# مروری بر روش­های متن­کاوی در حوزه­ی اقتصاد رفتاری

سعیده انبایی فریمانی[[1]](#footnote-1)

مجید وفایی جهان[[2]](#footnote-2)

# چکیده

استخراج دانش از عظیم داده­های متنی موجود در رسانه­های اجتماعی و گروه­های خبری توسط تکنیک­های متن کاوی در بسیاری از زمینه­ها از جمله حوزه­های مالی کاربرد دارد. استفاده گسترده­ی ­افراد از شبکه­های اجتماعی و انتشار اخبار در رسانه­های مبتنی بر وب، سبب تولید روز افزون منابع عظیمی از داده­های موثر بر رفتار سرمایه­گذاران بازارهای مالی شده است لذا مطالعه­ی چگونگی جمع­آوری و تحلیل عظیم داده­های متنی در بررسی تاثیر آن­ها بر رفتار سرمایه­گذاران بازارهای مالی در راستای تدوین سیستم­های پشتیبان تصمیم امری مهم و مطلوب است. مرور پیش رو به مطالعه­ی 90مرجع منتشر شده از سال 2006 تا ژانویه 2019 با موضوع بررسی برهم­کنش بازارهای مالی و وقایع خبری منتشر شده در وب و اطلاعات موجود در رسانه­های اجتماعی با رویکرد متن کاوی می­پردازد. وجه تمایز پژوهش انجام شده با سایر مرور­های موجود توجه به مسئله­ی تحلیل رفتار سرمایه­گذاران بازارهای مالی از دیدگاه علم کامپیوتر و هوش مصنوعی است، لذا در این مرور به مطالعه­ی انواع منابع اطلاعاتی مورد استفاده در این حوزه، روش­های مختلف بازنمایی متن جهت سازماندهی به داده­های متنی، روش­های تحلیل احساس و واکشی اطلاعات و انواع روش­های یادگیری ماشین در تحلیل عظیم داده­ها پرداخته می­شود. هدف این پژوهش مرور مرز دانش با رویکرد توجه به تحلیل عظیم داده­ها و معرفی مسیرهای آینده تحقیق در روش­های متن کاوی و هوش مصنوعی مورد استفاده در تکنیک­های تحلیل همبستگی، سیستم­های پیشگو و توصیه­گر در بازارهای مالی نظیر بورس و فارکس می­باشد.

# کلمات کلیدی

متن کاوی- تحلیل احساس- پیش­بینی بورس- تحلیل عظیم داده­ها- اخبار- شبکه­های اجتماعی

# مقدمه

با گسترش استفاده از اینترنت و رشد شبکه های اجتماعی و همچنین اهمیت ویژه­ی اقتصاد بر رشدیافتگی ملت­ها، مطالعه­ی بر هم­کنش آن­ها همواره دارای اهمیت می­باشد. بعد از این که در سال 1965 Fama نظریه­ی بازار کارامد را ارائه داد و نشان داد رفتارسری­های زمانی مالی تحت تاثیر اطلاعاتی است که در دسترس سرمایه­گذاران قرار می گیرد، 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 اخبار مربوط به وقایع سیاسی، اقتصادی، اجتماعی و یا احساسی که کاربران شبکه­های اجتماعی در پیام­های ارسالی منعکس می­کنند، می­باشند. منابع اطلاعاتی مانند گروه­های خبری در وب، شبکه­های اجتماعی نظیر توئیتر[[3]](#footnote-3)، بوردهای گفتگو تخصصی بورس مانند استوک توئیت[[4]](#footnote-4) به یک منبع مهم تصمیم­گیری برای سرمایه­گذاران تبدیل شده­اند و استخراج دانش موجود در آن­ها می­تواند به سرمایه­گذاران در اتخاذ تصمیم­های تجاری بهتر کمک کننده باشد. درحوزه­ی اقتصاد رفتاری، از سال­های گذشته تا کنون تحقیق­های زیادی از طریق تحلیل­های فنی[[5]](#footnote-5) بر اطلاعات ساختارمند و سپس تحلیل های بنیادین[[6]](#footnote-6) بر داده­های بدون ساختار متنی با استفاده از تکنیک­های متن کاوی[4] به استخراج دانش موجود در چنین منابع اطلاعاتی برای مدسازی سیستم­های پشتیبان تصمیم[[7]](#footnote-7) و بهبود دقت در مدل­های پیشگو مالی [[8]](#footnote-8) و طراحی سیستم­های توصیه­گر استراتژی تجارت[[9]](#footnote-9) پرداخته­اند.

روش پیشنهادی کوتلر[[10]](#footnote-10) در سال 1988 را می­توان جزء اولین پژوهش­ها با رویکرد تحلیل فنی در بررسی تاثیر اخبار بر بازار بورس دانست[5]. این تحقیق با رویکرد تحلیل فنی به بررسی میزان تغییر در بازگشت سهام[[11]](#footnote-11) 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, 6-9] در حالیکه در تحلیل­های بنیادین از روش­های متن­کاوی برای سازماندهی به داده­ها استفاده می­شود. در طی سال­های اخیر با گسترش شبکه­های اجتماعی و گروه­های خبری موجود در اینترنت، حجم منابع متنی در اینترنت به سرعت در حال افزایش می­باشد و سرمایه­گذاران هر روز با حجم زیادی از منابع متنی مواجه می­شوند. بر این پایه تمرکز تحقیق­های انجام شده در سال­های اخیر بر استفاده از تکنیک­های متن کاوی[[12]](#footnote-12)، واکشی اطلاعات[[13]](#footnote-13) و روش­های هوش مصنوعی[[14]](#footnote-14)[4]، برای ساختار بخشیدن به داده­های وب واستخراج دانش موجود در اخبار و رسانه­های اجتماعی با در نظر گرفتن مسئله­ی عظیم داده­ها[[15]](#footnote-15) می­باشد. تکنیک­های متن کاوی نظیر تحلیل احساس[[16]](#footnote-16)[10-12] ، تعیین نقش نحوی واژه­ها[[17]](#footnote-17)[13, 14] و بازنمایی واژه­ها نظیر روش Word Embedding [15, 16]، روش­های یادگیری ماشین نظیر یادگیری عمیق[17-19] در این حوزه استفاده شده­اند.

در این مرور به بررسی پژوهش های انجام شده در جهت سازماندهی متون خبری وبی و شبکه­های اجتماعی برای بررسی تاثیر اخبار و رسانه­های اجتماعی بر سرمایه­گذاران و بهبود دقت مدل­های پیشگو پرداخته می شود. روش­های ارائه شده اغلب در سه حوزه­ی ارائه­ی مدل­های پیشگوی میزان بازگشت در بورس[[18]](#footnote-18)، تکنیک­های تحلیل همبستگی[[19]](#footnote-19) میان اخبار منتشر شده و رفتار سرمایه­گذاران و همچنین سیستم­های پیشنهاددهنده­ی استراتژی تجارت می­باشند. بخش زیادی از مقالات بررسی شده دارای تحلیل بنیادین می باشند و این ویژگی وجه تمایز مرور پیش رو با سایر مقالات مروری[4, 20, 21] موجود در این حوزه می­باشد. وجه تمایز دیگر مرور پیش­رو بررسی مسئله­ی عظیم داده­ها و معرفی ابزارها و روش­های تحلیل آن­ها در این حوزه است. شکل 1- نمودار بخش­های مرور نمودار بخش­های مختلف مرور را نشان می­دهد. در بخش دوم به بررسی بنیان مرور که بر اساس مطالعه­ی اقتصاد رفتاری بنا نهاده شده است می­پردازیم. بخش سوم پیکره­ی مرور را شرح می­دهد. در بخش چهارم عظیم داده­ها، ابزارها و چالش­ها بیان می­شود. در بخش پنج نتیجه گیری و مسیرهای آینده­ی تحقیق بیان خواهد شد.

منابع اطلاعاتی

سازماندهی متن

تجمیعی از واژگان

مفاهیم معنایی

جاسازی واژه­ها

سازماندهی اطلاعات

تحلیل احساس

شناسایی رویداد

شناسایی آشوب

استخراج دانش

مدل­های پیشگو

تحلیل رفتار سرمایه­گذاران

سیستم­های پیشنهاددهنده استراتژی تجارت

پیکره­ی مرور

عظیم داده­ها، ابزارها و چالش­ها

مقدمه

بنیان مرور

روش مرور

منبع انتشار

نتیجه­گیری و مسیرهای آینده

شکل 1- نمودار بخش­های مرور

# بنیان مرور

بنیان مرور بر مطالعه­ی انواع روش­های ارائه شده در حوزه­ی اقتصاد رفتاری بنا نهاده شده است. لذا هدف مرور در مطالعه­ی پل ارتباطی بین بازارهای مالی و منابع متنی در بستر اینترنت می­باشد. از این رو ابتدا به بیان مراحل متن کاوی جهت سازماندهی داده­های متنی و سپس به معرفی انواع بازارهای مالی پرداخته می­شود.

متن کاوی فرایند استخراج اطلاعات مفید از منابع داده ای با استفاده از تکنیک های شناسایی و استخراج الگوهای جذابی است که از تحلیل داده های بدون ساختار استخراج می گردند. [22]. فرایند متن کاوی دارای سه فاز پیش پردازش متن، سازماندهی متن و استفاده از تکنیک های داده کاوی جهت استخراج الگو می باشد. فرایند پیش پردازش شامل جداسازی توکن ها[[20]](#footnote-20)، حذف کلمات توقف[[21]](#footnote-21)و یکسان کردن ریشه لغات[[22]](#footnote-22) و تعیین نقش نحوی واژه­ها می باشد. در اغلب روش­های بررسی شده مراحل پیش­پردازش به جهت آماده سازی داده­های متنی برای سازماندهی انجام گرفته­ است. در برخی از روش­ها از فاز تعیین نقش نحوی به عنوان یک روش برای انتخاب ویژگی استفاده شده است. در فاز تعیین نقش نحوی اغلب از ابزار NLTK POS Tagger استفاده شده است. با استفاده از این ابزار به هر واژه بر اساس نقش نحوی آن یک برچسب مانند فعل، اسم، صفت داده می­شود. در روش پیشنهادی [23] تنها افعال و صفات به عنوان ویژگی انتخاب شده­اند. در روش پیشنهادی [24] هر جمله در عنوان خبر، به صورت سه تایی ( فاعل، فعل، مفعول) تبدیل شده­است. در روش پیشنهادی[25] به بررسی تاثیر انتخاب کلمات در متن اخبار و رفتار مشتریان بازار پرداخته شده است. در این روش علاوه بر دسته بندی افعال موجود در عناوین خبری، اسامی هم به دو دسته ی فاعل و مفعول با دسته بندی شده اند.

بر اساس نظریه ی بازار کارامد[[23]](#footnote-23) که در سال 1965 توسط فاما[[24]](#footnote-24) مطرح شد، قیمت سرمایه منعکس کننده­ی تمام اطلاعات در دسترس راجع به آن سرمایه است. این مسئله امری طبیعی است زیرا با انتشار اطلاعات جدید، واکنش بازار نسبت به آن، سبب تغییر قیمت سرمایه خواهد شد و بازار مداوم بر مبنای ریسک سرمایه­گذاران تغییر نمی کند[1]. 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 2- انواع بازارهای مالی را نشان می دهد[26].

شکل 2- 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 نظیرمایکروسافت[[25]](#footnote-25) و اپل[[26]](#footnote-26) در این لیست قراردارند. در بازار تبادل خارجی فارکس که در آن به مبادله ی پول بر اساس انواع جفت ارزهای خارجی مانند EUR/USD پرداخته می شود. شکل 3- انواع بازارهای مالی مورد مطالعه در منابع مرور شده را نشان می­دهد.

شکل 3- انواع بازارهای مالی مورد مطالعه در منابع مرور شده

## 1-2. روش مرور

در مرور انجام شده 90 مرجع از سال 2006 تا ژانویه 2019 بررسی شده است. شکل 4- نمودار پراکندگی روش­ها بر اساس سال انتشار را نشان می­دهد. اکثر مراجع مورد بررسی دارای کلید واژه هایی نظیر Sentiment Analysis, Stock Market Prediction, Text Mining, Financial News, می­باشند. شکل 5- نمودار ابر کلمات کلیدی در مراجع مرور شده را نشان می­دهد.

|  |  |
| --- | --- |
| شکل 4- نمودار پراکندگی روش­ها بر اساس سال انتشار | شکل 5- نمودار ابر کلمات کلیدی در مراجع مرور شده |

## 2-2. منبع انتشار مراجع

جدول 1- منبع انتشار روش­های مرور شده را نشان می­دهد. منابع دو دسته­ی کنفرانسی و ژورنالی می­باشند. منابع ژورنالی که دارای ضریب تاثیر بیشتر از دو می­باشند در جدول هایلایت شده­اند.

