به نام خدا

تعریف پروژه خبرکاوی

روزانه اخبار بیشماری در تلکس خبرگزاری‌های مختلف قابل مشاهده می‌باشد، حجم این اخبار و گوناگونی آن‌ها به قدری است که از توان پردازش یک فرد یا حتی خود جامعه فراتر می‌باشد. یکی دیگر از معظلات این اخبار ضد و نقیض بودن بعضی از آن‌ها با بعضی دیگر می‌باشد. رایانه در این مسأله می‌تواند به کمک انسان‌ها بیاید. یکی از راه‌های متداول استفاده از سایت‌های خبرخوانی(دیده‌بان) می‌باشد به گونه‌ای که خبرها را از سایت‌های مختلف تجمیع می‌کنند و در یک ساختار مشخص به کاربر نمایش می‌دهند. این سایت‌ها از تکنیک‌هایی از جمله دسته‌بندی و خوشه‌بندی اخبار استفاده می‌کنند تا بتوانند به بهترین نحو ممکن اخبار را به کاربر خود نمایش دهند.

پروژه‌ی خبرکاوی می‌خواهد پای خود را فراتر از این ابزارها بگذارد و بتواند با استفاده از علم داده دانش برخواسته از اخبار را به نمایش بگذارد. این دانش با تجمیع تمامی اخبار و پردازش بر روی آن‌ها حاصل می‌شود و می‌تواند اطلاعات بسیار مفیدی را در اختیار کاربران قرار دهد. در ادامه به صورت موضوعی خدماتی که این پروژه می‌تواند در اختیار کاربران قرار دهد را می‌آوریم.

شناسایی انواع مختلف ارتباط بین اخبار یکی از این خدمات می‌باشد. هر دو خبر می‌تواند از طرق مختلف به یکدیگر مرتبط باشند. شباهت دو خبر، هم‌زمانی، دارا بودن تناقض خبری و ... می‌تواند انواعی از ارتباطات بین دو خبر باشند.

شناسایی یک رویداد خبری و تجمیع تمامی اخبار مرتبط با آن نیز یکی دیگر از خدماتی می‌باشد که در این پروژه در نظر گرفت شده است. رویدادهای خبری را می‌توان به صورت صریح یا ضمنی در متن اخبار مختلف مشاهده کرد. تکنیک‌های هوش مصنوعی می‌تواند به ما کمک کند تا اخباری که دارای اطلاعات جدیدی می‌باشند را به عنوان یک رویداد خبری جدید شناسایی کنند. پس از شناسایی رویدادهای خبری می‌توان تمامی اخبار مرتبط با آن رویداد خبری را تجمیع کرد و در اختیار کاربر قرار داد تا کاربر بتواند از ابعاد مختلف آن رویداد خبری مطلع شود. علاوه بر این با اضافه کردن بعد زمان به اخبار کاربر می‌تواند سیر اخبار و تغییرات آن رویداد خبری را متوجه شود.

یافتن ارتباط بین اخبار مختلف همچنین این امکان را می‌دهد تا نحوه‌ی انتشار یک خبر در خبرگزاری های مختلف نیز به دست آید. این کار کمک می‌کند مرجع یک خبر به خوبی پیدا شود.

تعریف پروژه

پروژه تحلیل خبر دارای چند قسمت کلی می‌باشد:

