ردیابی رسانه‌های اجتماعی متعدد برای پیش‌بینی رویدادها در بازار سهام

چکیده

مسئله مدلسازی روندهای مداوم در حال تغییر در بازارهای مالی و ایجاد پیش‌بینی‌های معنی‌دار و زمان واقعی در مورد تغییرات قابل توجه در این بازارها از جانب اقتصاددانان و دانشمندان علاقمندی‌های چشمگیری را معطوف به خود کرده است. علاوه بر شاخص‌های سنتی بازار، رشد رسانه‌های اجتماعی متنوع، اقتصاددانان را قادر می‌سازد تا شاخص‌های میکرو و زمان واقعی را در مورد عوامل احتمالی تاثیرگذار بر بازار همچون احساسات و عواطف عمومی، پیش‌بینی‌ها و رفتارها را تحت نفوذ قرار دهند. چندین ویژگی مربوط به بازار خاص را ارائه می‌کنیم که از منابع متنوعی همچون اخبار، حجم جستجوی گوگل و توئیتر کاوش می‌شود. علاوه بر این همبستگی بین این ویژگی‌ها و نوسانات بازار مالی را بررسی می‌کنیم. در این مقاله، رویکرد دلتا نایو بیز (DNB) را به منظور ایجاد پیش‌بینی درباره بازارهای مالی ارائه می‌کنیم. آنالیز آینده‌نگرانه دقیق دقت پیش‌بینی تولید شده از منابع متعدد، منابع ترکیب شده با منابع تولید شده از منبع مستقل را ارائه می‌کنیم. در می‌یابیم که پیش‌بینی‌های منابع متعدد عملکرد بهتری نسبت به پیش‌بینی‌های تک منبع دارد، حتی اگر با برخی از محدودیت‌ها همراه باشد.

کلمات کلیدی: پیش‌بینی بازار؛ رسانه‌های اجتماعی؛ ترکیب مشخصات؛ گوگل ترندز؛ burst توئیتر؛ تمایلات اخبار.

1. مقدمه

پیش‌بینی‌های مربوط به بازار مالی به دلیل نوسانات ذاتی خودش پیچیده شده است. دستیابی به سیگنال‌های این نوسان ارائه برآوردهای مناسب درباره «نوسانات قیمت بازار» اولین علاقمندی اقتصاددانان می‌باشد. این مسئله موجب علاقمندی شگرف محققان در رشته‌های گوناگونی همچون اقتصادیات، آمار و علم اطلاعات شده است. بر این اساس این امر منجر به طیف وسیعی از روش‌ها با هدف مدلسازی بازارهای سهام می‌شود [11، 16، 17، 20، 23].

در بیشتر رویکردهای قدیمی، محققان بازار سهام را با سوابق تاریخی قیمت‌ها مشخص می‌کنند و امضاهائی را پیدا می‌کنند که نشان‌دهنده‌ی افزایش یا کاهش قیمت‌ها بر اساس این سری‌های زمانی تاریخی است. با اینحال، این روش‌های سری زمانی مالی در کل از شاخص‌های انسانی، مانند واکنش عمومی نا آگاه است و اغلب تمایل به دقت آنها در پیش‌بینی تغییرات ناگهانی، بزرگ در ارزش بازار یافت شده است [4]. به تازگی، با رشد فراگیر رسانه اجتماعی [6، 14] که به افراد اجازه می‌دهد تا به آسانی احساسات خودشان را بیان کنند [21]، دیدگاه‌ها و نگرانی‌ها، کاوش زمان واقعی این عوامل امکانپذیر می‌شود. علاوه بر این، اکنون جنبه‌های مختلف احساسات عمومی را می‌توان با آنالیز شبکه‌های اجتماعی متعدد استخراج نمود. در این مقاله، داده‌های روند جستجوی جهانی از گوگل، مقالات آرشیو اخبار از اخبار بلومبرگ و توئیت‌های مرتبط از توئیتر را جمع‌آوری و آنالیز می‌کنیم. با استفاده از روش‌های بدون ناظر، ویژگی‌های را از این منابع موجود در دسترس عموم استخراج می‌کنیم. به کمک این ویژگی‌ها، مجموعه ای از آزمایش‌ها را طراحی می‌کنیم تا ارتباط و همبستگی‌های میان رفتار انسان و نوسانات بازار در بازارهای آمریکای جنوبی را بررسی کنیم. با این آنالیز، مدل‌هایی را ارائه می‌کنیم که تغییرات بزرگ (رخدادها) در ارزش بازار را با استفاده از مهمترین عوامل استخراج کننده اطلاعات پیش‌بینی می‌کند. بطور خاص، با توجه به این سه منبع داده در روز d و قیمت‌های تاریخی سهام برای یک بازار، مدل‌های ارائه شده خودمان در تلاش است تا ارزش بازار سهام را در حداقل روز $d + 1$ پیش‌بینی کند.

مشارکت یا تسهیمات مهم این مقاله عبارت است از:

- یک آنالیز از روندهای جستجوی گوگل، اخبار بلومبرگ و توئیتر را ارائه می‌کنیم تا اطلاعات در مورد روندهای بازار را جمع‌آوری کنیم و این روندهای رسانه اجتماعی را تعیین کمیت کنیم.
- ویژگی‌های burst از توئیتر را شناسایی می‌کنیم و این ویژگی‌ها را در رخدادهای burst گروه‌بندی می‌کنیم. همچنین همبستگی‌های این رخدادهای ترتیبی یا پشت سر هم را با روندها بازار بررسی میکنیم و پیدا می‌کنیم.
- مدل دلتا نایو بیز را به منظور پیش‌بینی نوسانات بازار مالی با استفاده از منابع متعدد رسانه اجتماعی ارائه می‌کنیم. اگرچه تلاش‌های قبلی در بررسی ترکیبات منابع برای کاربردهای مربوط به دارایی یا مالیه وجود دارد [12، 19]، با

اینحال بیشتر آثار بر روی بررسی مجموعه داده‌ها متمرکز است. در این راستا، با توجه به بهترین دانش خودمان، اثر ما در وهله اول نه تنها برای محاسبه منابع است بلکه مدل‌های پیش‌بینی را ارائه می‌کند که منابع متعدد را مورد استفاده قرار می‌دهد.

- در نهایت، یافته‌های خودمان را درباره همبستگی متقابل بین شاخص‌های بازار استخراج شده از منابع متعدد رسانه‌های اجتماعی ارائه می‌کنیم و اطلاعات حاصله از هر منبع داده را بطور گسترده آنالیز می‌کنیم.

2. اثر مرتبط

در این بخش چندین اثر مربوط به زمینه‌های آنالیز سری‌های زمانی مالی، مدلسازی بازارهای مالی و استخراج ویژگی‌ها از جریانات آنلاین داده‌ها را بررسی می‌کنیم.

آنالیز مالی سری‌های زمانی. آنالیز سری‌های زمانی مالی یکی از محبوب‌ترین رویکردها در مدلسازی بازار بوده است. مدل‌های ناهمسان *شرطی* خودرگرسیو تعمیم یافته (GARCH) [5] در حوزه مالی از دهه 1980 بطور وسیع به کار گرفته می‌شود. الگوریتم‌های خوشه‌بندی به منظور تشریح مجدد سری‌های زمانی [11] و شناسایی سهام موقت همبسته مورد استفاده قرار می‌گیرد [1]، روش‌هایی هستند که در پردازش داده‌های خودمان استفاده می‌کنیم. به تازگی، [17] طول تلفات اطلاعات بعنوان یک شاخص پیشرو در اندازه‌گیری بی‌ثباتی یا ناپایداری جهانی ارائه می‌شود.