جدول 1- منبع انتشار روش­های مرور شده

|  |  |
| --- | --- |
| Asia Pacific Management Review | 1 |
| Automatic Documentation and Mathematical Linguistics | 1 |
| Cluster Computing | 2 |
| Conference - ACM | 6 |
| Conference - EMNLP | 1 |
| Conference - IEEE | 9 |
| Conference - wiley | 2 |
| Decision Support Systems | 6 |
| Economic Modelling | 1 |
| Economic Systems | 2 |
| Energy Economics | 1 |
| Expert Systems with Applications | 8 |
| Information Systems | 1 |
| International Journal of Data Science and Analytics | 1 |
| International Review of Economics and Finance | 2 |
| International Review of Financial Analysis | 1 |
| Journal of Banking & Finance | 1 |
| Journal of Behavioral and Experimental Finance | 1 |
| Journal of Computational Science | 1 |
| Journal of Economic Behavior and Organization | 1 |
| Journal of Economic Psychology | 1 |
| Journal of Empirical Finance | 1 |
| Journal of Finance | 3 |
| Journal of Financial Economics | 1 |
| Journal of Intelligent Information Systems | 1 |
| Knowledge Based Systems | 2 |
| Knowledge Information Systems | 1 |
| Multimedia Tools Application | 1 |
| Neural Computing & Application | 2 |
| Neurocomputing | 3 |
| North American Journal of Economics and Finance | 1 |
| Personal and Ubiquitous Computing | 1 |
| PhysicA | 2 |
| Procedia - Social and Behavioral Sciences | 2 |
| Procedia Computer Science | 1 |
| Progress in Artificial Intelligence | 1 |
| Research in International Business and Finance | 1 |
| SCIENCE CHINA Information Sciences | 1 |
| The Journal of Finance and Data Science | 2 |
| The Review of Financial Studies | 2 |

# 3. پیکره­ی مرور

بر اساس نظریه­ی بازار کارامد، اطلاعاتی که از طریق اینترنت در اختیار سرمایه­گذاران قرار می­گیرد، شامل گزارش­های فنی بانک­جهانی، اخبار سیاسی و اقتصادی بر رفتار سرمایه­گذاران موثر می­باشند. در طی سال­های اخیر، روش­های مختلفی با استفاده از تکنیک­هایی نظیر واکشی اطلاعات و پردازش زبان طبیعی به بررسی تاثیر این اطلاعات بر سرمایه­گذاران بورس و بازارهای مالی پرداخته­اند، در حالیکه در مطالعات مروری انجام شده در این حوزه، اغلب روش­های مورد بررسی مربوط به قبل از سال 2017 و با تکیه بر تعداد اخبار منتشر شده به بررسی تاثیر آن­ها بر سرمایه­گذاران پرداخته­اند. وجه تمایز مرور انجام شده با مرورهای[4, 20, 21, 27] در بررسی روش­هایی است که از تکنیک­های جدید در متن کاوی و واکشی اطلاعات استفاده کرده­اند.

در این بخش به بررسی روش­های ارائه ­شده با توجه به نوآوری اصلی آن­ها در هر کدام از چهار بخش ارائه شده در شکل 6- نمودار چهارچوب تحلیل داده­های وب بر پیش­بینی بورس می پردازیم. شکل 6- نمودار چهارچوب تحلیل داده­های وب بر پیش­بینی بورس را نشان می­دهد.

Bag of words

Topic modeling

World embedding

Stock market prediction

Forex market prediction

Investor behavior analysis

News

Social Media

Online Search

Wikipedia pages

Stock message boards

Sentiment Analysis

Anomaly detection

Event modeling

شکل 6- نمودار چهارچوب تحلیل داده­های وب بر پیش­بینی بورس

## 1-3. منابع داده­ای

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 sina wibo، 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. جدول 2- مقایسه­ی روشهایی که از چندین منبع داده­ای استفاده کرده­اند. را نشان می­دهد.

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

همزمان با وقوع یک رویداد خبری، آمار مراجعه به صفحات ویکی­پدیا و تعداد جستجوها در گوگل با کلید واژه­های مرتبط با آن رویداد بیشتر می شود و افراد به بیان نظرات خود در شبکه های اجتماعی می پردازند[28]. این نظریه ای است که منجر به استفاده­ی همزمان از چنین منابع داده­ای برای بهبود دقت پیش­بینی می گردد. چالشی که در استفاده­ی همزمان از چنین منابع داده­ای در کنار داده­های بورس وجود دارد ناهمگن بودن داده­ها نسبت به یکدیگر است. در روش پیشنهادی [15] یک روش مبتنی بر فضای سه­تایی[[27]](#footnote-27) برای غلبه بر ناهمگون بودن داده­های بورس، تحلیل احساس اخبار و نظرات کاربران توئیتر ارائه شده است. با توجه به استفاده­ی همزمان از اخبار و نظرات کاربران توئیتر فضای ویژگی بسیار بزرگ و خلوت خواهد بود، لذا یک روش تجزیه­ی سه­گانه[[28]](#footnote-28) ارائه کرده­اند. روش پیشنهادی [29] یک روش پیشنهاد استراتژی تجارت بر اساس اخبار و توئیت­ها ارائه داده است. در این روش به تفاوت داده­ها در شبکه­های اجتماعی و اخبار توجه شده است. آن­ها نشان داده­اند داده های شبکه­های اجتماعی نظیر توئیتر زودتر از اخبار رویدادها را گزارش می­دهند ولی اخبار قابلیت اعتماد بیشتری دارند. آن­ها با استفاده از این دو ویژگی، در نتایج نشان داده­اند زمانی که تحلیل احساس اخبار و داده­های شبکه اجتماعی در یک جهت حرکت می­کنند بازار نیز در همان جهت حرکت می کند و می توان یک استراتژی تجاری بلندمدت اتخاذ کرد. در روش پیشنهادی[28] برای بهبود دقت پیش­بینی یک پایگاه دانش[[29]](#footnote-29) از منابع داده­ای مختلف نظیر تعداد دفعات بازدید از صفحات ویکی­پدیا[[30]](#footnote-30) مرتبط با سهام شرکت اپل[[31]](#footnote-31) از وب سایت WikipediaTrends.com، حجم اطلاعات آنلاین تولید شده از قبیل اخبار و نظرات افراد در google News راجع به کمپانی اپل، شاخص های فنی متداول در بورس مانند قیمت بستن[[32]](#footnote-32) از وب سایت yahoo finance و شاخص­های مالی فنی[[33]](#footnote-33) نظیر RSI , LW ,… تشکیل داده­اند. در روش پیشنهادی [30, 31] از ترکیبی از داده های شبکه اجتماعی، اخبار و داده­های بورس چین برای پیش­­بینی قیمت سهام در بورس شانگهای استفاده شده است. در روش پیشنهادی [32] از ترکیب داده­های بورس، google trends و تحلیل احساس اخبار برای پیش­بینی سهام اپل استفاده شده است. در روش پیشنهادی [33] یک فرهنگ لغات از واژه­های مربوط به کمپانی­های بورس سهام نیویورک[[34]](#footnote-34) تشکیل شده است و به ازای هر کمپانی تعدادی مفهوم معنایی با استفاده از روش مدلسازی مفاهیم معنایی ال­دی­ای [[35]](#footnote-35)از اخبار منتشر شده در وب سایت YAHOO Finance در بازه ی 2012 تا 2016 استخراج شده است .

جدول 2- مقایسه­ی روشهایی که از چندین منبع داده­ای استفاده کرده­اند.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Literature | Social media | News Source | Other | Market | Dataset interval | Machine Learning | correlation Analysis | TimeFrame |
| Qing Li, 2016[34] | eastmoney post bar | Sina news |  | CSI100 | 2011 | High order tensor regression |  | Minute |
| Steve Y. Yang, 2017[29] |  | Northern Light Single Point business news |  | S&P 500 NASDAQ | 2012-2015 | generic programming optimization | lag estimation | Daily |
| Gaowei Zhang, 2017[31] | Eastmoney post bar | Sina news |  | Shanghai 50ETF | 2008-2015 | naïve bayes, LSTM | Granger causality |  |
| Aditi Kaushal, 2017[32] |  | Thomson Reuters | Google Trends | AAPL stock | 2011-2016 | regression |  | Daily |
| Bin Weng, 2017[28] |  | google news | Wikipedia Trends | AAPL stock | 2012-2015 | svm, decision tree,Neural Net |  | Daily |
| Tomoki Ito,2018[35] | Yahoo!Finance | Reuters |  |  | Sep-15 | gradient interpretable NN |  |  |
| Xi Zhang,2018[15] |  | Xueqiu,Guba,Sina,Hexun |  | HK and CSI100 | 2015 | MFC | Correlation | Daily |
| Shiming Deng, 2018[36] | Weibo microblogging | Sina,Hexun | Baidu and 360 search engine | Shanghai and Shenzhen stock | 2013–2016 | regression |  | Daily |
| Mattia Atzeni, 2018[37] | twitter, stock tweet | AP News, Reuters, Forbes |  |  |  | Lasso regression,SVR | Correlation |  |
| José Gildo de Araújo Júnior,2018[38] | Twitter, Facebook, LinkedIn and GooglePlus, | Handelsblatt |  | FBOVESPA | 2000-2015 | DT,RF, AdaBoost, | Kendall correlation | 15 minutes |

## 2-3. سازماندهی متن

اولین روش­های منتشر شده در حوزه­ی اقتصاد رفتاری اغلب از یک روش رگرسیونی برای بررسی ارتباط میان انتشار اخبار و تاثیر آن بر قیمت در بورس، حجم معاملات ویا بازگشت­های غیر عادی استفاده کرده­اند. در این روش­ها، تحلیل فنی بر اساس تعداد اخبار منتشر شده و یا تحلیل احساسی که توسط یک خبره­ی انسانی انجام گرفته است. روش­های [39-42] با ارائه­ی مدل­های رگرسیونی مبتنی بر پارامترهایی نظیر تعداد اخبار منتشر شده به بررسی میزان بازگشت در بورس پرداخته­اند. روش پیشنهادی [8] که در سال 1998 توسط ویتریچ[[36]](#footnote-36) ارائه شد، را می­توان جزء اولین پژوهش ها در استفاده از تکنیک­های متن کاوی در واکشی اطلاعات سودمند از متون خبری برای پیش­بینی بورس دانست. به دنبال پیشرفت در تکنیک­های متن کاوی و افزوده شدن بر حجم داده­ی متنی در اینترنت، روش­های گوناگونی در نحوه­ی سازماندهی اسناد و آموزش مدل­های پیشگو در این حوزه به کار گرفته شد. در این میان می­توان به سه روش تجمیعی از واژه­ها ، مدلسازی مفاهیم معنایی و جاسازی واژه­ها اشاره کرد. جدول 3- مقایسه­ ی روش ها بر اساس نحوه­ی سازماندهی متن را نشان­می دهد. در این جدول منبع وبی و مدت زمانی که داده­ی متنی از آن استخراج شده است بیان شده است. 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[43]. روش جاسازی واژه­ها در سال 2013 توسط میکولو یکی از محققان گوگل برای شرایطی که تعداد نمونه های آموزشی بسیار زیاد باشد و در واقع با مسئله­ی عظیم داده­ها مواجه باشیم ارائه شد[44]. این تکنیک با استفاده از دو روش مبتنی بر شبکه یادگیری عمیق اسکیپ گرام[[37]](#footnote-37) و تجمیعی از واژه­های پیوسته[[38]](#footnote-38) به ازای هر واژه در پیکره­ی متنی یک بردار تولید می کند. نمودار شکل 8- روش سازماندهی متن بر اساس سال انتشار منبع مورد استفاده را نشان می­دهد. این نمودار نشان می­دهد از سال 2017 به بعد جهت گیری روش­ها بر استفاده از تکنیک جاسازی واژه­ها می­باشد. در ادامه به بررسی هر کدام از روش­های ذکر شده در شکل 8- روش سازماندهی متن بر اساس سال انتشار منبع مورد استفاده می­پردازیم.

شکل 8- روش سازماندهی متن بر اساس سال انتشار منبع مورد استفاده

جدول 3- مقایسه­ ی روش ها بر اساس نحوه­ی سازماندهی متن

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Category | literature | Text Mining attribute | | | | Analysis | | |
|  |  | Media | Media Source | Media Source Interval | Feature Selection /Extraction | Market | Time Frame | Machine Learning |
| BOW | R. P. Schumaker, 2012 | news | Yahoo Finance | 10/2005 - 11/2005 | POS | S&P500 | 20 minutes | SVR |
| M.Hagenau, 2013 | news | DGAP and EuroAdhoc | 19997 - 2011 | 2 word combination | DGAP | Intra Day | SVM |
| A. K. Nassirtoussi,2015 | news | MarketWatch.com | 2008 - 2011 | Ontology | Forex | 2 hours | SVM |
| M. Dang,2016 | news | Vietnam News Wires | 2014 - 2015 | POS | Vietnam stock market | 2 hours | SVM |
| S. Seifollahi,2018 | news | MarketWatch.com | 2008- 2011 | POS | Forex | Intra Day | SVM |
| LDA | J. CuiQing ,2014 | social media | Yahoo! Finance |  | LDA | 2008 America bank crisis | 20 minutes | regression |
| Y.RAO,2016 | news | Hexun | 1/2015 - 12/2015 | LDA | CSI100 | daily | LP-LDA |
| A. Atkin,2018 | news | Reuters | 9/2011 to 09/2012 | LDA | NASDAQ | minutes | naïve bayes |
| H. Petr,2018 | news | Reuters database | 2012-2016 | LDA | NYSE | daily | DNN |
| Word Embedding | H. D. Huynh, 2017 | news | Thomson Rueters, bloomberg | 2006-2014 | word2vec | S&P500 | day-week | LSTM |
| M. R. Vargas, 2018 | news | Reuter | 2006-2013 | word2vec | S&P500 | daily | LSTM,CNN |
| Q. Liu, 2018 | news | Reuters | 2007-2012 | word2vec | S&P 500 | daily | HCAN. |
| X. Zhang, 2018 | news+social media | Xueqiu,Guba,Sina,Hexun | 1/2015 - 12/2015 | word2vec | HK and CSI100 | daily | MFC |

