ظاهر شده‌اند. تکنیک جاسازی واژه‌ها به دنبال بازنمایی برداری برای هر واژه است به طوریکه در بازنمایی برداری، نقش نحوی و معنایی آن واژه براساس احتمال رخداد متقابل آن در یک همسایگی از سایر واژه‌ها، منعکس شود. در صورتی که V را مجموعه تمام واژه‌های موجود در اسناد خبری پیکره‌ی D در نظر گرفته شود، روش مبتنی بر شبکه عصبی skip-gram [۱۸]، به دنبال پیش‌بینی واژه‌های موجود در یک همسایگی به طول k از واژه‌ی ورودی شبکه به صورت بدون ناظر و محاسبه‌ی بازنمایی برداری جاسازی شده‌ی واژه‌ها در فضای m بعدی $\Re^m$ به فرم $\overrightarrow{w_i}$، می‌باشد. در این صورت از رابطه‌ی ۱ برای محاسبه‌ی احتمال رخداد هر واژه در همسایگی از بردار متناظر واژه‌ی ورودی به شبکه‌ی عصبی استفاده می‌شود. در این رابطه احتمال رخداد کلمه‌ی $w_{c.j}$ در یک همسایگی به طول d از کلمه‌ی w_i بر اساس آنتروپی متقابل آن‌دو محاسبه می‌شود و در نهایت بردار جاسازی شده برای هر واژه‌ی w_i به فرم $\overrightarrow{w_i}$ محاسبه می‌شود.

$$p(w_{c.j}|w_i) = \frac{\exp(w_{c.j}^T \cdot w_i)}{\sum_i \exp(w_j^T \cdot w_i)} \quad (۱)$$

بازار تبادل ارزهای خارجی[۴] که به اختصار آن را فارکس۵ می‌نامند، یکی از انواع بازارهای مالی است که در آن سرمایه‌گذاران به معامله پول بر اساس انواع جفت ارزها می‌پردازند. این بازار نیز مانند سایر بازارهای مالی تحت تاثیر اخبار واکنش نشان می‌دهد [۴۳]. در این بازار هر معامله بر اساس نسبت دو ارز خارجی به یکدیگر انجام می‌شود. ارزهایی مانند دلار، یورو، ین از این جمله هستند. در این بازار معامله بر اساس بازه‌های زمانی۶ یک دقیقه، ۱۵ دقیقه، ۳۰ دقیقه، ۴ ساعتی، روز، هفته و ماه انجام

۱ Word2vec

۲ Google

۳ Cross Entropy معیاری پایه‌ای در نظریه‌ی اطلاعات و مبتنی بر احتمال شرطی که به عنوان یک کاربرد از آن می‌توان میزان میزان تنوع رخداد دو واژه در یک پیکره‌ی متنی را سنجید.

۴ Foreign Exchange Market

۵ FOREX

۶ Time frame

می‌گیرد. شاخص‌های پایه‌ای شامل قیمت باز[۱]، قیمت بستن[۲]، حجم معاملات، قیمت فروش و قیمت خرید می‌باشند که بر اساس مبنای زمانی محاسبه می‌شوند. روش پیشنهادی [۱۲] جزء اولین کارها با رویکرد متن کاوی در بورس فارکس می باشد. آن ها اخبار مربوط به هر جفت ارز را بر اساس مهر زمانی از دو منبع Marketwatch.com و google RSS reader API جمع آوری کرده اند، تمرکز خود را بر جفت ارز EUR/ USD قرار داده است، سپس پیش بینی بازگشت در بورس به صورت کوتاه مدت در بازه های زمانی ۲ ساعتی انجام گرفته است. در نهایت از دسته بند SVM جهت پیش بینی بازگشت در بورس کرده است.

۳- روش پیشنهادی

در این بخش به بیان مراحل مختلف روش پیشنهادی پرداخته شده است. روش پیشنهادی مورد مطالعه در پژوهش، شامل مراحل ساخت فضای برداری جاسازی شده‌ی واژه‌ها، مدلسازی مفاهیم معنایی پنهان اقتصادی، برداری سازی اسناد، یافتن بهترین زمان بررسی تاثیر خبر بر بازار و در نهایت پیش‌بینی بازار مبتنی بر اخبار با استفاده از ترکیبی از سه دسته‌بند ماشین بردار پشتیبان ، جنگل تصادفی و XGBoost می‌باشد. نمودار شکل ۱ شمایی از روش‌شناسی پژوهش را نشان می‌دهد.

۳-۱. مدلسازی مفاهیم معنایی پنهان اقتصادی

در روش پیشنهادی، ابتدا پیکره‌ای از اخبار اقتصادی مرتبط با بازار فارکس تشکیل شد، سپس مراحل پیش‌پردازش نظیر حذف اعداد و کلمات توقف، یکسان سازی ریشه‌ی افعال انجام گردید و در نهایت، با استفاده از کتابخانه‌ی Gensim در پایتون در محیط Spyder 3.02 نمایش برداری برای تمام واژه‌ها در پیکره استخراج شد (بلاک ۱ از شکل شماره ۱). پس از استخراج بازنمایی برداری برای واژه‌ها مبتنی بر روش جاسازی واژه‌ها، می‌توان آن‌ها را بر اساس مفهوم پنهانی که در کنار هم گزارش می‌دهند، دسته‌بندی[۳] کرد [۴۴]. برای این منظور، با خوشه‌بندی بردارهای جاسازی شده، می‌توان واژه‌هایی که مفهوم مشترکی را در کنار هم می‌سازند، در یک دسته قرار داد. این واژه‌ها دارای نمایش برداری با میزان مشابهت زیادی می‌باشند. این شباهت از این رو ایجاد شده است که این دو واژه، اغلب در سند خبری در کنار هم ظاهر می‌شده‌اند و دارای آنتروپی متقابل بیشتری می‌باشند. الگوریتم k-means [۴۵] به عنوان یک الگوریتم ساده و پایه‌ای بدون ناظر در خوشه‌بندی بردارهای جاسازی شده،برای تولید مفاهیم معنایی مطرح است[۱۵, ۴۴]. در روش پیشنهادی برای تولید مفاهیم پنهان اقتصادی نیز از الگوریتم K-means[۴] استفاده شده است، بطوریکه بردارهای ساخته‌شده $\vec{w_i} \in \Re^m$، به عنوان ورودی این الگوریتم، در نظر گرفته شدند و خروجی مجموعه K تا مفوم پنهان به فرم $C^K = \{c_1, \dots, c_k\}$ است و هر خوشه به فرم $\mathbf{c_i} = \{\overrightarrow{w_1}. \dots . \overrightarrow{w_n}\}$ شامل تعدادی

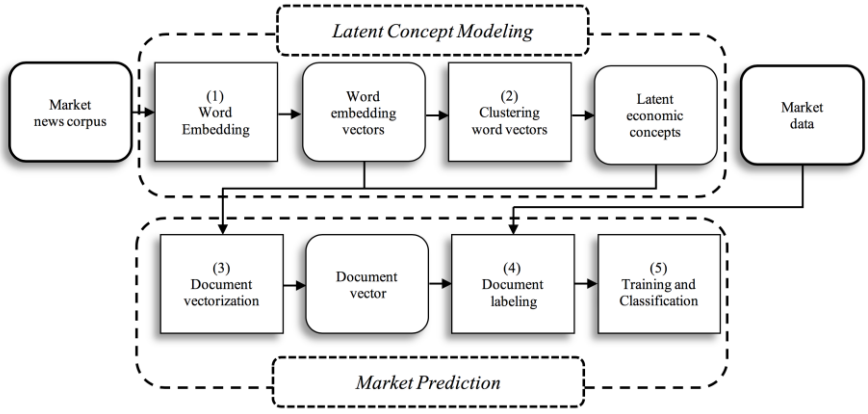
۱ قیمتی که معامله در ابتدای بازه‌ی زمانی مورد معامله با آن شروع می‌شود.

۲ قیمتی که معامله در انتهای بازه‌ی زمانی مورد معامله با آن تمام می‌شود.

