**بسمه تعالی**

****

**پیشنهاده رساله دکتری**

**دانشکده مهندسی کامپیوتر، واحد مشهد، دانشگاه آزاد اسلامی**

**عنوان:**

**ارائه­ی مدل تطبیقی رفتار سرمایه­گذاران بازارهای مالی مبتنی بر شناسایی رویدادهای خبری**

**دانشجو:**

**سعیده انبایی فریمانی**

**استاد راهنما:**

**دکتر مجید وفایی جهان**

**فهرست مطالب**

1. [مقدمه 3](#_Toc14107634)
2. [بیـان مسـأله 4](#_Toc14107635)
3. [**ضرورت و اهميت پژوهش** 5](#_Toc14107636)
4. [اهداف تحقیق 7](#_Toc14107637)
5. [**پرسش‌های تحقیق** 7](#_Toc14107638)
6. [**فرضيه‌هاي پژوهش** 8](#_Toc14107639)
7. [**متغیرهای پژوهش** 8](#_Toc14107640)
8. [**ادبيات نظري و پیشینه پژوهش** 8](#_Toc14107641)

[1-8. ادبیات نظری 9](#_Toc14107642)

[2-8. بنیان مرور](#_Toc14107643) 25

1. [**روش‌شناسی تحقيق** 36](#_Toc14107644)
2. [**جنبه نوآوری پژوهش** 38](#_Toc14107645)
3. [نرم افزارها و بستر سخت افزاری مورد نیاز 39](#_Toc14107646)
4. [زمانبندی پژوهش 39](#_Toc14107647)

**فهرست شکل­ها**

[شکل 1 - نمودار پراکندگی روشها بر اساس سال انتشار 5](#_Toc14101590)

[شکل 2 - نمودار ابر کلمات کلیدی در مرور ادبیات 5](#_Toc14101590)

[شکل 3- پراکندگی روشها بر اساس نحوه ی واکشی اطلاعات 6](#_Toc14101591)

[شکل 4- انواع بازارهای مالی 9](#_Toc14101592)

[شکل 5- انواع بازارهای مالی مورد مطالعه در منابع مرور شده 10](#_Toc14101593)

[شکل 6 - پراکندگی شاخص معامله گری مورد مطالعه در مرور ادبیات 11](#_Toc14101594)

[شکل 7 - پراکندگی مبنای زمانی معامله مورد مطالعه در مرور ادبیات 11](#_Toc14101595)

[شکل 8 - پراکندگی منابع اطلاعاتی مورد استفاده در روشهای مرور شده 12](#_Toc14101596)

[شکل 9- نمودار روش سازماندهی متن در مرور ادبیات 13](#_Toc14101597)

[شکل 10- ابزارها و فرهنگ لغات مورد استفاده در تحلیل احساس 15](#_Toc14101598)

[شکل 11 - انواع روشهای تحلیل همبستگی میان دو سری زمانی در ادبیات تحقیق 16](#_Toc14101599)

[شکل 12 - انواع روش های یادگیری ماشین 21](#_Toc14101600)

[شکل 13- مراحل تحلیل عظیم دادهها و ارائه ی یک مدل پیشگو 22](#_Toc14101601)

[شکل 14- خواص داده های عظیم 23](#_Toc14101602)

[شکل 15- دسته بندی روش های تشخیص رویداد در بازارهای مالی 25](#_Toc14101603)

[شکل 16 - فلوچارت روش پیشنهادی 35](#_Toc14101604)

# مقدمه

با گسترش استفاده از اینترنت و رشد شبکه­های اجتماعی و همچنین اهمیت ویژه­ی اقتصاد بر رشدیافتگی ملت­ها، مطالعه­ی بر هم­کنش آن ها همواره دارای اهمیت می باشد. تا قبل از سال 1956 تصور می­شد سری­های زمانی مالی رفتاری تصادفی دارند. فاما در سال 1965، نظریه­ی بازار کارآمد را ارائه داد. او در این نظریه، تصادفی بودن رفتار سری­های زمانی مالی را رد کرد و نشان داد رفتار سری­های زمانی مالی تحت تاثیر اطلاعاتی است که در دسترس سرمایه­گذاران قرار می گیرد و قیمت سرمایه منعکس کننده تمام اطلاعات راجع به آن است. در پی ارائه­ی این نظریه، اقتصاد رفتاری پایه­گذاری شد[1, 2]. اقتصاد رفتاری به مطالعه­ی روانشناختی رفتارهای سرمایه­گذاران و نقش عوامل اجتماعی، فرهنگی و احساسی بر تصمیم­های تجاری آن­ها برای توجیه ناهنجاری­های بازار می­پردازد[3]. بر این اساس اغلب سرمایه­گذاران تحت تاثیر وقایع سیاسی، اقتصادی، اجتماعی و یا احساسی که کاربران شبکه­های اجتماعی در پیام­های ارسالی منعکس می­کنند، برای تجارت تصمیم­گیری می­کنند[4]. منابع اطلاعاتی مانند گروه­های خبری در وب، شبکه­های اجتماعی نظیر توئیتر[[1]](#footnote-1)، بوردهای گفتگو تخصصی بورس مانند استوک توئیت[[2]](#footnote-2) به یک منبع مهم تصمیم­گیری برای سرمایه­گذاران تبدیل شده­اند و استخراج دانش موجود در آن­ها می تواند به سرمایه­گذاران در اتخاذ تصمیم­های تجاری بهتر کمک کننده باشد[5].

رشد شبکه­های اجتماعی و حجم عظیمی از داده­های متنی که روزانه در اخبار و رسانه­های اجتماعی مالی منتشر می­شوند به یک منبع مناسب برای شناسایی عوامل اطلاعاتی موثر بر بازار تبدیل شده­اند[6]. از دیدگاه علم کامپیوتر، چالشی که در استخراج اطلاعات از چنین منابعی وجود دارد، بدون ساختار بودن داده­ها و حجم زیاد آن­هاست. آخرین فنون ارائه شده در متن­کاوی نظیر تکنیک جاسازی واژه­ها[[3]](#footnote-3) روش­هایی برای سازماندهی به این داده­ها و استخراج اطلاعات از آن­ها با توجه به حجم عظیم داده­ها ارائه داده­اند[7]. روش­های تحلیل احساس[[4]](#footnote-4) به عنوان یک فن مناسب از کاوش عقیده[[5]](#footnote-5) راجع به بازار به کار گرفته شده­اند[8]. در واقع این فنون به عنوان ابزاری برای استخراج اطلاعات موجود در رسانه­ها به عنوان ورودی مدل­های پیشگو مالی مبتنی بر یادگیری ماشین تبدیل شده­اند[9].

در طی سال­های گذشته تا کنون، استخراج دانش با هدف کشف الگو از یک مجموعه عظیم از داده­های وب به یک زمینه جذاب تحقیقاتی از دیدگاه صنعت تبدیل شده است[10]. تا کنون استخراج اطلاعات از متون وبی جهت کشف دانش در سه حوزه­ی تحلیل رفتار سرمایه­گذاران بر پایه­ی آزمون­های آماری، ارائه­ی سیستم­های توصیه­گر مبتنی بر روش­های هوش­ مصنوعی و ارائه­ی مدل­های پیشگو با استفاده از روش­های یادگیری ماشین به کار گرفته شده است. اما ماهیت پیچیده­ی بازارهای مالی پیش­بینی آینده بازار را بسیار دشوار نموده است، بطوریکه روش­های پیشگو مبتنی بر یادگیری ماشین دارای دقت مطلوبی نمی­باشند و با توجه به نیاز سرمایه­گذاران به سیستم­های پشتیبان تصمیم، تحقیق در این حوزه ادامه دار است.

# بیـان مسـأله

مدلسازی رفتار یک مسئله­ی پیچیده با ویژگی­های متغیر در زمان است[11]. یکی از عواملی که ممکن است سبب تغییرات شدیدی در بازارهای مالی شود، رفتار سرمایه­گذاران در پی رخداد رویدادهایی با موضوع­های متفاوت سیاسی، ملّی، اجتماعی و حتی فرهنگی است. از این رو ارائه­ی سیستمی مبتنی بر اقتصاد رفتاری که بتواند با شناسایی خودکار رویدادها از اخبار منتشر شده در گروه­های خبری، دقت پیش­بینی در بازارهای مالی را بالا ببرد مطلوب است. تا­کنون پژوهش­های زیادی با استفاده از روش­های متن­کاوی سعی بر سازماندهی متون بدون ساختار وبی و تحلیل احساس اسناد خبری برای ارائه­ی سیستم­های پیشگوی مالی داشته­اند [12-14]. روش­های زیادی به تحلیل احساس اخبار و پیش­بینی بر اساس آن پرداخته­اند[12-14] در حالیکه به اطلاعات پنهان میان اخبار مرتبط توجهی نداشته­اند. دسته­ی کمی از روش­ها با درنظر داشتن ارتباط محتوایی میان اسناد خبری سعی در شناسایی رویدادهای مهم داشته­اند[15-17]، در صورتیکه نتایج تحقیق­های[18-20] نشان می­دهد، علاوه بر رویدادهای مهم، رویدادهای متداول مانند گزارش­های ماهانه­ی بانک جهانی تاثیر خود را دارند و همچنین همه­ی خبرهای منتشر شده وزن تاثیر یکسانی ندارند و سرمایه­گذاران به اخبار کهنه[[6]](#footnote-6) و اخبار منفی واکنش مضاعفی نشان می­دهند. بنابراین بهتر است برای بهره بردن از اطلاعات موجود در محتوای خبر و ارتباط آن با سایر اخباری که به صورت جریانی[[7]](#footnote-7) رویداد­ها را گزارش می­دهند، اسناد خبری به صورت برخط[[8]](#footnote-8) دسته­بندی[[9]](#footnote-9) شوند [21].

در حقیقت پس از وقوع رویدادها، رسانه­ها با توجه به میزان اهمیت آن رویداد، رفتار مشخصی برای انتشار اسناد مربوط به آن دارند و سپس سرمایه­گذاران با دسترسی به این اخبار تحت تاثیر رفتار رسانه قرار می­گیرند و تجارت می­کنند. لذا برای مدلسازی رفتار رسانه­ها علاوه بر احساسی که یک سند خبری دارد میزان ارتباط آن با سایر اسنادی که در قبل با موضوع یکسانی منتشر شده­اند اهمیت دارد. آنچه که می توان در روش­های پیشگو مالی مبتنی بر اخبار از آن به عنوان یک شکاف تحقیقاتی یاد کرد عدم توجه به بعد زمان و پردازش جریانی دنباله­ای از اسناد خبری با موضوع یکسان است که رویدادی را با یک قطبیت احساس و الگوی انتشار[[10]](#footnote-10) مشخص گزارش می­دهند. بنابراین می­توان یک رویداد موثر بر بازار را یک دنباله از اسناد خبری مرتبط تعریف کرد بطوریکه این دنباله بر اساس موضوع، تحلیل احساس[[11]](#footnote-11) و الگوی انتشار، به اندازه­ی مشخصی سرمایه­گذاران را تحت تاثیر قرار دهد، سپس دنباله محو شود و تاثیرش نیز پس از مدتی محو شود. روش پیشنهادی در مطالعه­ی پیش رو در راستای بهبود دقت پیش­بینی بازار، سعی در مدلسازی رفتار سرمایه­گذاران مالی بر اساس شناسایی رویدادهای خبری دارد. طرح پیشنهادی سعی دارد از تحلیل احساس و فنون متن­کاوی به عنوان یک ابزار برای تحلیل محتوا و شناخت ارتباط میان اسناد خبری استفاده نماید بطوریکه بتواند میزان انرژی رویدادها برای تاثیر بر بازار را با توجه به رفتار رسانه­ها مدلسازی کند. و سپس بر اساس مدل ارائه­شده به پیش­بینی بپردازد.

# **ضرورت و اهميت پژوهش**

یکی از زمینه­های نوظهور کاربرد تحلیل داده­های عظیم در واکشی و استخراج اطلاعات[[12]](#footnote-12) مورد نیاز از داده­های متنی وبی و شبکه­های اجتماعی است[22]. بنا بر پیشگفته، اهمیت تدوین سیستم­های پشتیبان تصمیم با توجه به ماهیت پیچیده­ی بازارهای مالی با استفاده از تحلیل اخبار، بر هیچ کس پوشیده نیست. مرور انجام شده در این پژوهش مبتنی بر مطالعه­ی 90مرجع داخلی و خارجی در بازه­ی زمانی سال 2006 تا ژانویه 2019، انجام شده ­است. شکل 1 - نمودار پراکندگی روش­ها بر اساس سال انتشار را نشان می­دهد. و شکل 2 - نمودار ابر کلمات کلیدی در مراجع مرور شده را نشان می­دهد. تعداد پژوهش­های ارائه شده در استفاده از فنون متن­کاوی در اقتصاد رفتاری، که در شکل 1 قابل مشاهده است، در طی دو سال گذشته، رشد زیادی داشته است.

|  |  |  |
| --- | --- | --- |
| **شکل 1 - نمودار پراکندگی روش­ها بر اساس سال انتشار** | شکل 2 - نمودار ابر کلمات کلیدی در مراجع مرور شده |  |

از میان روش­های مرور شده دسته­ای از روش­ها با استفاده از تحلیل احساس به استخراج اطلاعات مورد نیاز در جهت طراحی سیستم­های پیشگو پرداخته­اند. در برخی از روش­ها فرایند تحلیل احساس از متون خبری استفاده شده است و در برخی تحلیل احساس بر داده­هایی با منشا شبکه­های اجتماعی انجام گرفته است. دسته­ی کمی از روش­ها با در نظر گرفتن این نکته که معمولا در پی وقوع یک رویداد مهم مانند خروج انگلستان از اتحادیه اروپا، سرمایه­گذاران در یک بازه­ی زمانی مشخص با طیف گسترده­ای از اخبار پیرامون این رویداد مواجه می­شوند، به استخراج اطلاعات از یک سلسله اخبار مرتبط با یکدیگر جهت شناسایی رویداد پرداخته­اند. در شکل 2 پراکندگی روش­ها بر اساس نحوه­ی واکشی اطلاعات گویای این نکته است.

شکل 2 پراکندگی روش­ها بر اساس نحوه­ی واکشی اطلاعات

بر مبنای این فرضیه که رویدادهای مهمی نظیر مخالفت کنگره با یک طرح نجات اقتصادی تا مدت زمان زیادی بازار را تحت تاثیر خود خواهند داشت[[13]](#footnote-13)و بخش بزرگی از برایند انرژی تاثیر گذار بر بازار را به خود اختصاص می­دهند و رویدادهای متداولی مانند گزارش­های ماهانه­ی بانک جهانی در کنار رویدادهای متداول دیگر بر رفتار سرمایه­گذاران تاثیر گذار خواهند بود، سوالی که به ذهن متبادر می­شود این نکته است که آیا با شناسایی برخط رویدادها، می­توان منبعی از اطلاعات را مدلسازی کرد که در کنار میزان بازگشت در بازار بتواند سبب بهبود دقت پیش­بینی در بازارهای مالی شود. در واقع آیا با شناسایی برخط رویدادها می­توان مدلسازی تطبیقی تاثیر آن­ها را به نحوی انجام داد بطوریکه همزمان اثر رویدادهایی با زمان تاثیر طولانی­تر[[14]](#footnote-14) را در کنار رویدادهایی با زمان تاثیر کوتاه­تر[[15]](#footnote-15) در کنار یکدیگر مدل­سازی کرد. محققین در طرح پژوهشی پیش­رو امیدوار هستند بتوانند با شناسایی رویدادها از جریان داده­ها به مدلسازی تطبیقی تاثیر آن­ها در مدت زمانی که توسط اخبار گزارش می­شوند، بپردازند.