### 3-2-1. تجمیعی از واژه­ها

از جمله روش­های قدیمی در بازنمایی متن روش تجمیعی از واژه‌‌ها[[39]](#footnote-39) می­باشد[45] که در آن تعدادی از واژگان موجود در متون در یک پیکره متنی، بعنوان یک فرهنگ لغت در نظرگرفته می‌شوند. در این روش، فرایند متن‌کاوی شامل ساختن ماتریسی از میزان فراوانی واژه‌‌ها در هر کدام از اسناد می­باشد. در روش پیشنهادی [25] را می­توان جزء اولین روش­های ارائه شده با تحلیل بنیادین دانست. در این روش به بررسی تاثیر انتخاب کلمات در متن اخبار و رفتار مشتریان بازار پرداخته شده است و متون خبری به روش تجمیعی از واژه­ها سازمان یافته­اند. در شرایطی که تعداد واژه‌‌های موجود در فرهنگ لغات بسیار زیاد باشد چالش­هایی از جمله ابعاد بالای ماتریس‌ استخراج شده و یا مسئله‌ی ماتریس­ خلوت بوجود می‌آید که سبب کاهش دقت در تحلیل احساس خواهد شد. در میان روش­های مرور شده در تحلیل احساس متون خبری در پیش­بینی بورس، نوآوری دسته­ای از روش­ها در نحوه­ی انتخاب ویژگی[[40]](#footnote-40) جهت غلبه بر مسئله بوده است. روش پیشنهادی[46] در فرایند انتخاب ویژگی تکنیک ترکیب دو واژه[[41]](#footnote-41) را ارائه داده است که تعمیم یافته­ی 2-gram است با این تفاوت که فاصله ی بیشتر از صفر تا 5 را مجاز می داند. در فاز انتخاب ویژگی از روش محاسبه ی قدرت توصیف گری کلمات به دو روش chi-square و Bi-Normal Separation بر اساس میزان مثبت بودن ترم­ها استفاده شده است. آن ها برای آموزش مدل خود، بر اساس نوسانات قیمت بورس یک فرهنگ لغات بصورت ترکیب دو واژه، دارای برچسب تولید کرده اند. در روش پیشنهادی [23] در فاز انتخاب ویژگی، تنها نهادهای فعلی و صفات موجود در متن خبر بعد از فرایند تعیین نقش نحوی واژه­ها به عنوان ویژگی انتخاب می شوند. در روش پیشنهادی [26] در فرایند انتخاب ویژگی از عناوین خبری پس از پیش پردازش های متداول، تنها کلماتی استخراج شده اند که عینا در هستان­شناسی[[42]](#footnote-42) وردنت وجود داشته اند و مترادف آن ها نیز به مجموعه ویژگی ها افزوده شده است. سپس مقدار تعداد دفعات تکرار در سند نسبت به فراوانی کل[[43]](#footnote-43) برای هر ویژگی محاسبه و وزنی بر اساس احساس مستخرج از هستان­شناسی سنتی­وردنت[[44]](#footnote-44) به عنوان ضریب در آن لغت اثر کرده است که این وزن بیانگر مثبت یا منفی بودن آن لغت می باشد ، لغاتی که دارای وزن کمتر از حد آستانه هستند از فضای ویژگی ها حذف شده اند. از جمله معایب استفاده از هستان­شناسی­ها در فرایند انتخاب ویژگی، وجود واژه­های دارای ابهام [[45]](#footnote-45)می­باشد. در روش پیشنهادی [13] به حذف نقش واژه­های دارای ابهام از مجموعه ویژگی­ها جهت بهبود تحلیل احساس پرداخته شده است. روش ارائه شده با بررسی نقش نحوی کلمه و کلمات مجاور لغت به محاسبه ی میزان احساس لغات مجموعه ویژگی ها پرداخته است.

در تمام روش­های ذکر شده، فرایند انتخاب ویژگی تا زمانی پاسخگو خواهد بود که با چالش عظیم داده­ها[[46]](#footnote-46) مواجه نباشیم. در حالیکه با گسترش استفاده از اینترنت، سرمایه­گذاران با حجم زیادی از داده­ها مواجه هستند.

### 3-2-2. مدلسازی مفاهیم معنایی

یکی از راه­حل­های ارائه شده جهت کاهش ابعاد ویژگی­ها، روش استخراج ویژگی مدلسازی مفاهیم معنایی با استفاده از تکنیک تجزیه­ی نامنفی ماتریس­ها می­باشد. در سال 2003 بلی[[47]](#footnote-47) و سایرین در[47] روش ال­دی­ای، که یک روش تجزیه ماتریس می باشد برای مدلسازی مفاهیم معنایی ارائه دادند. در روش آن ها هر سند به صورت توزیع روی تعدادی مفهوم معنایی مدل می شود. در واقع هدف اصلی این روش کاهش ابعاد ویژگی های مسئله، غلبه بر مسئله ی خلوت بودن ماتریس ها در مدل تجمیع واژه­گان و مدلسازی مفاهیم معنایی بوده است. در روش پیشنهادی [48] برای بررسی فرضیه­ی موثر بودن اخبار اقتصادی بر پیش بینی پارامتر بازگشت در سهام[[48]](#footnote-48) با استفاده از روش تجزیه ی نا منفی ماتریس ال­دی­ای به استخراج تعدادی مفهوم پرداخته و هر سند با توزیع رو هر کدام از مفاهیم برداری شده است و سپس پیش­بینی با استفاده از تحلیل احساس مفاهیم استخراج شده انجام گرفته است. در روش پیشنهادی [33] یک فرهنگ لغات از واژه­های مربوط به کمپانی­های سهام بورس نیویورک تشکیل شده است و به ازای هر کمپانی تعدادی مفهوم معنایی با استفاده از روش ال­دی­ای از اخبار منتشر شده در YAHOO! Finance در بازه ی 2012 تا 2016 استخراج شده است . سپس برای هر مفهوم با استفاده از روش مبتنی بر فرهنگ لغات یک رتبه­ی احساس نسبت داده شده است. سپس با استفاده از روش شبکه­ی عصبی به پیش­بینی پرداخته است. در روش پیشنهادی[49] به بررسی و تحلیل پویای رفتار ذینفع­های سهام در وقوع رویداد بحرانی ورشکستگی بانکداری در آمریکا در سال 2008 در سه بازه ی قبل، حین و بعد از وقوع پرداخته شده است. در روش پیشنهادی با تکیه بر ویژگی های آماری متن مانند طول متن، تعداد جملات آن؛ خوشه بندی این ویژگی ها به شناسایی افراد ذینفع در یک سهام پرداخته شده است. سپس مفاهیم معنایی مرتبط با استفاده از الگوریتم ال­دی­ای استخراج شده است. در نهایت از یک مدل رگرسیونی خطی مبتنی بر ویژگی های آماری لغات در هر رویداد، تحلیل احساس هر رویداد ورتبه­ی کلمات کلیدی در هر خوشه، برای پیش بینی حجم و بازگشت سهام استفاده شده است. روش پیشنهادی [50] پس از استخراج مفاهیم معنایی مرتبط با سهام خاص، به محاسبه­ی میزان تاثیر آن­ها پرداخته است.

### 3-2-3. جاسازی واژه­ها

روش جاسازی واژه­ها[[49]](#footnote-49) در سال 2013 توسط میکولو[[50]](#footnote-50) یکی از محققین گوگل برای سازماندهی متون وبی در عظیم داده­ها ارائه شد که در آن به ازای هر واژه در سند بر اساس لغاتی که معمولا همراه آن لغت در متن ظاهر می­شوند با استفاده از یک شبکه یادگیری عمیق یک بردار[[51]](#footnote-51) تولید می­شود [44]. در روش پیشنهادی[15] برای شناسایی رویدادهایی نظیر خرید یا فروش، به ازای هر سند خبری، افعال موجود در عنوان هر خبر با کمک یک ابزار تعیین نقش نحوی لغات شناسایی شده­اند. سپس با استفاده از تکنیک بردار برای هر واژه، 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، بطوریکه هر خوشه نماینده ی افعالی است که یک مفهوم مانند صعود یا نزول قیمت را مدل می کنند. در این روش با استفاده از مفاهیم استخراج شده، تحلیلی احساس داده­های شبکه ی اجتماعی و داده­های بورس چین به پیش­بینی پرداخته است. در روش­های پیشنهادی [16, 51, 52] برای بهبود دقت پیش­بینی، اسناد خبری با استفاده از روش بردار برای هر واژه، برداری شده­اند و پیش­بینی با استفاده از شبکه­ یادگیری عمیق انجام گرفته است.

## 3-3. سازماندهی اطلاعات

یکی از زمینه­های نوظهور کاربرد تحلیل داده­های عظیم در واکشی و استخراج اطلاعات[[52]](#footnote-52) مورد نیاز از داده­های متنی وبی و شبکه­های اجتماعی است[53]. امروزه تکنیک­های پردازش زبان طبیعی مانند تحلیل احساس به واکشی و استخراج اطلاعات از منابع موجود در شبکه جهانی اینترنت کمک می­کنند. استخراج چنین اطلاعاتی می­تواند در بسیاری از زمینه­ها از جمله تدوین سیستم­های پشتیبان تصمیم سرمایه­گذاران مالی بسیار سودمند باشد. از میان روش­های مرور شده دسته­ای از روش­ها با استفاده از تحلیل احساس به استخراج اطلاعات مورد نیاز در جهت طراحی چنین سیستم­هایی 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 دیگری 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. شکل 9 پراکندگی روش­ها بر اساس نحوه­ی واکشی اطلاعات را نشان می­دهد. در ادامه به بررسی هر کدام از روش­های واکشی اطلاعات پرداخته می­شود.

شکل 9 پراکندگی روش­ها بر اساس نحوه­ی واکشی اطلاعات

### 1-3-3. تحلیل احساس

فرایند تحلیل احساس متون منابع متنی شامل دسته بندی منابع به سه دسته احساس مثبت و منفی و خنثی می باشد. اغلب از دو روش مبتنی بر فرهنگ لغات[[53]](#footnote-53) دارای برچسب و روش مبتنی بر یادگیری ماشین[[54]](#footnote-54) 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. در روش مبتنی بر فرهنگ لغات، لغات موجود در یک سند متنی بر اساس برچسبی که در فرهنگ لغات دارند، برچسبی در بازه­ی [-1,1] می­گیرند و در نهایت قطبیت احساس سند بر اساس حاصلجمع برچسب­ها محاسبه می­شود[54]. شکل 10پراکندگی روش ها بر اساس انواع نرم افزارها یا فرهنگ لغات مورد استفاده در تحلیل احساس را نشان می­دهد .­

شکل 10پراکندگی روش ها بر اساس انواع نرم افزارها یا فرهنگ لغات مورد استفاده در تحلیل احساس

در روش مبتنی بر یادگیری ماشین، ابتدا معمولا عناوین خبری بر حسب زمان انتشار و نوسان نرخ بازگشت در سهام در همان زمان، برچسب میگیرند. برای مثال در صورتی که بازگشت در سهام در زمان t نسبت به زمان t-1 نزول داشته باشد خبر منتشر شده در زمان t برچسب منفی می گیرد. سپس به روش کیسه ای از لغات فرهنگ لغاتی از کلمات دارای برچسب تشکیل می گردد که در فاز تحلیل احساس، از این پیکره استفاده می گردد.

از این رو که جنس داده­های متون خبری با پست­های افراد در شبکه­های اجتماعی متفاوت است و معمولا داده­ها با منشا شبکه­های اجتماعی به زبان گفتار نزدیک تر هستند و دارای حجم زیاد و سرعت بالای تولید و نویز زیاد می­باشند، فرایند تحلیل احساس اخبار و شبکه­های اجتماعی جداگانه بحث می­شود. نوآوری روش­های ارائه شده در هر دو بخش تحلیل احساس اخبار یا شبکه­های اجتماعی ارئه­ی یک روش جدید در تحلیل احساس با توجه به چالشی که در آن زمینه بوده است می­باشد.