۳ classification

۴ یک الگوریتم خوشه‌بندی متداول که بر اساس فاصله‌ی اقلیدسی نقاط از یکدیگر در فضا آنها را به صورت بدون ناظر خوشه‌بندی می‌کند.

واژه با زمینه مشترک می‌باشد و هر لذا می‌توان هر خوشه را نماینده‌ی یک مفهوم معنایی پنهان در نظر گرفت (بلاک شماره ۲ از شکل ۱).



شکل ۱- فلوچارت روش پیشنهادی

۳-۲. بازنمایی اسناد

بازنمایی یک سند متنی به عنوان یک کار زیربنایی در دسته‌بندی متن[۱] مطرح است [۴۶]. شیوه‌ی بازنمایی با توجه به فضای گسترده‌ی ویژگی‌های متنی، چالشی اساسی در سازماندهی به متن است. مدلسازی مفاهیم معنایی یکی از فنون کاهش ابعاد ویژگی‌ها می‌باشد [۴۷]. یکی از روش‌هایی که می‌توان از مدلسازی مفاهیم معنایی پنهان بهره جست، محاسبه‌ی بازنمایی برداری اسناد خبری بر اساس توزیع واژه‌های آن در مفاهیم پنهان است [۱۵]. واژه‌های عنوان و متن خبر در کنار هم موضوع سند را بر اساس یک ترکیب از واژه‌های خوشه‌بندی شده در مفاهیم معنایی پنهان، می‌سازند. این شیوه‌ی بازنمایی سبب می‌شود در عین حال که ابعاد ویژگی‌ها نسبت به روش متداول تجمیعی از واژه‌ها کم شود، موضوع هم در بازنمایی منعکس گردد و اسناد با موضوع مشابه، بازنمایی برداری مشابهی پیدا کنند. در حقیقت، می‌توان بیان کرد اسنادی با موضوعی مشابه، دارای توزیع واژه‌ی مشابهی روی مفاهیم پنهان هستند؛ و لذا این شیوه‌ی سازماندهی متن، به عنوان یک گام مهم در درک ارتباط موضوعی میان اسناد خبری مطرح می‌باشد. شبکه کد زیر، الگوریتم تولید بردار برای اسناد را بیان می‌کند. روش پیشنهادی برای بازنمایی

الگوریتم ۱. برداری سازی اسناد خبری Bag-of-Economic-Concepts (BoEC)[42]

ورودی: دنباله اخبار $N_l = \{title, content, timestamp\}$ ، مفاهیم پنهان اقتصادی $C^K = \{c_1, \dots, c_k\}$ ، فضای R^m بازنمایی بردار به واژه کلمات پیکره متنی، تعداد بیشترین تعداد واژه‌های مشابه برای هر توکن

خروجی: بردار $x_l \epsilon C^K$ حاوی بازنمایی برداری lامین خبر

/* مرحله۱ – پیش‌پردازش */

1- $titleKeywords \leftarrow \boldsymbol{Preprocess}(title)$
2- $descKeywords \leftarrow \boldsymbol{Preprocess}(content)$
3- $titleVectors \leftarrow titleKeywords\ corresponding\ vectors\ from\ R^m$

۱ Text classification

4- $extWords \leftarrow \boldsymbol{TOP\_N\_Similar}(\text{titleVectors}, \text{n}, R^m)$
5- $totalExtWords \leftarrow \boldsymbol{Append}(titleKeywords, descKeywords, extWords\ )$

/* مرحله۲ – برداری‌سازی */

6- for $i = 1$ $\boldsymbol{to}$ K $\boldsymbol{do}$
7- $x_l[i] \leftarrow number\ of\ tokens\ in\ totalExtWords\ \in\ \boldsymbol{c}_i$
8- *end for*
9- *return* x_l

معنایی، یک سند بر اساس فراوانی تکرار واژه‌هایش برداری اسناد خبری، شیوه‌ای مبتنی بر تجمیعی از مفاهیم معنایی پنهان روش‌های مبتنی بر مفاهیم در تعداد مشخصی خوشه برداری می‌شود[۱] (خط شماره ۷ از الگوریتم ۱).

در این الگوریتم با هدف تقویت موضوع سند، تعدادی از واژه‌ها که مشابهت معنایی زیادی با واژه‌های عنوان خبر دارند به مجموعه ویژگی‌ها افزوده شد. بدین صورت که، ابتدا برای کلمات عنوان خبر، تعدادی از کلمات با مشابهت مفهومی از فضای برداری جاسازی شده، انتخاب و به مجموعه ویژگی‌ها افزوده می‌گردد (خط شماره ۳ از الگوریتم ۱)و بعد، فراوانی تکرار هر واژه در مفاهیم معنایی به دو شیوه تجمیعی از مفاهیم معنایی بسط یافته[۲] و تجمیعی از مفاهیم معنایی بسط‌یافته نرمال‌شده[۳] محاسبه گردید[۴]. نرمال سازی بر اساس محاسبه‌ی فراوانی تکرار مفاهیم در اسناد انجام گردید[۵](بلاک شماره ۳ از شکل ۱). در بخش ارزیابی به تحلیل نتایج، زمانی که برداری سازی بر اساس فراوانی واژه‌های سند در مفاهیم انجام گیرد و در حالتی که از روش پیشنهادی تجمیعی از مفاهیم معنایی بسط یافته انجام پذیرد، بیان می‌گردد.

۳-۳. برچسب‌دهی به اخبار

پس از انتشار یک سند خبری ممکن است سرمایه‌گذاران در بازه‌ی زمانی مشخصی تحت تاثیر قرار بگیرند و در بازار یکی از دو وضعیت صعود و نزول رخ دهد و یا خبر دارای تاثیر چندانی بر جفت ارز مبدا نباشد و مقدار قیمت بستن[۶]، تقریبا بدون تغییر باقی بماند. بنابراین مسئله‌ی پیش‌بینی بازار مبتنی بر اخبار، به صورت یک مسئله‌ی دسته‌بندی چند کلاسه فرموله می‌شود [۱۲, ۴۸]. به منظور برچسب‌دهی اسناد خبری، تاکنون مطالعات مختلفی بر بازه‌ی زمانی برای بررسی تاثیر خبر انجام گرفته است. محققین در تحقیق‌هایی نظیر [۱۲, ۲۳, ۴۹] معتقد هستند، واکنش سرمایه‌گذاران در بازه‌ی زمانی یک ساعت پس از انتشار خبر، در بازار قابل بررسی است. در این تحقیق مبنای زمانی Δtرا در نظر گرفتیم و بهترین مقدار آن را می‌یابیم. از این رو، شیوه‌ی برچسب‌دهی به هر خبر، بر اساس تغییر قیمت بستن در بازه‌ی زمانی Δtمدت قبل از انتشار خبر و Δtمدت بعد از

۱ Concept Frequency (CF)

۲ Extended Concept Frequency(ECF)

۳ Extended Concept Frequency Inverse Document Frequency(ECF-IDF)

۴ Extended Concept Frequency(ECF)

۵ Concepts inverse Document Frequency(IDF)

۶ Close price

انتشار خبر انجام می‌گیرد. بر همین اساس، به محاسبه‌ی میزان تغییر قیمت بستن معاملات جفت ارز EUR/USD (close price)[1] بر اساس زمان انتشار خبر و بازه‌ی زمانی Δt، بر اساس رابطه‌ی (۲) پرداخته می‌شود.

$$label = \begin{cases} \boldsymbol{Up} & if \quad close_{\,timestamp\ of\ news\ +\Delta t} - close_{\,timestamp\ of\ news\ -\Delta t} > th \\ \boldsymbol{Down} & if \quad close_{\,timestamp\ of\ news\ +\Delta t} - close_{\,timestamp\ of\ news\ -\Delta t} < th \\ \boldsymbol{Neutral} & if \quad close_{\,timestamp\ of\ news\ +\Delta t} - close_{\,timestamp\ of\ news\ -\Delta t} = th \end{cases} \qquad (۲)$$