# اهداف تحقیق

**ـ هدف کلی**

بهبود دقت پیش­بینی میزان بازگشت در بازار­های مالی مبتنی بر مدلسازی تطبیقی انرژی رویدادهای خبری

**ـ اهداف ويژه**

1. شناسایی برخط رویداد از جریان­داده­های خبری جهت کشف ارتباط پنهان میان اسناد خبری
2. مدلسازی تطبیقی انرژی رویدادها بر بازار با توجه به شناسایی برخط رویدادها، مبتنی بر تحلیل احساس و الگوی انتشار هر رویداد
3. تاثیر دیدگاه استخراج رویدادها بر وزن­دهی به اخبار برای مدلسازی میزان انرژی آن­ها
4. موثر بودن انرژی رویدادها بر بازار هدف در بازار مالی مورد بررسی
5. تاثیر رابطه­ی علّی رویدادها بر میزان بازگشت در بازار
6. اطلاع متغیر انرژی رویداد از آینده­ی بازگشت در بازار

# **پرسش‌های تحقیق**

1. آیا شناسایی رویداد باید متکی بر عنوان خبر انجام شود یا تحلیل بر کل متن خبر انجام شود؟
2. آیا دیدگاه استخراج رویدادها بر وزن­دهی به اخبار برای مدلسازی میزان انرژی آن­ها موثر است؟
3. میزان تاثیر خطای دسته­بندی برخط اسناد خبری بر مدلسازی میزان انرژی رویدادها چقدر است؟
4. آیا انرژی رویدادها در جامعه­ی مورد بررسی با میزان بازگشت در آن بازار دارای همبستگی بلندمدت است؟
5. آیا رابطه­ی علّی رویدادها بر میزان بازگشت در بازار وجود دارد؟
6. میزان اطلاعی که انرژی رویدادها از آینده­ی بازار هدف دارد چقدر است؟
7. آیا مدلسازی تطبیقی انرژی رویدادها بر بازار با توجه به شناسایی برخط رویدادها، مبتنی بر تحلیل احساس و الگوی انتشار هر رویداد می­تواند سبب بهبود دقت پیش­بینی گردد؟

# **فرضيه‌هاي پژوهش**

1. دیدگاه استخراج رویدادها بر وزن­دهی به اخبار برای مدلسازی میزان انرژی آن­ها موثر است.
2. میزان تاثیر خطای دسته­بندی برخط اسناد خبری بر مدلسازی میزان انرژی رویدادها ناچیز است.
3. انرژی رویدادها در جامعه­ی مورد بررسی با میزان بازگشت در آن بازار دارای همبستگی بلندمدت است.
4. در کوتاه مدت رابطه­ی علّی رویدادها بر میزان بازگشت در بازار وجود دارد.
5. انرژی رویدادها از آینده­ی بازار هدف اطلاع دارد.
6. مدلسازی تطبیقی انرژی رویدادها در کنار بازگشت در بازار سبب بهبود دقت پیش­بینی می­گردد.

# **متغیرهای پژوهش**

* متغیر مستقل : انرژی رویدادها که بر اساس برایند انرژی هر رویداد محاسبه خواهد شد.
* متغیر وابسته : میزان بازگشت در بازار
* عدم قطعیت­ها: خشه[[16]](#footnote-16) حاصل از رفتار بنیادگراها، خشه حاصل از روندجویان، خشه حاصل از خطای خوشه­بندی برای شناسایی رویدادها

# **ادبيات نظري و پیشینه پژوهش**

در این بخش، ابتدا به بیان مبانی نظری پژوهش و سپس مرور ادبیات می­پردازیم. در بخش مبانی نظری پژوهش، ابتدا به معرفی بازارهای مالی و شاخص­ها در آن پرداخته می­شود و سپس فنون سازماندهی به متون وبی بیان می­شود. در بخش بعد مبانی تحلیل سری­های زمانی مالی بیان می­شود و در نهایت تعاریف عظیم داد­ه­ها و ابزارهای تحلیل آن­ها از این جهت ارائه شده است که محقق برای مدلسازی عنوان نیازمند استفاده از ابزارهای تحلیل عظیم داده­[[17]](#footnote-17) است. در مرور ادبیات تحقیق، برخی از روش­ها سعی در ارائه­ی یک روش انتخاب ویژگی و یا بازنمایی متن داشته­اند. برخی یک روش نو در تحلیل احساس ارائه داده­اند و دسته­ای دیگر از فنون متن­­کاوی به عنوان یک روش بازیابی اطلاعات برای پیش­بینی و تحلیل رفتار استفاده کرده­اند. با توجه به این که دیدگاه محقق به مسئله­ی مدلسازی رفتار، توجه به ماهیت رویدادی اخبار و ارتباط میان اسناد خبری است، در این بخش تنها به مراجعی که دیدگاهی همسو با دیدگاه محقق داشته­اند استناد خواهد شد، و سپس روش­هایی که به بررسی رفتار سرمایه­گذاران و ارائه­ی مدل­های پیشگو با استفاده از فنون متن کاوی پرداخته­اند، بیان می­شوند و در نهایت به بررسی پژوهش­هایی که به زبان فارسی در کشور منتشر شده­اند، پرداخته می­شود.

## 1-8. ادبیات نظری

در این بخش مبانی نظری پژوهش تعریف و در هر مورد آماری مبتنی بر تحلیل ادبیات ارائه خواهد شد. با توجه به تاریخچه­ی کوتاه روش­های ارائه شده[[18]](#footnote-18)، هدف از بیان آمار نشان دادن گذشته و مسیر آینده­ی تحقیق در هر کدام از موارد ذکر شده است. آمار استخراج شده بر مبنای مرور ادبیات انجام شده در این تحقیق است و غالباً روش­های بررسی شده از دید علم کامپیوتر و فنون متن­کاوی به پیش­بینی و تحلیل رفتار در اقتصاد رفتاری پرداخته­اند. باید این نکته را یادآور شد که تاریخچه­ی کاربرد تحلیل اخبار در پیش­بینی بازارهای مالی به سال 1988 بر می­گردد. روش پیشنهادی کوتلر[[19]](#footnote-19) در سال 1988 را می­توان جزء اولین پژوهش­ها با رویکرد تحلیل فنی در بررسی تاثیر اخبار بر بازار بورس دانست[23]. این تحقیق با رویکرد تحلیل فنی به بررسی میزان تغییر در بازگشت سهام[[20]](#footnote-20) تحت تاثیر انتشار اخبار با موضوع­های مختلف پرداخته است. در تحلیل­های فنی اغلب از تعداد اخبار، گزارش­های دارای ساختار منابع بین­المللی نظیر بانک جهانی استفاده ­شده است و یا تحلیل اخبار توسط یک خبره­ی انسانی انجام گرفته است[2, 24-27] در حالیکه در تحلیل­های بنیادین از روش­های متن­کاوی برای سازماندهی به داده­ها استفاده می­شود و از آن­جا که در این تحقیق هدف تحلیل بنیادین اخبار می­باشد تنها به ارائه­ی آمار بر اساس روش­هایی که از فنون متن­کاوی استفاده نموده­اند بسنده شده است.

#### 1-1-8. بازارهای مالی

در میان پژوهش­های انجام شده، بازارهای مالی متفاوتی مورد بررسی قرار گرفته­اند. در هر کدام از بازارهای مالی نوع خاصی از سرمایه مورد داد و ستد قرار می گیرد. شکل 3- انواع بازارهای مالی را نشان می دهد[28].

شکل 3- انواع بازارهای مالی

به عنوان مثال نمونه ای ازCapital markets ، میتوان به بازار بورس Standard & Poor's 500که به طور خلاصه آن را S&P 500 می نامند و متشکل از سهام 500 شرکت بزرگ در آمریکا می باشد نام برد. شرکت هایی نظیر مایکروسافت[[21]](#footnote-21)و اپل[[22]](#footnote-22) در این لیست قراردارند. در بازار تبادل خارجی فارکس[[23]](#footnote-23) به مبادله ی پول بر اساس انواع جفت ارزهای خارجی مانند EUR/USD پرداخته می شود. در این بازار معاملات بر اساس نسبت دو جفت ارز به یکدیگر انجام می­گردد. شکل 4- انواع بازارهای مالی مورد مطالعه در منابع مرور شده را نشان می­دهد. بیشتر مطالعات انجام شده بر بورس نیویورک که شامل سه سهام عمده­ی DOW Jones, S&P 500, NASDAQ می­باشد، انجام شده­است.

شکل 4- انواع بازارهای مالی مورد مطالعه در منابع مرور شده

#### 2-1-8. اصطلاحات بازارهای مالی

مبنای معامله یک سرمایه­گذار می­تواند به صورت دقیقه، 15 دقیقه، 30 دقیقه، یک ساعت، چهار ساعت، روز، هفته و یا ماه باشد. اگر فرض کنیم مبنای معامله روز باشد به بیان تعریف اصطلاحات رایج در بازار پرداخته می­شود.

* قیمت باز[[24]](#footnote-24) : قیمتی که معاملات در ابتدای هر روز بر مبنای آن شروع می­شود.
* قیمت بستن[[25]](#footnote-25): قیمتی که معاملات در انتهای هر روز بر مبنای آن تمام می­شود.
* نرخ بازگشت[[26]](#footnote-26): میزان تغییر قیمت بستن نسبت به روز گذشته بر حسب درصد
* نوسان بازگشت[[27]](#footnote-27): انحراف معیار نرخ بازگشت
* حجم معاملات[[28]](#footnote-28): تعداد معاملات در یک روز

|  |  |
| --- | --- |
| شکل 5 پراکندگی شاخص معامله­گری مورد مطالعه در مرور ادبیات | شکل 6 پراکندگی مبنای زمانی معامله مورد مطالعه در مرور ادبیات |

با توجه به شکل 5 پراکندگی شاخص معامله­گری مورد مطالعه در مرور ادبیات را نشان می­دهد، بیشترین مطالعه بر نرخ بازگشت در بازار انجام شده است. و اغلب تاثیر رسانه بر رفتار سرمایه­گذارانی که بر مبنای روز معامله می­کنند، مطالعه شده است.

#### 3-1-8. منابع اطلاعاتی

اخبار، داده­های شبکه­های اجتماعی نظیر توییتر، آمار جستجوهای آنلاین در موتورهای جستجو، آمار مراجعه به صفحات ویکی پدیا و همچمنین بوردهای تخصصی گفتگو در بورس نظیر سینا[[29]](#footnote-29) ، انواع منابع داده­ای هستند که از طریق اینترنت در اختیار سرمایه­گذاران قرار می­گیرند. در میان روش­های بررسی شده برخی تنها از یک منبع داده­ای استفاده نموده­اند و برخی از ترکیبی از چند منبع داده­ای استفاده نموده­اند. شکل 7- پراکندگی منابع اطلاعاتی مورد استفاده در روش­های مرور شده را به تفکیک سال نشان می­دهد. همزمان با وقوع یک رویداد خبری، آمار مراجعه به صفحات ویکی­پدیا و تعداد جستجوها در گوگل با کلید واژه­های مرتبط با آن رویداد بیشتر می­شود و افراد به بیان نظرات خود در شبکه­های اجتماعی می­پردازند[29] و همین مسئله دلیلی بر روند رو به رشد مطالعات بر استفاده از چندین منبع داده­ای در طی چند سال گذشته می­باشد.

شکل 7- پراکندگی منابع اطلاعاتی مورد استفاده در روش­های مرور شده

#### 4-1-8. فنون سازماندهی به متن

متن­کاوی فرایند استخراج اطلاعات مفید از منابع داده­ای با استفاده از تکنیک­های شناسایی و استخراج الگوهای جذابی است که از تحلیل داده­های بدون ساختار استخراج می­گردند[30]. فرایند متن­کاوی دارای سه فاز پیش­پردازش متن، سازماندهی متن و استفاده از تکنیک­های داده کاوی جهت استخراج الگو می­باشد. فرایند پیش­پردازش شامل جداسازی نشانه­ها[[30]](#footnote-30)، حذف کلمات توقف[[31]](#footnote-31)و یکسان­کردن ریشه لغات[[32]](#footnote-32) و تعیین نقش نحوی واژه­ها می باشد. به دنبال پیشرفت در فنون متن­کاوی و افزوده­ شدن بر حجم داده­ی متنی در اینترنت، روش­های گوناگونی در نحوه­ی سازماندهی اسناد بدون ساختار متنی در این حوزه به کار گرفته شده است. در این میان می­توان به سه روش تجمیعی از واژه­ها[[33]](#footnote-33)[31] ، مدلسازی عنوان به روش تخصیص دریکله به مفاهیم پنهان[[34]](#footnote-34)[32] و جاسازی واژه­ها[[35]](#footnote-35) [7]اشاره کرد. نحوه­ی سازماندهی داده­ها پس از پیش­پردازش باید به گونه­ای انجام شود که در مواجهه با حجم زیاد نمونه­ها، روش یادگیری ماشین قادر به استخراج اطلاعات سودمند باشد[33]. روش مبتنی بر یادگیری عمیق با عنوان بردار برای هر واژه[[36]](#footnote-36) یا جاسازی واژه­ها در سال 2013 توسط میکولو یکی از محققان گوگل برای شرایطی که تعداد نمونه­های آموزشی بسیار زیاد باشد و در واقع با مسئله­ی عظیم داده­ها مواجه باشیم ارائه شد، در حالیکه روش تجمیعی از واژه­ها در چنین شرایطی با چالش­ ابعاد بالا و فضای خلوت ویژگی­ها مواجه می­شود و روش تخصیص دریکله به مفاهیم پنهان نیز با استفاده از تجزیه­ی ماتریس ویژگی­ها در چنین شرایطی بار محاسباتی بالایی پیدا خواهد کرد. شکل 8 نمودار روش سازماندهی متن در مرور ادبیات را نشان می­دهد. این نمودار نشان می­دهد از سال 2017 به بعد جهت­گیری روش­ها بر استفاده از تکنیک جاسازی واژه­ها می­باشد.

شکل 8 نمودار روش سازماندهی متن در مرور ادبیات

#### 5-1-8. مدلسازی عنوان[[37]](#footnote-37)

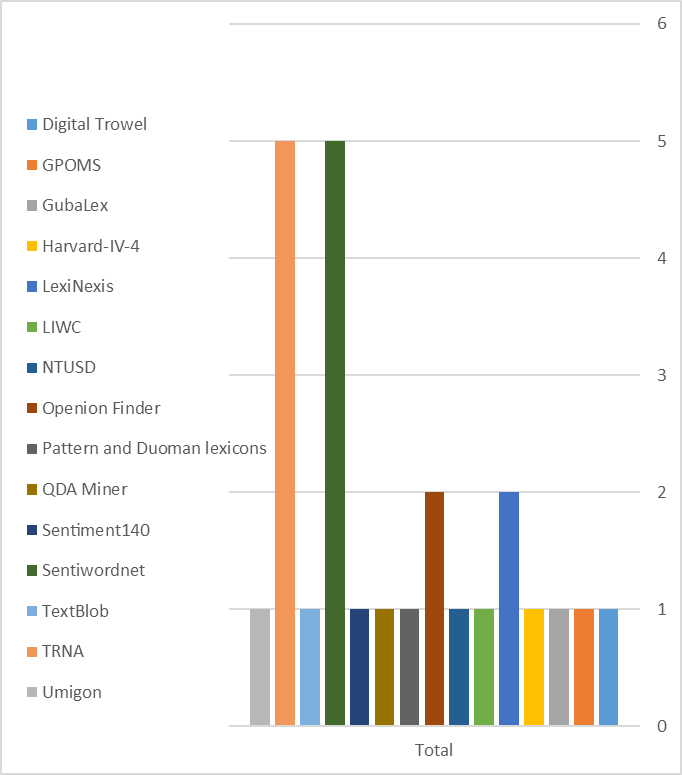
در یادگیری ماشین و پردازش زبان طبیعی، مدلسازی عنوان به معنایی شناسایی موضوع­هایی است که در یک پیکره­ی متنی موجود هستند. مدلسازی عنوان برای یافتن روابط معنایی پنهان در متون نیز به کار گرفته می­شود و از آن با عنوان مدلسازی مفاهیم معنایی[[38]](#footnote-38) نیز یاد می­شود. روش مدلسازی عنوان متکی بر فن بازنمایی متن است. معمولا از دو روش پایه­ی بازنمایی متن به صورت تجمیعی از واژه­ها و همچنین جاسازی واژه­ها برای آماده سازی متن جهت مدلسازی عنوان­ها استفاده می­شود. روش مدلسازی عنوان با توجه به ویژگی­های متفاوت این دو تکنیک را می­توان به دو دسته­ی مدلسازی عنوان با استفاده از روش­های آماری و یا مدلسازی عنوان احتمالی تقسیم کرد.