همبستگی با رسانه اجتماعی. با شیوع و توسعه اخیر پلتفرم‌های داده بزرگ [25، 24]، اتخاذ وظائف داده‌کاوی در شبکه‌های اجتماعی آنلاین تولید نتایج حالت هنر را نشان می‌دهد. آنجا تعدادی از آنالیزهای اکتشافی مربوط به رسانه اجتماعی با بازارهای سهام وجود دارد. [16] یک همبستگی بین حجم معاملات شرکت‌های برتر و حجم جستجوی گوگل از این اسامی شرکت‌ها را پیدا می‌کند. [12] همبستگی یا ارتباط بین حجم پرس‌وجوی جستجو و میانگین صنعتی داو جونز (DJIA)[1] را بررسی می‌کند و حجم بالاتری از جستجوی اصطلاحات مالی مشخص را پیدا می‌کند که قیمت‌های پائین‌تر DJIA را نشان می‌دهد. علاوه بر این، [15] در می‌یابد که راهبردهای تجاری مبتنی بر حجم 98 کلمه کلیدی از جستجوی گوگل ترندز سرمایه‌گذاری تصادفی را با توجه به کل گردش مالی انجام می‌دهد. به

[1] یک شاخص اقتصادی بازار آزاد است که توسط وال استریت ژورنال منتشر و ارائه می‌گردد.

تازگی، [13] «تحرکات مشترک» بین قیمت‌های سهام و مقالات خبری برای پیشبینی بازار سهم را مطالعه می‌کند. [4] عواطف عمومی از توئیتر را محاسبه می‌کند و در می‌یابد که منحنی احساس و عاطفه «آرام» دارای یک همبستگی بسیار قدرتمند با ارزش‌های میانگین صنعتی داو جونز است. [20] استدلال بر این دارد که تعدادی از اجزای متصل در یک زیر نمودار محدود در نمودارهای زمان محدود دارای همبستگی بالا با حجم معامله شده دارد. در این مقاله، تحقیقی را ایجاد می‌کنیم که حاکی از این است که جمع‌آوری حجم جستجو و تغییرات ناگهانی در احساسات در بین شبکه‌های اجتماعی با عملکرد مالی بازار توسط ترکیب این عوامل در یک چارچوب پیش‌بین شده متحد مرتبط است.

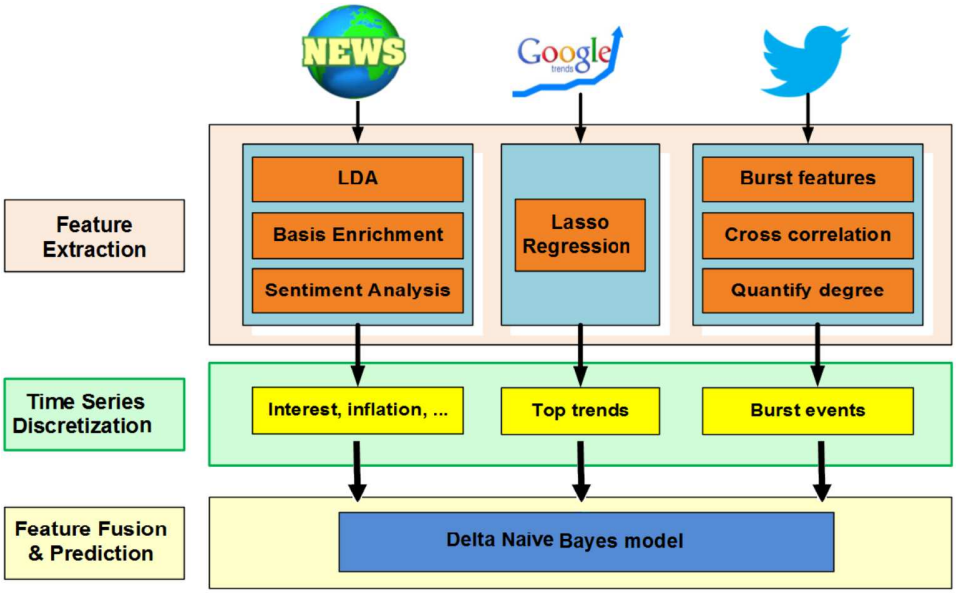
فیوژن ویژگی با توجه به روش‌های فیوژن، [10] رویکرد محبوب فیلتر کالمن برای مسائل پیش‌بینی و فیلترینگ خطی را ارائه می‌کند که حسگرهای ترتیبی متعدد را اندازه‌گیری می‌کند تا حالات دینامیکی سیستم را برآورد کند. طبقه‌بندی کننده بیز نایو بعنوان مدل موثری شناخته می‌شود که برچسب‌های کلاس برای ویژگی‌های چند بعدی مبتنی بر حداکثر احتمالات خلفی یا پسین را برآورد می‌کند. مدل‌های مارکوف پنهان بطور موفقیت‌آمیزی برای تشخیص الگوی زمانی در مناطقی همچون تصاویر ترتیبی به کار می‌رود [27]. در آزمایش خودمان، مدل کلاسیک نایو بیز را، برای فیوژن ویژگی و پیش‌بینی مالیه استفاده و اصلاح می‌کنیم.

3. استخراج ویژگی

مرحله نخست در روش خودمان استخراج ویژگی‌ها از رسانه‌های اجتماعی است. همانطور که در شکل 1 نشان داده شده است، هر منبع داده مستلزم پردازش است تا یک سری زمانی شود که از آن برای پیش‌بینی استفاده می‌کنیم. برای مقالات از اخبار بلومبرگ، مدلسازی موضوع، پردازش زبان طبیعی و آنالیز عواطف را استفاده می‌کنیم تا ارزش شاخص‌های اقتصادی را برآورد کنیم. برای داده‌های گوگل، رگرسیون لاسو را استفاده می‌کنیم تا مشخص کنیم که چه اصطلاحاتی از مجموعه سفارش خودمان از اصطلاحات مربوط به دارائی آموزنده‌تر است. برای توئیتر، «ویژگی‌های burst» را در «رخدادهای burst » زنجیر می‌کنیم و مواردی را در نظر می‌گیریم که دارای بالاترین درجه داغی است.

3.1. جستجوی گوگل ترندز

حجم و میزان جستجوی گوگل را بعنوان منبعی برای اطلاعات جایگزین درباره احساسات عمومی نسب به شاخص‌های سهام و نگرانی‌ها درباره آینده امور مالی در نظر می‌گیریم. این روندهای پویا بطور بالقوه عوامل مهمی هستند که می‌توانند



شکل 1. چارچوب کلی سیستم.

چارچوب آنالیز مالیه یا دارایی را تحت تاثیر قرار دهند. چندین مولف حجم جستجوی گوگل را به منظور پیش‌بینی تغییرات بازار سهام استفاده کرده‌اند [12، 15، 16]. با اینحال، از آنجا که رویکردهای آنها میزان جستجو برای کلمات کلیدی و عبارات را با استفاده از بازده‌های تجمعی در طول یک دوره زمانی ثابت محاسبه می‌کند، آنها مستعد به از دست دادن توزیع پویای کلمات کلیدی در طول یک چرخه جستجو هستند نظر به این که اهمیت این کلمات دچار تغییر می‌شود. برای یافتن اصطلاحات بسیار مفهومی در هر چرخه، مقررات- L1، یا لاسو [22] را استفاده می‌کنیم تا لغت‌نامه خودمان را از اصطلحات مالی فیلتر کنیم. این رویکرد موجب می‌شود روش خودمان به اندازه کافی قدرتمند باشد تا بتواند با آسانی فیدبک متخصصین را با اضافه نمودن کلمات کلیدی در صورت لزوم ترکیب یا یکی کند.