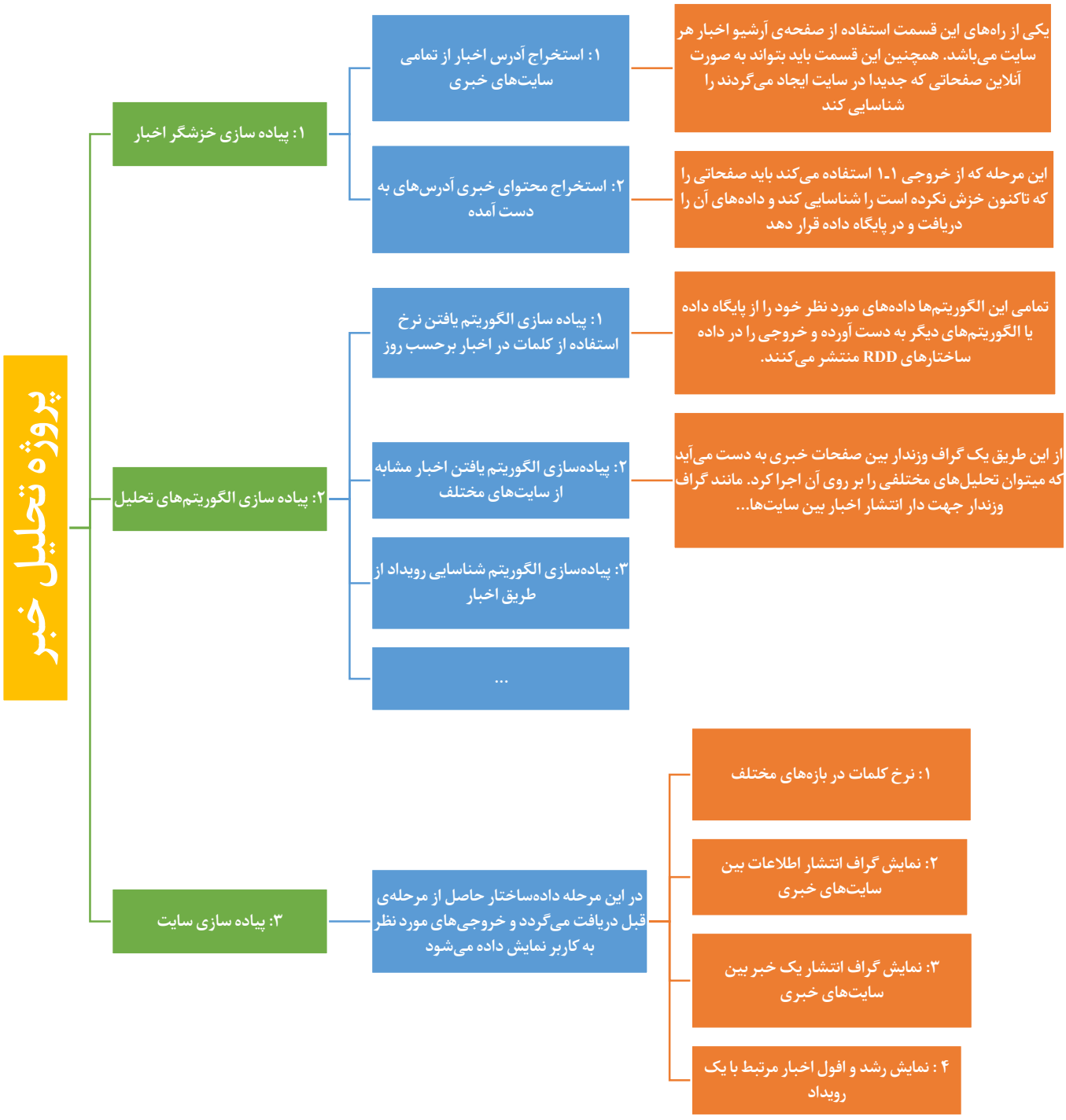
در مرحله‌ی pre-processing تمامی داده‌های خبری استخراج می‌گردند و برای تحلیل آماده می‌گردند. هر خبر به اجزای تشکیل دهنده‌ی آن شامل عنوان، خلاصه، متن خبر، تاریخ خبر و ویژگی‌های دیگری که در هر سایت متفاوت می‌باشد تقسیم می‌شود. این قسمت در نسخه‌ی اول پروژه به صورت آفلاین انجام شده است و خبرهای تعدادی سایت خبری به صورت یکجا[1] دریافت گردیده است ولی در نسخه‌ی آنلاین پروژه به صورت آنلاین با انتشار هر خبر به مجموعه‌ی پایگاه داده‌ی خبری اضافه می‌گردد.

مرحله‌ی processing شامل پیاده سازی و اجرای تمامی الگوریتم‌های تحلیل متن کاوی و خبری می‌باشد و خروجی‌های مورد نیاز برای پروژه را به دست می‌آورد. از جمله کشف ارتباطات بین اخبار مختلف، استخراج کلیدواژه‌ها از آن‌ها، استخراج گراف ارتباط اخبار و سایت‌های خبری و ... که در ادامه‌ی پروژه می‌تواند دربردارنده‌ی قابلیت‌های مختلفی باشد. شناسایی شباهت متن اخبار منتشر شده، شناسایی تناقضات خبری در متن‌هایی که مشابه یکدیگر می‌باشند، استخراج کلیدواژه ها و شناسایی رویدادهای خبری از جمله‌ی تحلیل‌هایی هستند که در این مرحله انجام می‌گردند.