یکی از روش­های متداول در مدلسازی عنوان با تکیه بر ویژگی­های آماری رخداد واژه­ها در اسناد، روش تخصیص دریکله به مفاهیم پنهان یا به اختصار ال­دی­ای است. این روش در سال 2003 توسط بلی[[39]](#footnote-39) و سایرین در[34] ارائه شد. هدف اصلی این روش کاهش ابعاد ماتریس ویژگی­های و غلبه بر چالش خلوت بودن ماتریس ویژگی­ها در مدل تجمیع واژه­گان بوده است. آن­ها با تجزیه­ی ماتریس اسناد-لغات، به دو ماتریس لغات-عنوان­ها و عنوان­ها-اسناد، توزیع احتمال رخداد واژه­ها در هر عنوان یا مفهوم پنهان را به همراه توزیع احتمال رخداد هر عنوان در اسناد محاسبه نموده­اند. در صورت کاربرد این روش در تحلیل متن فرض می‌شود هر متن دارای ترکیبی از چند مفهوم پنهان است و هر مفهوم پنهان نیز دارای توزیع احتمالی روی واژه‌‌های آن پیکره است بنابراین، واژه‌‌های مشاهده شده در هر مفهوم پنهان متغیرهای چندجمله‌ای هستند و مفاهیم متناسب نیز متغیرهای ال‌دی‌ای با توزیع دریکله هستند . در این روش با استفاده از توزیع دریکله و توزیع چندجمله‌ای به تعیین میزان تاثیر هر واژه از پیکره در هر مفهوم پنهان و همچنین توزیع احتمال رخداد مفاهیم پنهان در متون پرداخته می‌شود. روش ال‌دی‌ای به واژگان موجود در یک متن وبی با توجه به فراوانی تکرار آنها اهمیت داده و آن­ها را با توجه به فراوانی تکرار در مفاهیم توزیع می‌کند و ترتیب رخداد واژه­ها در کنار هم لحاظ نمی­شود.

دسته­ی دوم روش­هایی هستند ک با استفاده از تکنیک یادگیری عمیق و مبتنی بر ویژگی احتمالی رخداد واژه­ها در کنار یکدیگر به مدلسازی عنوان پرداخته­اند. روش پیشنهادی [35-37] پس از بازنمایی واژه­ها به روش جاسازی واژه­ها توسط یک شبکه یادگیری عمیق، به خوشه بندی فضای برداری حاصل شده از خروجی فاز بازنمایی لغات پرداخته­اند و هر خوشه را معادل یک عنوان در نظر گرفته­اند. از جمله ویژگی­های روش جاسازی واژه­ها تولید بردارهایی با شباهت زیاد برای لغاتی است که معمولا در کنار یکدیگر در یک سند متنی ظاهر می­شوند. این دسته روش­ها با استفاده از این ویژگی روش جاسازی واژه­ها به خوشه بندی بردارهای تولید شده پرداخته­اند. نتایج ارزیابی عناوین استخراج شده بر اساس این روش­ها نشان می­دهد معیار F1 نسبت به ال­دی­ای دارای بهبود است.

#### 6-1-8. تحلیل احساس

فرایند تحلیل احساس شامل طبقه­بندی[[40]](#footnote-40) سند متنی به سه دسته احساس مثبت و منفی و خنثی می باشد. اغلب از دو روش مبتنی بر فرهنگ لغات[[41]](#footnote-41) دارای برچسب و روش مبتنی بر یادگیری ماشین[[42]](#footnote-42) جهت برچسب دهی خودکار به منابع متنی مانند اخبار یا پست­های افراد در شبکه­های اجتماعی استفاده می­شود. در روش برچسب دهی دستی هر خبر توسط یک خبره در اقتصاد برچسب داده می­شود، اما در روش­های خودکار معمولا از یک فرهنگ لغات دارای برچسب و یا یک روش یادگیری ماشین فرایند برچسب دهی طی می­شود. شکل 10 ابزارها و فرهنگ لغات دارای برچسب که در حوزه­ی اقتصاد رفتاری استفاده شده­اند را نشان می­دهد.



شکل 9 ابزارها و فرهنگ لغات مورد استفاده در تحلیل احساس

#### 7-1-8. تحلیل همبستگی[[43]](#footnote-43)

در سری­های زمانی مالی تحلیل همبستگی ممکن است بر روی یک سری زمانی با گذشته­ی خودش[[44]](#footnote-44) و یا دو سری زمانی با یکدیگر انجام شود. از آن­جا که هدف پژوهش مطالعه­ی رابطه­ی میان دو سری زمانی انرژی رویدادها و میزان بازگشت در بازار است در اینجا به بیان مبانی نظری تحلیل همبستگی دو سری زمانی اشاره می­شود. روش­های متداول تحلیل همبستگی مانند همبستگی پیرسون[[45]](#footnote-45) که میزان همبستگی خطی[[46]](#footnote-46) دو سیگنال را تعیین می­کند و یا روش­های تحلیل همبستگی متقابل[[47]](#footnote-47) دو سری زمانی که میزان جابجایی مکانی یک سیگنال نسبت به سیگنال دیگر را می­سنجد از این دسته هستند. شکل 10 انواع روش­های تحلیل همبستگی میان دو سری زمانی در ادبیات تحقیق را نشان می­دهد. بیشترین استفاده از روش­های تحلیل همبستگی خطی متداولی مانند پیرسون و تست علّیت گرنجر[[48]](#footnote-48) که یک رابطه­ی خطی علّی بین دو سری پیدا می­کند، انجام شده است.

شکل 10 انواع روش­های تحلیل همبستگی میان دو سری زمانی در ادبیات تحقیق

در مطالعه­ی رابطه­ی وابستگی خطی بین دو سری زمانی از تحلیل همبستگی استفاده می­شود در حالیکه در راستای مطالعه­ی نوع رابطه­ی غیرخطی دو سری زمانی (هم­رفتاری[[49]](#footnote-49)، رابطه­ی علّی[[50]](#footnote-50)، رابطه­ی علّی متقابل) می­توان از معیار­های مبتنی بر نظریه­ی اطلاعات استفاده نمود. این مزیت با محاسبه­ی توزیع احتمال توام دو متغیر تصادفی مبتنی بر نظریه­ی اطلاعات و نظریه­ی احتمال شرطی امکان پذیر می­شود.

#### 8-1-8. تحلیل استقلال و وابستگی در نظریه­ی اطلاعات

در راستای بهبود دقت در مدل­های پیشگو، درک ارتباط (وابستگی(علّی، متقابل، هم­رفتاری) یا استقلال) میان متغیرها اهمیت فراوانی دارد[38]. در نظریه­ی آمار و احتمال، یک رابط[[51]](#footnote-51) به عنوان یک تابع توزیع تجمعی چندمتغیره[[52]](#footnote-52) مطرح می­شود بطوریکه توزیع حاشیه­ای هر کدام از متغیرها در آن یکنواخت باشد. از این تابع توزیع تجمعی برای توصیف وابستگی میان متغیرها استفاده می­شود بطوریکه متغیرهایی مستقل دارای توزیع حاشیه­ای یکنواخت در مقابل یکدیگر می­باشند و با هم یک رابط را تشکیل می­دهند[39]. در مورد متغیرهایی با ماهیت پیوسته درک نوع توزیع حاشیه­ای داده­ها در مثال­های واقعی یک مسئله­ی پیچیده است، اما در مورد متغیرهایی با ماهیت گسسته که از یک توزیع چندجمله­ای[[53]](#footnote-53) پیروی می­کنند، تحلیل وابستگی ساده­تر است بطوریکه می­توان از توزیع شرطی آن­ها در مقابل هم استفاده کرد[40]. در راستای پاسخگویی به وابستگی میان بازار مورد مطالعه و انرژی محاسبه شده از رویدادهای خبری در بلند مدت می­توان از تست آماری رابط استفاده نمود[41]. پس از تعیین وابستگی میان دو سری زمانی، مطالعه­ی نوع ارتباط آن­ها در بهبود دقت پیش­بینی موثر است. گسسته­سازی سری­­های زمانی، فرایند مطالعه­ی نوع رابطه­ی آن­ها که می­تواند به صورت یک رابطه­ی علّی یک طرفه و یا حتی متقابل[[54]](#footnote-54) و یا یک رابطه­ی هم­رفتاری باشد را با استفاده از فنون مبتنی بر نظریه­ی اطلاعات ساده­تر می­کند. البته ممکن است در فرایند گسسته سازی، میزان دقت در توزیع اولیه داده­ها کمتر شود که تا کنون راه­حل­هایی برای رفع این مسئله ارائه شده است.

نظریه­ی اطلاعات، در سال 1949 توسط شانون مطرح شد[42]. معیار کلیدی در نظریه­ی اطلاعات آنتروپی است که میزان عدم قطعیت در یک متغیر تصادفی را می­سنجد. برای مثال در یک سکه با میزان احتمال برابر رخداد پشت و رو، هیچ میزانی از عدم قطعیت وجود ندارد. از جمله معیارهای مهم در بررسی میزان اطلاع دو متغیر تصادفی از یکدیگر، معیار اطلاعات متقابل[[55]](#footnote-55) و اطلاعات متقابل نقطه­ای­­[[56]](#footnote-56) بر مبنای آنتروپی است. معیار واکشی اطلاعات متقابل با محاسبه­ی مقدار متوسط اطلاعات متقابل نقطه­ای بر توزیع حاشیه­ای آن­ها در قبال یکدیگر، به بررسی میزان وابستگی متقابل بین دو متغیر تصادفی و وجود رابطه­ی تقابل میان آن­دو می­پردازد در حالیکه معیار واکشی اطلاعات متقابل نقطه­ای به محاسبه­ی میزان اطلاع یک متغیر از دیگری و وجود یک مسیر علّی از یکی به دیگری می­پردازد. تا کنون از این دو معیار در محاسبه­ی میزان اطلاع در سری­های زمانی چند متغیره نیز استفاده شده است[43].

یکی از مطالعات مهم در سری­های زمانی مالی مطالعه­ی صعود و نزول در سری زمانی است. یک صعود به معنای منتقل شدن از یک می­نیموم محلی تا ماکزیموم محلی بعدی است و نزول بلعکس می­باشد. .در ادامه به بیان برخی شاخص­ها در تحلیل وابستگی دو سری زمانی مالی بر اساس صعود و نزول در آن می­پردازیم. لازم است یادآور شویم با توجه به نوع مطالعه، تا کنون محققین فنونی را برای گسسته سازی سری زمانی در مطالعه­ی سطوح مختلف صعود و نزول ارائه داده­اند. برای مثال در روش [43]، به بررسی این فرضیه که آیا وجود چند نوسان کوچک پشت سر هم دلیلی بر بروز یک شوک شدید در در نوسان بازار خواهد شد، پرداخته است، و گسسته سازی با توجه به همین هدف با 3 الفبای صفر،یک و دو انجام پذیرفته است. صفر برای نوسان منفی، یک برای نوسان مثیت و 2 برای نوسان غیر عادی ارائه شده است.

1. آنتروپی

(1)

(2)

(3)

1. آنتروپی فرایندی

برای یک فرایند مارکوفی دارای توزیع سکون آنتروپی فرایندی به شکل زیر تعریف می­شود.

(4)

(5)

1. تخمین آنتروپی

روش­های مختلفی از جمله تخمین ماکزیموم درستنمایی برای تخمین میزان آنتروپی ارائه شده است . یکی از روش­های تخمین آنتروپی در نظر گرفتن توزیع Binned روی متغیرها و گسسته سازی بر اساس همین توزیع است.

(6)

(7)

ثابت شده است در نظر گرفتن این توزیع یک مقدار بایاس منفی روی مقدار تخمینی آنتروپی نسبت به مقدار واقعی قرار می­دهد که گرسبرگر[[57]](#footnote-57) یک تابع تخمین آنتروپی برای این منظور ارائه داده است[44].

( (8)

(9,10)

1. اطلاعات متقابل

میزان کاهش عدم قطعیت راجع به متغیر تصادفی x با وجود y را می­سنجد. به عبارتی این معیار تعیین می­کند برای دانستن y به چه اندازه­ای از تعداد سوالات با پاسخ بله یا خیر برای دانستن مقدار متغیر تصادفی x کاسته می­شود.

(11)

از این معیار می­توان با در در نظر گرفتن یک طول بلاک مشخص، برای مطالعه­ی Lag در سری­های زمانی نیز استفاده نمود. البته نمی­توان از این معیار برای تعیین جهت جریان اطلاعات در سری­های زمانی استفاده کرد.

(12)

1. آنتروپی منتقل شده

در مطالعه­ی جهت جریان اطلاعات میان دو سری زمانی بلاک بندی شده از معیار آنتروپی منتقل شده استفاده می­شود. با استفاده از این معیار می­توان میزان اطلاعات بدست­ آمده از گذشته­ی x و گذشته­ی y منتقل شده به x را سنجید. ثابت شده است در متغیرهایی با توزیع گوسی میزان آنتروپی منتقل شده از گذشته مرتبه t با میزان علّیت گرنجر یکی می­شود.

(12)

(13)

1. آنتروپی منتقل شده­ی موثر

برای حذف تاثیر خشه موجود در سری­های زمانی بر میزان آنتروپی منتقل شده می­توان با درهم ریزی داده­ها هر گونه وابستگی میان گذشته و حال سری زمانی را تخریب کرد و از معیار آنتروپی منتقل شده­ی موثر استفاده کرد.

(14)

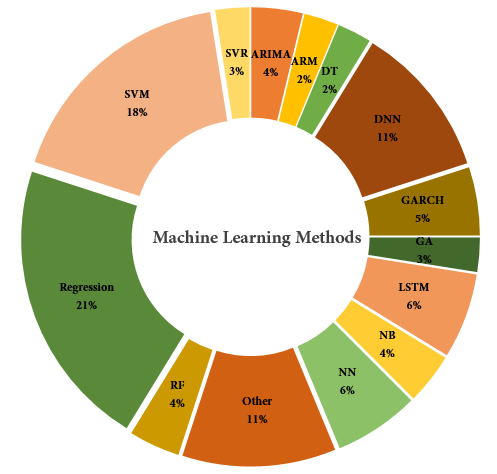
1. مقدار اطلاعات نسبی بدست آمده در X از گذشته­ی Y

(15)

معیارهای معرفی شده در این بخش، می­توانند در راستای پاسخگویی به سوالات تحقیق استفاده شوند.