بطور رسمی، برای هر یک از کلمات کلیدی k خودمان، میزان جستجوی آن $x^k(t)$ را در زمان t محاسبه می‌کنیم. نشان می‌دهیم که این حجم و میزان نرخ‌های آتی سهام $P(t+1)$ را در زمان t + 1 تحت تاثیر قرار می‌دهد. تغییرات در قیمت بازار $\Delta P(t+1)$، در زمان t + 1 را با تغییرات در میزان جستجوی کلمات کلیدی در زمان t، $\Delta x^k(t)$

مقایسه می‌کنیم. این همبستگی و رابطه را بعنوان یک معادله خطی تعریف می‌کنیم همانطوری که در معادله 1 نشان داده شده است. a_k هر وزن هر اصطلاح است.

$$\Delta P(t+1) = \sum_{k=1}^{K} a_k \Delta x^k(t) \qquad (1)$$

اوزان با استفاده از معادله 2 برای هر زمان محاسبه می‌شود که در آن بهترین کلمات کلیدی کلماتی هستند که دارای بالاترین اوزان هستند و همچنین دارای اوزان غیر صفر می‌باشند. λ را از ارزیابی‌های تجربی مبتنی بر دقت پیش‌بینی-های نهایی فیکس می‌کنیم.

$$\begin{aligned} a &= (a_1, a_2, \ldots, a_K) \\ &= argmin_a(\sum_t(\Delta P(t) - \sum_{k=1}^{K} a_k \Delta x^k(t))^2 + \lambda \sum_k |a_k|) \end{aligned} \qquad (2)$$

از آنجا که حجم و میزان جستجوی گوگل بصورت هفتگی گزارش می‌شود، پس می‌توان در گرایش کلی از پرس‌وجوهای عمومی نهان‌سازی را اعمال نمود. این مسئله را با در نظر گرفتن $z-score(4)$ از میزان جستجوی گوگل برای هر هفته را بعنوان مقدار نمایه‌ای برای هر روز از آن هفته ارائه می‌کنیم. برای قیمت‌های بازار، که هر روز با تغییر همراه است، $z-score(30)$ از هر روز نزدیک به قیمت را بعنوان مقداری برای آن روز استفاده می‌کنیم. یک امتیاز-z تعریف می‌شود بعنوان

$$z-score(n) = (X - \mathcal{M})/\Sigma \qquad (3)$$

که در آن X اختلاف یک روزه است، M متوسط حرکت n-روز عقبی از اختلافات یک روز است، و Σ انحراف استاندارد آن روزهای عقبی از اختلافات تک روزه متحرک n-روز است. این مرحله داده‌های هفتگی جستجوی گوگل ترندز را با داده‌های روزانه قیمت سهام هم‌تراز می‌کند. همچنین تظمین می‌کند که واریانس‌ها تقریبا بر حسب همان مقدار زمان: یعنی حدودا یک ماه اندازه‌گیری می‌شوند. با ایجاد توابعی که گوگل ترندز را به ارزش‌های بازار مالی متصل می‌سازد، می‌توانیم ببینیم که چگونه تغییرات در میزان جستجو تغییرات در بازار را منعکس می‌کند و همچنین ویژگی‌های آموزنده و مفهومی را برای هر پیش‌بینی انتخاب می‌کند

3.2. نظرکاوی از مقالات خبری

برای ایجاد این پیش‌بینی‌ها، به مشخصات اقتصادی مانند نرخ بهره، نرخ تورم، GDP[۲]، اعتماد مصرف کننده، و سرمایه-گذاری خارجی علاقمند هستیم. متاسفانه ارزش‌های رسمی برای این ویژگی‌ها فورا در دسترس قرار نمی‌گیرد که همین امر موجب می‌شود آنها برای پیش‌بینی اقتصادی در زمان واقعی ناپایدار باشند. به جای تکیه بر ارزش‌های موسسات دولتی یا تحقیقاتی، زبان طبیعی و پردازش آماری در اخبار اقتصادی آنلاین منتشر شده توسط بلومبرگ نیوز را استفاده می‌کنیم و محتویات پردازش را جایگزین داده‌های واقعی و فاقد دسترسی در نظر می‌گیریم.

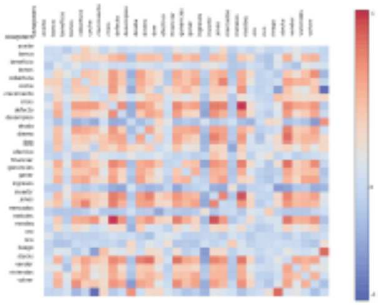
مقالات بلومبرگ 401,923 (به شکل آزاد که با استفاده از API خودشان د دسترس هستند) را از ماه آوریل 2010 تا ژوئن 2013 جمع‌آوری می‌کنیم. برای فیلتر این مجموعه با تنها مقالاتی درباره اخبار اقتصادی، الگوریتم مدلسازی موضوع تشخیص پنهان دیریکله محبوب [2] را استفاده می‌کنیم تا موضوعات مورد علاقه را شناسایی کنیم. بصورت تصادفی تعداد 25, 041 مقاله را بعنوان مجموعه آموزشی انتخاب می‌کنیم، تعداد موضوعات و عناوین در 30 ثابت است و مجموعه‌ای از موضوعات محاسبه می‌شود. سپس بصورت دستی موضوعاتی را که بیشتر مربوط به مالیه و اقتصادیات مبتنی بر آنالیزی از 30 کلمه کلیدی برتر از هر موضوع است، انتخاب می‌کنیم. هر سندی در هر سندی در هر مجموعه از نوشته‌جات متشکل از هر یک از موضوعات انتخاب شده برای پیش‌بینی مالی در نظر گرفته می‌شود.

سپس، مجموعه‌ای از ابزارهای پردازش زبان طبیعی توسعه یافته توسط فن‌آوری پایه را استفاده می‌کنیم تا اصل موضوع‌سازی، تشخیص مرز جمله، بخشی از برچسب زبان گفتار و شناسایی اصطلاح هر اسم در هر مقاله را انجام دهیم. از آنجا، فرهنگ لغت سفارشی سطح کشور را برای مقالات بنیادی کشور مورد استفاده قرار می‌دهیم. این پردازش در زبان طبیعی این اجازه را به ما می‌دهد تا عباراتی را که به یک کشور خاص اشاره می‌کند شناسایی کنیم چرا که مباحث مقالات ممکن است در مورد چندین کشور باشد. در نهایت، فرهنگ لغت سفارشی ویژگی اقتصادی را همراه با خروجی-های تشخیص اصطلاح نام استفاده می‌کنیم تا امتیازات عاطفی برای هر ویژگی برای هر کشوری را محاسبه کنیم. این امتیازات یا نمرات احساسی بعنوان ورودی‌هایی برای پیش‌بینی‌های اقتصادی خودمان به کار گرفته می‌شود.

3.3. تشخیص burst توئیتر

[2] **تولید ناخالص داخلی**

با نفوذ و هیاهوی شبکه‌های اجتماعی، توئیتر سریعا بعنوان منبعی ارزشمند در بازتاب جنبش‌های اجتماعی و نشان دادن عواطف پیچیده افراد تکامل یافت. تحقیق [4، 20] نشان می‌دهد که بازارهای مالی



شکل. 2. ویژگی همبستگی متقابل از رهدادهای burst توئیتر.

ارتباط تنگاتنگی با تحرکات اجتماعی دارند و علاوه بر این، احساسات انسانی نیز برآمده از آن است یعنی توسط توئیتر هدایت می‌شود. با استفاده از توئیتر، می‌توانیم تازه‌ترین تحرکات اجتماعی را در زمان واقعی تشخیص دهیم که بعنوان burstهای توئیتر ارائه می‌شود [7]. با آنالیز همبستگی بین این burstهای و آشفتگی یا نوسان بازار، بررسی می‌کنیم که چگونه آنها با بازارهای مالی همبسته و مرتبط هستند. چارچوب کلی ما برای تشخیص burst توئیتر شامل مراحل زیر است:

1. ایجاد شبکه‌ای برای ویژگی‌های burst : برای یک پنجره مشخص زمانی، TF-IDF را به منظور شناسایی ویژگی‌های BURST و تخصیص هر ویژگی با یک امتیاز burst به کار میبریم.
2. خوشه‌بندی ویژگی‌های burst در رخدادهای burst: هر جفت از گره‌های همبسته با وزنی برابر با امتیاز همبستگی خودشان متصل به لبه می‌شوند.
3. محاسبه داده‌های جایگزین: حجم و میزان توئیتر و امتیازدهی عواطف را به منظور اندازه‌گیری هر اثرگذاری رخداد burst استفاده می‌کنیم.