جدول ۱ یک نمونه خبر را نشان می‌دهد. تمامی اخبار از یک خبرگذاری جمع‌آوری شده‌اند و مهر زمانی بر اساس زمان استاندارد گرینویچ می‌باشد. یادآور می‌شود که بروکرها اطلاعات شاخص‌های مالی مربوط به جفت ارزها را در بازه زمانی متفاوتی از یک دقیقه تا روز در اختیار قرار می‌دهند. به عنوان مثال با فرض $\Delta t = 1\ hour$ ، برای برچسب‌دهی به خبری که در ساعت ۸:۵۰ دقیقه روز 2018.08.09 منتشر شده است، میزان تغییر قیمت بستن جفت ارز EUR/USD در ساعت ۷:۵۰ دقیقه نسبت به ساعت ۹:۵۰ دقیقه محاسبه می‌شود. در صورتی که این مقدار از حد آستانه بیشتر باشد، این خبر دارای برچسب صعود خواهد بود و در صورتی که از حد آستانه کمتر باشد، به این خبر برچسب نزول داده می‌شود و در غیر این صورت برچسب خبر، بدون تغییر در نظر گرفته می‌گردد. در روزهای تعطیل خبرگزاری خبری منتشر نمی‌کند و شاخص‌ها نیز با توجه به تعطیلی بازار بدون تغییر باقی می‌مانند. لذا برای برچسب‌دهی به خبری که در ابتدای روز دوشنبه منتشر می‌شود، میزان تغییر قیمت در ساعات پایانی روز جمعه گذشته نسبت به Δt مدت پس از انتشار خبر محاسبه خواهد شد.

جدول ۱ – یک نمونه خبر مرتبط با EUR/USD

Title	Content	timestamp	pair
EUR/USD pulls back on Friday, still heads for highest weekly close since June	The Euro opens the new week with the bears firmly in control. Long weekends will be evaporating the early week already-thin volume offering.' The EUR/USD is trading into 1.1510 ahead of the European continent … …Monday morning market open,	Mon, 08 Oct 2018 04:10:50 Z	EUR/USD

۳-۴. آموزش دسته‌بند[2]

با توجه به این نکته که یک مجموعه داده‌ی اخبار برداری شده و دارای برچسب است، تنها مرحله باقی‌مانده آموزش مدل پیشگو با استفاده از الگوریتم‌های یادگیری ماشین خواهد بود. در اینجا از الگوریتم ماشین بردار پشتیبان، روش‌های ترکیبی جنگل

۱ EUR/USD جفت ارز مبدا همان جفت ارز مورد مطالعه در این پژوهش است.

۲ classifier

تصادفی و XGBoost برای پیش‌بینی بازار استفاده شده است [۳۷, ۵۰]. برای آموزش دسته‌بند، پس از برداری سازی اسناد و برچسب‌دهی به آن‌ها، ابتدا داده‌ها درهم‌ریزی می‌شوند تا ترتیب انتشار اخبار تغییر کند، سپس از ٪۸۰ مجموعه داده برای آموزش دسته‌بند استفاده شد. ٪۲۰ باقیمانده به عنوان مجموعه داده‌ی آزمون استفاده شدند. در بخش ارزیابی همچنین به بررسی تاثیر تعداد نمونه‌های آموزشی و افزایش بازه‌ی دیتاست اخبار نیز پرداخته خواهد شد.

۴- شبیه‌سازی و تنظیم پارامترهای مدل

در این بخش جزئیات مربوط به مجموعه داده و ویژگی‌های آماری آن ارائه می‌شود و به بیان شیوه‌ی تنظیم پارامترهای مختلف روش پیشنهادی پرداخته شده است. در دسته‌بندی چندکلاسه، معیارهایی نظیر صحت[۱]، دقت[۲] و میزان خطای recall و معیار F1 به عنوان معیارهای ارزیابی مطرح هستند [۵۱, ۵۲]. از این رو، در این بخش به بررسی و ارزیابی پارامترها بر اساس مقادیر این معیارها پرداخته شده است.

۴-۱. مجموعه داده و ویژگی‌های آماری

مجموعه داده‌ی اخبار مرتبط با بازار فارکس از طریق یک اسکرپر[۳] از وب سایت www.Fxstreet.com از تاریخ ۲۸ آگوست ۲۰۱۸ تا کنون به صورت آنلاین در حال دریافت است. در روزهای تعطیل خبرگزاری خبری منتشر نمی‌کند و در سایر روزها به طور متوسط ۲۵۰ خبر به صورت شبانه‌روزی(با توجه به ماهیت شبانه‌روزی بودن بازار فارکس) منتشر می‌شود. در بین اخبار برخی خبرها دارای برچسب تعیین کننده جفت ارز مرتبط با آن خبر هستند و برخی خبرها برچسب ندارند. جدول ۲ شاخص‌های مالی نظیر قیمت باز، بستن، خرید و فروش و حجم معاملات رابرای جفت ارز EUR/USD نشان می‌دهد. در بازار فارکس معاملات بر اساس نسبت دو جفت ارز به یکدیگر انجام می‌شود. ارزهایی نظیر GBP, JPY,EUR,USD از این جمله هستند. در مجموعه داده‌ی جمع‌آوری شده، اخبار دارای برچسبی هستند که مشخص می‌کند، خبر منتشر شده به کدام جفت ارز ارتباط دارد. از این رو که جفت‌ارز EUR/USD یک جفت ارز پایه در بازار فارکس می‌باشد، در این پژوهش به عنوان مطالعه‌ی موردی انتخاب شد.

جدول۲ – چند نمونه شاخص‌های مالی جفت ارز EUR/USD در برش زمانی یک دقیقه

Date&time	Ask	Bid	Close	Open	Trade Volume
2018.05.15 01:59	1.19294	1.19296	1.19285	1.19286	39
2018.05.15 2:00	1.19286	1.19297	1.19286	,1.19296	35

۱ accuracy

۲ precision

۳ Scraper

جدول ۳- مشخصات مجموعه داده‌ی اخبار

Corpus	Total news document	Words count	Size
ForexNews Dataset	40,341	63,612	50MB
EUR/USD News	2,860	13,638	5MB

در این‌تحقیق، از این رو که هدف پیش‌بینی جفت ارز EUR/USD بوده است، از اسناد خبری با برچسب EUR/USD به عنوان مجموعه داده‌ی آموزش استفاده شده است. این مجموعه داده دارای ۲۸۶۰ سند خبری از تاریخ ۸ اکتبر ۲۰۱۸ تا تاریخ۷ مارچ ۲۰۲۰ می‌باشد. برای هر سند خبری ویژگی‌های عنوان، محتوای خبر (شامل متن خبر و گزارشی از خلاصه‌ی خبر)، مهر زمانی، جفت ارز مرتبط با خبر وجود دارد. برای بهبود دقت این روش، ابتدا مراحل پیش‌پردازش روی پیکره انجام شده است. مراحل پیش‌پردازش شامل حذف اعداد، کلمات توقف، عبارات باقاعده‌ای نظیر URL و تگ‌های IMG و یکسان سازی ریشه‌ی افعال با استفاده از کتابخانه‌های NLTK , RE در پایتون انجام شده‌است.

۴-۲. تنظیم پارامترهای روش پیشنهادی

به منظور بررسی تاثیر پارامترهای مختلف بر نتایج مدل پیشگو، پارامترهای زیر به ازای دسته‌بند ماشین بردار پشتیبان آزموده شده‌اند.