#### 9-1-8. یادگیری ماشین

وضعیت بازارهای مالی با منتشر شدن اطلاعات و اخبار جدید در وب تغییر می کند و بسیاری از سیستم­های پیشگو ارائه شده با استفاده از روش­های یادگیری ماشین و تحلیل متن در جهت بهبود دقت مدل­های پیشگو و کاهش ریسک تجارت ارائه شده­اند. مدل­های پیشگو در حوزه­ی اقتصاد رفتاری با استفاده از روش­های یادگیری ماشین مختلفی به پیش­بینی آینده بازار بر اساس اخبار و پارامترهایی نظیر بازگشت در بورس، قیمت و یا حجم معاملات پرداخته­اند. شکل 11 انواع روش های یادگیری ماشین در مرور را نشان می­دهد.



شکل 11 انواع روش های یادگیری ماشین

در این حوزه اغلب از روش­های پیشگو­ی ایستا، نظیر رگرسیون خطی و یا رگرسیون پویا نظیر ARIMA و GARCH استفاده شده است و بعد از آن ماشین بردار پشتیبان و شبکه­های عصبی و شبکه­های یادگیری عمیق نظیر LSTM استفاده شده است. روش­هایی نظیر شبکه­ی بیزی پویا که قادر به مدلسازی رفتار متغیر با زمان سیستم­های پویا هستند در این حوزه بسیار کم استفاده شده است که می­توان آن را ناشی از پیچیدگی روش­های تخمین پارامترهای شبکه دانست اما از آن­جا که این روش قادر به مدلسازی ارتباط غیرخطی متغیر در زمان میان متغیر­ها با استفاده از توزیع احتمال توام آن­هاست، برای محقق جذاب است.

#### 10-1-8. عظیم داده­ها، ابزارها و چالش­ها

در عصر حاضر با فراگیر شدن استفاده از شبکه­های اجتماعی و منابع موجود در وب، عظیم داده­ها به یک سرمایه ارزشمند برای سازمان­ها، دولت­ها و افراد تبدیل شده اند. شناخت خواص[[58]](#footnote-58) وکیفیت داده­های عظیم و چگونگی تحلیل[[59]](#footnote-59) آن­ها دو عامل مهم در ارزشمند کردن داده می­باشد. تحلیل داده­های عظیم به عنوان مرز دانش امروزی، در بسیاری از زمینه­ها از جمله مدل­های پیشگو مالی مبتنی بر داده­های شبکه­های اجتماعی و متنی در وب کاربرد دارد.

شکل 12مراحل تحلیل عظیم داده­ها و ارائه ی یک مدل پیشگو را نشان می­دهد. مدیریت پایگاه داده­ی حجیم شامل انتخاب نوع پایگاه داده و یا نگهداری داده­ها در بستر توزیع شده و تعیین سلسله مراتب برای ذخیره سازی داده­ها می­باشد. لحاظ کردن سطحی از هر کدام از خواص عظیم داده­ها، با توجه به نوع نیاز، شناسایی صفات داده­های ناهمگن در هر کدام از منابع داده­ای مانند داده­های بورس و اخبار، پیش پردازش و تمیز سازی داده ها شامل حذف خشه، اعتبارسنجی داده ها و یکسان سازی داده­ها و در نهایت یکپارچه سازی داده­ها با استفاده از طراحی و بهینه سازی پرس­وجو­هایی از منابع داده­ای مختلف، پنج گام قبل از آموزش مدل پیشگو می­باشند[45].

شکل 12مراحل تحلیل عظیم داده­ها و ارائه ی یک مدل پیشگو

از جمله ویژگی­های داده ها در شبکه­های اجتماعی و بازارهای مالی تطابق با خواص داده های عظیم می­باشد. در [46] خواص داده­های عظیم تحت 5V شناخته شده است که این ویژگی­ها شامل حجم[[60]](#footnote-60) داده، سرعت[[61]](#footnote-61) تولید، ارزش[[62]](#footnote-62) داده، تنوع[[63]](#footnote-63) داده ، راستی[[64]](#footnote-64) می­باشند. می توان ویژگی ارتباطات[[65]](#footnote-65) را نیز به این خواص افزود و آن را تحت 6V شناخت، که در بندهای زیر هر کدام به همراه ابزار مورد نیاز توضیح داده شده است.

شکل 13خواص داده­های عظیم

* **حجم** : با گسترش شبکه های اجتماعی منابع داده­ای به سرعت در حال افزایش هستند و حجم داده­ی تولیدی در سازمان­ها در حال رشد می باشد. و لذا سازمان­ها برای مقابله با چالش گسترش حجم داده باید به سراغ بستر توزیع شده­ای نظیر هادوپ[[66]](#footnote-66) به عنوان یک اکوسیستم توزیع شده مدیریت داده­های حجیم بروند و یا از بستر تنسورفلو برای مدیریت پردازه­ها بر محیط چندپردازه­ای GPU و زبان برنامه­نویسی پایتون استفاده کرد.
* **سرعت** : داده­ها در پلت­فرم­های مالی نظیر بورس به سرعت تولید می­شوند و به سرعت تغییر می­کنند. در این راستا باید سرعت ذخیره سازی و پردازش داده ها نیز افزایش پیدا نماید. لذا به جای پردازش دسته ای[[67]](#footnote-67) پردازه­ها، باید به سراغ ابزارهای پردازش جریانی داده­ها برویم که به عنوان مثال می­توان به Storm، Spark و Flink اشاره کرد. ابزارهای ذکر شده در کنار مکانیزم های پردازش دسته­ای یک موتور پردازش جریانی نیز ارائه داده­اند.
* **تنوع**: با گسترش منابع داده­ای، داده با توجه به منبعی که از آن استخراج می­شود ممکن است به صورت دارای ساختار مانند یک فایل تصویر و یا بدون ساختار مانند یک سند متن باشد، لذا روش تحلیل داده باید قادر به سازماندهی بر اساس تنوع آن باشد. در نحوه­ی ذخیره سازی داده­های بدون ساختار، پایگاه داده­های بدون ساختاری مانند MongoDB، Cassandra ارائه شده­اند که از ساختار کلید-مقدار در ذخیره سازی داده­های عظیم استفاده می­کنند.
* **راستی**: به کیفیت داده و کیفیت منابع داده­ای بر می­گردد. داده می تواند دارای خشه یا بایاس باشد همچنین ممکن است مقادیر گم شده در آن وجود داشته باشد و یا دارای عدم قطعیت باشد. در تحلیل داده اولین گام بررسی اصالت منابع تولید داده است. این که آیا منبعی که داده را تولید کرده است مورد اعتماد است یا خیر. یکی از زمینه­های تحقیقاتی مهم آینده در اقتصاد رفتاری بررسی تاثیر شایعات و اخبار جعلی بر بازار و سرمایه­گذاران می­باشد.
* **ارتباطات**: به میزان ارتباطات مستقیم به ارتباطات ممکن در یک شبکه گویند. مسئله ی مهم، این نکته است که با گذشت زمان و برقراری ارتباط بین افراد در شبکه­های اجتماعی این نسبت رشد می کند. افزایش این نسبت سبب گسترده شدن و داغ شدن یک رفتار در شبکه می شود. مطالعه­ی رفتار کاربران شبکه­های اجتماعی که با یک الگوی مشخص اطلاعات را نفوذ می­دهند، مطلوب است[47].
* **ارزش**: توجه به هر کدام از ویژگی­های ذکر شده به نوع کسب و کاری که با داده­ی عظیم مواجه است بستگی دارد و این سازمان است که سطح تنوع را با توجه به هزینه و نیازمندی­های خودش تعیین می­کند و با در نظر­گرفتن نیاز تجاری، عظیم داده را به یک منبع ارزشمند تبدیل می­کند

# 2-8. بنیان مرور

در 29 سپتامبر 2008، بعد از آن که کنگره­ی آمریکا طرح نجات اقتصادی رئیس­جمهور جرج بوش را رد کرد، در عرض یک روز ارزش سهام DOW Jones Industrial به مقدار 777.68 پوینت سقوط کرد[[68]](#footnote-68). نزول ارزش سهام شرکت­های آمریکایی تحت تاثیر این رویداد تا حدود 18 ماه به طول انجامید. این رویداد و انتشار اخبار مربوط به آن و نظرات افراد در شبکه­های اجتماعی، سبب تحت تاثیر قرار دادن سرمایه­گذاران و چنین نزول شدیدی در ارزش سهام شرکت­های آمریکایی گردید. در واقع اخبار، رویدادهایی با موضوعات سیاسی، فرهنگی، اجتماعی را گزارش می­دهند و به دنبال وقوع این رویدادها افراد نظرات خود را در شبکه­های اجتماعی منعکس می­کنند.

#### 1-2-8. شناسایی رویداد

در مرور روش­های موجود در شناسایی رویداد­های موثر بر بازارهای مالی با دو دسته­ی کلی از روش­ها مواجه می شویم. شکل 14- دسته بندی روش های تشخیص رویداد در بازارهای مالی را نشان می­دهد. دسته­ی اول فنونی هستند که با اتکا بر الگوی متمایز فراوانی تکرار دسته­ای از واژه­ها و یا شاخص احساس و حجم و تعداد توئیت منتشر شده در هر برش زمانی به شناسایی رویدادهای مهم پرداخته­اند. دسته­ی دوم به ارتباط میان اسناد خبری توجه نموده­اند و بر اساس تحلیل محتوای اسناد سعی در دسته­بندی اسناد داشته­اند. اغلب روش­ها در این حوزه به تحلیل رفتار مشتریان بازارها پرداخته­اند و تعداد کمی مدل پیشگو ارائه ­داده­اند. این نکته را می­توان ناشی از این موضوع دانست که در ارائه­ی مدل پیشگو نمی­توان تنها به نقش رویداد­های مهم اکتفا کرد، و وضعیت موجود بازار و رویدادهای متداول نیز در پیش­بینی اهمیت دارند.

شکل 14- دسته بندی روش های تشخیص رویداد در بازارهای مالی

در دسته ی اول، روش مبتنی بر تحلیل احساس [16] با استفاده از یک شاخص تجمعی بر اساس تحلیل احساس به شناسایی رویداد و مدلسازی تاثیر آن با استفاده از تست علّیت گرنجر[[69]](#footnote-69) پرداخته است. آن­ها یک سری زمانی از 8 سطح از احساس را در هر روز با استفاده از اخبار چندین گروه خبری، به ازای هر کمپانی کره­ای محاسبه نموده­اند و با استفاده از روش گرنجر به بررسی فرضیه­ی موثر بودن اخبار بر تغییرات بازار پرداخته­اند. نتایج این تحقیق نشان می­دهد در هر کدام از تغییرات شدید احساس، یک رویداد مهم رخ داده است. همچنین در روش پیشنهادی [17] با این فرضیه که اغلب افراد رویدادهای واقعی را در توئیتر گزارش می­کنند به شناسایی رویدادها از داده­های گروه­های مالی[[70]](#footnote-70) در توئیتر پرداخته است. با توجه به وجود خشه زیاد در داده­های توئیتر، هدف این مقاله حذف توئیت­های خشه و استخراج وقایع مرتبط با سهام DOW Jones از توئیت­ها می­باشد. روش پیشنهادی آن ها از جمله روش­های تحلیل احساس مبتنی بر فرهنگ لغات می باشد که بر اساس آن توئیت­ها با توجه به تعداد لغات مثبت یا منفی، برچسب مثبت یا منفی گرفته­اند و شناسایی رویداد بر اساس تعداد توییت منتشر شده در هر روز انجام گرفته است بطوریکه در صورتی که رتبه­ی اختصاص یافته به توییت­های هر روز از حد آستانه ای بیشتر باشد، به عنوان یک رویداد تلقی می­گردد. در واقع در این دسته بر اساس یک حد آستانه از رتبه­ی احساسی استخراج شده، رویداد­ها شناسایی شده­اند و به نوع رویداد و الگوی انتشار آن توجهی نشده است. در روش مبتنی بر شناسایی رویداد با توجه به الگوی متمایز فراوانی واژه­ها [20] به شناسایی رویداد­های اصلی و فرعی بر بازار گوشت گاو در ایرلند پرداخته­اند. آن ها با بیان این مطلب که لغات در رویدادهای اصلی نظیر زلزله و جنگ تا قبل از وقوع رویداد در پیکره­ی متنی فرکانس تکرار بسیار کم یا نادری دارند اما بعد از وقوع رویداد، در بازه ای کوتاه به شدت تکرار می شوند معیاری با استفاده از نسبت درستنمایی ماکزیموم ارائه دادند، و از معیار اطلاعات متقابل نقطه­ای[[71]](#footnote-71) برای دسته بندی لغات هم توزیع برای شناسایی رویدادهای اصلی پرداخته­اند. و با این فرضیه که رویدادهای فرعی با لغاتی مرتبط با بازار گزارش می شوند که این لغات در پیکره ی متنی دارای الگوی تکرار متمایزی نمی­باشند، پس از محاسبه­ی فراوانی تکرار برای هر لغت در پیکره و برچسب­دهی بر اساس وضعیت بازار در بازه یک هفته؛ با استفاده از دسته­بند ماشین بردار پشتیبان به تشخیص رویدادهای فرعی پرداخته­اند. همچنین در روش پیشنهادی [48] یک روش تشخیص رویداد­های آشوبناک[[72]](#footnote-72) در بازار با استفاده از تحلیل داده­های توئیتر ارائه شده است. آن ها از فاکتورهایی نظیر حجم توئیت، متن توئیت، تحلیل احساس توئیت­ها و منطقه­ی جغرافیایی که توئیت در آن منتشر شده است برای تعیین میزان نفوذپذیری توئیتر بر سرمایه­گذاران استفاده نموده­اند. بررسی خود را بر رویداد الحاق دو کمپانی Tesco و Booker محدود نموده­اند. در این دسته تنها رویدادهایی که الگوی انتشار آشوبناک دارند شناسایی می شوند.

دسته­ی دوم روش­های مبتنی بر شباهت محتوایی می­باشند. در این دسته روش­ها یک روش بررسی شباهت محتوایی اخبار ارائه شده­ است. از آن­جا که داده­های خبری نسبت به داده­ها در شبکه­های اجتماعی قابلیت اطمینان بیشتری دارند، تحلیل محتوایی آن­ها نسبت به داده­های توئیتر ارجعیت دارد و البته داده­ در شبکه­های اجتماعی زودتر از اخبار منتشر می­شوند و تحلیل الگوی انتشار در آن­ها ارجعیت دارد[49]. در روش پیشنهادی [50] از دو هستی­شناسیWordNet و VerbNet برای شناسایی رویداد استفاده شده است. از اخبار ویژگی­هایی به فرم (فاعل، فعل، مفعول) با استفاده از روش تعیین نقش نحوی واژه­ها[[73]](#footnote-73) استخراج شده است و یکسان سازی افعال با استفاده از هستی­شناسی VerbNet سبب شده است اخباری که با افعال مترادف گزارش می شوند در یک دسته قرار گیرند و به عنوان رویداد تلقی شوند. چالشی که در استفاده از هستی­شناسی­ها وجود دارد وجود کلمات با معانی مختلف در حوزه­های متفاوت است. همچنین در روش پیشنهادی [51] با استفاده از تکنیک پاراگراف به بردار[[74]](#footnote-74) به دسته­های سیاسی، اقتصادی، ملی تعیین می گردد، سپس با استفاده از روش تست علّیت گرنجر به بررسی فرضیه ی موثر بودن رویدادهای هر دسته بر نوسانات بورس با توجه به الگوی تکرار شده در گذشته پرداخته است. روش پیشنهادی [52] برای شناسایی رویداد به ازای هر سند خبری، افعال موجود در عنوان هر خبر با کمک یک ابزار تعیین نقش نحوی شناسایی شده­اند. سپس با استفاده از تکنیک جاسازی واژه­ها برای هر فعل یک بردار 100 بعدی تولید کرده است و با خوشه بندی بردارهای استخراج شده، هر خوشه را معادل یک رویداد (خرید کردن، بدست آوردن و ...) در نظر گرفته­اند. در این روش به ارتباط بین آیتم های خبری پرداخته نشده است و در حقیقت عنوان­ها با خوشه بندی استخراج شده­اند، بطوریکه هر خوشه نماینده­ی افعالی است که یک مفهوم مانند صعود یا نزول قیمت را مدل می کنند و البته در این روش نقش سایر عوامل نحوی در نظر گرفته نشده است. روش پیشنهادی [53] یک کرنل بر پایه­ی­ ماشین بردار پشتیبان ارائه دادند که در آن ارتباط بین اسناد خبری با استخراج کلمات کلیدی با تکنیک 2جفت واژه[[75]](#footnote-75) برداری شده اند و فراوانی تکرار آن ها در هر سند خبری محاسبه و ارتباط کلمات کلیدی در اسناد خبری توسط یک گراف مدل شده است. این روش با مدلسازی گراف، اخباری که به نوعی یک رویداد را گزارش می دهند مرتبط می­سازد در حالیکه در تکنیک 2 جفت واژه سازماندهی اسناد به روشی مبتنی بر ماتریس انجام گرفته است و در متن با حجم زیاد با چالش ابعاد زیاد فضای ویژگی­ها مواجه خواهیم بود. جدول1 به مقایسه­ی روش­های بررسی شده می­پردازد.