شناسایی ویژگی‌های burst. از آنجا که علاقمند به پیش‌بینی‌ها بازار سهام در کل کشور هستیم، باید مشخص کنیم که از کدام کشور یک توئیت داده می‌شود. اگرچه توئیت‌های حاوی طول و عرض جغرافیائی کاربر هستند، با اینحال بیشتر توئیت‌ها فاقد این داده‌ها هستند و بایستی با استفاده از سایر روشها دارای طول و عرض جغرافیائی باشند. برای

دور زدن این موضوع، الگوریتم ژئوغنی‌سازی (غنی‌سازی جغرافیائی) دیگری را برای مجموعه داده‌های خودمان با استفاده از ابزار کدبندی جغرافیائی تشریح شده در [18] به کار می‌بریم. زمانی که برای یک کشور مورد علاقه دارای مجموعه‌ای از توئیت‌ها باشیم «امتیاز burst» را برای هر یک از دوره‌های زمانی در طول زمان محاسبه می‌کنیم. هر امتیاز burst از دوره‌های زمانی با استفاده از «فرکانس سند معکوس فرکانس» (TF-IDF) محاسبه می‌شود. در زمان معین t، زمانی که $|z - score(30)|$ دوره بزرگتر از آستانه باشد، این بعنوان یک مشخصه burst برچسب می‌خورد یعنی تلقی می‌شود.

گروه‌بندی به رخدادهای burst. پس از محاسبه امتیازات burst برای هر دوره، مجموعه‌ای از بردارهای مشخصه burst ذکر شده توسط $B = \{b_0, b_1, \ldots, b_l\}$ را داریم. قصد داریم تا یک نمودار، $G = (B, E, C)$، را ایجاد کنیم به طوری که گره‌ها به یک لبه در E متصل می‌شود اگر همبستگی به اندازه کافی قدرتمندی، C، بین دو گره وجود داشته باشد [26]. معادله 4 را استفاده می‌کنیم تا امتیازات همبستگی را محاسبه کنیم که در آن f^* ساختمان پیچیده f را نشان می‌دهد و m تاخیر زمانی است. نتایج را میتوان در شکل 2 مشاهده نمود.

$$(f * g)[n] = \sum_{m=-\infty}^{\infty} f^*[m]g[n+m] \qquad (4)$$

تنها بردارهایی را در نظر می‌گیریم که امتیاز همبستگی متقابل آنها بیشتر از آستانه پیش تعریف شده باشد، که برای اهداف ما این مقدار 0.1 است، که بعنوان مقدار همبستگی بالا می‌باشد. یک لبه، e، را بین بردارهای بسیار همبسته با امتیاز همبستگی c ایجاد می‌کنیم. توجه داشته باشید که چون تنها بردارهای بسیار همبسته (ویژگیهای burst) به لبه‌ها متصل هستند، پس گروه‌بندی ویژگی‌های burst در رویدادهای burst به مسئله شناسایی جوامع در شبکه‌های توئیتر با هر جامعه‌ای بعنوان یک رویداد burst متشکل از مجموعه‌ای از مشخصات burst تبدیل می‌شود. روش لووین [3] را برای شناسایی جوامع در شبکه‌های توئیتر مورد استفاده قرار می‌دهیم.

اندازه‌گیری درجه داغی رخداد. رخداد را در نظر می‌گیریم که در یک دوره زمانی کوتاه صورت می‌گیرد که با حجم بالای توئیتر داغ می‌شود. برای هر رویداد E، فهرست ویژگی‌های burst f_i و حجم توئیتری مرتبط با آن را به دست می‌آوریم و میزان داغی رویداد را محاسبه می‌کنیم. باترکیب واکنش عمومی با این رخداد، رویدادی را خوش‌بینانه یا

بلعکس بدبینانه تلقی می‌کنیم. از منظر ریاضیاتی، برای یک موضوع با رویدادهای m burst، حجم توئیتر v_i و محاسبات عاطفی منفی s_{ni}، میزان داغی آن، $H_e(t)$، در اسلات زمانی t را می‌توان مطابق با معادله 5 محاسبه نمود.

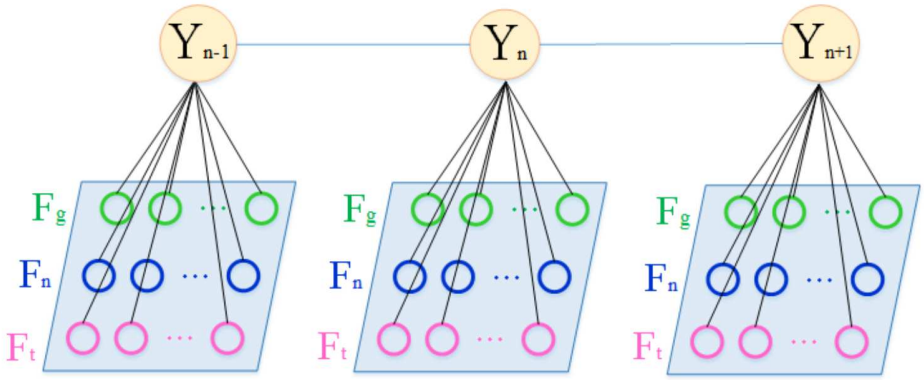
$$H_e(t) = \sum_{i=1}^{m} \frac{v_i * s_{ni}}{t} \qquad (5)$$

4. اثر کلی و پیش‌بینی

حجم جستجوی گوگل بعنوان یک جو اتمسفری از هویت مصرف کننده و انتظارات عمومی درباره وضعیت‌های اقتصادی در نظر گرفته می‌شود، که بعنوان یک اشکارساز پس‌زمینه برای پیش‌بینی‌های بازار مورد استفاده قرار می‌گیرد [15، 16]. مقالات اینترنت از شرکت‌هایی مانند بلومبرگ نیوز منبع خوبی برای نظرات داخلی نظرات کارشناسان، گزارشات رخدادهای مهم و آنالیز وضعیت است، که حاوی اطلاعات زیاد و تا حدودی بسیار حرفه‌ای و همچنین سیگنال‌های دقیق نسبت به سایر منابع رسانه اجتماعی است [8]. در نهایت، توئیتر یکی از محبوب‌ترین رسانه‌های اجتماعی در دستیابی به تحرکات اجتماعی در حال ظهور است. بخصوص در اخبار فوری، توئیتر تحرکات و خبردهی سریع‌تری نسبت به دیگر رسانه‌های قدیمی دارد [9]. برای استفاده کامل از این منابع، یک چارچوب فیوژن ویژگی را ایجاد می‌کنیم تا ویژگی‌های استخراج شده از هر: عوامل خبر (F_n)، عوامل میزان جستجوی گوگل (F_g)، و عوامل رویداد توئیتر (F_t) را ترکیب کنیم. انتظار داریم عملکرد بهبود یافته‌ای را با ترکیب این سه منبع در روشی هوشمندانه به دست آوریم.

4.1. فیوژن ویژگی

از آنجا که ویژگی مشاهده ما برآوده از سه منبع متفاوت است، آنها را مستقل از همدیگر در نظر می‌گیریم. همانطور که در شکل 3 نشان داده شده است، $Y = \{y_n\}$ را بعنوان برچسب تاریخی قیمت سهام ذکر می‌کنیم، که در آن y_n رسته قیمت سهام در روز n را نشان می‌دهد. همچنین ویژگی‌های مشاهده از جستجوی گوگل ترندز، بلومبرگ نیوز و توئیتر را به ترتیب بعنوان F_g, F_n و F_t اشاره می‌کنیم. راهبردی را برای ترکیب این ویژگی‌ها به کار می‌بریم، که بعنوان $x_n = (F_g, F_n, F_t)_n$ تعریف می‌شود، که در آن x_n ترکیب سه ویژگی مشاهده شده در روز $n - 1$ است. علاوه بر این، عادی‌سازی پیش از فیوژن ویژگی شرط است چون F_g, F_n و F_t ممکن است از مقیاس‌های مختلفی باشد.