**تحلیل احساس اخبار** : روش پیشنهادی [55] در سال 2015 با بیان این مسئله که روش های پیشین به تحلیل اخبار به روش تجمیعی از واژه­ها پرداخته­اند و به تحلیل احساس آن توجهی نداشته­اند، با استفاده از روش مبتنی بر فرهنگ لغات به پیش­بینی بورس بر اساس تحلیل احساس و داده­های مالی پرداخته است. از دو فرهنگ لغت Harvard psychological dictionary وLoughran–McDonald استفاده نموده­اند. روش پیشنهادی[56] با ارائه ی یک روش مبتنی بر یادگیری ماشین به استخراج یک پیکره از لغات دارای برچسب بر اساس فرکانس تکرار در اخبار برچسب گرفته بر اساس مدل پیشگوی GARCH پرداخته است. در این روش اخبار به دو دسته ی جذاب و غیر جذاب دسته بندی شده اند. آن­ها با در نظر داشتن سری زمانی نوسان با استفاده از مدل پیشگوی GHARCH پیش بینی بازار برای لحظه ی بعد انجام داده­اند. در صورتی که تفاوت بیشتر از حد آستانه بین مقدار واقعی و پیش بینی شده وجود داشته باشد به اخباری که چند دقیقه قبل از این تایم فریم منتشر شده اند برچسب جذاب داده شده است و بدین صورت یک مجموعه داده دارای برچسب از اخبار تشکیل شده است. با این حال، اخباری دارای برچسب جذاب هستند که بتواند تاثیر معناداری در بازار داشته باشند. در این روش از اخبار و اطلاعات اقتصادی مربوط به کمپانی های تجاری در بورس کشور پرتقال استفاده شده است. روش پیشنهادی [57] با ارائه­ی یک روش تحلیل احساس مبتنی بر فرهنگ لغات، سعی در شناسایی احساس های صریح و ضمنی در متون خبری اقتصادی وابسته به چند کمپانی هلندی داشته­اند. آن­ها با بیان چالشی که در برخی جملات متون خبری احساس به صورت صریح توسط لغات دارای بار احساسی بیان نشده است به ارائه­ی روش تحلیل احساس بر اساس نقش نحوی کلمات در جمله برای تحلیل احساس ضمنی پرداخته­اند. در روش پیشنهادی [26] به تحلیل احساس متون خبری بر اساس نقش معنایی کلمات در عنوان خبر پرداخته شده است. در فرایند استخراج ویژگی از عناوین خبری پس از پیش پردازش های متداول، تنها کلماتی استخراج شده اند که عینا در آنتولوژی وردنت وجود داشته اند و مترادف آن ها نیز به مجموعه ویژگی ها افزوده شده است. سپس مقدار فراوانی تکرار در سند نسبت به کل متن[[55]](#footnote-55) برای هر ویژگی محاسبه و وزنی بر اساس آنتولوژی سنتی­وردنت به عنوان ضریب در آن لغت اثر کرده است که این وزن بیانگر مثبت یا منفی بودن آن لغت می باشد ، لغاتی که دارای وزن کمتر از حد آستانه هستند از فضای ویژگی ها حذف شده اند. در روش پیشنهادی [14] دسته­بندی متون خبری به سه دسته مثبت، منفی و خنثی به روش مبتنی بر فرهنگ لغت[[56]](#footnote-56) و تکنیک کاوش در قواعد انجمنی[[57]](#footnote-57) انجام گرفته است. از مجموعه داده­ی LM استفاده شده است. در این روش به کلمات بعد از تعیین نقش نحوی به کمک ابزار porterStemmer 6 نوع برچسب یا شاخص کارایی اختصاص می­یابد و سپس با استفاده از تکنیک کاوش در قواعد انجمنی طبقه­بندی انجام گرفته­ است. روش پیشنهادی [58] یک فرهنگ لغات دارای بار احساسی چند زبانه از پیکره متنی چند زبانه­ای در حوزه­ی اقتصادی ساخته­اند. روش پیشنهادی [59] با بیان این نکته که اغلب افراد علاوه بر توجه به عنوان خبر در بخش­هایی از متن خبر نیز به دنبال اطلاعات مورد نیاز خود می­گردند با بیان یک رابطه ریاضی بخش­های جذاب از متن و عنوان خبر را استخراج نموده و سپس به تحلیل احساس آن­ها پرداخته است. در روش [60] علاوه بر فرایند تحلیل احساس اخبار به دو دسته ی مثبت و منفی ، برای هر خبر منبع انتشار آن خبر و نقطه نظر مربوط به آن نیز بررسی شده است. منبع خبر می تواند یکی از 4 دسته ی (رسانه ی دولتی مرکزی ، رسانه ی دولتی محلی، رسانه ی غیر دولتی و رسانه های اقتصادی چینی ) باشد و برای هر خبر تحلیل آن از نقطه نظر دانشگاهی ، دولتی و صنعت نیز وجود دارد. روش­های [61, 62] علاوه بر تحلیل احساس اخبار از ویژگی­هایی نظیر میزان ارتباط هر خبر به هر کدام از شرکت های صاحب سهام و میزان جدید بودن آن خبر را نیز استفاده کرده­اند.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Category | Litrature | Media source | | SA[[58]](#footnote-58) Method | | Tools | | Feature Selection | | | Market | correlation | | | Machine Learning | |
| News Based Sentiment Analysis | P.S.M. Nizer, 2012 | Portuguese news wires | | Machine Learning | |  | | BOW | | | BOVESPA |  | | | GARCH,SVM, Naïve Bayes | |
| Xiaodong Li,2014 | FINET | | Lexicon based | |  | | BOW | | | S&P 500 |  | | | SVM | |
| A. K. Nassirtoussi,2015 | MarketWatch.com | | lexicon based | | Sentiwordnet | | Ontology | | | Forex |  | | | SVM | |
| Marjan Van de Kauter, 2015 | a corpus of Dutch sentences | | lexicon based | | Pattern and Duoman lexicons | | POS | | | Dutch Company |  | | | SVM | |
| S.Krishnamoorthy , 2017 | financial phrase-bank dataset | | lexicon based | |  | | POS | | |  |  | | | ARM- SVM,NB | |
| Social Media Based Sentiment Analysis | Johan Bollen, 2011 | Twitter | | software | | Openion Finder , GPOMS | | news count | | | Dow Jones | Granger cross correlation | | | fuzzy neural net | |
| Xue Zhang, 2011 | Twitter | | hashtag | |  | | tweet count | | | DJIG, NASDAQ , S&P 500 | Pearson | | |  | |
| Andrew Sun, 2016 | Stocktweet | | hashtag | |  | | matrix factorization | | | S&P 500 |  | | | regression | |
| Mariana Daniel , 2017 | Twitter | | lexicon based | | Sentiment140 | |  | | | Dow Jones |  | | |  | |
| Chi-San Ho,2017 | Twitter | |  | |  | |  | | | Dow Jones | correlation | | | Regression-DBN | |
| Mattia Atzeni,2018 | | Twitter, StockTweet | | Ontology | | Sentiwordnet | | n-gram , Ontology |  | | | correlation | Lasso Regression,SVR | |

* **تحلیل احساس داده­ها­ی شبکه اجتماعی** : افراد اغلب رویدادهای واقعی را در توئیتر گزارش می کنند[10]. نظرات افراد مختلف در توئیتر و شبکه های اجتماعی اینچنینی می­تواند منبع مناسبی برای تصمیم­گیری سرمایه­گذاران باشد. در برخی از پژوهش­های با رویکرد پیش بینی مالی به بررسی نقش چنین رسانه هایی پرداخته شده است. از جمله ویژگی­های رسانه­های اجتماعی حجم عظیم داده­ها و سرعت بالای تولید آن­ها، وجود نویز زیاد در نگارش متن پیام­های منتشر شده، ابهام در احساس بیان شده در جملات، خلوت بودن ماتریس ویژگی­ها و پویا بودن نظرات در زمان می باشد. در هر کدام از پژوهش های زیر به بخشی از این چالش­ها توجه شده است. روش پیشنهادی [63] که تا کنون بیش از 1000 ارجاع به آن شده است، به بررسی تاثیر توئیتر بر سرمایه­گذاران پرداخته است. آن­ها توئیت­ها را با استفاده از ابزارهای Opinion Finder و GPOMS به ترتیب به دو و شش سطح احساس دسته­بندی کرده­اند سپس با استفاده از تست علیت گرنجر[[59]](#footnote-59) به بررسی همبستگی متقابل[[60]](#footnote-60) بین سری زمانی احساس­های استخراج شده و بازگشت در بازار پرداخته­اند. روش پیشنهادی [64] از تگ­های بیانگر احساس نظرات افراد در توئیتر استفاده نموده است و به بررسی همبستگی بین تگ هایی نظیر ترس، اضطراب، خوشحالی و روند سهام DOW Jones، S&P 500 و NASDAQ پرداخته است. نتایج آن­ها نشان می­دهد سرمایه­گذاران در زمانی که عدم قطعیت وجود دارد بیشتر از داده­های احساسی استفاده می کنند. روش پیشنهادی[12] با در نظر گرفتن حجم زیاد داده های مربوط به شبکه های اجتماعی، با ارائه ی یک روش جدید در تجزیه ی ماتریس ها برای غلبه بر مشکل فضای خلوت[[61]](#footnote-61) سعی در مدلسازی تاثیر نظرات کاربران در بازگشت در سهام S&P 500 داشته است. در روش پیشنهادی [37] یک روش با نظارت تحلیل احساس با استفاده از ویژگی­های لغوی و معنایی استخراج شده از عناوین خبری و داده­های شبکه اجتماعی ارائه داده است. در این روش برچسب عددی پیوسته در بازه­ی [-1,1] می باشد. در این روش انتخاب ویژگی با استفاده از ویژگی های نحوی(UniGram ,2 Gram , 3 gram)، ویژگی­های معنایی و ترکیبی از هر دو حالت انجام گردیده است. سپس از یک روش تخمین همبستگی لغات با هر سند برای تعیین قطبیت[[62]](#footnote-62) اسناد مجموعه ی آموزشی استفاده کرده است برای غلبه بر پردازش داده های عظیم آزمایش ها بر بستر محیط Spark انجام شده است. در روش پیشنهادی[65] به بررسی پویای نظرات سرمایه گذاران در زمان پرداخته شده است . دراین روش، نظرات در هر لحظه به نظر و احساس سرمایه گذاران در تایم فریم قبل وابسته در نظر گرفته شده اند. جهت پیش بینی بازگشت در بورس از دو دسته داده های مالی مربوط به بازار و تحلیل احساس نظرات کاربران YahooFinance استفاده شده است.

### 3-3-2. شناسایی رویداد

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[[63]](#footnote-63). 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. روش پیشنهادی [66] برای بررسی نقش رویدادهای سیاسی، اجتماعی و یا اقتصادی در بورس به گروه بندی اخبار بر اساس نوع جهت گیری گروه خبری که خبر را منتشر می کند پرداخته است. آن­ها اخبار را از گروه­های خبری مختلفی که به طور تخصصی رویدادهای با موضوعات فقط سیاسی، اجتماعی و یا ملی را گزارش می کنند جمع آوری کرده­اند و یک مجموعه داده­ی خبری دارای برچسب بر اساس نوع گروه خبری ساخته­اند. سپس با استفاده از روش آزمون علیت [[64]](#footnote-64) گرنجر به بررسی فرضیه ی موثر بودن رویدادهای هر دسته بر نوسانات بورس با توجه به الگوی تکرار شده در گذشته پرداخته­اند. روش های مختلفی سعی در شناسایی خودکار رویدادها از اخبار یا شبکه­های اجتماعی داشته­اند. در روش پیشنهادی [67] رویدادها به دو دسته ی اصلی و فرعی تاثیر­گذار بر بازار تقسیم شده­اند . در روش آن ها رویداد اصلی رویدادی تعریف شده است که لغات موجود در اخبار گزارش­دهنده­ی آن رویداد، به یکباره ظهور پیدا می کنند و سپس در مدت کوتاهی محو می شوند و رویداد فرعی رویدادهایی تعریف شده­اند که با لغات متداول در بازار سهام گزارش می شوند و در کنار یکدیگر سبب تغییر احساس سرمایه­گذاران و تغییر در بازار می شوند. رویدادهای اصلی با توجه به الگوی متمایزی که در رخداد کلماتی که خبر با آن­ها گزارش می­شود با استفاده از معیار واکشی اطلاعات متقابل[[65]](#footnote-65) و احتمال درستنمایی[[66]](#footnote-66) آن­ها شناسایی شده­اند و رویدادهای فرعی با اعمال روش نظارتی راهنمای بردار پشتیبان[[67]](#footnote-67) بر ماتریس فراوانی لغات- اسناد شناسایی شده­اند. همچنین در روش پیشنهادی [68] یک روش تشخیص رویداد های آشوبناک[[68]](#footnote-68) در بازار با استفاده از تحلیل داده­های توئیتر ارائه شده است. آن ها از فاکتورهایی نظیر حجم توئیت، متن توئیت، تحلیل احساس توئیت­ها و منطقه ی 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 روش پیشنهادی [10] شناسایی رویداد بر اساس تعداد توییت منتشر شده در هر روز و تحلیل احساس آن­ها 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 می گردد. روش پیشنهادی [69] با استفاده از تحلیل احساس و معرفی یک شاخص تجمعی بر اساس آن به شناسایی رویداد و مدلسازی تاثیر آن با استفاده از تست علیت گرنجر پرداخته است. فرضیه ی این تحقیق بر این پایه استوار است که چه نوع احساسی سبب ایجاد یک حرکت در بورس می شود. 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 آزمون علیت گرنجر به بررسی فرضیه­ی تحقیق خود پرداخته­اند. نتایج این تحقیق نشان می­دهد در هر کدام از تغییرات شدید احساس، یک رویداد مهم رخ داده است.

دسته­ای از روش­ها با توجه به این نکته که یک رویداد اغلب توسط تعدادی خبر منتشر می­شود به شناسایی رویداد بر اساس ارتباط محتوایی اخبار پرداخته­اند. بر اساس تعریف ارائه شده در [70] یک رویداد توسط دسته اخباری با موضوع یکسان در یک بازه زمانی مشخص گزارش می­شود. در روش پیشنهادی [24] از دو آنتولوژی وردنت و ورب­نت برای شناسایی رویداد استفاده شده است. از اخبار ویژگی­هایی به فرم (فاعل، فعل، مفعول) با استفاده از روش تعیین نقش­های نحوی[[69]](#footnote-69) استخراج شده است و یکسان سازی افعال با استفاده از آنتولوژی ورب­نت انجام شده است اخباری که با افعال مترادف گزارش می شوند در یک دسته قرار گیرند و به عنوان رویداد تلقی شوند روش پیشنهادی [71] رویدادهای استخراج شده بر مبنای روش [24] را تحت عنوان جاسازی رویدادها به عنوان ویژگی­های ورودی یک شبکه یادگیری عمیق به کار برده است. همچنین در روش پیشنهادی [66] با استفاده از تکنیکی مبتنی بر شبکه عصبی هر پاراگراف به برداری نگاشت شده و سپس برچسب دهی با مقادیر سیاسی، اقتصادی، ملی تعیین می گردد، سپس با استفاده از تست علیت گرنجر به بررسی فرضیه ی موثر بودن رویدادهای هر دسته بر نوسانات بورس با توجه به الگوی تکرار شده در گذشته پرداخته است. همچنین روش پیشنهادی [15] 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 حقیقت مفاهیم معنایی با خوشه بندی استخراج شده­اند، بطوریکه هر خوشه نماینده ی افعالی است که یک مفهوم مانند صعود یا نزول قیمت را مدل می کنند. روش پیشنهادی . [72] یک کرنل بر پایه­ی روش ماشین بردار پشتیبان ارائه دادند که در آن ارتباط بین آیتم های خبری با استخراج کلمات کلیدی با تکنیک 2gram برداری شده اند و فرکانس تکرار آن ها در هر آیتم خبری محاسبه و ارتباط کلمات کلیدی در آیتم های خبری توسط یک گراف مدل شده است. روش پیشنهادی سعی دارد با مدلسازی گراف، اخباری که به نوعی یک رویداد را گزارش می دهند مرتبط سازد. در روش­های ارائه شده در این دسته به تحلیل احساس و الگوی انتشار توجهی نشده است. جدول 4 مقایسه­ی روش­های شناسایی رویداد را نشان می­دهد.