جدول 1- مقایسه روش­های شناسایی رویداد

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Correlation Analysis | Prediction method | Contribution | Method | Reference | Number |
| Granger Causality | NO | Proposed aggregated Sentiment Factor for 8 level sentiment time Series Analysis, | Sentiment Analysis | Chayanin Wong 2016, [16] | 1 |
| NO | No | Dow Jones related Tweet noise removing and Sentiment Score Analysis for event Detection | Sentiment Analysis | Mariana Daniel 2017, [17] | 2 |
| No | MNB[[76]](#footnote-76) and SGD[[77]](#footnote-77) prediction Method | Extraction of same distribution words with log Likelihood and PMI based Clustering for major event detection and SVM for Minor event detection | Anomaly Detection Analysis | Yuriy Gurin 2017 [20] | 3 |
| NO | NO | Tweeter volume, text, hashtag, geographical and sentiment Analysis announcement on Tesco PLC and Booker Group PLC | Anomaly Detection Analysis | Fern´andez Vilas 2018 [48] | 4 |
| NO | LSTM Deep Neural Network | WordNet and VerbNet anthology and POS tagging for subject, object and verb detection | Content based Analysis | Xiao Ding 2014 [50] | 5 |
| NO | LSTM Deep Neural Network | Using EventtoVec method for to Prediction Stock Price Movement with deep Neural Network | Content based Analysis | Xiao Ding 2015 [54] | 6 |
| Granger Causality | LSTM Deep Neural Network | Paragraph2Vec similarity based classification on political, economical and ... category | Content based Analysis | Ishan Verma 2017 [51] | 7 |
| Yes | High dimensional tensor regression | Clustering word Embedding representation of verbs in title of news , sentiments on social media and Market Data tensor representation and decomposition techniques | Content based and semantic Analysis | Xi Zhang 2018 [52] | 8 |
| NO | Enhanced Kernel SVM | Bag of words for news modeling and Graph based 2Gram representation for relayed Article Extraction | Content based Analysis | Wen Long 2019 [53] | 9 |

#### 2-2-8. تحلیل رفتار

تحلیل رفتار سرمایه­گذاران شامل پژوهش­هایی با محوریت نحوه­ی تعیین قیمت در بازار با مدلسازی رفتار مشتریان، واکنش سرمایه­گذاران نسبت به اخبار دسته دوم، تاثیر انتخاب کلمات در اخبار بر مشتریان بازار، رفتار سرمایه­گذاران در بحران­های مالی و عوامل موثر در ریسک پذیری مشتریان می باشد. در روش پیشنهادی[55] با بیان این فرضیه که وضعیت بازار و قیمت محصولات مالی در بازار توسط مشتریانی تعیین می شود که تحت تاثیر اخبار قرار می گیرند، یک مدل ریاضی برای مدلسازی رفتار بازار با وقوع اخبار جدید ارائه شده است. آن­ها برای بررسی فرضیه­ی خود با تمرکز بر بورس فارکس[[78]](#footnote-78) از داده­های جفت ارز USD/JPY در بازه ی ساعت 15:30 10 مارس تا 16:30 11 مارس که زلزله­ی شدیدی در Tohoku-Kanto رخ داده است و داده­های اخبار منتشر شده در همین بازه استفاده کرده­اند. در مدل ارائه شده مشتریان بازار به دو دسته­ی بنیادگراها[[79]](#footnote-79) و پی گیران روند[[80]](#footnote-80) تقسیم شده­اند و نقش هر دو در نوسان قیمت مدل شده است. بنیادگراها کسانی هستند که به ارزش واقعی سهام معتقد هستند و در صورت نوسان قیمت سهام با معامله­گری سعی در بازگرداندن قیمت به مقدار واقعی دارند و پی گیران روند کسانی هستند که با هدف موافق با روند بازار معامله می کنند. همچنین خشه حاصل از نوسان[[81]](#footnote-81) نیز در این مدل لحاظ شده است.

در روش پیشنهادی [56] به بررسی فرضیه ی مطلع بودن موسسات مالی از اخبار قبل از وقوع آن­ها پرداخته است. آن­ها برای بررسی فرضیه ی خود با ارائه­ی یک مدل رگرسیونی متکی بر ساعت وقوع خبر، تحلیل احساس خبر و جریان معاملات (خرید – فروش) ارائه داده­اند. با بررسی میزان جریان معاملات و زمان وقوع خبر به این نتیجه دست یافته­اند که تغییر قیمت بر اثر خبر قبل از تاریخ انتشار اخبار رخ می دهد و موسسات مالی از وقوع اخبار مطلع هستند.

در [57] یک روش آماری برای مقایسه ی نحوه ی تاثیر اخبار خوب و بد بر سرمایه­گذاران ارائه شده است. آن ها برای بررسی این فرضیه که سرمایه گذاران چه زمانی به اخبار توجه می کنند از تست آماری رابط[[82]](#footnote-82) استفاده کرده­اند و نتایج نشان می دهد بازار به اخبار منفی واکنش مضاعفی به صورت نامتقارن نشان می دهد. روش پیشنهادی [58] به تحلیل رفتار سرمایه­گذاران سهام شرکت­های سهامی بزرگ هندی تحت تاثیر داده­های چهار شبکه اجتماعی Facebook ,tweeter ,Youtube ,Linkedin پرداخته­اند و از فاکتورهایی نظیر تعداد توئیت منتشر شده در هر برش زمانی و قیمت سهام استفاده کرده­اند. روش پیشنهادی [59] به تحلیل همبستگی داده­های توئیتر با سه هشتک کلیدی در بازه­ی زمانی حین و بعد از انتخابات انگلستان در سال 2016 و بورس FTSE100 پرداخته­اند، در واقع فرضیه­ی آن­ها بررسی تاثیر رویدادهای سیاسی در سهام FTSE100 با استفاده از تحلیل داده­های شبکه­های اجتماعی بوده­ است. نتایج آن­ها نشان می­دهد در کوتاه­ مدت همبستگی مثبت وجود دارد.

روش­های [18, 19] برای بررسی واکنش سرمایه گذاران بر اخبار دست دوم[[83]](#footnote-83) ارائه شد. بر اساس یافته­های [18, 19] سرمایه­گذاران به اخبار کهنه به صورت مضاعف واکنش نشان می دهند و ارتباط مستقیمی بین تاثیر اخبار اقتصادی راجع به بیکاری و بازگشت سهام شرکت های S&P 500 در هفته­ی بعد وجود دارد و این روند در طی هفته­های بعد بصورت معکوس ادامه می یابد. در این تحقیق از مجموعه داده ی LexisNexis database استفاده شده است که اخبار اقتصادی را از 398 روزنامه جمع آوری کرده است و برای هر خبر رتبه ی نو بودن آن نیز وجود دارد. روش پیشنهادی [60] به بررسی همبستگی بین آمار جستجوی کلمات کلیدی مرتبط با بورس مستخرج از وب سایت google trends و بازگشت در بازار Bitcoin پرداخته شده است. یک معیار متوسط شاخص قیمت Bitcoin بر اساس ارزهای مختلف و حجم معاملات آن ها معرفی کرده است و همچنین یک شاخص google Trend Index روی جستجوی واژه های مربوط به Bitcoin نیز از سایت google trend برای مدلسازی رفتار سرمایه گذاران بیت کوین محاسبه کرده است. بر اساس این دو معیار دو سری زمانی تشکیل شده است که به روش MF-DCCA که یک روش چندفرکتالی تحلیل همبستگی بین دو سری زمانی می­باشد به بررسی همبستگی این دو سری زمانی پرداخته است.

روش پیشنهادی [61] به بررسی عوامل تاثیرگذار بر ریسک­پذیری افراد در سرمایه­گذاری روی بورس کشور آلمان پرداخته است. در مجموعه داده­ی استفاده شده، از تعدادی از افراد خواسته شده اطلاعاتی از شرایط اقتصادی خودشان مانند خرید خانه، ارتقا شغلی و .. را وارد کنند. سپس به بررسی دو دسته عوامل فردی و محیطی بر میزان ریسک پذیری افراد پرداخته است. اطلاعات محیطی از طریق فرایند تحلیل احساس روی اخبار اقتصادی منتشر شده در بازه ی زمانی تحقیق انجام شده است. لذا یک معادله­ی تخمین ریسک سرمایه­گذاری در بازه­های کوتاه مدت دو روزه – یک هفته­ای – ماهانه و سالانه انجام گردیده است. در روش پیشنهادی[62] به بررسی همبستگی بین شاخص احساس سرمایه­گذاران و میزان بازگشت در بورس آمریکا به صورت متغیر با زمان در بازه ی بلند­مدت پرداخته شده است.آن ها با استفاده از روش DCC-MIDAS به بررسی متغیر با زمان همبستگی پرداخته­اند در حالیکه سایرین مدل خود را ایستا در نظر گرفته اند.

جدول 2 یافته ها در روش های تحلیل رفتار

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Literature | sentiment Analysis | Market | Analysis Method | Time Frame | Parameter | Finding |
| T. Ochiai, 2011[55] | Yes | Forex | mathematical model | 5 minutes | volatility | Their findings show that the model with multiplicative noise can reproduce the dynamics observed in the real financial market affected by the arrival of high impact news |
| T. Hendershott ,2011[56] | Yes | NYSE | regression | daily | trading volume, volatility | Their results suggest that significant price discovery related to news stories occurs through institutional trading prior to the news announcement date |
| I. Medovikov, 2016[57] | No | US | Copulas Statics |  | equity returns | Their Finding shows market reacts strongly and negatively to the most unfavourable macroeconomic news, but appears to largely discount the good news. |
| B. Kaushik, 2017[58] | No | NSE stock | correlation | daily | stock prices | there is no satisfactorily explainable or mathematically expressible relationship between the stock prices of firms and their respective SM activities |
| G. Birz, 2017[18] | Yes | S&P 500 | correlation | week | volatility | Their findings show statistically and economically significant relationship between stale news stories on unemployment and next week’s S&P 500 returns. This effect is then completely reversed during the following week. |
| W. Zhang, 2018[60] | No | Bitcoin market | MF-DCCA | daily | Trading volume | By employing the Multifractal Detrended Cross-correlation Analysis method, they find that the change of Google Trends (CGT) and Bitcoin market, i.e., returns and changes of volume, are overall higher degree of multifractal in the long-term and weak multifractal in the short-term. |
| F. Tausch, 2018 [61] | Yes | DAX | correlation | day-week-month | inflation | They find that while a short term increase in good news is positively related to the willingness to take risks, the relation is negative if we consider a long term increase. An increase in negative economic news coverage is negatively related to individuals’ willingness to take risks, irrespective of the time frame |
| L. Fang, 2018 [62] | Yes | S&P 500 | DCCA-MIDAS | daily | return | The results show that the composite index of investor sentiment has a significantly positive influence on the long-term stock-bond correlation, and the shock of crises significantly decrease the average correlation but the effect of sentiment does not change significantly |
| Tahir M. Nisar, 2018 [59] | Yes | FTSE 100 | correlation | daily | close price | Their findings show there is evidence of causation between public sentiment and the stock market movements, in terms of the relationship between MOOD and the daily closing price, and the time lag findings of MOOD and PRICE. |

#### 8-3-2. روش­های پیشگو

مدل­های پیشگو در حوزه­ی اقتصاد رفتاری با استفاده از روش­های یادگیری ماشین مختلفی به پیش­بینی آینده بازار بر اساس اخبار و پارامترهایی نظیر بازگشت در بورس، قیمت و یا حجم معاملات پرداخته­اند. روش پیشنهادی [63] در سال 2007 را می­توان جزء اولین روش­های پیشگو با رویکرد متن­کاوی دانست. این روش به تحلیل فرکتال بازار تحت تاثیر اخبار خوب و بد پرداخته است. تحلیل احساس اخبار با Openion Finder انجام شده است. سپس با استفاده از شاخصی که بر روی تحلیل احساس تعریف شده است بعد فرکتال محاسبه می شود و بر این اساس از آریما[[84]](#footnote-84) جهت پیش­بینی استفاده شده است. [64] به بررسی میزان تاثیر جریان اطلاعات روی بازگشت سهام S&P100 stocks برای بازه ی 2000 تا 2010 به صورت ساعتی پرداخته اند. اخبار از مجموعه داده ی RavenPack News Analytics Database استخراج شده است. در این مجموعه داده اخبار علاوه بر دارابودن برچسب مثبت یا منفی دارای چند برچسب دیگر نظیر رتبه ی میزان ارتباط خبر به هر شرکت[[85]](#footnote-85)، رتبه ی احساس هر خبر[[86]](#footnote-86)، رتبه ی جدید بودن یک خبر[[87]](#footnote-87) و متوسط رتبه ای که با استفاده از فنون مختلف تحلیل احساس برای یک خبر بدست آمده است[[88]](#footnote-88)، نیز هستند. این رتبه ها توسط یک خبره تعیین شده است. در روش پیشنهادی[65] پیش بینی بازگشت در بورس بصورت روزانه با توجه به تحلیل احساس روی اخبار با در نظر گرفتن یک گروه بندی بر شرکت­ها انجام گردیده است. بر اساس استاندارد Global Industry Classification Standard(GICS) ایجاد شده توسط S&P شرکت­هایی که پیش­بینی بر آن­ها انجام شده است به 5 گروه، گروه­بندی شده­اند و از مجموعه داده­ی LexiNexis برای اخبار استفاده کرده است. در این مجموعه داده اخبار دارای برچسب مثبت، منفی و خنثی می­باشند. آن ها در فرایند انتخاب ویژگی از روش chi-square استفاده کرده­اند. در پیش پردازش، اخباری که در طی چند روز متوالی تکرار شده اند فقط یک بار در نظر گرفته شده­اند. در روش پیشنهادی[66] به بررسی و تحلیل احساس داده­های شبکه­ی اجتماعی www.psychsignal.com در بازگشت و حجم معاملات در بازار سهام 5 شرکت معروف و در مبناای 2 دقیقه پرداخته است. در تحلیل احساس از جمله روش­های مبتنی بر فرهنگ لغات استفاده کرده است و لغات موجود در نظرات به صورت دو دسته ی صعودی و نزولی رتبه اختصاص داده­اند و از یک مدل رگرسیونی برای پیش­بینی استفاده شده است. نوآوری روش پیشنهادی [15] معرفی یک شاخص آماری به نام شاخص تجمعی احساس اخبارANSI[[89]](#footnote-89) به نام ANSI بوده است. این شاخص برای سهام هر شرکت با توجه به تحلیل احساس اخبار همان شرکت برای هر روز/هفته/ ماه محاسبه شده است. برای شرکت­های بزرگ که تاثیر بیشتری بر بازار دارند وزن بیشتری اختصاص داده است. تاثیر این شاخص بر حجم معاملات و بازگشت به صورت روزانه هفتگی و ماهانه بررسی شده است. روش پیشنهادی [67] با استفاده از روش یادگیری تقویتی به پیش­بینی بر اساس اخبار پرداخته است. روش­های [51, 68-71] از شبکه عصبی یادگیری عمیق مدلسازی توالی دنباله­ها [[90]](#footnote-90) [72] برای پیش­بینی بر اساس اخبار و قیمت در بازار بورس استفاده کرده­اند.