- **طول همسایگی**: یکی از پارامترها در آموزش مدل بردار‌به‌واژه، طول همسایگی میان واژه‌هاست که بر اساس نتایج آزمایش‌ها بهترین مقدار F1 زمانی بدست آمد که طول همسایگی ۳ در نظر گرفته شده بود. معیارF1 به دلیل وجود برچسب‌های نامتوازن در دو دسته‌ی صعود و نزول و بی تاثیر انتخاب شده است.
- **بعد بردار بازنمایی واژه‌ها**: همچنین بهترین نتایج زمانی بدست آمد که بعد بازنمایی جاسازی‌شده واژه‌ها ۱۰۰ در نظر گرفته شد و واژه‌هایی با فراوان تکرار کمتر از ۳ از مجموعه کلمات کلیدی حذف شدند.
- **تعداد مفاهیم پنهان**: مطالعه تعداد مفاهیم استخراج شده به عنوان یک زمینه‌ی باز تحقیقاتی در طی سال‌های اخیر مطرح است [۵۳]. بهترین مقدار F1 در روش پیشنهادی را به ازای تعداد $k = 210$ مفهوم پنهان بدست آمده است.
- **حجم پیکره‌ی متن**: به منظور بررسی تاثیر حجم پیکره‌ی متنی بر نتایج مدل پیشگو، در فاز آموزش مدل بردار به واژه آزمایشی ترتیب دادیم. در حالت اول تنها از اخبار مرتبط با جفت ارز EUR/USD و در حالت دوم از تمامی خبرهای جمع‌آوری شده، برای ساخت مفاهیم پنهان استفاده شد. بر اساس این داده‌ها بهترین مقدار F1 در حالت اول ۷۳٪ و در حالت دوم ۷۶٪ کسب شد، که در نهایت بازنمایی برداری واژه‌ها بر اساس مدل آموزش دیده روی کل پیکره اخبار محاسبه گردید و در فاز بررسی تاثیر خبر بر بازار و پیش‌بینی تنها از اخبار مرتبط با جفت‌ارز EUR/USD استفاده شد. حجم پیکره‌ی متنی که برای آموزش روش جاسازی واژه‌ها و استخراج مفاهیم پنهان اقتصادی استفاده شد، یک معیار مهم در

محاسبه‌ی بهتر بازنمایی برداری جاسازی شده‌ی واژه‌هاست. هر چه حجم پیکره بیشتر باشد، بازنمایی برداری دقیق‌تر خواهد بود. ارزیابی نتایج آزمایش‌ها نشان داد، با افزوده شدن بر حجم پیکره‌ی متنی، معیار صحت افزایش می‌یابددر الگوریتم بردار‌به‌واژه هر چه تعداد دفعات ظاهر شدن کلمات با مفهوم مشابه در پیکره بیشتر باشد، صحت روش در تولید بازنمایی برداری مشابه بالاتر خواهد بود.

۵- تجزیه و تحلیل یافته‌ها

در این بخش ابتدا به بررسی تحلیل زمان تاثیر خبر بر بازار و انتخاب دسته‌بند پرداخته می‌شود. سپس مقایسه نتایج با روش‌های پایه بیان گردیده و در نهایت به بیان تفسیرپذیری مفاهیم و قدرت روش در مدلسازی ارتباط موضوعی میان اسناد مرتبط و غیر مرتبط پرداخته می‌شود.

۵-۱. تحلیل تاثیر خبر در زمان

در این پژوهش به اخبار بر اساس میزان تغییرات قیمت (رابطه ۲) در Δt مدت قبل و بعد از انتشار خبر برچسب داده شد. به منظور بررسی تاثیر مقدار Δt بر نتایج روش پیشنهادی، مقادیر مختلفی از ۵ دقیقه تا ۲ ساعت را آزموده و نتایج دسته‌بند بردار پشتیبان را به ازای هر دور از آزمایش و معیار F1 در جدول ۴ گزارش شده است. نتایج نشان می‌دهد که بهترین مقدار F1 در $\Delta t = ۶۰$ دقیقه کسب گردیده است. به عبارتی می‌توان نتیجه گرفت که در جفت‌ارز EUR/USD زمان تقریبی تاثیر خبر بر بازار حدود یک ساعت می‌باشد، لذا سایر آزمایش‌ها در آموزش مدل پیشگو بر مبنای یک ساعت انجام پذیرفته است.

جدول ۴ – نتایج بررسی صحت به ازای مقادیر مختلف Δt

Δt	Accuracy	F1 Macro
10	0.80	0.45
20	0.82	0.46
30	0.83	0.45
40	0.75	0.65
50	0.84	0.46
60	0.76	0.71
70	0.75	0.70
80	0.75	0.69
90	0.75	0.70
100	0.75	0.68
110	0.75	0.68

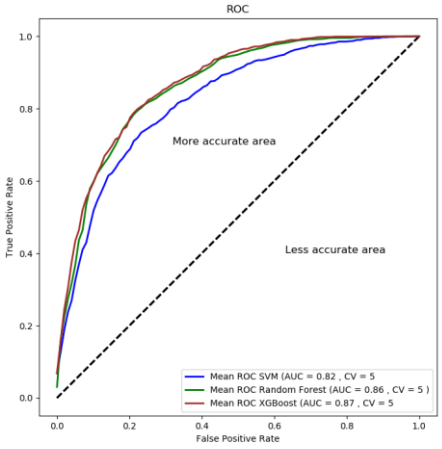
به منظور بررسی بیشتر تاثیر خبر بر بازار جفت ارز EUR/USD، آزمایش را در پنجره‌های زمانی مختلف نیز تکرار کردیم. بطوریکه در اولین آزمایش بازه‌ای کوتاه شامل ۳۵۰ خبر را در نظر گرفتیم و در هر دور از آزمایش ۳۵۰ خبر افزوده شد تا در نهایت کل مجموعه داده پوشش داده شد. نتایج در جدول ۵ بیان گردیده است. بهبود معیار F1 در فرایند آزمایش نشان از اهمیت در نظر گرفتن تاثیر اخبار بر پیش‌بینی بازارهای مالی دارد.

جدول ۵- بررسی تاثیر پنجره‌ی زمانی و تعداد اخبار بر نتایج روش پیشنهادی

Time interval	Number of news documents	F1 Macro
2018-10-08 to 2019-03-26	1,600	0.49
2018-10-08 to 2019-05-17	1,950	0.56
2018-10-08 to 2019-07-02	2,300	0.69
2018-10-08 to 2019-08-22	2,650	0.71
2018-10-08 to 2020-02-07	2,860	0.71

۵-۲. ارزیابی طبقه‌بند

به منظور انتخاب بهترین طبقه‌بند، روش پیشنهادی با اعمال بهترین تنظیمات بدست آمده از مراحل قبل آزموده شده است. برای این منظور، سه دسته‌بند ماشین بردار پشتیبان، جنگل تصادفی و XGBoost را با اعمال تنظیمات یکسان به ازای تمامی پارامترهای روش پیشنهادی، آزموده شده است. شکل ۲ نمودار معیار AUC/ROC را به ازای هر سه دسته‌بند نشان می‌دهد. همانطور که در جدول ۵ نشان داده شده است، بهترین نتیجه زمانی بدست آمده است که از طبقه‌بند XGBoost استفاده شده است.



شکل ۱- نمودار ROC به ازای طبقه‌بندهای مختلف

جدول ۶- معیارهای ارزیابی به ازای طبقه‌بندهای مختلف

Classifier	AUC	Accuracy	F1 Macro	F1 Micro
SVM	0.82	0.76	0.71	0.76
Random Forest	0.86	0.79	0.75	0.78
XGBoost	0.87	0.79	0.76	0.79

۵-۳. مطالعات فرسایشی

به منظور ارزیابی کارآمدی روش پیشنهادی، به مطالعات فرسایشی روش تجمیعی از مفاهیم پنهان اقتصادی بسط یافته بر مبنای تعداد واژه‌هایی که در بسط عنوان هر خبر استفاده شده است و بررسی تاثیر عنوان و محتوای خبر در بهبود دقت پرداخته شده است. در هر دور از آزمایش، دو حالت تنها عنوان خبر و حالتی که عنوان و محتوا به عنوان ویژگی استفاده شده باشد در نظر گرفته شده و برای هر کلمه‌ی کلیدی در عنوان خبر، تعداد ۱ تا ۵ واژه که بیشترین مشابهت معنایی را با آن داشته‌اند، به مجموعه ویژگی‌ها اضافه شده است. جدول ۵ نتایج مقایسه‌ی بهترین مقدار متوسط وزن‌دار[۱] F1-Score را به ازای حالات مختلف مطالعات فرسایشی نشان می‌دهد. بر این اساس، بهترین مقدار F1-Score در حالتی که از کلمات عنوان و محتوا به همراه ۷ واژه که بیشترین مشابهت کسینوسی با واژه‌های عنوان داشته‌اند بدست آمده است. لذا می توان به تاثیر کلمات موجود در محتوای خبر و اهمیت بسط موضوع خبر بر اساس کلمات عنوان خبر پی برد.