جدول 4 مقایسه­ی 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, [69] | 1 |
| NO | No | Dow Jones related Tweet noise removing and Sentiment Score Analysis for event Detection | Sentiment Analysis | Mariana Daniel 2017, [10] | 2 |
| No | MNB[[70]](#footnote-70) and SGD[[71]](#footnote-71) 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 [67] | 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 [68] | 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 [24] | 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 [71] | 6 |
| Granger Causality | LSTM Deep Neural Network | Paragraph2Vec similarity based classification on political, economical and ... category | Content based Analysis | Ishan Verma 2017 [66] | 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 [15] | 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 [72] | 9 |

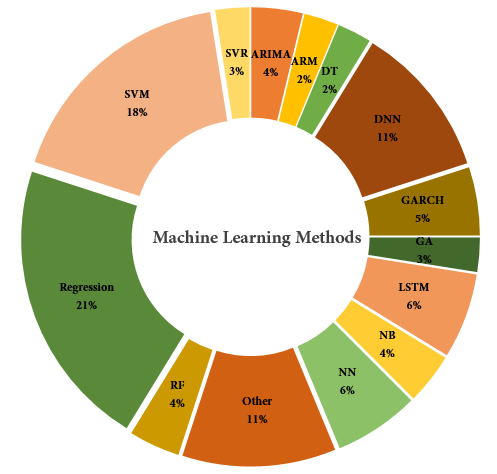
## 4-3. استخراج دانش

در طی سال­های اخیر استخراج دانش با هدف کشف الگو از یک مجموعه عظیم از داده­های وب به یک زمینه جذاب تحقیقاتی از دیدگاه صنعت تبدیل شده است[73]. تحلیل داده، یادگیری ماشین و کشف دانش را می توان یک زمینه­ی مطالعاتی حاصل از اشتراک علم کامپیوتر، هوش مصنوعی و آمار و ریاضی دانست[74]. در مرور انجام شده استخراج اطلاعات از متون وبی جهت کشف دانش در سه حوزه­ی تحلیل رفتار سرمایه­گذاران بر پایه­ی آزمون­های آماری، ارائه­ی سیستم های توصیه گر مبتنی بر روش های هوش مصنوعی و ارائه ی مدل­های پیشگو با استفاده از روش­های یادگیری ماشین به کار گرفته شده است. شکل 11- تکنیک­های کشف دانش و یادگیری ماشین در روش­های مرور شده را نشان می­دهد. در ادامه به بیان هر کدام از سه بخش ارائه شده می­پردازیم.

شکل 11- تکنیک­های کشف دانش و یادگیری ماشین در روش­های مرور شده

### 1-4-3 سیستم­های پیشگو

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 12 انواع روش های یادگیری ماشین انواع تکنیک­های یادگیری ماشین که در روش­های مرور شده به کار گرفته شده است، را نشان می­دهد.



شکل 12 انواع روش های یادگیری ماشین

روش پیشنهادی [75] 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[[72]](#footnote-72) جهت پیش­بینی استفاده شده است. [76] 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[[73]](#footnote-73)، رتبه ی احساس هر خبر[[74]](#footnote-74)، رتبه ی جدید بودن یک خبر[[75]](#footnote-75) و متوسط رتبه ای که با استفاده از تکنیک های مختلف تحلیل احساس برای یک خبر بدست آمده است[[76]](#footnote-76)، نیز هستند. این رتبه ها توسط یک خبره رتبه دهی شده است. در روش پیشنهادی[77] پیش بینی بازگشت در بورس بصورت روزانه با توجه به تحلیل احساس روی اخبار با در نظر گرفتن یک گروه بندی بر شرکت ها انجام گردیده است. بر اساس استاندارد Global Industry Classification Standard(GICS) ایجاد شده توسط S&P شرکت هایی که پیش بینی بر آن ها انجام شده است به 5 گروه، گروه بندی گردیده اند و از مجموعه داده ی LexiNexis برای اخبار استفاده کرده است. در این مجموعه داده اخبار دارای برچسب مثبت، منفی و خنثی می باشند. آن ها در فرایند انتخاب ویژگی از روش chi-square استفاده کرده اند. در فرایند پیش پردازش، اخباری که در طی چند روز متوالی تکرار شده اند فقط یک بار در نظر گرفته شده اند. در روش پیشنهادی[78] به بررسی و تحلیل احساس داده های شبکه ی اجتماعی www.psychsignal.com در بازگشت و حجم معاملات در بازار سهام 5 شرکت معروف و در تایم فریم 2 دقیقه پرداخته است. در تحلیل احساس از جمله روش های مبتنی بر فرهنگ لغات استفاده کرده است و لغات موجود در نظرات به صورت دو دسته ی صعودی و نزولی رتبه اختصاص داده اند و از یک مدل رگرسیونی برای پیش بینی استفاده شده است. نوآوری روش پیشنهادی [79] معرفی یک شاخص آماری به نام شاخص تجمعی احساس اخبارANSI[[77]](#footnote-77) 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 شرکت های 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 [80] با استفاده از روش یادگیری تقویتی به پیش­بینی بر اساس اخبار پرداخته است. روش­های [16, 31, 51, 52, 66] از شبکه عصبی یادگیری عمیق مدلسازی توالی دنباله­ها [[78]](#footnote-78) [81] برای پیش­بینی بر اساس اخبار و قیمت در بازار بورس استفاده کرده­اند.

بازار تبادل ارزهای خارجی[[79]](#footnote-79) که به اختصار آن را فارکس[[80]](#footnote-80) 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[82]، که آن را می­توان تحت تاثیر کمبود مجموعه اخبار مرتبط با این بازار دانست. روش پیشنهادی [26] جزء اولین کارها با رویکرد متن کاوی در بورس فارکس می باشد. آن ها اخبار مربوط به هر جفت ارز را بر اساس مهر زمانی از دو منبع Marketwatch.com و google RSS reader API جمع آوری کرده اند، تمرکز خود را بر جفت ارز EUR/ USD قرار داده است . 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 [83] تاثیر انتشار گزارش­های سازمان یافته اقتصاد کلان و اخبار منتشر شده بر جهش در نرخ بازگشت در بازار Forex بررسی شده است. در این روش پس از حذف نویز از سیگنال بازگشت در ارزهایی نظیر EUR,CHF,JPY,GBP از یک تست آماری ناپارامتری برای تعیین جهش ها استفاده شده است. سپس با استفاده از یک روش مبتنی بر گارچ[[81]](#footnote-81) به بررسی میزان تاثیر رویدادهای خبری که توسط خبره­ی انسانی شناسایی شده­اند و گزارش های ماهانه ی Bloomberg بر جهش های شناسایی شده می پردازد. روش پیشنهادی از تکنیک­های متن کاوی برای شناسایی رویداد استفاده نکرده است و میزان غافلگیری هر گزارش و خبر توسط خبره ی انسانی تعیین می شود. روش پیشنهادی [84] به بررسی تاثیر اخبار بر جفت ارز EUR/ USD در بحران ورشکستگی در سال 2008 پرداخته است. در روش پیشنهادی [85] به بررسی تاثیر اخبار اقتصادی با منشاء FED , ECB بر بازار نوظهور ارزهای جدید CZK, HVF, PLN در اتحادیه ی اروپا در سال های 2011 تا 2015 پرداخته است. آن ها با ارائه ی یک پارامتر به نام بازگشت غیر عادی به بررسی تحلیل احساس اخبار بر اندیکاتورهای مالی در بازه ی زمانی 5 دقیقه ای پرداخته­اند. نتایج نشان می دهد تاثیر خبر نه تنها بعد از وقوع خبر بلکه قبل از ان هم دیده می شود، همچنین میزان بازگشت بسته به منشاء و کیفیت خبر متفاوت است و جفت ارزهایی با پایه ی USD بیشتر از ارزهای اروپایی تحت تاثیر قرار می گیرند.

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 دسته­بندی دو دسته­ای[[82]](#footnote-82) تبدیل شده­است. این دسته روش­ها از معیارهایی نظیر صحت[[83]](#footnote-83)، دقت و recall و همچنین میانگین دو پارامتر دقت و recall تحت عنوان معیار F1 استفاده نموده­اند[86]. 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[[84]](#footnote-84) و یا مجذور خطای مطلق متوسط[[85]](#footnote-85) جهت ارزیابی روش خود استفاده کرده­اند.

جدول 5 مقایسه­ی مدل­های پیشگو

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Literature | Machine Learning | Market | TimeFrame | Parameter | evaluation Metrics | Contribution |
| V.P. Romanov, 2007[75] | ARIMA | Forex | day-week-month | volatility |  | they used Fractal Model for Estimating News Influence on Market Volatility |
| A. K. Nassirtoussi, 2015[26, 86] | 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[83] | 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[76] | 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[51] | 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[77] | 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[16] | 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[84] | 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[85] | 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 |

### 2-4-3 سیستم­های توصیه­گر تجاری

سیستم­های پیشنهاددهنده یکی از عنوان­های داغ تحقیقاتی در حوزه­ی اقتصاد می­باشند که به سرمایه­گذاران در جهت کسب سود بیشتر و جلوگیری از زیان کمک می­کنند. اغلب روش­های ارائه شده در این بخش دارای ساختاری مطابق شکل 13 ساختار سیستم­های پیشنهاددهنده تجاریمی­باشند.

Textual Data

Market Data

Sentiment Analysis

Sliding window

and

Forecasting Model

Recommendation Module

شکل 13 ساختار سیستم­های پیشنهاددهنده تجاری

روش پیشنهادی [87] یک سیستم پیشنهاد دهنده­ی استراتژی تجارت طراحی کرده است که همیشه لیستی از پر فروش­ترین و پرخریدترین سهام­ها را در اختیار سرمایه­گذاران قرار می­دهد. آن­ها برای این منظور از الگوریتم RankNet و ListNet بر تحلیل احساس اخبار بر اساس دو شاخص احساس تجمعی و شوک در احساس استفاده نموده­اند. روش پیشنهادی [88] یک سیستم پیشنهاددهنده­ی استراتژی تجارت بر اساس تحلیل احساس اخبار در بازه­ی زمانی 5 دقیقه­ای با استفاده از یک ابزار تجاری ارائه داده­ است. روش پیشنهادی [89] یک سیستم پشتیبان تصمیم مالی با استفاده از اخبار طراحی کرده است. این سیستم بر اساس اخبار یک استراتژی کوتاه مدت خرید-فروش با استفاده از روش یادگیری ماشین جنگل تصادفی[[86]](#footnote-86) پیشنهاد می­دهد. روش پیشنهادی [29] با استخراج یک جامعه از سرمایه گذاران شناخته شده از توئیتر، یک روش پیشنهاد استراتژی تجارت بر اساس اخبار و توئیت­ها با معرفی دو شاخص فنی بر اساس تحلیل احساس اخبار و توئیتر ارائه داده است. بر طبق نتایج این تحقیق اخبار سه روز قبل، کمک می کند تا نظرات کاربران در روز جاری را توجیه کنیم. آن­ها با استفاده از این ویژگی، در نتایج نشان داده­اند زمانی که تحلیل احساس اخبار و داده­های شبکه اجتماعی در یک جهت حرکت می­کنند بازار نیز در همان جهت حرکت می کند و می توان یک استراتژی تجاری بلندمدت اتخاذ کرد. روش پیشنهادی[90]یک سیستم پیشنهاددهنده­ی سرمایه­گذاری بر اساس تحلیل احساس داده­های شبکه اجتماعی سرمایه­گذاران Guba در بورس چین ارائه شده است. آن­ها با استفاده از روش متداول مبتنی بر فرهنگ لغت یک شاخص توافق برای هر سهام در هر تایم فریم بر اساس تعداد نظرات مثبت و منفی تعریف کرده­اند. نتایج این تحقیق نشان می­دهد بین روند بازار سهام و شاخص احساس یک همبستگی مثبت وجود دارد، احساس منفی طولانی مدت خبر از وقوع یک نزول در بازار می­دهد.