بازار تبادل ارزهای خارجی[[91]](#footnote-91) که به اختصار آن را فارکس[[92]](#footnote-92) 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. در اغلب روش­های پیشگو در بازار فارکس، از تکنیک های متن کاوی برای استخراج وقایع تاثیرگذار استفاده نشده است و تنها بر اساس زمان انتشار و نرخ بازگشت در بازار به بررسی و تحلیل پرداخته شده است[73]، که آن را می­توان تحت تاثیر کمبود مجموعه اخبار مرتبط با این بازار دانست. روش پیشنهادی [28] 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 ارزEUR/USD قرار داده است . پیش بینی بازگشت در بورس به صورت کوتاه­ مدت در بازه های زمانی 2 ساعتی انجام گرفته است. از دسته بند SVM جهت پیش بینی بازگشت در بورس استفاده شده است. در روش پیشنهادی [74] تاثیر انتشار گزارش­های سازمان یافته اقتصاد کلان و اخبار منتشر شده بر جهش در نرخ بازگشت در بازار فارکس بررسی شده است. در این روش پس از حذف خشه از سیگنال بازگشت در ارزهایی نظیر EUR,CHF,JPY,GBP از یک تست آماری ناپارامتری برای تعیین جهش­ها استفاده شده است. سپس با استفاده از یک روش مبتنی بر گارچ[[93]](#footnote-93) به بررسی میزان تاثیر رویدادهای خبری که توسط خبره­ی انسانی شناسایی شده­اند و گزارش­های ماهانه­ی بلومبرگ[[94]](#footnote-94) بر جهش­های شناسایی شده می­پردازد. روش پیشنهادی از تکنیک­های متن کاوی برای شناسایی رویداد استفاده نکرده است و میزان غافلگیری هر گزارش و خبر توسط خبره­ی انسانی تعیین شده است. روش پیشنهادی [75] به بررسی تاثیر اخبار بر جفت ارز EUR/ USD در بحران ورشکستگی در سال 2008 پرداخته است. در روش پیشنهادی [76] به بررسی تاثیر اخبار اقتصادی با منشاء FED , ECB بر بازار نوظهور ارزهای جدید CZK, HVF, PLN در اتحادیه­ی اروپا در سال­های 2011 تا 2015 پرداخته است. آن ها با ارائه­ی یک پارامتر به نام بازگشت غیر عادی به بررسی تحلیل احساس اخبار بر شاخص­مالی در بازه ی زمانی 5 دقیقه­ای پرداخته­اند. نتایج نشان می­دهد تاثیر خبر نه تنها بعد از وقوع خبر بلکه قبل از ان هم دیده می­شود، همچنین میزان بازگشت بسته به منشاء و کیفیت خبر متفاوت است و جفت ارزهایی با پایه­یUSD بیشتر از ارزهای اروپایی تحت تاثیر قرار می­گیرند.

در ارزیابی روش­هایی یادگیری ماشین مورد مطالعه در این پژوهش، اغلب روش­ها به بررسی صعود یا نزول در بازگشت در بورس پرداخته­اند. از این رو مسئله به صورت طبقه­بندی دو دسته­ای[[95]](#footnote-95) تبدیل شده­است. این دسته روش­ها از معیارهایی نظیر صحت[[96]](#footnote-96)، دقت و recall و همچنین میانگین دو پارامتر دقت و recall تحت عنوان معیار F1 استفاده نموده­اند[5]. دسته*ی* دیگری از روش­های مورد مطالعه با در نظر گرفتن یک توزیع پیوسته روی داده­ها از روش­های رگرسیونی روی داده­ها جهت پیش­بینی استفاده نموده­اند و لذا از معیار متوسط خطای مطلق[[97]](#footnote-97) و یا مجذور خطای مطلق متوسط[[98]](#footnote-98) جهت ارزیابی روش خود استفاده کرده­اند.

جدول 2 مقایسه­ی مدل­های پیشگو

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Literature | Machine Learning | Market | TimeFrame | Parameter | evaluation Metrics | Contribution |
| V.P. Romanov, 2007[63] | ARIMA | Forex | day-week-month | volatility |  | they used Fractal Model for Estimating News Influence on Market Volatility |
| A. K. Nassirtoussi, 2015[5, 28] | SVM | Forex | 2 hours | close | Precision / Recall / Accuracy | their Model predict intraday directional-movements of a currency-pair in the foreign exchange market based on the text of breaking financial news-headlines |
| I. E. Ouadghiri , 2016[74] | Tobit GARCH Model | Forex | daily | Volatility | MAE | examine the intraday effects of surprises from scheduled macroeconomic announcements and unscheduled event news on six major exchange rate excess returns (jumps) using a Tobit model |
| Y. Shi, 2016[64] | MRS-GARCH | s&p500 | daily | Volatility | MAE | their model analyzes the effects of news and its sentiment on the idiosyncratic volatility by Tobit GARCH Model |
| H. D. Huynh, 2017[68] | BGRU | S&P500 | day-week | stock price movements | Accuracy | their predictive model used BGRU based on both online financial news and historical stock prices data to predict the stock movements in the future |
| Y. Shynkevich, 2016[65] | SVM-KNN | S&P 500 | daily | stock price movements | Accuracy | studies how the concurrent, and appropriately weighted, usage of news articles, having different degrees of relevance to the target stock, can improve the performanceof financial forecasting |
| M. R. Vargas, 2018[71] | LSTM,CNN | S&P500 | daily | volatility | Accuracy | they use deep learning models for daily directional movements prediction of a stock price using financial news titles and technical indicators as input |
| W. B. Omrane, 2017[75] | AR(2),GARCH(1,1) | Forex | 5 minutes | volatility | MAE | they study the LOB response to macro news announcements and variation of ‘surprise’ with different states of economy. |
| E. K. cenda,2018[76] | ESA | Forex | 5 minutes | return close prices | MAE | they analyze the impact of euro zone/German and U.S. macroeconomic news announcements and the communication of the monetary policy settings of the ECB and the Fed on the forex markets of new EU members |

#### 4-2-8. پژوهش­های داخلی

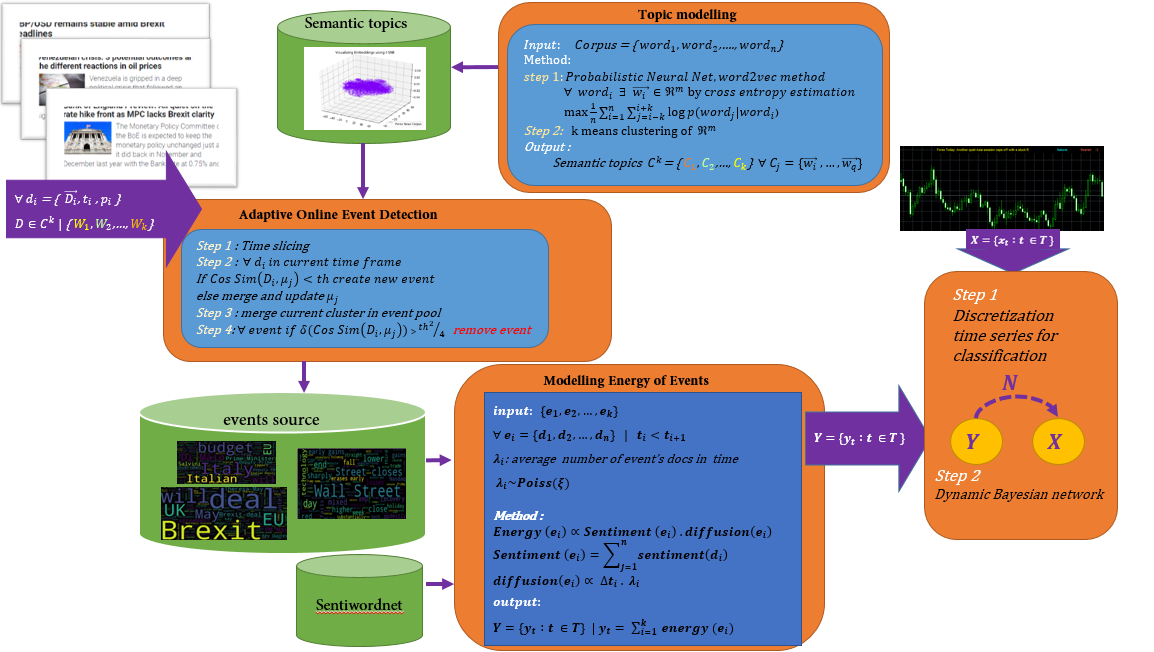
تا کنون پژوهش­های کمی به زبان فارسی در مسئله­ی کاربرد متن­کاوی در پیش­بینی بورس منتشر شده است. روش پیشنهادی[77] به جمع­آوری نظرات افراد و تحلیل احساس آن جهت دسته­بندی به دو دسته­ی صعود یا نزول با استفاده از ماشین بردار پشتیبان پرداخته­است. روش پیشنهادی [78] با رویکرد تحلیل فنی به بررسی تاثیر اخبار خوب و بد بر نوسانات بازدهی سهام در ایران پرداخته است. نتایج این تحقیق نشان می­دهد در دوران تحریم تاثیر اخبار بد بر نوسانات بازدهی سهام بیشتر ازتاثیر اخبار خوب است اما در دوران مذاکرات هسته­ای و برجام تاثیر اخبار خوب بر نوسانات بازدهی سهام بیشتر از اخبار بد است. بر اساس یافته­ها در این پژوهش، ازاین رو که بازار بورس از اخبار خوب و بد تاثیر می­پذیرد، توصیه می­شود که سیاستگذاران با انجام سیاستهایی که منجر به اخبار خوب می­شود، زمینه رشد بازار سهام را فراهم کنند. در روش پیشنهادی [79] تغییرات قیمت شاخص کل بورس ایران به وسیله اخبار 10 سال گذشته روزنامه همشهری و داده های تاریخی بورس ایران پیش بینی شده است. در این روش مجموعه داده­های متنی فارسی پس از پیش پردازش با استفاده از روش­های متن­کاوی مورد تحلیل قرار گرفته­اند. ابتدا مجموعه کلمات کلیدی موثر بر شاخص­های اقتصادی به عنوان بردار مشخصه­های مورد استفاده برای الگوریتم­های یادگیری ماشین مورد استفاده قرار گرفته است. سپس پیش بینی الگوریتمی آینده بورس با تکیه بر متن­کاوی اخبار انجام شده است. در روش [80] با رویکرد تحلیل فنی، به بررسی تاثیر اخبار مربوط به طلا و مسکن بر نوسانات شاخص بورس اوراق بهادار تهران پرداخته­است. بدین منظور اخباری که به صورت روزانه منتشر می­شوند از منابع خبری جمع آوری گردیده است. سپس اخبار مربوط به طلا و مسکن در این بازه زمانی استخراج و با استفاده از مفهوم ارزش خبری رتبه­بندی شده­اند. نتایج به دست آمده از آزمونهای آماری در این تحقیق، رابطه معنا­داری بین اخبار مربوط به طلا و مسکن با نوسانات شاخص بورس اوراق بهادار تهران در 6ماهه پایانی سال 89 نشان نداده است.

در روش پیشنهادی [81] فرض اساسی این است که ویژگی­ها در بازار بورس در بازه زمانی خاصی بسیار رایج شده و پس از مدتی ناپدید می شوند. علت این امر، وقوع رخدادی خبرساز در دنیای واقعی است که پس از مدتی خبرساز بودن، خبرها در مورد آن متوقف می­شود. این مقاله با در نظر گرفتن این فرض، روشی جدید برای انتخاب ویژگی ارائه می­کند که با وزن­دهی بیشتر به ویژگی­هایی که به تازگی رؤیت شده­اند، آنها را به نسبت بیشتری در ویژگی­های نهایی دخالت می­دهد. با این روش آن­ها توانسته­اند در بازار ایران به صحت 74,7% و سود شبیه سازی 18,1% ، در مقابل سود متوسط بازار 14,1% برسند

.

# **روش‌شناسی تحقيق**

روش پیشنهادی یک چهارچوب برای شناساییبرخط رویداد­ها، مدلسازی تطبیقی یک سری زمانی از انرژی رویدادهای خبری ویک مدل پیشگو مبتنی بر شبکه بیزی پویا برای پیش­بینی میزان بازگشت[[99]](#footnote-99) در بازار ارائه می­دهد. شکل 15 فلوچارت روش پیشنهادی را نشان می­دهد.



شکل 15 فلوچارت روش پیشنهادی

در روش پیشنهادی برای شناسایی رویدادها، ابتدا اسناد به صورت برخط[[100]](#footnote-100) با استفاده از روش مبتنی بر جاسازی واژه­ها[[101]](#footnote-101)[21]، دسته­بندی[[102]](#footnote-102)می­شوند بطوریکه هر دسته نماینده­ی یک رویداد باشد. در این روش ابتدا مدلسازی عنوان برای تولید توزیع لغات پیکره­ی متنی در مفاهیم پنهان انجام گردیده است و سپس توزیع اسناد حول عنوان­های استخراج شده به صورت برخط محاسبه شده است.­ در گام بعد برای مدلسازی تطبیقی انرژی رویدادها ابتدا باید رویدادهایی که در حال محو شدن هستند شناسایی و حذف شوند و سپس برایند انرژی رویداد محاسبه گردد. برای شناسایی رویدادهایی که در حال محو شدن می­باشند، شاخصی بر اساس ماکزیموم واریانس شباهت هر سند خبری در یک رویداد با مرکز خوشه همان رویداد ارائه می­گردد. سپس برای هر رویداد پارامتر انرژی مبتنی بر دو معیار احساس و الگوی انتشار[[103]](#footnote-103) تعریف می­شود، بطوریکه تحلیل احساس متن اخبار مثبت یا منفی بودن انرژی را تعیین کند و الگوی انتشار که بر اساس تعداد و فواصل زمانی انتشار اسناد مربوط به هر رویداد محاسبه خواهد شد، بزرگی انرژی را تعیین نماید. در مدل پیشگو بیزی پویا پیشنهادی، پس از گسسته سازی سری­های زمانی با استفاده از یک سیستم استنتاج فازی [82] روشی مبتنی بر نظریه­ی اطلاعات برای تعیین ساختار شبکه­ی ناهمگن پیشنهادی ارائه خواهد شد.