مرحله‌ی post-processing پروژه شامل مورد کاربردهای مختلف می‌باشد که در آن‌ها کاربر می‌تواند به نتایج تحلیل به دست آمده از مرحله‌ی processing دسترسی پیدا کند. از ابزارهای مختلف برای نمایش نتایج استفاده می‌شود. مانند نمودارهای یک، دو یا چند بعدی و گراف‌های مختلف ارتباطی و نمودارهای مختلف آماری.

[1] Batch

برای انجام این فازها در ابتدا یک ساختار شکست بین فازهای مختلف پروژه در نظر گرفته شده است:



در گراف بالا اجزای مختلف هر مرحله قابل مشاهده می‌باشد. سه مرحله‌ی استخراج، تحلیل و نمایش در خانه‌های سبز رنگ نشان داده شده است و اجزای آن با توضیحاتی در خصوص نحوه‌ی پیاده سازی آن در ادامه‌ی آن‌ها آمده است.

نتایج حاصل از پروژه و مدل‌های درآمدی پیشنهادی برای آن

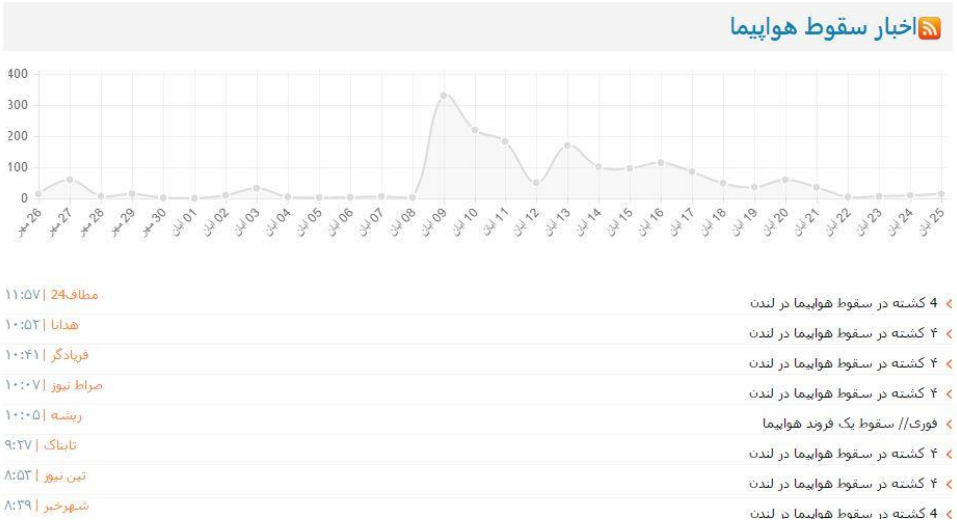
یکی از اصلی ترین خروجی‌های این پروژه تولید یک سایت خبرکاوی می‌باشد که می‌تواند تحلیل‌های مختلف از اخبار منتشر شده را در قالب نمودارها و گراف‌های جذاب در اختیار کاربر قرار دهد. یکی از مثال‌های خارجی سایت eventregistry.org می‌باشد.

امکاناتی که این سایت در اختیار کاربران قرار می‌دهد:

- مانیتور بر خط از اخبار منتشر شده در دنیا
- جستجوی رویدادها
- کسب اطلاعات کامل از یک رویداد براساس اخبار منتشر شده از آن
- خبرخوانی و مشاهده‌ی خبرهای در محدوده‌ی علائق و نیازمندی‌های کاربر
- مشاهده‌ی اخبار دستکاری شده و منحرف کننده
- مشاهده‌ی انتشار اخبار و کپی برداری خبری سایت‌ها
- مشاهده‌ی موضوعاتی که دارای روند بسیار زیادی شده‌اند
- پیش‌بینی آینده

نمونه‌ی ایرانی :

- تی نیوز



این سایت خبری برچسب‌های مهم را شناسایی می‌کند و اخبار مرتبط با آن‌ها نمایش می‌دهد. و همچنین نمودار توزیع برچسب را در دوره‌ی زمانی مشخص می‌کند.

مدل‌های درآمدی برای این گونه پروژه‌ها:

- ارائه سرویس تحلیل خبر خاص منظوره برای سازمان‌ها و نهادهای تصمیم گیرنده

امکاناتی که این ابزار می‌تواند در اختیار سازمان‌ها و نهادهای تصمیم گیرنده بگذارد عبارتند از :

- ✓ ارائه گزارش از موضوعات مختلف در بازه‌ی زمانی مشخص. به این وسیله سازمان می‌تواند در مواردی که مرتبط به فعالیت‌های خود و یا تاثیر گذار بر تصمیمات خود می‌باشد گزارشی دریافت کند که قابل تحلیل و تصمیم سازی می‌باشد.
- ✓ مشاهده‌ی افراد و سازمان‌های مرتبط با یک موضوع و یا یک رویداد خبری. بدین وسیله سازمان می‌تواند این ارتباطات را تحلیل کند و با برقراری ارتباط با آن‌ها دسترسی خود را به موضوع افزایش دهد.