جدول 6مقایسه­ی سیستم­های توصیه­گر

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Literature | Media Source | Sentiment Analysis | Machine Learning | Market | Contribution |
| Q. Song , 2014 [87] | Thomson Rueters | TRNA | ListNet , RankNet | S&P500 | They apply learning-to-rank algorithms to design trading strategies using relative performance of a group of stocks based on investors’ sentiment toward these stocks. |
| T. Geva , 2014[88] | Reuters | Digital Trowel | NN,DT,SLR | S&P500 | they build an end-to-end recommendation process including data preprocessing, modeling, validation, trade recommendations and economic evaluation based on News Sentiment Analysis |
| S. Feuerriegel , 2016 [89] | DGAP | lexicon Based | reinforcement learning | CDAX | they design trading strategies that utilize textual news in order to obtain profits on the basis of novel information entering the market and propose approaches for automated decision making based on supervised and reinforcement learning |
| S. Y. Yang, 2017[29] | Northern Light SinglePoint business news | sentiwordnet | generic programming optimization | S&P 500 NASDAQ | they propose a superior trading strategy based on the sentiment feedback strength between the news and tweets using generic programming optimization method |
| Y. Sun , 2018[90] | Guba | GubaLex |  | CSI300 | a novel stock recommendation system is developed based on a proposed theory concerning the correlation between Guba-based sentiment of the retail investors and the stock market trends in China. |

### 3-4-3 تحلیل رفتار سرمایه­گذاران

تحلیل رفتار سرمایه­گذاران می تواند در اتخاذ تصمیم­های بهتر به آن ها موثر باشد. تحلیل رفتار سرمایه­گذاران شامل پژوهش هایی با محوریت نحوه ی 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[91] 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[[87]](#footnote-87) 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[[88]](#footnote-88) و پی گیران روند[[89]](#footnote-89) 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 نویز حاصل از نوسان[[90]](#footnote-90) نیز در این مدل لحاظ شده است. در روش پیشنهادی [92] به بررسی فرضیه ی مطلع بودن موسسات مالی از اخبار قبل از وقوع آن­ها پرداخته است. آن­ها برای بررسی فرضیه ی خود با ارائه ی 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 [93] 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 روش آماری Copula استفاده کرده­اند و نتایج نشان می دهد بازار به اخبار منفی واکنش مضاعفی نشان می دهد. روش پیشنهادی [94] 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 تایم فریم و قیمت سهام استفاده کرده­اند. روش پیشنهادی [95] 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 [96, 97] برای بررسی واکنش سرمایه گذاران بر اخبار دست دوم[[91]](#footnote-91) ارائه شد. بر اساس یافته های [96, 97] مقاله 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 هفته ی بعد وجود دارد و این روند در طی هفته های 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 [98] 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 تحلیل همبستگی بین دو تایم سری می باشد به بررسی همبستگی این دو سری زمانی پرداخته است. روش پیشنهادی [99] به بررسی عوامل تاثیرگذار بر ریسک پذیری افراد در سرمایه گذاری روی بورس کشور آلمان پرداخته است. در مجموعه داده ی استفاده شده، از تعدادی از فرد 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 معادله ی تخمین ریسک سرمایه­گذاری در بازه­های کوتاه مدت دو روزه – یک هفته ای – ماهانه و سالانه انجام گردیده است. در روش پیشنهادی[100] به بررسی همبستگی بین شاخص احساس سرمایه گذاران و میزان بازگشت در بورس آمریکا به صورت متغیر با زمان در بازه ی بلند مدت پرداخته شده است.آن ها با استفاده از روش DCC-MIDAS به بررسی متغیر با زمان همبستگی پرداخته اند در حالیکه سایرین مدل خود را ایستا در نظر گرفته اند.

جدول 7 یافته ها در روش های تحلیل رفتار

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Literature | sentiment Analysis | Market | Analysis Method | Time Frame | Parameter | Finding |
| T. Ochiai, 2011[91] | 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[92] | 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[93] | 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[94] | 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[96] | 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[98] | 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 [99] | 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 [100] | 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 [95] | 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. |

# 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[[92]](#footnote-92) وکیفیت داده­های عظیم و چگونگی تحلیل[[93]](#footnote-93)[[94]](#footnote-94) آن­ها دو عامل مهم در ارزشمند کردن داده می­باشد. تحلیل داده های 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 14مراحل 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 داده ها 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 نویز، اعتبارسنجی داده ها و یکسان سازی داده ها و در نهایت یکپارچه سازی داده ها با استفاده از طراحی و بهینه سازی پرس وجو هایی از منابع داده ای مختلف ذخیره­سازی پنج گام قبل از آموزش مدل پیشگو می­باشند[101].

شکل 14مراحل 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 [102] خواص داده­های عظیم تحت 5V شناخته شده است که این ویژگی­ها شامل حجم[[95]](#footnote-95) داده، سرعت[[96]](#footnote-96) تولید، ارزش[[97]](#footnote-97) داده، تنوع[[98]](#footnote-98) داده ، راستی[[99]](#footnote-99) می­باشند. می توان ویژگی ارتباطات[[100]](#footnote-100) 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 15خواص داده­های عظیم

* حجم : با گسترش شبکه های اجتماعی منابع داده ای به سرعت در حال افزایش هستند و حجم داده ی تولیدی در سازمان ها در حال رشد می باشد. و لذا سازمان ها برای مقابله با چالش گسترش حجم داده باید به سراغ پلت فرم های توزیع شده ای نظیر هادوپ[[101]](#footnote-101) به عنوان یک اکوسیستم توزیع شده مدیریت داده های حجیم بروند.
* سرعت : داده ها در پلت­فرم های 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[[102]](#footnote-102) پردازه ها، باید به سراغ ابزارهای پردازش جریانی داده ها برویم که به عنوان مثال به Storm، Spark و Flink اشاره کرد. ابزارهای ذکر شده در کنار مکانیزم های پردازش دسته ای یک موتور پردازش جریانی نیز ارائه داده اند.
* **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 نویز 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.
* **ارتباطات**: به میزان ارتباطات مستقیم به ارتباطات ممکن در یک شبکه گویند. مسئله ی مهم در این ویژگی این است که با گذشت زمان وبرقراری 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.
* **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 نیازمندی های خودش تعیین می کند و با در نظر گرفتن نیاز تجاری، عظیم داده را به یک منبع ارزشمند تبدیل می­کند

درروش­های مرور شده با توجه به نوع داده و هدف پژوهش به برخی از صفات داده­های عظیم توجه شده است. روش پیشنهادی [103] حدود 31 میلیون پیام در توئیتر را در بازه­ی زمانی 22 دسامبر 2012 تا 29 اکتبر 2015 برای پیش­بینی حجم معاملات و بازگشت در سهام 3800 شرکت در بورس آمریکا جمع­آوری نموده است. با توجه به حجم زیاد نمونه­های داده­ای و تنوع آن­ها از لحاظ بدون ساختار بودن متن توئیت­ها، آن­ها داده­های توئیتر را در پایگاه داده­ی بدون ساختار MongoDB ذخیره سازی نموده­اند و با استفاده از تحلیل احساس این داده­ها و دسته بند ماشین بردار پشتیبان دسته­بندی را انجام داده­اند. روش پیشنهادی [37] با در نظر سرعت تولید داده­ها در توئیتر و استوک توئیت به پردازش جریانی داده با استفاده از موتور پردازش جریانی Apache Spark پرداخته­است. در این روش با بیان این نکته که امروزه داده­های تولید شده توسط افراد در شبکه­های اجتماعی و میکروبلاگ­ها به یک منبع ارزشمند برای سازمان­های تحقیقاتی تجاری و نظر سنجی عمومی جهت تحلیل احساس در حوزه­های مالی تبدیل شده­اند، یک روش تحلیل احساس عظیم داده­ها با استفاده از هستان شناسی مبتنی بر قاب[[103]](#footnote-103) به صورت پیوسته ارائه داد­ه­اند. روش پیشنهادی [104] با در نظر گرفتن ویژگی تنوع در عظیم داده­ها یک چهارچوب در جهت تجمیعی[[104]](#footnote-104) از عظیم داده­های ناهمگن برای پیش­بینی سهام شرکت­های چینی ارائه داده است. آن­ها با بیان این نکته که با افزایش حجم و تنوع داده­ها در حوزه­ی مالی، یک فرصت بی سابقه برای بهره بردن از این منبع غنی برای فهم بهتر بازار سهام و پیش­بینی قیمت فراهم شده است، یک چهارچوب جهت بهره بردن از داده­های متنوعی چون مقادیر عددی[[105]](#footnote-105)، داده های ترکیب شده[[106]](#footnote-106) و داده­های تابعی[[107]](#footnote-107) ارائه در راستای بنیان یک مدل پیشگو ارائه داده­اند. در روش پیشنهادی [15] با در نظر گرفتن تنوعی از عظیم داده­های ناهمگن یک فضای سه بعدی[[108]](#footnote-108) و یک مدل رگرسیون ابعاد بالا برای پیش­بینی بورس ارائه داده است. روش­های [51, 66] با استفاده از یک شبکه یادگیری عمیق از اخبار و داده­های بورس برای بهبود دقت مدل پیشگو استفاده کرده­اند. آن­ها از ابزار Tensorflow که یک کتابخانه یادگیری ماشین و به خصوص شبکه یادگیری عمیق به زبان پایتون می­باشد استفاده کرده­اند.

روش پیشنهادی [29] در مواجهه با مسئله­ی استخراج داده­های مورد نیاز و مرتبط با حوزه­ی مالی از توئیتر، با بهره بردن از ویژگی ارتباطات[[109]](#footnote-109) در عظیم داده­ها بین کاربران با علایق مشابه در حوزه­ی مالی، به استخراج جامعه از کاربران توئیتر و واکشی پیام­های منتقل شده بین افراد آن جامعه پرداخته­ است. برای این منظور آن­ها ابتدا 50 حساب شناخته شده کارشناسان سرمایه گذاری را در توییتر شناسایی کرده­اند و از کلمات کلیدی مشترک برای ایجاد لیستی از علایق جامعه سرمایه گذاری مالی استفاده کرده­اند. با ساختن دو لایه پیروان کارشناسان، تعدادی از معیارهای فیلترینگ دقیق را برای ایجاد مرز جامعه مالی بر اساس منافع ماندگار خود در موضوع سرمایه گذاری مالی اعمال کرده­اند و از معیارهایی نظیر مرکزیت بینابینی[[110]](#footnote-110) برای استخراج جامعه هدف خود جهت تحلیل احساس توئیت­های مرتبط و پیش­بینی بورس استفاده نموده­اند.

در آوریل سال 2017 سازمان بورس و اوراق بهادار آمریکا بیان کرد 27 شرکت به توسعه دهندگان وب سایت­ها مبالغی جهت انتشار اخبار کذب مثبت راجع به سرمایه­گذاری در آن شرکت­ها پرداخت کرده­اند[[111]](#footnote-111). در این سال حدود 250 سند حاوی اطلاعات جعلی راجع به این شرکت­ها منتشر شده بود. به دنبال انتشار این اسناد و با توجه به استفاده گسترده افراد از شبکه­های اجتماعی، اخبار و شایعات به سرعت در وب منتشر می­شود و سرمایه­گذاران را تحت تاثیر قرار می­دهد لذا اولین گام در تحلیل عظیم داده، اطمینان از اصالت داده و منبع تولید آن است[105]. یکی از زمینه­های تحقیقاتی مهم اقتصاد رفتاری بررسی تاثیر شایعات و اخبار جعلی بر بازار و سرمایه­گذاران می­باشد. روش پیشنهادی [106] در سال 2001 ارائه شد و به بررسی تاثیر شایعات منتشر شده در روزنامه­های ترکیه بر بازار بورس استامبول[[112]](#footnote-112) پرداخت. نتایج این تحقیق با رویکرد تحلیل بنیادین نشان می­دهد درست در هر چهار روز قبل از انتشار شایعه بازده غیرعادی مثبت قابل توجهی در سهام مشاهده می شود و بازده منفی غیرمعمولی در دوره پس از انتشار شایعه وجود دارد. شناسایی شایعه در این تحقیق توسط خبره انسانی انجام گرفته است درحالیکه در طی سال­های اخیر تحقیق­هایی در زمینه­ی شناسایی شایعه و اخبار جعلی در شبکه­های اجتماعی انجام شده است[107] و البته شناسایی خودکار شایعه و اخبار جعلی در حوزه­ی اقتصاد رفتاری یک زمینه­ی باز تحقیقاتی می­باشد.

# زمینه­های باز تحقیقاتی پیش­رو

گسترش استفاده از روش­های متن کاوی و هوش مصنوعی به ویژه یادگیری ماشین استخراج دانش پنهان در رابطه­ی بین شبکه­های اجتماعی و بازارهای مالی را جهت تجارت خودکار[[113]](#footnote-113) ممکن ساخته است. همچنین با پیشرفت تکنولوژی امکان پردازش حجم زیادی از داده­ها که روزانه در اخبار یا شبکه­های اجتماعی و ویکی­ها تولید می­شوند، در مدت زمان بسیار کوتاهی به صورت بلادرنگ[[114]](#footnote-114) فراهم گردیده است. در طی سال­های اخیر تعداد مطالعات انجام شده در زمینه­ی بررسی ارتباط رسانه و بازارهای مالی با استفاده از روش­های متن­کاوی رو به رشد بوده است[[115]](#footnote-115) و می­تواند نشان از این باور داشته باشد که زمینه­های باز تحقیقی در هر کدام از جهار بخش رسانه، بازنمایی متن، روش­های واکشی اطلاعات و مدل­های تحلیل رفتار وجود دارد و تحقیق در این زمینه ادامه خواهد داشت. در ادامه به بیان مسیر­های تحقیقاتی ممکن پرداخته­ می­شود.

* **رسانه**

با گسترش رشد شبکه­های اجتماعی و سهولت دسترسی افراد به چنین منابع اطلاعاتی زمینه­ی بازیابی اطلاعات موجود در چندین رسانه­ی اطلاعاتی فراهم شده­ است. در طی سال­های گذشته روش­های ارائه شده اغلب از یک منبع اطلاعاتی به همراه داده­های مالی برای بهبود دقت پیش­بینی استفاده کرده­اند درحالیکه تمرکز روش­های ارائه شده در طی دو سال گذشته بر استفاده از چندین منبع اطلاعاتی بوده است. برای مثال روش [108] با بیان این نکته که استفاده از جندین منبع اطلاعاتی در بهبود دقت پیش­بینی موثر است، به تجمیع داده­های اخبار و شبکه­های اجتماعی با یکدیگر پرداخته است بطوریکه توانسته یک مدل پیشگو رگرسیونی مستقل از از نوع داده­های جانبی ارائه دهد. در نتیجه توجه به مسئله­ی تنوع داده­ها به عنوان یکی از ویژگی­های عظیم داده­ها و استفاده از منابع داده­ای متنوع به عنوان یک زمینه­ی باز تحقیقاتی مطرح می­باشد.