نوآوری طرح پیشنهادی در تحلیل اخبار، مدلسازی یک سری زمانی از میزان انرژی رویدادها برای تاثیر بر بازار با توجه به محتوا و همچنین ارتباط میان اسناد خبری است. از این رو می­توان یک رویداد موثر بر بازار را یک دنباله از اسناد خبری مرتبط تعریف کرد بطوریکه این دنباله بر اساس موضوع، تحلیل احساس[[104]](#footnote-104) و الگوی انتشار، به اندازه­ی مشخصی سرمایه­گذاران را تحت تاثیر قرار دهد، سپس دنباله محو شود و تاثیرش نیز پس از مدتی محو شود. مزیت شناسایی آنلاین رویدادها در مدلسازی تطبیقی تاثیر آن­ها می­باشد بطوریکه می­توان به طور همزمان اثر رویدادهایی با زمان تاثیر طولانی­تر[[105]](#footnote-105) را در کنار رویدادهایی با زمان تاثیر کوتاه­تر[[106]](#footnote-106) در کنار یکدیگر مدل­سازی کرد. نوآوری طرح پیشنهادی در شبکه بیزی پویا، پیشنهاد ساختار شبکه بر اساس تست آماری جفت­های رابط و میزان آنتروپی منتقل شده است. در شبکه بیزی پویا وجود یک یال با مرتبه تاثیر n میان دو متغیر، دلیل بر وجود مسیر علّی میان دو گره است[83]. استفاده از تست آماری جفت­های رابط شرطی[[107]](#footnote-107) [44] برای مطالعه­ی رابطه­ی وابستگی میان انرژی رویدادها و بازگشت، در بلندمدت پیش­نهاد می­شود و سپس در تعیین مرتبه­ی زمانی رابطه­ی علی رویداد بر بازگشت در کوتاه­مدت، معیار آنتروپی منتقل­شده موثر از گذشته­ی مرتبه­ی n رویداد بر بازگشت پیش­نهاد می­شود[84].

فرضیه اول این تحقیق کارامد بودن روش ارائه شده در شناسایی برخط رویداد در فارکس می­باشد. برای بررسی این فرضیه پیشنهاد می شود در هر برش زمانی رویداد­های استخراج شده بر اساس میزان بازگشت برچسب صعود یا نزول بگیرند و با استفاده از روش ماشین بردار پشتیبان[[108]](#footnote-108)، به مقایسه­ی صحت[[109]](#footnote-109)پیش­بینی نسبت به حالتی که شناسایی رویداد به صورت غیربرخط [50, 52-54] انجام گردد، پرداخته شود. سپس از میزان اطلاعات متقابل برای تخمین اطلاعات موجود در سری زمانی محاسبه شده استفاده شود و برای تست مقاومت روش نسبت به خشه حاصل از خطای خوشه بندی، با روش درهم ریزی داده­ها جهت تخریب همبستگی­ سری زمانی با گذشته­اش میزان خشه تخمین زده شود. فرضیه­ی دوم بررسی موثر بودن اطلاعات مدل شده در انرژی رویدادها، بر بهبود دقت[[110]](#footnote-110) پیش­بینی نوسان بازگشت[[111]](#footnote-111) می­باشد. برای این منظور با توجه به امکان حضور خشه حاصل از خطای خوشه­بندی بر مدلسازی سری زمانی رویدادها وجود دارد، پیش­نهاد می­شود از معیار آنتروپی منتقل شده­ی موثر استفاده شود [85]. در بررسی کارامدی روش پیشنهادی ، کیفیت[[112]](#footnote-112) روش پیشگو را می­توان با استفاده از معیارهای ارزیابی طبقه بندی[[113]](#footnote-113) بررسی کرد.

در مطالعه­ی موردی، طرح پیش­رو با پیاده­سازی یک ماژول[[114]](#footnote-114)، اخبار مرتبط با بازار فارکس[[115]](#footnote-115) را از وب سایت­های مهم خبری[[116]](#footnote-116) از آگوست 2018 تا کنون جمع­آوری نموده است. از جمله ویژگی­های مجموعه داده­ی در حال جمع آوری وجود متادیتاهایی است که نشان می­دهد خبر منتشر شده به کدام جفت ارز ارتباط دارد و از جمله رویدادهای مهمی که در این بازه زمانی رخ داده است مسئله خروج انگلستان از اتحادیه اروپا و جنگ تجاری میان چین و آمریکا می­باشد. در مورد انتخاب مبنای زمانی معامله در روش پیشگو، با توجه به تعداد متوسط اخباری که خبرگزاری در هر روز در مورد هر جفت ارز منتشر می­کند[[117]](#footnote-117)، مطالعه بر مبنای معامله 4 ساعت و طول بلاک یک روز پیشنهاد می­شود.

در جمع­آوری داده­های بورس فارکس یک برنامه به زبان پایتون مبتنی بر REST API طراحی شد که قادر به گرفتن برخط داده­های جفت ارز مبدا از بروکر FXCM می­باشد.

# **جنبه نوآوری پژوهش**

1. مدلسازی تطبیقی سری زمانی انرژی رویدادها با توجه به شاخص احساس و ارتباط میان اسناد
2. استفاده از تست آماری رابط شرطی در تحلیل وابستگی دو سری زمانی
3. ارائه­ی یک روش مبتنی بر نظریه­ی اطلاعات در پیشنهاد ساختار شبکه بیزی پویا
4. ارائه­ی یک سیستم استنتاج فازی در گسسته­سازی سری زمانی
5. استفاده از معیارهایی مبتنی بر نظریه­ی اطلاعات در ارزیابی میزان اطلاعات بدست آمده از میزان انرژی رویدادها

## نرم­افزارها و بستر سخت­افزاری مورد نیاز

* محیط غیر تجاری ابر Anaconda برای پیاده سازی برنامه به زبان برنامه­نویسی پایتون
* سخت افزار GPU برای پیاده­سازی برنامه بر بستر کتابخانه­ی tensorflow
* با توجه به تحریم گسترده کشور در زمینه­های مالی، برای برقراری ارتباط با سرور بروکر و خبرگزاری نیاز به استفاده از پروکسی می­باشد و این مسئله برای محقق به یک چالش تبدیل شده است.

# زمانبندی پژوهش

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **رديف** | **شرح فعاليت** | **زمان كل**  **(ماه)** | **زمان اجرا به ماه** | | | | |
| **1** | **2** | **3** | **4** | **5** |
| 1 | پیاده­سازی فضای معنایی واژه­ها مستخرج از پیکره­ی خبری جمع­آوری شده بر بستر  Tensorflow | **1** |  |  |  |  |  |
| 2 | پیاده­سازی خوشه­بندی آنلاین داده­ها برای شناسایی رویدادها و مطالعه ­بر شاخص شناسایی رویدادهای محو شده |  |  |  |  |  |  |
| 3 | ارزیابی صحت پیش­بینی در فرضیه­ی اول تحقیق و نگارش مقاله­ی اول |  |  |  |  |  |  |
| 4 | پیاده­سازی سری زمانی انرژی رویدادها و تدوین سیستم فازی گسسته سازی سری های زمانی  و طراحی یک واسط برای مشاهده­ی سری زمانی انرژی و بازار |  |  |  |  |  |  |
| 5 | مطالعه­ی رابطه وابستگی بین رویداد و بازگشت در بازار با تست رابط شرطی |  |  |  |  |  |  |
| 6 | تدوین مقاله دوم |  |  |  |  |  |  |
| 7 | تدوین شبکه بیزی پویا در مدلسازی رفتار |  |  |  |  |  |  |
| 8 | ارزیابی دقت روش پیشگو و میزان اطلاعات بدست آمده |  |  |  |  |  |  |
| 9 | تدوین مقاله سوم |  |  |  |  |  |  |
| 10 | طراحی یک عامل معامله­گر خودکار به زبان پایتون در بروکر FXCM |  |  |  |  |  |  |

# **منابع**

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

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

3. Vikash Ramiah, X.X., Imad A. Moosa,, *Neoclassical finance, behavioral finance and noise traders: A review and assessment of the literature,.* International Review of Financial Analysis,, 2015. **41**: p. 89-100.

4. Tetlock, P.C., *Does Public Financial News Resolve Asymmetric Information?* The Review of Financial Studies, 2010. **23**(9): p. 3520–3557.

5. 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.

6. 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.

7. Tomas Mikolov , K.C., Greg Corrado , Jeffrey Dean, *Efficient Estimation of Word Representations in Vector Space.* CoRR, 2013. **abs/1301.3781**: p. 1301-3781.

8. Johan Bollen, H.M., Xiaojun Zeng, *Twitter mood predicts the stock market.* Journal of Computational Science, 2011. **2**: p. 1-8.

9. Zhang, X., et al., *Improving stock market prediction via heterogeneous information fusion.* Knowledge-Based Systems, 2018. **143**: p. 236-247.

10. Bruha, I., *From machine learning to knowledge discovery: Survey of preprocessing and postprocessing.* Intelligent Data Analysis 2000. **4**(3): p. 363-374.

11. Dabrowski, J.J. and J.P. de Villiers, *A unified model for context-based behavioural modelling and classification.* Expert Systems with Applications, 2015. **42**(19): p. 6738-6757.

12. Qikai Liu, X.C., Sen Su, Shuguang Zhu, *Hierarchical Complementary Attention Network for Predicting Stock Price Movements with News.*, in *In The 27th ACM International Conference on Information and Knowledge Management (CIKM ’18)*. 2018, ACM: Torino, Italy.

13. Marjan Van de Kauter, D.B., Véronique Hoste *Fine-grained analysis of explicit and implicit sentiment in financial news articles.* Expert Systems with Applications, 2015. **42**.

14. Minh Dang, D.D., *Improvement Methods for Stock Market Prediction using Financial News Articles*, in *3rd National Foundation for Science and Technology Development Conference on Information and Computer Science*. 2016, IEEE.

15. Y.C. Wei, Y.C.L., J. N. Chen, Y. J. Hsu *Informativeness of the market news sentiment in the Taiwan stock market.* North American Journal of Economics and Finance, 2017. **39**: p. 158–181.

16. Chayanin Wong, I.Y.K., *Predictive Power of Public Emotions as Extracted from Daily News Articles on the Movements of Stock Market Indices*, in *International Conference on Web Intelligence*. 2016, IEEE/WIC/ACM.

17. Mariana Daniel, R.F.N., Nuno Horta, *Company event popularity for financial markets using Twitter and sentiment analysis.* Expert Systems With Applications 2017. **71**: p. 111-124.

18. Birz, G., *Stale economic news, media and the stock market.* Journal of Economic Psychology, 2017. **61**: p. 87–102.

19. Tetlock, P.C., *All the News That’s Fit to Reprint: Do Investors React to Stale Information?* The Review of Financial Studies, 2011. **24**(5): p. 1481–1512.

20. Yuriy Gurin, T.S., Mark T. Keane, *Discovering News Events that Move Markets*, in *Intelligent Systems Conference*. 2017: London, UK.

21. Linmei Hu, B.Z., Lei Hou ,Juanzi Li, *Adaptive online event detection in news stream.* Knowle dge-Base d Systems, 2017. **135**: p. 105–112.

22. Federica Bisio, C.M., Paolo Gastaldo, Rodolfo Zunino and a.E. Cambria, *Sentiment-Oriented Information Retrieval: Affective Analysis of Documents Based on the SenticNet Framework*. Studies in Computational Intelligence. Vol. 639. 2016.

23. David Cutler, J.P., Lawrence Summers, *What moves stock prices?* The Journal of Portfolio Management Spring, 1988. **15**(3): p. 4-12.

24. P. L. Davies , M.C., *Stock prices and the publication of second-hand information.* J. Bus, 1978. **51**(1): p. 43–56.

25. B. M. Barber, D.L., *The “Dart board” column: Secondhand information and price pressure.* J. Financ. Quant. Anal, 1993. **28**(2): p. 273-284.

26. al., B.W.u.a.e. *Daily stock market forecast from textual Web data*. in *in Proc. IEEE Int. Conf. Syst., Man, and Cybern*. 1998. IEEE.

27. M. L. Mitchell , J.H.M., *The impact of public information on the stock market.* J. Finance, 1994. **49**(3): p. 923–950.

28. Arman Khadjeh Nassirtoussi , S.A., Teh Ying Waha, David Chek Ling Ngo, *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**: p. 306–324.

29. Bin Weng, M.A.A., Fadel M. Megahed, *Stock Market One-Day Ahead Movement Prediction Using Disparate Data Sources.* Expert Systems With Applications, 2017. **79**: p. 153-163.

30. R. Feldman, J.S., *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*, ed. C.U. Press. 2006: Cambridge University Press.

31. Salton, G., *Automatic text processing: the transformation, analysis, and retrieval of information by computer*. 1989: Addison-Wesley Longman Publishing Co., Inc. 530.

32. David M. Blei, A.Y.N., Michael I. Jordan, *Latent Dirichlet Allocation.* Journal of Machine Learning Research 2003. **3**: p. 993-1022.

33. Long Ma , Y.Z., *Using Word2Vec to process big text data*, in *International Conference on Big Data (Big Data)*. 2015, IEEE.

34. David M. Blei, A.Y.N., Michael I. Jordan, *Latent Dirichlet Allocation.* Journal of Machine Learning Research, 2003. **3**: p. 993-1022.

35. Zhang, Y., et al., *Does deep learning help topic extraction? A kernel k-means clustering method with word embedding.* Journal of Informetrics, 2018. **12**(4): p. 1099-1117.

36. Hu, L., et al., *Adaptive online event detection in news streams.* Knowledge-Based Systems, 2017. **138**: p. 105-112.

37. Zhang, W., Y. Li, and S. Wang, *Learning document representation via topic-enhanced LSTM model.* Knowledge-Based Systems, 2019. **174**: p. 194-204.

38. Hobæk Haff, I., et al., *Structure learning in Bayesian Networks using regular vines.* Computational Statistics & Data Analysis, 2016. **101**: p. 186-208.

39. Nasri, B.R. and B.N. Rémillard, *Copula-based dynamic models for multivariate time series.* Journal of Multivariate Analysis, 2019. **172**: p. 107-121.

40. Kiran Karra, L.M., *Hybrid Copula Bayesian Networks*, in *JMLR: Workshop and Conference Proceedings , 2016*. 2016. p. 240-251.

41. Medovikov, I., *When does the stock market listen to economic news? New evidence from copulas and news wires.* Journal of Banking & Finance, 2016. **65**: p. 27-40.

42. Shannon, C.E., & Weaver, W. . . : . *The mathematical theory of communication*. 1949, Urbana: University of Illinois Press.

43. Schwill, S., *Entropy Analysis of Financial Time Series*, in *Business Administration*. 2015, University of Manchester.

44. Grassberger, P., *Entropy Estimates from Insufficient Samplings*. 2003.

45. Wei Ji, L.W., *Big data analytics based fault prediction for shop floor scheduling.* Journal of Manufacturing Systems, 2018. **43**: p. 187–194.

46. Gema Bello-Orgaz, J.J.J., David Camacho, *Socialbigdata:Recentachievementsandnewchallenges.* Information Fusion 2016. **28**: p. 45–59.

47. Ghasemaghaei, M., *The role of positive and negative valence factors on the impact of bigness of data on big data analytics usage.* International Journal of Information Management, 2018.

48. Ana Fern´andez Vilas, R.P.D.ı.R., Keeley Crockett, Majdi Owda, Lewis Evans, *Twitter permeability to financial events: an experiment towards a model for sensing irregularities.* Multimed Tools Appl, 2018.

49. Steve Y. Yang , S.Y.K.M., Anqi Liu, Andrei A. Kirilenko, *Genetic programming optimization for a sentiment feedback strength based trading strategy.* Neurocomputing, 2017. **264**: p. 29–41.

50. Xiao Ding, Y.Z., Ting Liu, Junwen Duan. *Using Structured Events to Predict Stock Price Movement: An Empirical Investigation*. in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2014. Doha, Qatar: Association for Computational Linguistics.

51. Ishan Verma, L.D., and Hardik Meisheri. *Detecting, Quantifying and Accessing impact of News events on Indian Stock Indices*. in *In Proceedings of WI ’17*. 2017. Leipzig, Germany.

52. Xi Zhang , Y.Z., Senzhang Wang , Yuntao Yao , Binxing Fang , Philip S. Yu, *Improving stock market prediction via heterogeneous information fusion.* Knowle dge-Base d Systems, 2018. **143**: p. 236–247.

53. Wen Long , L.S., Yingjie Tian, *A new graphic kernel method of stock price trend prediction based on Þnancial news semantic and structural similarity.* Expert Systems with Applications, 2019. **118**: p. 411-424.