جدول ۵ - مقایسه‌ی F1Score متوسط وزن دار در مجموعه داده‌ی EUR/USD News Corpus

Model		N	Accuracy	F1 Macro	F1 Micro
BoEC-word2vec	Title	0	0.68	0.57	0.68
BoEC-word2vec	Title	3	0.66	0.51	0.66
BoEC-word2vec	Title	5	0.67	0.53	0.67
BoEC-word2vec	Title	7	0.67	0.52	0.67
BoEC-word2vec	Title	9	0.67	0.53	0.67
BoEC-word2vec	title and content	0	0.76	0.75	0.76
BoEC-word2vec	title and content	3	0.76	0.76	0.76
BoEC-word2vec	title and content	5	0.78	0.76	0.76
BoEC-word2vec	title and content	7	0.79	0.76	0.79
BoEC-word2vec	title and content	9	0.77	0.76	0.76

۱ Weighted average F1

چالش اول در پردازش داده‌های متنی، تعداد زیاد ویژگی‌ها و خلوت بودن فضای ویژگی‌ها در روش‌های مبتنی بر تجمیعی از واژه‌هاست. با توجه به این نکته که روش پیشنهادی، از جمله روش‌های مبتنی بر مفاهیم پنهان میان واژه‌هاست، لذا در روش ارائه شده با کم کردن ابعاد ویژگی‌ها و ساختن تعداد ۲۱۰ مفهوم پنهان، فضای ویژگی‌ها کمتر گردیده و در عین حال صحت بهبود پیدا کرده است. در بسیاری از روش‌های پیشگو مبتنی بر تحلیل اخبار، با هدف کاهش ابعاد ویژگی‌ها تنها از عنوان یک سند خبری استفاده شده است[۵, ۶, ۳۷, ۵۰, ۵۴, ۵۵]، در حالیکه بخش زیادی از اطلاعات مرتبط با موضوع خبر، در محتوا وجود دارد. در روش پیشنهادی با در نظر گرفتن این حقیقت که عنوان خبر به عنوان هسته‌ی موضوعی آن، حاوی واژه‌هایی است که بیشتر بیانگر موضوع هستند و محتوا به نوعی شرحی بر موضوع آن سند خبری است، با افزودن ۷ واژه‌ی مشابه با کلمات کلیدی در عنوان، علاوه بر عنوان بسط‌یافته‌ی خبر، از محتوا نیز استفاده شد، در حالیکه نتایج بهبود داشت. در روش‌های پیشنهادی [۴۰, ۵۶-۵۸] محققین با ارائه‌ی یک سیستم ترکیب اطلاعاتی[۱]، از تنسور و ترکیبی از چندین منبع اطلاعاتی استفاده شده است، در حالیکه در روش پیشنهادی تنها از خبر به استخراج ویژگی پرداخته شد و می‌توان به عنوان یک کار زیربنایی، بازنمایی ارائه شده را به صورت یک بعد از تنسور، در سیستم‌های اطلاعاتی ترکیبی افزود و دقت پیش‌بینی را افزایش داد.

۵-۴. مقایسه با روش‌های پایه

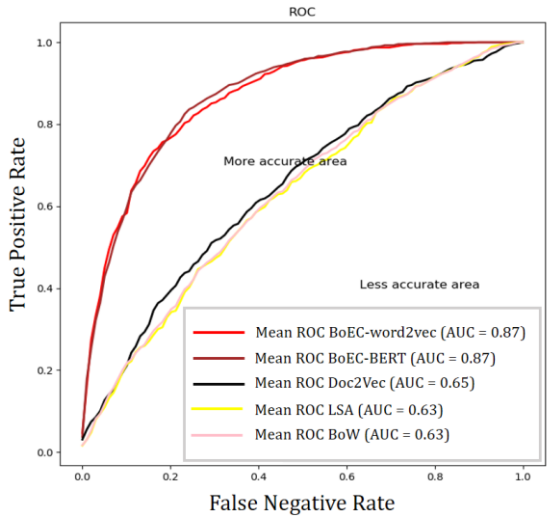
به منظور ارزیابی کارامدی روش پیشنهادی تجمیعی از مفاهیم پنهان اقتصادی در پیش‌بینی بازار مبتنی بر اخبار، این روش را با سایر روش‌های بازنمایی متن نظیر تجمیعی از واژه‌ها (BoW) و روش‌های مدلسازی عنوان (LDA, LSA)، و روش سندبه‌بردار مقایسه شده است. نتایج بیان شده در جدول ۶ و شکل ۳ برتری روش پیشنهادی را نسبت به سایر روش‌های مدلسازی عنوان نشان می‌دهد. روش‌های پایه‌ای بازنمایی متن که در این تحقیق با سایر روش‌ها مقایسه شده است به شرح زیر است:

- تجمیعی از واژه‌ها به روش TFIDF : در این روش بردار مربوط به هر سند بر اساس فراوانی تکرار واژه‌های هر سند نسبت به کل پیکره محاسبه می‌شود.
- مدلسازی تحلیل معنایی LSA: این روش [۱۷] یک تکنیک تجزیه ماتریس است. بردار هر خبر بر اساس تجزیه اجزای اصلی حول ۲۱۰ مفهوم پنهان محاسبه شده است.
- مدلسازی مفاهیم پنهان تخصیص دریکله LDA: این روش [۱۶] یک تکنیک تجزیه نامنفی ماتریس‌هاست که با استفاده از یک تحلیل بیزی سه سطحی بردار هر خبر حول تعداد ۲۱۰ مفهوم پنهان محاسبه می‌شود.
- مدل سند به بردار Doc2vec : این روش [59] یک تکنیک مبتنی بر شبکه عصبی برای بازنمایی برداری هر سند است که در اینجا بعد بازنمایی برداری هر سند ۲۱۰ در نظر گرفته شده است.

۱ Information Fusion

جدول ۶- مقایسه با روش‌های پایه

Model	AUC	Accuracy	F1 Macro	F1 Micro
BoW	0.63	0.67	0.40	0.52
LSA	0.63	0.70	0.40	0.54
Doc2vec	0.65	0.70	0.60	0.66
BoEC	0.87	0.79	0.76	0.79



شکل ۲- نمودار ROC/AUC روش پیشنهادی و روش‌های پایه

بر اساس نتایج بیان شده در شکل ۳، میزان سطح زیر نمودار به ازای دو پیاده‌سازی متفاوت از روش پیشنهادی نسبت به سایر روش‌های بازنمایی متن بیشتر است و نتایج کسب شده در روش پیشنهادی نسبت به سایر روش‌های مدلسازی عنوان قابل اعتمادتر است. نتایج آزمایش‌ها نشان می‌دهد که صحت پیش‌بینی روند برابر ٪۷۹ درصد بوده و نسبت به سایر روشهای پایه در مدلسازی عنوان به ترتیب در معیارهای AUC، صحت و F1 به میزان ٪۲۴، ٪۱۲ و ٪۳۶ بهبود داشته است، درحالی که از هیچ‌گونه ویژگی دیگری نظیر شاخص‌های مالی و تحلیل احساس، استفاده نشده است.

۵-۵. تفسیرپذیری روش پیشنهادی

در راستای ارزیابی شیوه‌ی سازماندهی به سند متنی ارائه شده در این پژوهش، و قدرت روش در بازنمایی موضوعی اسناد، در این بخش ابتدا به ارائه‌ی تفسیری از مفاهیم پنهان ساخته شده پرداخته شده است، سپس چند سند متنی با موضوع‌های مرتبط با ایران، برگزیت و جنگ را تفسیر خواهد شد. شیوه‌ی بازنمایی روش پیشنهادی، سبب می‌شود در عین حال که ابعاد ویژگی‌ها نسبت به روش متداول تجمیعی از واژه‌ها کم شود، موضوع هم در بازنمایی منعکس گردد و اسناد با موضوع مشابه، بازنمایی برداری مشابهی پیدا کنند. از این رو، در جدول ۷ برای تعداد ۲۰ مفهوم، تعداد کل واژه‌ها و چند نمونه از کلمات هر مفهوم بیان

شده است. وجود کلماتی نظیر Low, High, Open, pips, Trends, Levels, Pivot, Point در یک مفهوم، شهودی بر ادعای ما نسبت به قرار گرفتن کلمات مرتبط با یکدیگر و ساختن مفاهیم معنایی پنهان اقتصادی می‌باشد.