شکل. 3. فیوژن ویژگی: Y_n وضعیت بازار است. F_n ضریب اخبار است. F_t ضریب توئیتر است. F_g ضریب گوگل ترندز است.

4.2. گسسته‌سازی سهام

برای اطمینان از دوره‌های نوسان‌پذیری بال و تغییراتی که رخ می‌دهد بسیاری از هفته‌ها به دقت اندازه‌گیری می‌شوند، امتیازات-z در طول یک پنجره زمانی 30 روزه را محاسبه می‌کنیم. براساس آستانه‌های امتیاز-z، وضعیت سهام را به ترتیب زیر تعریف می‌کنیم. زمانی که $z - score(30) \leq -1$، آن را بعنوان وضعیت «غوطه‌ور» نامگذاری می‌کنیم، زمانی که $z - score(30) \geq 1$، آن را بعنوان وضعیت «مواج» می‌نامیم، بعبارتی دیگر، زمانی که $|z - score(30)| < 1$ ، آن را وضعیت «هموار» نامگذاری می‌کنیم.

4.3. مدل دلتا نایو بیز

نایو بیز یک طبقه‌بندی خوب شناخته شده‌ای است که، با فرض استقلال شرطی ویژگی‌ها، احتمال خلفی برچسب‌ها برای یک مجموعه مشخصه ورودی را محاسبه می‌کند. متاسفانه، مدل اصلی نایو بیز برای پیش‌بینی‌های بازار مالی به دلیل توزیع کلاس بسیار مبهم و منحرف پائین‌تر از سطح بهینه است. در فرایندهای یادگیری خودمان، در می‌یابیم که 78.6 درصد ویژگی‌های خودمان به همان کلاس اختصاص می‌یابد. نتیجه با توجه به ارزش‌های بازار سهام که اغلب زیاد نیست و تغییرات ناگهانی ندارد معقول می‌باشد.

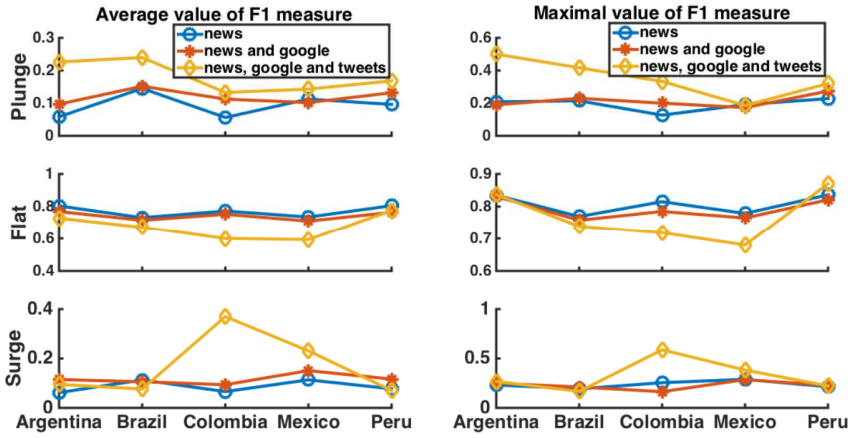
الگوریتم 1 مدل دلتا نایو بیز

ورودی‌ها: مجموعه داده‌های تاریخی $\{(x_{n-L}, y_{n-L}), ..., (x_n, y_n)\}$، اندازه پنجره یادگیری L، و آستانه احتمالی t، و مهمترین مشاهده اخیر

خروجی‌ها: $\hat{y}_{n+1}$

1. با استفاده از مدل نایو بیز، ماتریس شرطی احتمالی $P(Y|X)$ از $s_0 = \{(x_{n-L}, y_{n-L}), ..., (x_{n-1}, y_{n-1})\}$ را محاسبه کنید؛
2. با استفاده از مدل نایو بیز، ماتریس شرطی احتمالی $Q(Y|X)$ از $s_1 = \{(x_{n-L+1}, y_{n-L+1}), ..., (x_n, y_n)\}$ را محاسبه کنید؛
3. $\Delta P(Y|X) = \text{abs}(Q(Y|X) - P(Y|X))$؛ را محاسبه کنید؛
4. $p_{max} = \max_{y_j} \Delta P(Y = y_j | X = x_{n+1})$؛
5. $\text{Return } \hat{y}_{n+1} = \arg\max_{y_j} \Delta P(Y = y_j | X = x_{n+1}) \text{ when } p_{max} > t\text{; otherwise return } \hat{y}_{n+1}$ بعنوان «هموار».

برای رفع این معایب، مدل دلتا نایو بیز را در الگوریتم 1 ارائه می‌کنیم. برخلاف مدل نایو بیز اصلی، این طبقه‌بندی کننده مبتنی بر



شکل. 4. عملکرد با اخبار مستقل، دو منبع و سه منبع در کشورهای متعدد. ردیف بالاتر عملکرد پیش‌بینی حالت غوطه‌ور را نشان می‌دهد، ردیف‌های سمت چپ و راست نیز به ترتیب مقیاس متوسط F_1 و مقیاس بیشینه F_1 را نشان می‌دهد. ردیف وسط عملکرد وضعیت هموار را نشان می‌دهد. ردیف پائین عملکرد وضعیت مواج را ارائه می‌کند.

تغییر افزایشی احتمال خلفی به جای خود احتمال خلفی است. برای مرحله زمانی n^{th} در سری‌های گسسته، $P(Y|X)$ و $Q(Y|X)$ را در مجموعه داده‌های تاریخی $s_0 = \{(x_{n-L}, y_{n-L}), \ldots, (x_{n-1}, y_{n-1})\}$ و $s_1 = \{(x_{n-L+1}, y_{n-L+1}), \ldots, (x_n, y_n)\}$ محاسبه می‌کنیم. اختلاف بین s_0 و s_1 توسط $\Delta P(Y|X) = Q(Y|X) - P(Y|X)$ ارائه می‌شود. برای یک مشاهده جدید x_{n+1}، $P(Y|X)$ یا $Q(Y|X)$ را بعنوان طبقه-بندی کننده بطور مستقیم استفاده نمی‌کنیم. همچنین، در ابتدا مقدار بیشینه احتمال شرطی $\bar{p}_{max} = \max_{y_j} \Delta P(Y = y_j|X = x_{n+1})$ را بررسی میکنیم. اگر $\bar{p}_{max}$ بزرگتر از برخی از آستانه های پیش تعریف شده t باشد پس y_{n+1} پیش‌بینی می‌شود بعنوان

$$\hat{y}_{n+1} = \arg\max_{y_j} \Delta P(Y = y_j|X = x_{n+1}).$$

بعبارتی دیگر، y_{n+1} بعنوان حالت «هموار» پیش‌بینی می‌شود.

همانطور که می‌بینیم، تنها تفاوت بین s_0 و s_1 این است که نقطه (x_{n-L}, y_{n-L}) تازه‌ترین نقطه (x_n, y_n) جایگزین می‌شود. در نتیجه تفاوت دو معیار احتمال شرطی $P(Y|X)$ و $Q(Y|X)$ بایستی کاملا کوچک باشد. این کمک میکند تا تشریح کنیم که چرا مدل نایو بیز اصلی همیشه همان کلاس را تخصیص می‌دهد. علاوه بر این، تاثیر تازهترین نقطه (x_n, y_n) در پیش‌بینی خودمان توسط سایر نقاط $L - 1$ کاهش می‌یابد. لازم به ذکر است که استفاده از نقاط L به ما کمک می‌کند تا از اثرات چندین نقطه داده نویزدار اجتناب کنیم. همچنین انتخاب L مهم است. با استفاده از یک مقدار پائین پیچیدگی محاسباتی در محاسبه ماتریس احتمال شرطی $P(Y|X)$ کاهش می‌یابد اما خطرات تداخلات ناشی از نقاط نویزدار بیشتر می‌شود. از سوی دیگر، استفاده از یک مقدار بالاتر برای L منجر به نهان‌سازی بهتر تداخل می‌شود اما پیچیدگی محاسباتی بیشتر می‌شود.