54. Xiao Ding, Y.Z., Ting Liu, Junwen Duan. *Deep Learning for Event-Driven Stock Prediction*. in *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence*. 2015. IEEE.

55. T. Ochiai , J.C.N., *A model for the dynamic behavior of financial assets affected by news: The case Tohoku–Kanto earthquake.* Physics Letters A, 2011. **375**: p. 3552–3556.

56. Terrence Hendershott, D.L., Norman Schürhoff, *Are institutions informed about news?* Journal of Financial Economics, 2014.

57. Medovikov, I., *When does the stock market listen to economic news? New evidence from copulas and news wires.* Journal of Banking & Finance, 2016. **65**: p. 27–40.

58. Bhavya Kaushik, H.H., P. Vigneswara Ilavarasan. *Social media usage vs. stock prices: an analysis of Indian firms*. in *Information Technology and Quantitative Management (ITQM 2017)*. 2017. Procedia Computer Science

59. Tahir M. Nisar, M.Y., *Twitter as a tool for forecasting stock market movements: A short-window event study.* The Journal of Finance and Data Science 4, 2018. **4**: p. 101-119.

60. Wei Zhang , P.W., Xiao Li , Dehua Shen, *Quantifying the cross-correlations between online searches.* Physica A, 2018. **509**: p. 657–672.

61. Franziska Tausch , M.Z., *Stability of risk attitudes and media coverage of economic news.* Journal of Economic Behavior and Organization, 2018. **150**: p. 295–310.

62. Libing Fang, H.Y., Yingbo Huang, *The Role of Investor Sentiment in the Long-term Correlation between U.S. stock and bond markets.* International Review of Economics and Finance, 2018.

63. V.P. Romanov, O.A.N., E.A. Panteleeva, A.S. Federyakov, *Fractal Model of Estimating News and Insider Influence on Market Volatility.* Automatic Documentation and Mathematical Linguistics, 2007. **41**(4): p. 141–149.

64. Yanlin Shi , K.-Y.H., Wai-Man 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**.

65. Yauheniya Shynkevich , T.M.M., Sonya A. Coleman , Ammar Belatrech, *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.

66. M.S. Checkley , D.A.H., H. Alles *The hasty wisdom of the mob: How market sentiment predicts stock market behavior.* Expert Systems With Applications, 2017. **77**: p. 256–263.

67. X. Li , H.X., Ran Wang , Yi Cai , Jingjing Cao, Feng Wang ,Huaqing Min, Xiaotie Deng, *Empirical analysis: stock market prediction via extreme learning machine.* Neural Computing and Applications, 2016. **27**(1): p. 67–78.

68. Huy D. Huynh, L.M.D., and Duc Duong, *A New Model for Stock Price Movements Prediction Using Deep Neural Network*, in *SoICT ’17:Eighth International Symposium on Information and Communication Technology*. 2017, ACM: Nha Trang City, Viet Nam.

69. Gaowei Zhang, L.X., Yunlan Xue, *Model and forecast stock market behavior integrating investor sentiment analysis and transaction data.* **cluster computing**, 2017. **20**: p. 789–803.

70. C. Lee and V. Soo, *Predict Stock Price with Financial News Based on Recurrent Convolutional Neural Networks*, in *Conference on Technologies and Applications of Artificial Intelligence (TAAI)*. 2017: Taipei, Taiwan. p. 160-165.

71. Manuel R. Vargas , C.E.M.d.A., Gustavo L. G. Bichara, Alexandre G. Evsukoff, *Deep Leaming for Stock Market Prediction Using Technical Indicators and Financial News Articles*, in *International Joint Conference on Neural Networks (IJCNN)*. 2018, IEEE: Rio de Janeiro, Brazil.

72. Schmidhuber., S.H.a.J.u., *Long short-term memory.* Neural computation, 1997. **9**(8): p. 1735–1780.

73. B. E´ gert, E.e.K.e., *The impact of macro news and central bank communication on emerging European forex markets.* 2014. **38**: p. 73–88.

74. Imane El Ouadghiri, R.U., *Jumps in equilibrium prices and asymmetric news in foreign exchange markets.* Economic Modelling, 2016. **54**: p. 218–234.

75. Walid Ben Omrane, Y.T., Robert Welch, *Scheduled Macro-News Effects on a Euro/US Dollar Limit Order Book around the 2008 Financial Crisis.* Research in International Business and Finance, **2017**.

76. Evˇzen Koˇcenda, M.M., *Intraday effect of news on emerging European forex markets: An event study analysis.* Economic Systems, 2018.

77. كاروان, ل. and م. خادمي, *عقيده كاوي در تحليل شاخص هاي بورس جهت پيش بيني وضعيت سهام اوراق بهادار*, in *سومين كنگره بين المللي كامپيوتر، برق و مخابرات*. 1395, دانشگاه تربيت حيدريه: تربت حيدريه.

78. مرادي, س. and ح. حيدري, *بررسي تاثير مهمترين اخبار سياسي و اقتصادي بر نوسانات (تلاطم) بازدهي سهام در ايران*, in *سومين كنفرانس جامع و ملي اقتصاد مقاومتي*. 1396, دانشگاه يزد، دانشگاه بوعلي سينا همدان، دانشگاه صنعتي نوشيرواني بابل و شركت پژوهشي طرود شمال: بابلسر.

79. فرازنده نيا, م., ب. مجيدي, and ع. موقر, *پيش بيني شاخص هاي اقتصادي با استفاده از متن كاوي داده هاي خبري فارسي*, in *كنفرانس بين المللي مهندسي و علوم كامپيوتر*. 1395, دانشگاه آزاد اسلامي واحد نجف آباد: نجف آباد.

80 غلامي, م., *بررسي رابطه بين اخبار مربوط به طلا و مسكن با نوسانات بورس اوراق بهادار تهران*, in *نخستين كنفرانس ملي تحقيق و توسعه در مديريت و اقتصاد مقاومتي*. 1397, پژوهشگاه فرهنگ و هنر (پژوهشكده مديريت و توسعه ): تهران.

81. شايان فر, ن. and و. درهمي, *انتخاب ويژگي متغير با زمان: روشي جديد براي كاهش ويژگي در پيش بيني كننده هاي مبتني بر خبر در بازار بورس*, in *هفتمين كنفرانس بين المللي فناوري اطلاعات و دانش*. 1394, داشگاه اروميه: اروميه.

82. Jizba, P., H. Kleinert, and M. Shefaat, *Rényi’s information transfer between financial time series.* Physica A: Statistical Mechanics and its Applications, 2012. **391**(10): p. 2971-2989.

83. Murphy, K.P., *Dynamic Bayesian Networks: Representation, Inference and Learning*, in *Computer Science*. 2002, UNIVERSITY OF CALIFORNIA, BERKELEY.

84. Marschinski, R. and H. Kantz, *Analysing the information flow between financial time series.* The European Physical Journal B - Condensed Matter and Complex Systems, 2002. **30**(2): p. 275-281.

85. Darbellay, G.A. and D. Wuertz, *The entropy as a tool for analysing statistical dependences in financial time series.* Physica A: Statistical Mechanics and its Applications, 2000. **287**(3): p. 429-439.

1. twitter [↑](#footnote-ref-1)
2. Stock tweet [↑](#footnote-ref-2)
3. Word embedding [↑](#footnote-ref-3)
4. Sentiment Analysis [↑](#footnote-ref-4)
5. Opinion mining [↑](#footnote-ref-5)
6. اخباری که در چند روز متوالی عینا تکرار می­شوند [↑](#footnote-ref-6)
7. Stream of news [↑](#footnote-ref-7)
8. Online Event Detection [↑](#footnote-ref-8)
9. Cluster [↑](#footnote-ref-9)
10. هر چه یک رویداد مهم­تر باشد تعداد اخبار بیشتری پیرامون حواشی آن گزارش می گردد و سپس با محو شدن آن، تعداد اخبار کمتری نیز آن را پوشش خواهد داد. [↑](#footnote-ref-10)
11. Sentiment [↑](#footnote-ref-11)
12. Information Retrieval [↑](#footnote-ref-12)
13. 1 به نقل از ویکی پدیا در 29 سپتامبر 2008، بعد از آن که کنگره­ی آمریکا طرح نجات اقتصادی رئیس­جمهور جرج بوش را رد کرد، در عرض یک روز ارزش سهام DOW Jones Industrial به مقدار 777.68 پوینت سقوط کرد. نزول ارزش سهام شرکت­های آمریکایی تحت تاثیر این رویداد تا حدود 18 ماه به طول انجامید. [↑](#footnote-ref-13)
14. Long Term process [↑](#footnote-ref-14)
15. Short Term Process [↑](#footnote-ref-15)
16. noise [↑](#footnote-ref-16)
17. Big data [↑](#footnote-ref-17)
18. ارجاع به شکل 1 - نمودار پراکندگی روش­ها بر اساس سال انتشار [↑](#footnote-ref-18)
19. David M. Cutler [↑](#footnote-ref-19)
20. Stock Return [↑](#footnote-ref-20)
21. Microsoft [↑](#footnote-ref-21)
22. Apple [↑](#footnote-ref-22)
23. FOREX [↑](#footnote-ref-23)
24. Open Price [↑](#footnote-ref-24)
25. Close price [↑](#footnote-ref-25)
26. Return Rate [↑](#footnote-ref-26)
27. Volatility [↑](#footnote-ref-27)
28. Trading Volume [↑](#footnote-ref-28)
29. sina wibo [↑](#footnote-ref-29)
30. tokenization [↑](#footnote-ref-30)
31. Stop word remove [↑](#footnote-ref-31)
32. Stemming [↑](#footnote-ref-32)
33. Bag Of Words [↑](#footnote-ref-33)
34. Latent Dirichlet Allocation [↑](#footnote-ref-34)
35. Word Embedding [↑](#footnote-ref-35)
36. *Word2Vec* [↑](#footnote-ref-36)
37. Topic modelling [↑](#footnote-ref-37)
38. Semantic concepts [↑](#footnote-ref-38)
39. Beli [↑](#footnote-ref-39)
40. Classification [↑](#footnote-ref-40)
41. Lexicon Based Sentiment Analysis [↑](#footnote-ref-41)
42. Machine Learning Based [↑](#footnote-ref-42)
43. Correlation Analysis [↑](#footnote-ref-43)
44. Auto Correlation [↑](#footnote-ref-44)
45. Pearson Correlation Analysis [↑](#footnote-ref-45)
46. Linear correlation [↑](#footnote-ref-46)
47. Cross Correlation Analysis [↑](#footnote-ref-47)
48. Granger Causality Test [↑](#footnote-ref-48)
49. Co-Movements [↑](#footnote-ref-49)
50. Causal Relationship [↑](#footnote-ref-50)
51. Copula [↑](#footnote-ref-51)
52. Multivariate Cumulative probability distributions [↑](#footnote-ref-52)
53. Multinomial distribution [↑](#footnote-ref-53)
54. Mutual causal relationship [↑](#footnote-ref-54)
55. Mutual Information [↑](#footnote-ref-55)
56. Pointwise Mutual Information [↑](#footnote-ref-56)
57. Grassberger [↑](#footnote-ref-57)
58. Big Data Characteristics [↑](#footnote-ref-58)
59. Big Data Quality [↑](#footnote-ref-59)
60. Volume [↑](#footnote-ref-60)
61. Velocity [↑](#footnote-ref-61)
62. Value [↑](#footnote-ref-62)
63. Variety [↑](#footnote-ref-63)
64. Veracity [↑](#footnote-ref-64)
65. Valence [↑](#footnote-ref-65)
66. Hadoop Eco Systems [↑](#footnote-ref-66)
67. Batch processing [↑](#footnote-ref-67)
68. Wikipedia [↑](#footnote-ref-68)
69. Granger causality [↑](#footnote-ref-69)
70. Financial comunity [↑](#footnote-ref-70)
71. Pointwise Mutual Information [↑](#footnote-ref-71)
72. Anomaly detection [↑](#footnote-ref-72)
73. POSTagging [↑](#footnote-ref-73)
74. paragraph2vec [↑](#footnote-ref-74)
75. 2gram [↑](#footnote-ref-75)
76. Multi-dimensional Naïve Bayes [↑](#footnote-ref-76)
77. Stochastic Gradient Decent [↑](#footnote-ref-77)
78. FOREX [↑](#footnote-ref-78)
79. fundamentalist [↑](#footnote-ref-79)
80. 3 trend trader [↑](#footnote-ref-80)
81. Volatility [↑](#footnote-ref-81)
82. Copula [↑](#footnote-ref-82)
83. منظور از اخبار دست دوم اخباری است که به صورت دو روز متوالی به صورت عینا تکرار شده باشند. [↑](#footnote-ref-83)
84. ARIMA [↑](#footnote-ref-84)
85. relevance score (REL) [↑](#footnote-ref-85)
86. event sentiment score (ESS)

    رتبه ای بین صفر تا 100 از میزان مثبت یا منفی بودن یک خبر. [↑](#footnote-ref-86)
87. event novelty score (ENS)

    رتبه ای بین صفر تا 100 میزان جدید بودن خبر را در 24 ساعت گذشته تعیین می کند [↑](#footnote-ref-87)
88. composite sentiment score (CSS)

    رتبه ای بین صفر تا 100 . کمتر از 50 منفی و بیشتر از آن مثبت تلقی می شود [↑](#footnote-ref-88)
89. Aggregate News Sentiment Index [↑](#footnote-ref-89)
90. LSTM [↑](#footnote-ref-90)
91. Foreign Exchange Market [↑](#footnote-ref-91)
92. FOREX [↑](#footnote-ref-92)
93. GARCH [↑](#footnote-ref-93)
94. Bloomberg [↑](#footnote-ref-94)
95. Binary classification task [↑](#footnote-ref-95)
96. accuracy [↑](#footnote-ref-96)
97. Mean Absolute Error [↑](#footnote-ref-97)
98. Root Mean Squared Error [↑](#footnote-ref-98)
99. Return Rate [↑](#footnote-ref-99)
100. Online [↑](#footnote-ref-100)
101. Word Embedding

     یک روش مبتنی بر یادگیری عمیق که در آن به ازای لغات یک پیکره متنی عظیم، به هر لغت بر اساس رخداد آن در یک پنجره­ی همپوشان از سایر همسایه­های آن­ها یک بردار اختصاص می­دهد. [↑](#footnote-ref-101)
102. Cluster [↑](#footnote-ref-102)
103. Diffusion pattern

     هر چه یک رویداد مهم­تر باشد تعداد اخبار بیشتری در یک بازه­ی زمانی کوتاه پیرامون حواشی آن منتشر می­شود و سپس با محو شدن آن، تعداد اخبار کمتری نیز آن را پوشش خواهد داد [↑](#footnote-ref-103)
104. Sentiment [↑](#footnote-ref-104)
105. Long Term process [↑](#footnote-ref-105)
106. Short Term Process [↑](#footnote-ref-106)
107. Conditional Copulas Statistical Test [↑](#footnote-ref-107)
108. SVM [↑](#footnote-ref-108)
109. accuracy [↑](#footnote-ref-109)
110. precision [↑](#footnote-ref-110)
111. Volatility direction [↑](#footnote-ref-111)
112. quality [↑](#footnote-ref-112)
113. Precision, recall , accuracy, F1 (با توجه به گسسته سازی سری زمانی) [↑](#footnote-ref-113)
114. Web scraper [↑](#footnote-ref-114)
115. Foreign Stock Exchange FOREX [↑](#footnote-ref-115)
116. www. Fxstreet.com [↑](#footnote-ref-116)
117. به طور متوسط در هر روز 280 خبر اختصاصی راجع به بازار فارکس توسط خبرگزاری FXStreet منتشر می­شود [↑](#footnote-ref-117)