جدول۷- واژه‌های خوشه‌بندی شده در مفاهیم پنهان

#	Words
1	Hours, minutes, Monday, month, overnight, Thursday, today, week, yesterday
2	Low, High, Open, Last, pips, Trends, Levels, Pivot, Point, Pervious,
3	Moment, index, gold, Nikkei, metal, Asia, ,Jones, Street, equity, Dow, Nasdaq
4	Cycle, policy, interest, rat, Federal, Reserve, Fed, rate, hike, monetary, ECB, curve, Bank, Committee
5	Consumer, Confidence, Survey, MoM, YoY, Monetary, United, Kingdom, Japan, Australia, France
6	Analysis, Bulls, bear, break, trendline, breakout, uptrend, upper, candle, downtrend, Fib, descend
7	Conference, Commission, Reuters, official, White, President, Donald, Trump, Washington, Huawei
8	Wells, Danske, Rabobank, Nordea, Nomura, FXStreet, Bednarik, Commerzbank, Scotiabank, UOB
9	Octobers, disaster, pgrowth, momentumIn, roof, groundwork, homeowners, validity, underbelly ,fend
10	European, GBPUSD, UK, May, Italian, Brexit, GBP, Euro, vote, Pound, Sterling, EURGBP, British
11	Wake, space, dent, Australian, appetite, reaction, heavily, backdrop, upbeat, always, persistent,
12	USDJPY, USDCHF, cross, DXY, WTI, AUDUSD, NZDUSD, GBPJPY, USDCAD, EURJPY, YTD
13	Expansion, consistent, consumer, spend, increase, energy, production, job, earn, activity, industrial
14	Losses, worst, reverse, drop, weakest, lowest, decline, higher, lower, jump, highest, pick, hover,

15	eventually, safehaven, Loonie, assist, act, bout, capitalize, status, combination, build, upmove
16	Accept, possible, approve, leaders, Cabinet, proposals, Brussels, Brexiteers, Ireland, prepare, decide
17	Limit, challenge, need, find, period, still, stick, keep, cenario, rsquo, leave, intact, approach, suggest
18	Earlier, When, Oil, Market, Joint, OPEC, Ministerial, Monitoring, JMMC, compliance, original, group
19	Bearish, momentum, bias, technical, overbought, negative, indicators, oversold, slop, Momentum
20	Japanese, improve, Yen, assets, flow, tone, news, incoming, elevate, heighten, broad, light, softer

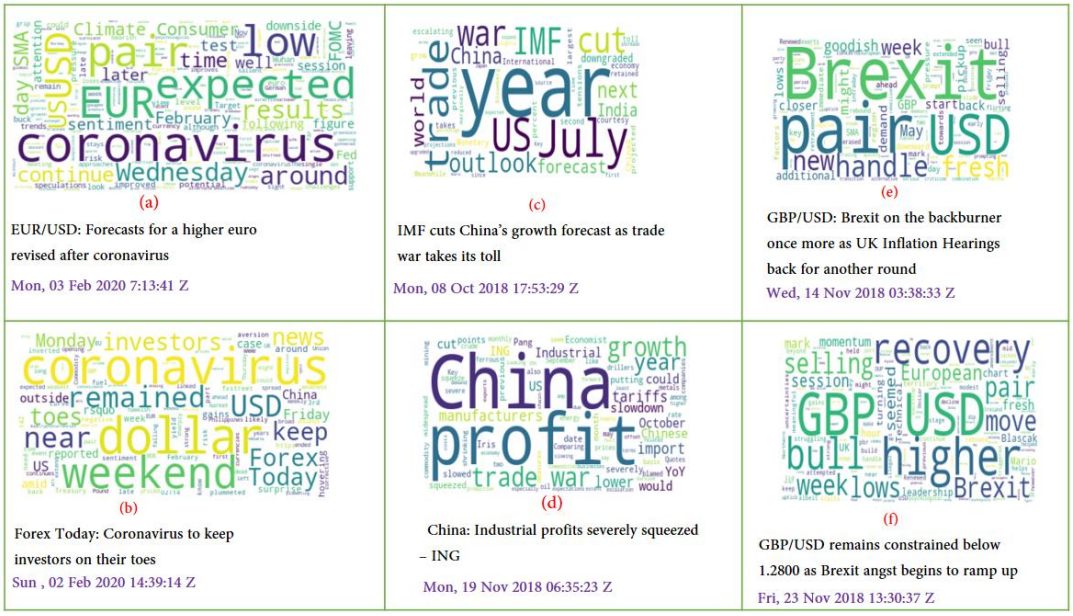
شکل ۴ چند نمونه خبر موجود در مجموعه داده را نشان می‌دهد. برای هر خبر، نمودار ابر، عنوان خبر و زمان انتشار در این شکل نشان داده شده است. همچنین بردار تولید شده توسط روش تجمیعی از مفاهیم بسط‌یافته به ازای ۵ واژه مشابه، در جدول ۸ بیان شده است. در شکل ۵ نمودار نقشه‌ی گرما[۱]، میزان شباهت کسینوسی این بردارها با هم را نشان می‌دهد. نمودارهای ابر کلمات[۲] دو خبر a,b در شکل ۴ مربوط به کرونا ویروس، خبرهای c,d مربوط به واژه‌ی جنگ تجاری چین و آمریکا و خبرهای e,f مربوط به مسئله‌ی برگزیت است. با توجه به بعد کم بردارهای تولید شده برای این خبرها، نمودار نقشه‌ی گرما، نشان می‌دهد بیشترین مشابهت کسینوسی بین خبرهایی با موضوع یکسان وجود دارد و میزان مشابهت خبرهایی با موضوع متفاوت کمتر از ۵۰ ٪ است. در روش پیشنهادی، بهترین صحت برای تعداد ۲۱۰ مفهوم پنهان بدست آمده، در حالیکه در این جدول بردارهای تولید شده بر اساس فراوانی تکرار واژه‌های اسناد در ۲۰ مفهوم پنهان گزارش شده است. با توجه به این حقیقت که اخبار با موضوع‌های مختلف، تاثیر متفاوتی بر بازار جفت‌ارزها دارند، این مثال شهودی بر ادعای ما مبنی بر بازنمایی موضوع اسناد در بردارهای تولید شده است و این شیوه‌ی بازنمایی بسط یافته، سبب بهبود دقت پیش‌بینی در روش پیشنهادی شده است.

۶. نتیجه‌گیری و پیشنهادهای آتی

در این پژوهش، روشی برای بازنمایی اسناد خبری و پیش‌بینی بازارهای مالی بر اساس میزان تشابه موضوعی اسناد، ارائه شد. در شیوه‌ی بازنمایی پیشنهادی، علاوه بر کاهش بعد ویژگی‌های متنی، ارتباط نحوی و معنایی میان واژه‌ها در نظر گرفته شد.

۱ heatmap

۲ worldcloud

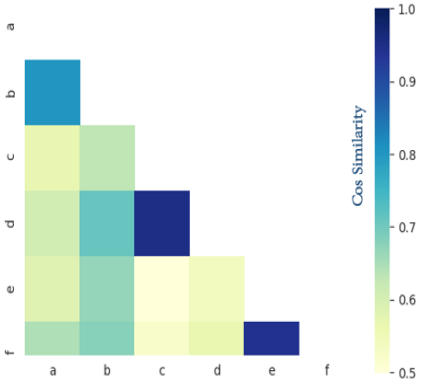


شکل ۴ – چند نمونه خبر در مجموعه داده‌ی Forex News Corpus. قسمت‌های a,b مربوط به کرونا ویروس، قسمت‌های c,d مربوط به جنگ تجاری چین و آمریکا و قسمت‌های e,f مربوط به واژه‌ی برگزیت، می‌باشند. برای هر خبر، عنوان و زمان انتشار نشان داده شده است[1].