- ارائه سرویس در حوزه‌ی Finance Technology برای مؤسسات مالی و سبدگردان سرمایه به منظور سرمایه گذاری کارآمد.

امکاناتی که این ابزار می‌تواند در اختیار این مؤسسات قرار دهد:

- ✓ بررسی وضعیت بازار به وسیله‌ی اخبار مرتبط با بازار (تجمیع اخبار مختلف مرتبط با قیمت‌های خبری)
- ✓ مشاهده‌ی روندهای داغ خبری که می‌توانند در قیمت‌ها تاثیر بگذارند. (پیش‌بینی تغییرات قیمت اقلام و کالاها)
- ✓ شناسایی تاثیر اخبار بر رشد و افول EPS[2] نمادهای بورس.

شناسایی تاثیر اخبار بر رشد و افول EPS نمادهای بورس.

به عنوان مثال درج خبر ادغام دو شرکت در خبرگزاری ها ، باید تحلیل شود که این ادغام بر EPS سهم خواهد افزود یا از آن خواهد کاست که هدف ، کسب سود غیر معمول یا جلوگیری از زیان توسط این خبر خواهد بود .

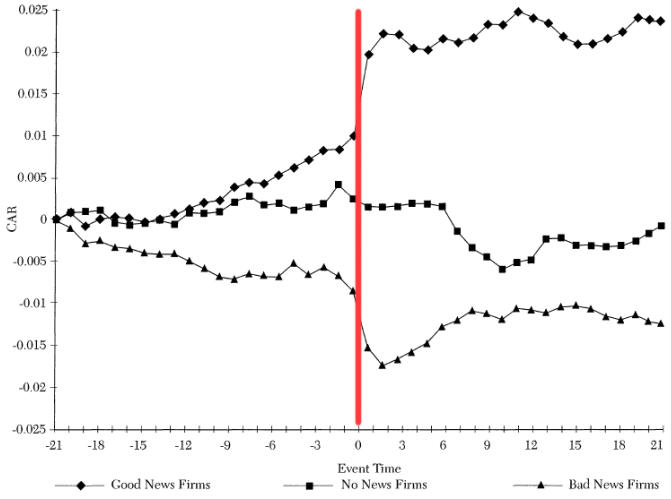
چون مطالعه ی رویداد ها به بررسی تاثیر یک خبر بر روی جهت و میزان تغییر قیمت یک سهم می پردازد ، در رشته های تحقیقاتی بسیاری می تواند کاربرد داشته باشد از جمله در حسابداری ، مالی ، مدیریت ، اقتصاد ، بازاریابی ، فناوری اطلاعات ، حقوق و علوم سیاسی .

برای بررسی تاثیر یک رویداد بر روی یک سهم باید پنجره بر روی گذشته سهم باز کنیم ، و تخمین بزنیم که بازده ی عادی سهم در روز خبر و چندین روز بعد از آن کجا خواهد بود . بعد از آن ، این شیوه به مقایسه بازده ی تخمینی و بازده ی واقعی برای تعیین میزان بازده ی غیر عادی می پردازد .

در این شیوه ، به گذشته ی 120 روزه ی سهم قبل از خبر نگاه می کنیم ، و با تحلیل رگرسیون یا استفاده از CAPM یا میانگین بازده تخمینی کوتاه مدت از آینده سهم بدست می آوردیم و با بازده ی واقعی مقایسه می کنیم .

برای مشخص کردن اینکه بازده ی غیر عادی (اختلاف بین بازده واقعی و تخمینی) غیر صفر است ، باید از آزمون ها آماری مختلف نظیر AR- , CAR- AAR- CAAR-استفاده کرد.یکی از پر کاربرد ترین تست ها ، تست t-student است که بازده ی غیر عادی را بر خطای جذر میانگین مربعات رگرسیون تقسیم می کند .

[2] Earnings per share



شکل1 سه نوع خبر که می‌توانند تاثیر مثبت ، منفی یا خنثی در ارزش یک سهم داشته باشند. این دسته ها را *Good News, Bad News* و *No News* مینامیم