* **سازماندهی اطلاعات**

شیوه­ی سازماندهی به متن با ارائه­ی روش جاسازی واژه­ها شکل متفاوتی نسبت به سایر روش­های سازماندهی متن به خود گرفته است. این روش با در نظر داشتن چالش حجم زیاد داده­ها در پردازش داده­های حجیم و با پشتیبانی گوگل ارائه شد. در طی سال­های گذشته اغلب روش­ها سعی در بهبود سازماندهی متن مبتنی بر روش تجمیعی از واژه­ها برای بهبود دقت پیش­بینی بازار داشته­اند درحالیکه بررسی­ها نشان می­دهد در طی دو سال گذشته استفاده از روش جاسازی واژه­ها در بازنمایی و سازماندهی متون در اقتصاد رفتاری رو به گسترش می­باشد[[116]](#footnote-116). در اغلب روش­های ارائه شده تحلیل احساس به شیوه­ی مبتنی بر فرهنگ لغات انجام گرفته است درحالیکه در تحلیل احساس روش­هایی مبتنی بر روش جاسازی واژه­ها ارائه شده­اند که در حوزه­ی اقتصاد رفتاری به کار گرفته نشده­اند. روش تحلیل احساس [109]، با جاسازی احساس واژه­ها یک روش بازنمایی از متن ارائه داده است که علاوه بر کدگذاری لغات در بردارها، احساس هم در همین شیوه­ی کدگذاری منعکس شود. با توجه به اهمیت تحلیل احساس در اقتصاد رفتاری یک زمینه باز تحقیقی در استفاده از جاسازی واژه­ها و تحلیل احساس مبتنی بر آن وجود دارد.

* **بازیابی اطلاعات**

علم نفوذ اطلاعات با استفاده از معیارهای واکشی اطلاعات به استخراج دانش می­پردازد. مطالعه­ی رفتار جمعی کاربران شبکه­های اجتماعی و چگونگی نفوذ و انتشار اطلاعات میان آن­ها در بسیاری از زمینه­ها از جمله شناسایی رویدادهای مهم با موضوعات مختلف و همچنین شایعه و اخبار جعلی کاربرد دارد. رویدادهای مختلف با موضوعات سیاسی، فرهنگی، طبیعی با واکنش­های متفاوتی در میان سرمایه­گذاران مواجه خواد بود و همچنین وقوع یک شایعه درست یا غلط ممکن است سبب سقوط ارزش سهام در بازه­ی زمانی بسیار کوتاهی گردد.

شناسایی رویداد به فرایند گروه­بندی اخباری که به نوعی یک موضوع یکسان را گزارش می­دهند اطلاق می­شود[110]. تا کنون در اغلب روش­های ارائه شده در اقتصاد رفتاری گروه­بندی اخبار اغلب توسط یک خبره­ی انسانی انجام شده است و معمولا شناسایی رویدادها به صورت ضمنی با اتکا بر تغییرات شدیدی در شاخص­ احساس و یا تکرار دسته­ای از واژه­ها، انجام گرفته است[10, 67-69, 79] در صورتیکه [72] در 2019 بیان کرد هم محتوا و هم اطلاعات ساختاری پنهان در ارتباط اخبار به پیش­بینی کمک می­کنند. اسناد خبری مختلف ممکن است وجوه گوناگونی از یک رویداد را گزارش دهند و سرمایه­گذاران با انتشار این اسناد واکنش مضاعفی بر این سلسله اخبار نشان دهند بطوریکه بر اساس نتایج [42, 97] سرمایه­گذاران به اخباری که با موضوع یکسان در چند روز متوالی منتشر می­شوند و همچنین اخبار با بار احساسی منفی واکنش مضاعفی نشان می­دهند. لذا پرداختن به اطلاعات پنهان میان اخبار مرتبط به هم و به دنبال آن شناسایی رویداد از جریان داده­های خبری به عنوان یک زمینه­ی باز تحقیقاتی در اقتصاد رفتاری مطرح می­شود.

وقوع یک شایعه­ی بااساس درست و یا یک شایعه­ی مبتنی بر اخبار جعلی ممکن است تاثیر شدیدی در بازارهای مالی داشته باشد. به عنوان مثال با انتشار خبر جعلی ورشکستگی شرکت هواپیمایی ایالات متحده آمریکا در سال 2008 در عرض نیم ساعت سهام این شرکت در بورس Nasdaq به میزان 27.7% سقوط کرد. شناسایی زودهنگام شایعه و اخبار جعلی با توجه به اهمیت فراوان آن در حوزه­ی اقتصاد رفتاری از میزان اهمیت بالایی برخوردار است. سرمایه­گذاران به شایعات منتشر شده در رسانه­ها خواه این شایعه دارای یک مبنای درست و یا مبتنی بر یک خبر بی اساس منتشر شده باشد، واکنش مضاعفی در طی یک بازه­ی زمانی بسیار کوتاه نشان می­دهند و همین نحوه­ی رفتار سرمایه­گذارانی که اخبار را در طی یک رفتار جمعی از طریق شبکه­های اجتماعی میان یکدیگر منتشر می­کنند می­تواند به عنوان یک منبع مطالعاتی بسیار ارزشمند در مورد شناسایی ویژگی­های زبانی[[117]](#footnote-117) متن شایعات و همچنین چگونگی نفوذ[[118]](#footnote-118) یک شایعه در شبکه­های اجتماعی در پراکنش شایعه محسوب شود. در حال حاضر تحقیق­های بسیار کمی در خصوص شناسایی شایعه و اخبار جعلی در حوزه­ی اقتصاد رفتاری انجام گرفته است که نشان از پیچیدگی بالای این مسئله و عدم وجود مجموعه داده­ی مناسب در این حوزه است. به عنوان یک منبع داده­ای از اخبار جعلی می­توان به منتشر شدن 127 سند جعلی مربوط به سهام 27 شرکت آمریکایی در سال 2017 اشاره کرد. بنا به گزارش سازمان بورس و اوراق بهادار ایالات متحده­ی آمریکا، ذینفعان در این شرکت­ها با دادن مبالغی به صاحبان برخی سایت­های خبری از آن­ها درخواست نموده بودند که اخبار جعلی مثبت راجع به سهام آن­ها منتشر کنند. همین امر دلیلی بر ادامه دار بودن تحقیق و مسیر طولانی پیش­رو در این حوزه می­باشد.

* **تحلیل رفتار**

یادگیری عمیق به عنوان یکی از روش­های جدید مورد توجه محققان، در بسیاری از زمینه­های پردازش عظیم داده­ها نظیر تحلیل متن، پردازش تصویر و ویدئو با موفقیت به کار گرفته شده است. این روش از طریق یادگیری نظارتی یا بدون نظارت به یادگیری بازنمایی چندسطحی و ساختار سلسله مراتبی ویژگی­ها برای طبقه­بندی و شناسایی الگو می­پردازد**[**111**].** تاکنونروش­هایی مبتنی بر یادگیری عمیق در مدلسازی توالی در داده­های بورس و اخبار منتشر شده ارائه شده است[112] اما تعداد مقالات منتشر شده طی چند سال اخیر نشان از ادامه­دار بودن پژوهش در این حوزه دارد. یکی از زمینه­های باز مطالعاتی دیگر تاثیر جریان اطلاعات ناشی از اخبار و شبکه­های اجتماعی بر انتقال سرمایه از یک بازار مالی به بازار دیگر می­باشد. در واقع سرمایه­گذاران با انتشار اخبار مربوط به برخی از رویدادها در جهت امنیت بیشتر، سرمایه­ی خود را از یک بازار به بازار دیگری منتقل می­کنند. همچنین مدلسازی تاثیر رویدادهای خبری بر بازارهایی که رفتار یکسانی[[119]](#footnote-119) دارند نیز به عنوان یک زمینه­ی مطالعاتی ممکن مطرح است. برای بررسی چنین مواردی که ممکن است ناشی از اخبار باشد، به­کارگیری روش­های مبتنی بر نظریه­ی اطلاعات جهت تخمین میزان تاثیر اطلاعات مستخرج از اخبار بر بازارهای مالی­­، مسیر تحقیق را تسهیل می­کند[113].

# نتیجه­گیری

آن­چه که به داده ارزش می­بخشد، شیوه­ی بازیابی اطلاعات از آن می­باشد. در پژوهشی که انجام شد به بررسی و مرور تحقیق­های انجام شده در بازیابی اطلاعات مستتر در منابع داده­ای متنی مبتنی بر وب نظیر شبکه­های اجتماعی، اخبار و گروه­های خبری و تاثیر این اطلاعات بر بازارهای مالی از دیدگاه علم کامپیوتر و با رویکرد متن­کاوی پرداخته شد. در این مطالعه­ی مروری 90 مرجع منتشر شده در طی سال­های 2006 تا ژانویه­ی 2019 بررسی شد. در طیفرایند مرور سعی شد تحلیلی کیفی و کمی از روش­های مبتنی بر اقتصاد رفتاری در چهار زمینه­ی رسانه­های مبتنی بر وب، شیوه­ی سازماندهی به داده­های متنی، شیوه­ی بازیابی اطلاعات و روش تحلیل اطلاعات جهت استخراج دانش ارائه شود. بطوریکه تحلیل کمی بتواند مسیر آینده­ی تحقیق در این حوزه را بیان کند و تحلیل کیفی مرز دانش در حوزه­­ی تاثیر رسانه بر سرمایه­گذاران بازارهای مالی را نشان دهد. هدف مرور انجام شده پوشش جنبه­های مختلف تعامل متن­کاوی و اقتصاد رفتاری و تعیین مسیرهای باز تحقیق در این حوزه بود بطوریکه سعی شد آخرین روش*­*­های متن­کاوی که در پیش­بینی بازارهای مالی، تحلیل رفتار سرمایه­گذاران و همچنین سیستم­های توصیه­گر استراتژی تجارت به کار گرفته شده­اند را پوشش دهد و با در نظر داشتن شیوه­های نو در متن کاوی و علم نفوذ اطلاعات مسیر تحقیق­های آینده را بیان کند. مرور انجام شده می­تواند به عنوان یک منبع مطالعاتی برای محققان علم کامپیوتر از دیدگاه کاربردی دانش آن­ها در سایر زمینه­ها ، دولت و سیاستگذاران اقتصادی در اتخاذ استراتژی های بلندمدت و بهره­بردن از دانش متن­کاوی و تحلیل رفتار و همچنین سرمایه­گذاران بازارهای مالی در جهت کاهش ریسک سرمایه­گذاری با استفاده از سیستم­های پشتیبان تصمیم، مطلوب باشد.

**Reference**

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. Arman Khadjeh Nassirtoussi, T.Y.W., Saeed Reza Aghabozorgi, David Ngo Chek Ling, *Text Mining for Market Prediction: A Systematic Review.* Expert Systems with Applications, 2014. **41**(16): p. 7653-7670.

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

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

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

8. 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.

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

10. 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.

11. Bing Li a , K.C.C.C., Carol Ou , Sun Ruifeng, *Discovering public sentiment in social media for predicting stock movement of publicly listed companies.* Information Systems, 2017. **69C**: p. 81-92.

12. Andrew Sun, M.L., Frank J. Fabozzi, *Trade the tweet: Social media text mining and sparse matrix factorization for stock market prediction.* International Review of Financial Analysis, 2016. **48**: p. 272–281.

13. Saeed Seifollahi, M.S., *Word sense disambiguation application in sentiment analysis of news headlines: an applied approach to FOREX market prediction.* Journal of Intelligent Information Systems, **2018**: p. 1-27.

14. Krishnamoorthy, S., *Sentiment analysis of financial news articles using performance indicators.* Knowledge and Information Systems, 2017. **56**(2): p. 373–394.

15. 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.

16. 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.

17. Ha´jek, P., *Combining bag-of-words and sentiment features of annual reports to predict abnormal stock returns.* Neural Computing and Applications, 2018. **29**: p. 343–358.

18. Hongping Hu , L.T., Shuhua Zhang , Haiyan Wang *Predicting the direction of stock markets using optimized neural networks with Google Trends.* Neurocomputing 2018: p. 1-8.

19. W. Chen, Y.Z., C. K. Yeo, C. T. Lau and B. S. Lee, *Stock market prediction using neural network through news on online social networks*, in *International Smart Cities Conference (ISC2)*. 2017, IEEE: Wuxi. p. 1-6.

20. Shweta Agarwal, S.K., Utkarsh Goel, *Stock market response to information diffusion through internet sources: A literature review.* International Journal of Information Management, 2019. **45**: p. 118–131.

21. B. Shravan Kumar , V.R., *A Survey of the applications of Text mining in Financial Domain.* Knowledge-Based Systems, 2016.

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

23. 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.

24. 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.

25. Robert P. Schumaker , Y.Z., Chun-Neng Huang , Hsinchun Chen *Evaluating sentiment in financial news articles.* Decision Support Systems, 2012: p. 458–464.

26. 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.

27. Qing Li, Y.C., Jun Wang,Yuanzhu Chen and Yuanzhu Chen, *Web Media and Stock Markets : A Survey and Future Directions from a Big Data Perspective* IEEE Transactions on Knowledge and Data Engineering, 2017.

28. 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.

29. 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.

30. *Model and forecast stock market behavior integrating investor sentiment analysis and transaction data.*

31. 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.