جدول 1. عملکرد مدل دلتا نایو بیز در شاخص سهام کشور کلمبیا COLCAP. N منابع اخبار را نشان می‌دهد، G نشان‌دهنده‌ی منبع گوگل ترندز است و T به معنای منبع توئیت‌ها است. t آستانه گذار را ارائه می‌کند، p دقت و r یادآوری را نشان می‌دهد.

Source	t	Plunge			Flat			Surge		
		p	r	F_1	p	r	F_1	p	r	F_1
N	0.02	0.147	0.074	0.097	0.69	0.78	0.731	0.146	0.132	0.138
	0.04	0.165	0.048	0.069	0.683	0.873	0.764	0.072	0.045	0.052
	0.06	0.133	0.028	0.044	0.686	0.909	0.78	0.088	0.035	0.047
	0.08	0.145	0.023	0.038	0.688	0.93	0.79	0.105	0.035	0.049
	0.1	0.195	0.018	0.033	0.691	0.949	0.799	0.112	0.03	0.043
NG	0.02	0.156	0.097	0.118	0.697	0.742	0.718	0.142	0.153	0.147
	0.04	0.178	0.097	0.124	0.702	0.817	0.754	0.138	0.094	0.11
	0.06	0.17	0.083	0.109	0.692	0.839	0.757	0.104	0.063	0.075
	0.08	0.186	0.083	0.112	0.688	0.856	0.762	0.117	0.054	0.07
	0.1	0.19	0.071	0.1	0.688	0.881	0.772	0.149	0.049	0.068
NGT	0.02	0.135	0.169	0.149	0.582	0.499	0.535	0.325	0.418	0.366
	0.04	0.134	0.146	0.139	0.593	0.594	0.588	0.379	0.418	0.394
	0.06	0.128	0.124	0.123	0.583	0.63	0.6	0.405	0.418	0.407
	0.08	0.15	0.124	0.134	0.587	0.681	0.627	0.374	0.358	0.36
	0.1	0.175	0.091	0.118	0.584	0.74	0.65	0.355	0.302	0.322

5 ارزیابی تجربی

هدف ما پیش‌بینی نوسانات در ارزش بورس اوراق بهادار است. بری این چالش، بر کشورهای آمریکای جنوبی از جمله آرژانتین، برزیل، کلمبیا، مکزیک و پرو متمرکز می‌شویم. نخست معیار ارزیابی خودمان را تشریح می‌کنیم سپس آنالیز دقیقی را از عملکرد ارائه می‌کنیم.

5.1. معیار ارزیابی

در نبود یک ماتریس عملکرد استاندارد، یک معیار ارزیابی مبتنی بر نرخ برخورد به شرح زیر را تعریف می‌کنیم. برای یک روز مستقل تنها سه وضعیت وجود دارد: غوطه‌ور یا سراشیب، هموار و یا مواج. برای محاسبه دقت هر وضعیت، تنها زمانی که وضعیت پیش‌بینی دقیقا مشابه است، پیش‌بینی صحیح را در نظر می‌گیریم.

یادآوری، پیش‌بینی و مقیاس F_1 در معیار به کار رفته در مسئله دسته‌بندی بطور وسیع استفاده می‌شود. تعریف رسمی از یادآوری $r = \frac{TP}{TP+FP}$ است، از پیشبینی $p = \frac{TP}{TP+FN}$ و از مقیاس F_1 نیز $F_1 = \frac{2rp}{r+p}$ است، که در آن TP مثبت واقعی را نشان می‌دهد، FP مثبت نادرست را نشان می‌دهد، و FN به معنای منفی نادرست است. در حالی که این موضوع برای محاسبه دقت و یادآوری برای یک مسئله دسته‌بندی دودوئی سر راست و مستقیم است، با اینحال این می‌تواند تماما گیج کننده باشد هم چنان که چگونگی محاسبه این مقادیر برای مسئله دسته‌بندی چند کلاسه مطرح

است. در مورد خودمان، مسئله چند طبقه سه کلاسه در میان است. برای یک کلاس منحصربفرد، مثبت‌های نادرست نمونه‌هایی هستند که بعنوان آن کلاس دسته‌بندی می‌شوند، اما در واقع این چنین نیست، و منفی‌های واقعی نمونه‌هایی هستند که آن کلاس و دسته نیستند، و همچنین بعنوان تعلقات موجود برای آن کلاس طبقه‌بندی نمی‌شوند (صف نظر از این که آنها به درستی طبقه‌بندی شده باشند).

5.2. آنالیز عملکرد

مجموعه داده‌های آزمایش خودمان از اول ژانویه سال 2012 تا 31 جولای 2013 است. اندازه پنجره آموزش L متغیر از 40 روز تا 200 روز را تنظیم می‌کنیم تا تلاش‌ها در شکل‌دهی توزیع کیفیت را تست و بررسی کنیم. یک ارزیابی جامع از مدل دلتا نایو بیز را در برابر جنبه‌های مختلف به شرح زیر ارائه می‌کنیم:

چگونه مدل خودمان برای 5 کشور مورد علاقه به کار می‌رود و پیش‌بینی برای وضعیت سرازیری، هموار و مواج چقدر خوب است؟ عملکرد کلی پیش‌بینی برای هر کشور را با توجه به سه وضعیت سهام در شکل 4 ترسیم می‌کنیم. می‌توانیم ببینیم در حالی که وضعیت غوطه‌ور ا سرازیری پیش‌بینی می‌شود، کشورهای برزیل و آرژانتین عملکرد قابل توجهی را به دست می‌آورند. زمانی که وضعیت مواج پیش‌بینی می‌شود، کلمبیا، با مقیاس F_1 به بزرگی 0.6 برتر از بقیه کشورها است. در حالی که وضعیت هموار پیش‌بینی می‌شود، کشورهای مکزیک و پرو وضعیت برجسته‌ای را به نمایش می‌گذارند.

منبع مستقل یا منابع متعدد، عملکرد کدام یک در دستیابی به وضعیت مواج، غوطه‌ور بهتر است؟ همانطور که می‌دانیم، بیشتر وضعیت‌های بازار سهام وضعیت هموار می‌باشد،که موجب می‌شود حتی این برای پیش‌بینی وضعیت مواج یا غوطه‌ور چالش برانگیز باشد. در آزایش خودمان، توجه ویژه‌ای را معطوف به پیش‌بینی وضعیت غوطه‌ور و مواج می‌کنیم، که در شکل 4 به ترتیب در ردیف بالاتر و پائین نشان داده شده است. می‌توانیم ببینیم که، سه منبع پیش‌بینی بطور مداوم عملکرد بهتری را نسبت به پیش‌بینی دو منبع دارند، در حالی که پیش‌بینی دو منبع با پیش‌بینی مشخصه مستقل برابر یا بزرگتر از آن است. یک دیگر از شواهد آشکار که در جدول 1 یافت می‌شود، این است که در آن عملکرد پیش‌بینی دقیق کلمبیا در یادآوری، دقت مقیاس F_1 نشان داده می‌شود. برای وضعیت مواج می‌توانیم ببینیم که،

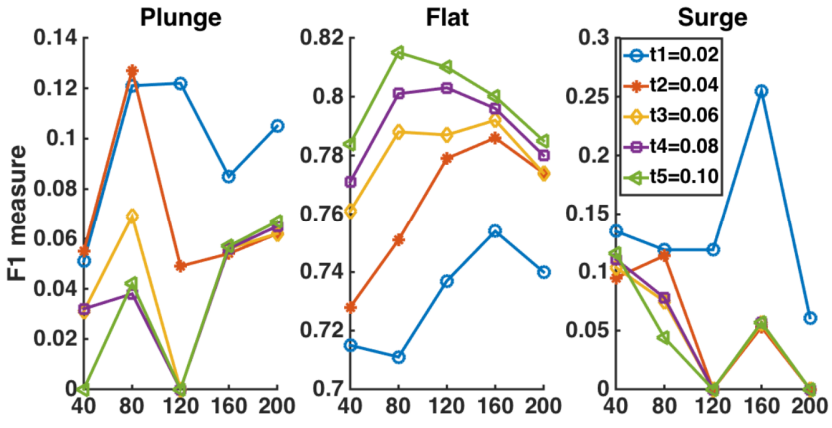
مقیاس F_1 پیش‌بینی ترکیب شده سه منبع به بزرگی 0.407 است در حالی که مقیاس F_1 پیش‌بینی دو منبع بیشینه برابر با 0.147 است، و امتیاز پیش‌بینی بیشینه اخبار مستقل تا حد پائین 0.138 است.