هم‌چنین تعدادی مفهوم معنایی پنهان که از رخداد واژه‌های مرتبط در یک همسایگی از هم، شکل می‌یابند، ساخته شد و در نهایت هر سند خبری بر اساس فراوانی تکرار واژه‌های عنوان بسط یافته و محتوا، مبنی بر همین مفاهیم پنهان برداری شد . همین امر سبب بازنمایی موضوعی شد که کنار هم قرار گرفتن مفاهیم معنایی مختلف در سند، آن را منعکس می‌کند. در شیوه‌ی بازنمایی پیشنهادی، با در نظر گرفتن اطلاعات پنهان میان اخبار با موضوع‌های مشابه، با استفاده از دسته‌بند XGBoost برای پیش‌بینی روند در بازار فارکس، آزموده شد. مجموعه داده‌ای از اخبار مرتبط با این بازار جمع‌آوری گردید. نتایج ارزیابی‌ها نشان داد هنگامی که برچسب‌دهی به اخبار مرتبط به جفت ارز EUR/USD، بر اساس میزان تغییر قیمت بستن یک ساعت قبل و بعد از انتشار خبر انجام شد، بهترین صحت ٪۷۹ بدست آمد.

در یک سند خبری مالی، بخشی از اطلاعات ارزشمند راجع به موضوع سند در محتوای خبر وجود دارد، از این رو صحت روش پیشنهادی زمانی که تحلیل تنها متکی بر واژه‌های عنوان انجام شود و در حالتی که از عنوان و محتوا برای بازنمایی برداری سند استفاده شود، آزموده شد. بهبود صحت پیش‌بینی در حالت دوم نشان داد، سرمایه‌گذاران علاوه بر عنوان به بخش‌هایی از محتوا هم توجه می‌کنند و این ویژگی، از جمله مزیت‌های روش پیشنهادی است. در روش‌های پیشگو مالی مبتنی بر اخبار، استفاده از شاخص‌های مالی و تحلیل‌های فنی و همچنین فنون تحلیل متن نظیر تحلیل احساس و تعداد خبر بسیار متداول می‌باشد. در روش پیشنهادی، تنها اسناد خبری بازنمایی شده‌اند و از هیچ شاخص مالی و ویژگی اضافی دیگری استفاده نشده است. با توجه به بعد کم و بازنمایی موضوعی سند، گنجاندن این شیوه‌ی بازنمایی به عنوان یک تحلیل بنیادین پایه‌ای در کنار استفاده از تحلیل‌های فنی، سبب بهبود دقت پیش‌بینی بازارهای مالی می‌شود. به عنوان پیشنهاد آینده، می‌توان این پژوهش را از جنبه‌های مختلفی بررسی کرد. با توجه به ماهیت متغیر بازار فارکس و

شرایط متلاطم جفت‌ارز EUR/USD، استفاده از شاخص‌های مالی نظیر pivot pointها و سایر اندیکاتورها بر چگونگی تعیین روند قیمت در این جفت‌ارز و بهبود شیوه‌ی برچسب‌دهی به اخبار موثر خواهد بود. همچنین، می‌توان با اعمال یک روش شناسایی رویداد، اسناد مرتبط را دسته‌بندی نمود و تاثیر موضوع‌های مختلف را بر جفت‌ارزهای همبسته بررسی کرد. استفاده از فنون مبتنی بر یادگیری عمیق نظیر مدل زبانی برت و شبکه‌های کانولوشن و LSTM به عنوان مدل پیشگو نیز می‌تواند سبب بهبود دقت روش پیشنهادی گردد.



شکل ۵ - نمودار heatmap میزان شباهت کسینوسی خبرهای a تا f

جدول ۷ – بردار تولید شده برای اسناد a تا f تولید شده توسط روش تجمیعی از مفاهیم پنهان بسط یافته

a	[2.0, 16.0, 25.0, 3.0, 12.0, 81.0, 19.0, 5.0, 11.0, 14.0, 0.0, 19.0, 0.0, 3.0, 13.0, 2.0, 2.0, 4.0, 36.0, 11.0]
b	[0.0, 19.0, 11.0, 9.0, 9.0, 82.0, 3.0, 5.0, 15.0, 16.0, 3.0, 10.0, 0.0, 10.0, 14.0, 3.0, 4.0, 2.0, 36.0, 5.0]
c	[0.0, 4.0, 14.0, 1.0, 15.0, 0.0, 3.0, 19.0, 1.0, 3.0, 0.0, 5.0, 0.0, 3.0, 2.0, 2.0, 5.0, 0.0, 0.0, 6.0]
d	[0.0, 5.0, 18.0, 0.0, 21.0, 5.0, 6.0, 22.0, 0.0, 12.0, 0.0, 9.0, 3.0, 1.0, 0.0, 3.0, 3.0, 1.0, 0.0, 3.0]
e	[9.0, 13.0, 37.0, 18.0, 3.0, 7.0, 5.0, 2.0, 21.0, 15.0, 25.0, 27.0, 6.0, 9.0, 6.0, 4.0, 8.0, 2.0, 0.0, 7.0]
f	[13.0, 9.0, 38.0, 32.0, 2.0, 5.0, 16.0, 15.0, 25.0, 21.0, 33.0, 28.0, 4.0, 1.0, 7.0, 6.0, 1.0, 2.0, 0.0, 12.0]

مراجع

[1] Fama, E.F., The behavior of stock-market prices. The Journal of Business,, 1965. 38(1): p. 34–105.

[2] C. W. Lin, T., A Behavioral Framework for Securities Risk. 2012.

[3] Shiller, R.J., From efficient markets theory to behavioral finance. Journal of Economic Perspectives, 2003. 17(1): p. 83–104.

[4] Liu, Q., et al., Hierarchical Complementary Attention Network for Predicting Stock Price Movements with News, in Proceedings of the 27th ACM International Conference on Information and Knowledge Management. 2018, ACM: Torino, Italy. p. 1603-1606.

[5] Van de Kauter, M., D. Breesch, and V. Hoste, Fine-grained analysis of explicit and implicit sentiment in financial news articles. Expert Systems with Applications, 2015. 42(11): p. 4999-5010.
[6] Dang, M. and D. Duong. Improvement methods for stock market prediction using financial news articles. in 2016 3rd National Foundation for Science and Technology Development Conference on Information and Computer Science (NICS). 2016.
[7] Salton, G., Automatic text processing: the transformation, analysis, and retrieval of information by computer. 1989: Addison-Wesley Longman Publishing Co., Inc. 530.
[8] Wuthrich, B., et al. Daily stock market forecast from textual web data. in SMC'98 Conference Proceedings. 1998 IEEE International Conference on Systems, Man, and Cybernetics (Cat. No.98CH36218). 1998.
[9] Johan Bollen, H.M., Xiaojun Zeng, Twitter mood predicts the stock market. Journal of Computational Science, 2011. 2: p. 1-8.
[10] Li, X., et al., Improving stock market prediction by integrating both market news and stock prices, in Proceedings of the 22nd international conference on Database and expert systems applications - Volume Part II. 2011, Springer-Verlag: Toulouse, France. p. 279-293.
[11] Nizer, P.S.M. and J.C. Nievola, Predicting published news effect in the Brazilian stock market. Expert Systems with Applications, 2012. 39(12): p. 10674-10680.
[12] Khadjeh Nassirtoussi, A., et al., Text mining of news-headlines for FOREX market prediction: A Multi-layer Dimension Reduction Algorithm with semantics and sentiment. Expert Systems with Applications, 2015. 42(1): p. 306-324.
[13] Seifollahi, S. and M. Shajari, Word sense disambiguation application in sentiment analysis of news headlines: an applied approach to FOREX market prediction. Journal of Intelligent Information Systems, 2019. 52(1): p. 57-83.
[14] Krishnamoorthy, S., Sentiment analysis of financial news articles using performance indicators. Knowledge and Information Systems, 2018. 56(2): p. 373-394.
[15] Kim, H.K., H. Kim, and S. Cho, Bag-of-concepts: Comprehending document representation through clustering words in distributed representation. Neurocomputing, 2017. 266: p. 336-352.
[16] Blei, D.M., A.Y. Ng, and M.I. Jordan, Latent Dirichlet Allocation. Journal of Machine Learning Research 2003. 3: p. 993-1022.
[17] Landauer, T.K., P.W. Foltz, and D. Laham, An introduction to latent semantic analysis. Discourse processes, 1998. 25(2-3): p. 259-284.
[18] Mikolov, T., et al., Efficient Estimation of Word Representations in Vector Space. CoRR, 2013. abs/1301.3781: p. 1301-3781.
[19] Li, Q., et al., Web Media and Stock Markets : A Survey and Future Directions from a Big Data Perspective. IEEE Transactions on Knowledge and Data Engineering, 2018. 30(2): p. 381-399.
[20] Arman Khadjeh Nassirtoussi, T.Y.W., Saeed Reza Aghabozorgi, David Ngo Chek Ling, Text Mining for Market Prediction: A Systematic Review. Expert Systems with Applications, 2014. 41(16): p. 7653-7670.
[21] Agarwal, S., S. Kumar, and U. Goel, Stock market response to information diffusion through internet sources: A literature review. International Journal of Information Management, 2019. 45: p. 118-131.
[22] Kumar, B.S. and V. Ravi, A survey of the applications of text mining in financial domain. Knowledge-Based Systems, 2016. 114: p. 128-147.