چگونه آستانه گذار t به شکل‌دهی توزیع کیفیت خودمان، در یادآوری و پیش‌بینی کمک می‌کند؟ برای حالت هموار، یک t بزرگتر اغلب F_1 بهتری را تولید می‌کند. در حالی که وضعیت غوطه‌ور و مواج است، t بزرگتر اغلب با F_1 کوچکتر همراه است. بنابراین انتخاب t یک مبادله بین معیار F_1 غوطه‌ور، مواج و هموار است.

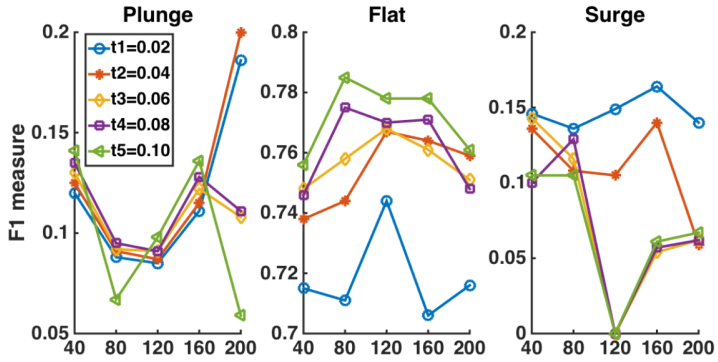
چگونه پنجره اندازه آموزش L با امتیازات مقیاس F_1 فرق دارد؟ تعریف می‌کنیم که آستانه گذار داده شده، پنجره اندازه یادگیری یا آموزش مختلف منجر به مقیاس F_1 مختلف با توجه به منابع مخلوط می‌شود، همانطور که در شکل 5(الف) دیده می‌شود. در مورد کلمبیا، بعنوان پیش‌بینی سه منبع، بهترین عملکرد در اندازه پنجره آموزش 70، هم در وضعیت غوطه‌ور و هم در وضعیت موی رخ می‌دهد. با اینحال، رابطه همیشه در میان منابع متصل مختلف متناسب و سازگار نیست. بنابراین یک بیانیه کلی درباره کارائی مقادیر پنجره آموزش شتاب‌زده است حتی اگر منابع ترکیب شده احتمالا برخی از روندهای بحث در بالا را نشان دهد، روابط احتمالا برای مثال خاص متفاوت است.

5.3 مطالعه موردی

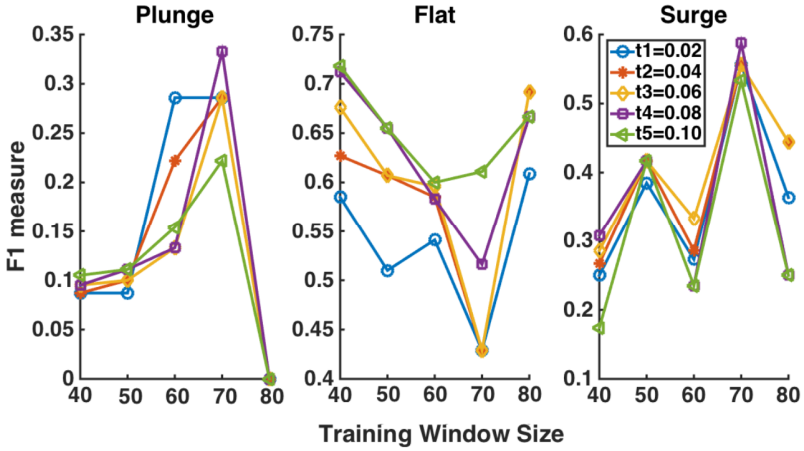
در این بخش اهمیت منابع را با نگاهی به مثال‌های مخصوص بعنوان مطالعات موردی اکتشاف می‌کنیم.



(الف) اخبار



(ب) اخبار و گوگل ترندز



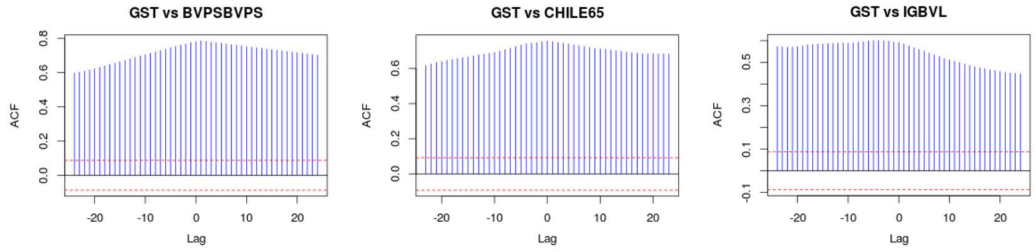
(ج) اخبار، گوگل ترندز و توئیت‌ها

شکل. 5. عملکرد پیش‌بینی کلمبیا، با مقادیر و اندازه‌های پنجره یادگیری و آستانه مختلف. محور X اندازه پنجره یادگیری و محور Y مقیاس F_1 را نشان می‌دهد. (الف) پیش‌بینی با استفاده از منبع مستقل اخبار. (ب) پیش‌بینی با استفاده از اخبار و گوگل ترندز. (3) پیش‌بینی با استفاده از اخبار، گوگل ترندز و توئیت‌ها.

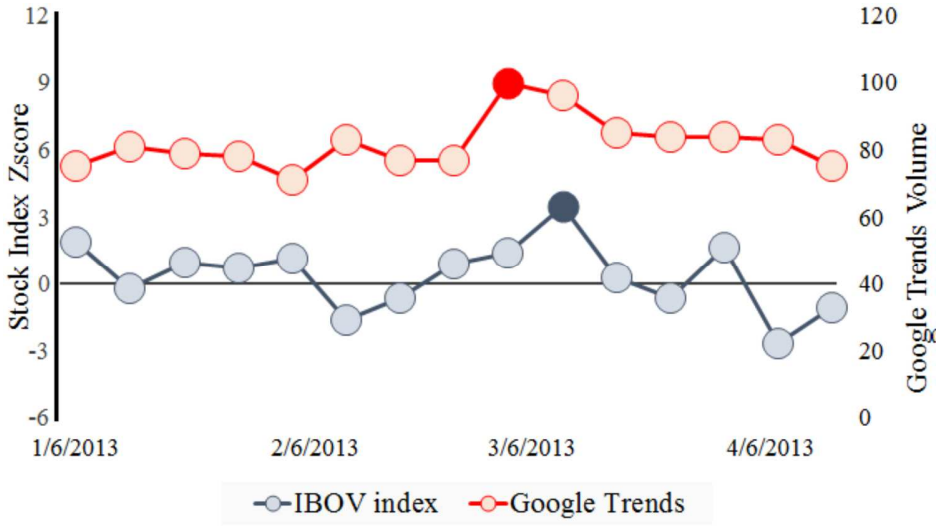
سرویس جستجوی گوگل ترندز. فرایند جستجوی گوگل ترندز گرایش به داشن یک همبستگی بالا با شاخص سهام سطح کشور دارد. چهار مثال از اصطلاحات روند جستجوی گوگل یا همان جستجوی گوگل ترندز (GST) را با شاخص سهام سطح کشور در شکل 6 ترسیم می‌کنیم. می‌توانیم امتیاز همبستگی بین GST و شاخص سهام را ببینیم مقداری که نزدیک به 0.8 است. همچنین اثبات شده است که گوگل ترندز یک شاخص پیشرو خوبی می‌باشد. شکل 7(الف)

همبستگی ویژگی گوگل ترندز با امتیاز- z (30) از شاخص سهام IBOV را نشان می‌دهد. همانطور که می‌بینیم، GST یک پیک محسوس را تنها در روزهای قبل از پیک واقعی در 27 ماه مارس سال 2013 را نشان می‌دهد.

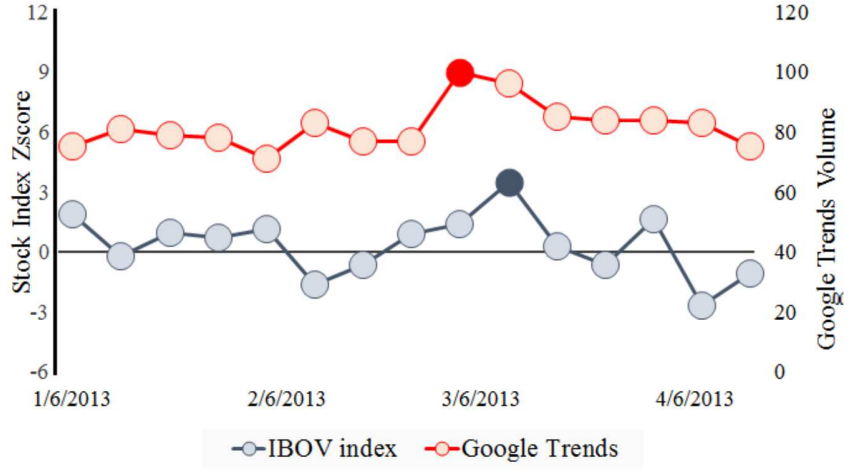
دیدگاه اخبار. دیدگاه‌های اخبار شاخص‌های بزرگی هستند و در برخی از موارد در تشخیص تغییرات ناگهانی در بازار سهام بسیار خوب هستند. همانطور که در شکل 7(ب) دیده می‌شود، به کمک یک سری از مقالات خبری از نهم ماه می سال 2012، از جمله این که «بولز برزیل تسلیم می‌شود چرا که مداخله دولت موجبات این امر را فراهم می‌کند»، رویکرد کاوش اخبار بصورت موفقیت‌آمیزی پیش‌بینی می‌کند که شاخص IBOV



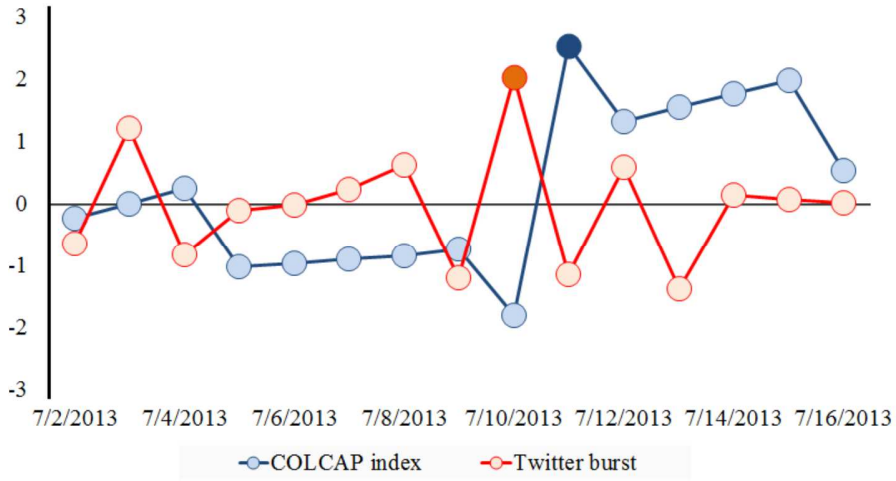
شکل. 6 همبستگی جستجوی گوگل ترندز (GST) با شاخص‌های سهام. محور-Y امتیاز همبستگی و محور-X زمان تاخیر (روز) را نشان می‌دهد.



(الف) گوگل ترندز



(ب) اخبار



(ج) burst توئیتر

شکل. 7. مثال‌های مطالعه موردی. خطوط آبی ارزش‌های واقعی بازار سهام را نشان می‌دهد. خطوط قرمز رنگ ارزش‌ها یا مقادیر پیش‌بینی شده را نشان می‌دهد. دایره‌های سیاه رویداد سیگا بالا را نشان میدهد. (الف) حجم جستجوی گوگل ترندز برای اصطلاحات انتخاب شده در حال گردش در اول ماه مارس سال 2003، روزهایی قبل از یک رویداد واقعی سیگما بالا در مقادیر امتیاز – z (30) برای شاخص IBOV برزیل در پنج ماه مارس. (ب) مقادیر IBOV را در برابر مقادیر اخبار پیش‌بینی شده بلومبرگ نشان می‌دهد. فهرست جعبات در نهم ماه می سال 2012 منتشر می‌شود، که به تولید رویداد ارزش رو به پائین پیش‌بینی شده در 14ام ماه می از 20ام ماه می کمک می‌کند که این در پیش‌بینی uptick در روز 21 ماه می استفاده می‌شود. (ج) ارزش‌ها برای burstهای توئیتر مواج در

10ام ماه جولای سال 2013 قبل از رویدا سیگما بالا در ارزش‌های امتیاز – z (30) برای شاخص COLCAP کلمبیا در 12ام ماه جولای را نشان می‌دهد.

تاثیرگذاری ناخوشایندی دارد. بطور مشابه، همچنین رویکرد قادر به پیش‌بینی uptick[3] در 21ام ماه می بر اساس اخباری با عنوان «برزیل در حال بازیابی رشد اقتصادی پس از نزول در سه ماه نخست است» منتشر شه در بیستم ماه می است. استفاده مهم از چندین مثال دیگر این موضوع بود که روش کاوش خبری ما برای پیش‌بینی بازار بسیار موثر است، بخصوص زمانی که اخبار کیفیت بالا در دسترس است.

رخدادهای burst توئیتر. در عین حال مطالعات موردی خودمان نشان می‌دهد که توئیتر یک منبع ارزشمند اطلاعاتی است. اخبار از طریق توئیتر نسبت به رسانه‌های خبری قدیمی بسیار سریع منتشر می‌شود. شکل 7(ج) مثالی را نشان می‌دهد که تاکید بر این دارد که چگونه burstهای توئیتر شاخص‌های تحرکات سهام را رهبری می‌کنند و، اگرچه نویزدار هستند، با اینحال اگر به درستی پردازش شوند می‌توانند یک شاخص زمان واقعی ارزشمند برای بازارهای مالی باشند. در آزمایش خودمان، روش تشخیص رخداد burst قادر به دستیابی به این تحرکات فوق‌العاده chatterتوئیتر و همچنین شناسایی موثر burstها از رویدادهای تاثیرگذار بر بازار است.

6. بحث

در این مقاله رویکرد جدیدی را برای پیش‌بینی بازار مالی با ترکیب منابع متعدد رسانه اجتماعی ارائه می‌کنیم. مدلسازی زبان، خوشه‌بندی موضوع و آنالیز عواطف را استفاده می‌کنیم تا نظرات و دیدگاه‌ها را از مقالات خبری استخراج کنیم، رگرسیون لاسو در داده‌های حجم جستجوی گوگل برای انتخاب ویژگی‌های بسیار آموزنده انتخاب می‌کنیم، و تکنیک-های گروه‌بندی رخداد و تشخیص پشت سر هم در داده‌های توئیتر را به کار می‌گیریم تا شاخص‌های بازار را از داده‌های توئیتر شناسایی کنیم. در نهایت، نتایج را با استفاده از مدل دلتا نایو بیز ارائه می‌کنیم. در این فرایند، همچنین نشان می‌دهیم که ترکیب منابع داده متعدد در مقایسه با هر ترکیبی از این منابع رسانه‌ای، عملکرد پیش‌بینی بهتری را به دست می‌آورد.

[3] چهره جدیدی از بازارهای معاملاتی تلفن همراه است.