[23] Romanov, V.P., et al., Fractal Model of Estimating News and Insider Influence on Market Volatility. Automatic Documentation and Mathematical Linguistics, 2007. 41(4): p. 141–149.
[24] El Ouadghiri, I. and R. Uctum, Jumps in equilibrium prices and asymmetric news in foreign exchange markets. Economic Modelling, 2016. 54: p. 218-234.
[25] Shi, Y., K.-Y. Ho, and W.-M. Liu, Public information arrival and stock return volatility: Evidence from news sentiment and Markov Regime-Switching Approach. International Review of Economics and Finance, 2016. 42.
[26] Nisar, T.M. and M. Yeung, Twitter as a tool for forecasting stock market movements: A short-window event study. The Journal of Finance and Data Science, 2018. 4(2): p. 101-119.
[27] Fang, L., H. Yu, and Y. Huang, The role of investor sentiment in the long-term correlation between U.S. stock and bond markets. International Review of Economics & Finance, 2018. 58: p. 127-139.
[28] Tausch, F. and M. Zumbuehl, Stability of risk attitudes and media coverage of economic news. Journal of Economic Behavior & Organization, 2018. 150: p. 295-310.
[29] Yang, S.Y., et al., Genetic programming optimization for a sentiment feedback strength based trading strategy. Neurocomputing, 2017. 264: p. 29-41.
[30] Sun, Y., M. Fang, and X. Wang, A novel stock recommendation system using Guba sentiment analysis. Personal and Ubiquitous Computing, 2018. 22(3): p. 575-587.
[31] Song, Q., A. Liu, and S.Y. Yang, Stock portfolio selection using learning-to-rank algorithms with news sentiment. Neurocomputing, 2017. 264: p. 20-28.
[32] Geva, T. and J. Zahavi, Empirical evaluation of an automated intraday stock recommendation system incorporating both market data and textual news. Decision Support Systems, 2014. 57: p. 212-223.
[33] Kaushal, A. and P. Chaudhary. News and events aware stock price forecasting technique. in 2017 International Conference on Big Data, IoT and Data Science (BID). 2017.
[34] Weng, B., M. A. Ahmed, and F. Megahed, Stock Market One-Day Ahead Movement Prediction Using Disparate Data Sources. Vol. 79. 2017.
[35] Farimani, S.A., et al., Investigating the informativeness of technical indicators and news sentiment in financial market price prediction. Knowle dge-Base d Systems, 2022.
[36] Gupta, K. and R. Banerjee, Does OPEC news sentiment influence stock returns of energy firms in the United States? Energy Economics, 2019. 77: p. 34-45.
[37] Long, W., L. Song, and Y. Tian, A new graphic kernel method of stock price trend prediction based on financial news semantic and structural similarity. Expert Systems with Applications, 2019. 118: p. 411-424.
[38] Zhang, G., L. Xu, and Y. Xue, Model and forecast stock market behavior integrating investor sentiment analysis and transaction data. Cluster Computing, 2017. 20(1): p. 789-803.
[39] Hajek, P. and A. Barushka, Integrating Sentiment Analysis and Topic Detection in Financial News for Stock Movement Prediction, in Proceedings of the 2nd International Conference on Business and Information Management. 2018, ACM: Barcelona, Spain. p. 158-162.
[40] Wang, H., S. Lu, and J. Zhao, Aggregating multiple types of complex data in stock market prediction: A model-independent framework. Knowledge-Based Systems, 2019. 164: p. 193-204.
[41] Shi, Y., W.-M. Liu, and K.-Y. Ho, Public news arrival and the idiosyncratic volatility puzzle. Journal of Empirical Finance, 2016. 37: p. 159-172.

[42] Farimani, S.A., et al. Leveraging Latent Economic Concepts and Sentiments in the News for Market Prediction. in 8th IEEE International Conference on Data Science and Advanced Analytics (DSAA). 2021. Portgual: IEEE.
[43] Égert, B. and E. Kočenda, The impact of macro news and central bank communication on emerging European forex markets. Economic Systems, 2014. 38(1): p. 73-88.
[44] Hu, L., et al., Adaptive online event detection in news streams. Knowledge-Based Systems, 2017. 138: p. 105-112.
[45] Shi, Z. Efficient online spherical k-means clustering. in Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005. 2005.
[46] Liu, B., Opinions, Sentiment, and Emotion in Text, ed. C.U. Press. 2015.
[47] Anbaee Farimani, S., H. Tabatabaee, and M. kaffashan kakhki, An Investigation into the Process of Organizing and Retrieving Web Texts Based on the Integration of Semantic Concepts In order to organize knowledge. IranDoc, 2019. 34(4): p. 1879-1904.
[48] Chen, W., et al. Stock market prediction using neural network through news on online social networks. in 2017 International Smart Cities Conference (ISC2). 2017.
[49] Jahan, M.V. and M. Akbarzadeh-Totonchi, From Local Search to Global Conclusions: Migrating Spin Glass-Based Distributed Portfolio Selection. IEEE Transactions on Evolutionary Computation, 2010. 14(4): p. 591-601.
[50] Shynkevich, Y., et al., Forecasting movements of health-care stock prices based on different categories of news articles using multiple kernel learning. Decision Support Systems, 2016. 85: p. 74-83.
[51] Jahan, M.V. and M.-R. Akbarzadeh-T, Composing local and global behaviors: Higher performance of spin glass based portfolio selection. Journal of Computational Science, 2012. 3(4): p. 238-245.
[52] Farimani, S.A. and M.V. Jahan. An HMM for online signature verification based on velocity and hand movement directions. in 2018 6th Iranian Joint Congress on Fuzzy and Intelligent Systems (CFIS). 2018.
[53] Koltcov, S., Application of Rényi and Tsallis entropies to topic modeling optimization. Physica A: Statistical Mechanics and its Applications, 2018. 512: p. 1192-1204.
[54] Ding, X., et al., Using Structured Events to Predict Stock Price Movement: An Empirical Investigation. 2014. 1415-1425.
[55] Jiang, C., et al., Analyzing market performance via social media: a case study of a banking industry crisis. Science China Information Sciences, 2014. 57(5): p. 1-18.
[56] Fernández Vilas, A., et al., Twitter permeability to financial events: an experiment towards a model for sensing irregularities. Multimedia Tools and Applications, 2019. 78(7): p. 9217-9245.
[57] Zhang, X., et al., Improving Stock Market Prediction via Heterogeneous Information Fusion. 2017.
[58] Zhang, W., et al., Quantifying the cross-correlations between online searches and Bitcoin market. Physica A: Statistical Mechanics and its Applications, 2018. 509: p. 657-672.
[59] Le, Q. and T. Mikolov, Distributed representations of sentences and documents, in Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32. 2014, JMLR.org: Beijing, China. p. II–1188–II–1196.