32. Aditi Kaushal, P.C., *News and Events Aware Stock Price Forecasting Technique*, in *International Conference on Big Data, IoT and Data Science (BID)*. 2017: Vishwakarma Institute of Technology, Pune,.

33. Hájek Petr, B.A. *Integrating Sentiment Analysis and Topic Detection in Financial News for Stock Movement Prediction*. in *International Conference on Business and Information Management*. 2018. Barcelona, Spain: ACM.

34. Qing Li, Y.C., Li Ling Jiang, Ping Li, and Hsinchun Chen, *A Tensor-Based Information Framework for Predicting the Stock Market.* ACM Trans. Inf. Syst., 2016. **34**(2).

35. Tomoki Ito, H.S., Kiyoshi Izumi, Kota Tsubouchi, Tatsuo Yamashita, *GINN: gradient interpretable neural networks for visualizing financial texts.* International Journal of Data Science and Analytics, 2018.

36. Shiming Deng, P.L., *The impact of attention heterogeneity on stock market in the era of big data.* Cluster Computing, 2018.

37. Mattia Atzeni, A.D., Diego Reforgiato Recupero, *Using frame-based resources for sentiment analysis within the financial domain.* Progress in Artificial Intelligence, 2018.

38. José Gildo de Araújo Júnior, L.B.M. *Using Online Economic News to Predict Trends in Brazilian Stock Market Sectors*. in *In Proceedings of Brazilian Symposium on Multimedia and theWeb*. 2018. Salvador-BA,Brazil: ACM.

39. Baker, M.P., Wurgler, J., *Investor Sentiment and the Cross‐Section of Stock Returns.* Journal of Finance, 2006. **61**(4): p. 1645-1680.

40. Tetlock, P.C., *Giving Content to Investor Sentiment: The Role of Media in the Stock Market.* The Journal of Finance and Data Science, 2007. **LXII**(3).

41. Paul C. Tetlock, M.S.T., SOFUS MACSKASSY *More Than Words: Quantifying Language to Measure Firms' Fundamentals.* The Journal of Finance and Data Science, 2008. **63**: p. 1437-1467.

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

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

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

45. Salton, G., *Automatic text processing*, ed. A.-. Wesley. 1989.

46. Michael Hagenau, M.L., Dirk Neumann, *Automated news reading: Stock price prediction based on financial news using context-capturing features.* Decision Support Systems, 2013. **55**: p. 685–697.

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

48. Adam Atkin, M.N., Enrico Gerding, *Financial news predicts stock market volatility better than close price.* The Journal of Finance and Data Science, 2018. **4**: p. 120e137.

49. JIANG CuiQing, L.K., CHEN Hsinchun, DING Yong, *Analyzing market performance via social media: a case study of a banking industry crisis.* SCIENCE CHINA Information Sciences, 2014. **57**(5): p. 1-18.

50. YUAN RAO, X.Z., Shumin Lu, *Research on News Topic-driven Market Flucatuation and Predication*, in *International Conference on Identification, Information and Knowledge in the Internet of Things*. 2016, IEEE.

51. 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.

52. 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.

53. 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.

54. Liu, B., *Opinions, Sentiment, and Emotion in Text*, ed. C.U. Press. 2015.

55. Xiaodong Li, H.X., Li Chen, Jianping Wang, Xiaotie Deng,, *News impact on stock price return via sentiment analysis.* Knowledge-Based Systems 2014. **69**: p. 14-23.

56. P.S.M. Nizer , J.C.N., *Predicting published news effect in the Brazilian stock market.* Expert Systems with Applications 2012. **39**: p. 10674–10680.

57. 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**.

58. Antonio Moreno-Ortiz, J.F.-C. *Identifying polarity in financial texts for sentiment analysis: a corpus-based approach*. in *Procedia - Social and Behavioral Sciences* 2015. Elsevier.

59. Q. Liu, X.C., S. Su, S. 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.

60. Li, K., *Reaction to news in the Chinese stock market: A study on Xiong’an New Area Strategy.* Journal of Behavioral and Experimental Finance, 2018. ***In Press***.

61. K. Gupta, R.B., *Does OPEC news sentiment influence stock returns of energy firms in the United States?* Energy Economics, 2018.

62. Y. Shia, W.M.L., K. Y. Ho, *Public news arrival and the idiosyncratic volatility puzzle.* Journal of Empirical Finance 2016. **37**: p. 159–172.

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

64. Xue Zhang, H.F., Peter A. Gloor,. *Predicting Stock Market Indicators Through Twitter “I hope it is not as bad as I fear”,*. in *Procedia - Social and Behavioral Sciences,*. 2011.

65. Chi-San Ho, P.D., Bin Gub, Prabhudev Konana, *The time-varying nature of social media sentiments in modeling stock returns.* Decision Support Systems, 2017. **101**: p. 69–81.

66. 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.

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

68. 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.

69. 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.

70. 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.

71. 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.

72. 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.

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

74. M. Schader, O.O., H. H. Bock, *Data Analysis, Machine Learning and Knowledge Discovery*. 2014: Springer.

75. 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.

76. 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**.

77. 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.

78. 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.

79. 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.

80. 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.

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

82. 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.

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

84. 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**.

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

86. 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.

87. Qiang Song, A.L., Steve Y. Yang,, *Stock portfolio selection using learning-to-rank algorithms with news sentiment,.* Neurocomputing, 2014. **246**: p. 20-28.

88. Tomer Geva, J.Z., *Empirical evaluation of an automated intraday stock recommendation system incorporating both market data and textual news.* Decision Support Systems 57 (2014) 212–223, 2014. **57**: p. 212-223.

89. Stefan Feuerriegel, H.P., *News-based trading strategies.* Decision Support Systems, 2016. **90**: p. 65-74.

90. Yunchuan Sun, M.F., Xinyu Wang, *A novel stock recommendation system using Guba sentiment analysis.* Personal and Ubiquitous Computing, 2018(3): p. 575–587.

91. 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.

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

93. 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.

94. 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

95. 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.

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

97. 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.

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

99. 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.

100. 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.

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

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

103. Nuno Oliveira, P.C., Nelson Areal, *The impact of microblogging data for stock market prediction: Using Twitter to predict returns, volatility, trading volume and survey sentiment indices.* Expert Systems With Applications, 2017. **73**: p. 125–144.

104. H.Wang, S.L., Jichang Zhao, *Aggregating multiple types of complex data in stock market prediction: A model-independent framework.* Knowledge-Based Systems, 2019. **15**: p. 193-204.

105. 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. **In Press**.

106. Kiymaz, H., *The effects of stock market rumors on stock prices: evidence from an emerging market.* Journal of Multinational Financial Management, 2001. **11**: p. 105–115.

107. Xichen Zhang, A.A.G., *An overview of online fake news: Characterization, detection, and discussion.* Information Processing & Management, 2019. **In Press**.

108. Wang, H., S. Lu, and J. Zhao, *Aggregating multiple types of complex data in stock market prediction: A model-independent framework.* Knowledge-Based Systems, 2019. **164**: p. 193-204.

109. Rezaeinia, S.M., et al., *Sentiment analysis based on improved pre-trained word embeddings.* Expert Systems with Applications, 2019. **117**: p. 139-147.

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

111. Zhang, Q., et al., *A survey on deep learning for big data.* Information Fusion, 2018. **42**: p. 146-157.

112. Vargas, M.R., et al. *Deep Leaming for Stock Market Prediction Using Technical Indicators and Financial News Articles*. in *2018 International Joint Conference on Neural Networks (IJCNN)*. 2018.

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

1. - دانشجوی دکتری، گروه مهندسی کامپیوتر، واحد مشهد، دانشگاه آزاد اسلامی، مشهد، ایران. anbaee@mashdiau.ac.ir [↑](#footnote-ref-1)
2. - دانشیار، گروه مهندسی کامپیوتر، واحد مشهد، دانشگاه آزاد اسلامی، مشهد، ایران [↑](#footnote-ref-2)
3. twitter [↑](#footnote-ref-3)
4. Stock tweet [↑](#footnote-ref-4)
5. Technical Analysis [↑](#footnote-ref-5)
6. Fundamental Analysis [↑](#footnote-ref-6)
7. Decision Support Systems [↑](#footnote-ref-7)
8. Financial predictive models [↑](#footnote-ref-8)
9. Recommender trading systems [↑](#footnote-ref-9)
10. David M. Cutler [↑](#footnote-ref-10)
11. Stock Return [↑](#footnote-ref-11)
12. Text Mining [↑](#footnote-ref-12)
13. Information retrieval [↑](#footnote-ref-13)
14. Artificial Intelligence(AI) [↑](#footnote-ref-14)
15. Big Data [↑](#footnote-ref-15)
16. Sentiment Analysis [↑](#footnote-ref-16)
17. Part Of Speech Tagging [↑](#footnote-ref-17)
18. Stock Return [↑](#footnote-ref-18)
19. Investor behavior correlation Analysis [↑](#footnote-ref-19)
20. tokenization [↑](#footnote-ref-20)
21. Stop word remove [↑](#footnote-ref-21)
22. Stemming [↑](#footnote-ref-22)
23. Efficient Market Hypothesis [↑](#footnote-ref-23)
24. [Eugene Fama](https://en.wikipedia.org/wiki/Eugene_Fama) [↑](#footnote-ref-24)
25. Microsoft [↑](#footnote-ref-25)
26. Apple [↑](#footnote-ref-26)
27. tensor based [↑](#footnote-ref-27)
28. tensor [↑](#footnote-ref-28)
29. Knowledge base [↑](#footnote-ref-29)
30. Wikipedia [↑](#footnote-ref-30)
31. Apple [↑](#footnote-ref-31)
32. close price [↑](#footnote-ref-32)
33. Indicators [↑](#footnote-ref-33)
34. NewYork Stock Exchange [↑](#footnote-ref-34)
35. LDA [↑](#footnote-ref-35)
36. B. Wiithrich [↑](#footnote-ref-36)
37. ScipGram [↑](#footnote-ref-37)
38. Continues bag of words [↑](#footnote-ref-38)
39. Bag Of Words [↑](#footnote-ref-39)
40. Feature Selection [↑](#footnote-ref-40)
41. 2 word combination [↑](#footnote-ref-41)
42. Ontology [↑](#footnote-ref-42)
43. TFIDF [↑](#footnote-ref-43)
44. Sentiwordnet [↑](#footnote-ref-44)
45. لغاتی که در حوزه­های مختلف معانی متفاوتی دارند. [↑](#footnote-ref-45)
46. Big Data [↑](#footnote-ref-46)
47. Beli [↑](#footnote-ref-47)
48. Stock return [↑](#footnote-ref-48)
49. Word Embedding [↑](#footnote-ref-49)
50. Mikolov [↑](#footnote-ref-50)
51. Word2vec [↑](#footnote-ref-51)
52. Information Retrieval [↑](#footnote-ref-52)
53. Lexicon Based Sentiment Analysis [↑](#footnote-ref-53)
54. Machine Learning Based [↑](#footnote-ref-54)
55. TFIDF [↑](#footnote-ref-55)
56. dictionary based [↑](#footnote-ref-56)
57. ARM [↑](#footnote-ref-57)
58. Sentiment [↑](#footnote-ref-58)
59. Granger Causality [↑](#footnote-ref-59)
60. cross correlation [↑](#footnote-ref-60)
61. sparsity [↑](#footnote-ref-61)
62. polarity [↑](#footnote-ref-62)
63. Wikipedia [↑](#footnote-ref-63)
64. یک روش آماری تحلیل همبستگی که به برسی تاثیر گذشته ی مرتبه ی n از یک سری زمانی بر آینده ی یک سری زمانی دیگر می پردازد [↑](#footnote-ref-64)
65. PMI [↑](#footnote-ref-65)
66. Likelihood [↑](#footnote-ref-66)
67. SVM [↑](#footnote-ref-67)
68. Anomaly detection [↑](#footnote-ref-68)
69. POSTagging [↑](#footnote-ref-69)
70. Multi-dimensional Naïve Bayes [↑](#footnote-ref-70)
71. Stochastic Gradient Decent [↑](#footnote-ref-71)
72. ARIMA [↑](#footnote-ref-72)
73. relevance score (REL) [↑](#footnote-ref-73)
74. event sentiment score (ESS) رتبه ای بین صفر تا 100 از میزان مثبت یا منفی بودن یک خبر. [↑](#footnote-ref-74)
75. event novelty score (ENS) رتبه ای بین صفر تا 100 میزان جدید بودن خبر را در 24 ساعت گذشته تعیین می کند. [↑](#footnote-ref-75)
76. composite sentiment score (CSS) بین صفر تا 100 . کمتر از 50 منفی و بیشتر از آن مثبت تلقی می شود [↑](#footnote-ref-76)
77. Aggregate News Sentiment Index [↑](#footnote-ref-77)
78. LSTM [↑](#footnote-ref-78)
79. Foreign Exchange Market [↑](#footnote-ref-79)
80. FOREX [↑](#footnote-ref-80)
81. GARCH [↑](#footnote-ref-81)
82. Binary classification task [↑](#footnote-ref-82)
83. accuracy [↑](#footnote-ref-83)
84. Mean Absolute Error [↑](#footnote-ref-84)
85. Root Mean Squared Error [↑](#footnote-ref-85)
86. Random Forest [↑](#footnote-ref-86)
87. FOREX [↑](#footnote-ref-87)
88. fundamentalist [↑](#footnote-ref-88)
89. [↑](#footnote-ref-89)
90. Volatility [↑](#footnote-ref-90)
91. منظور از اخبار دست دوم اخباری است که به صورت دو روز متوالی به صورت عینا تکرار شده باشند. [↑](#footnote-ref-91)
92. Big Data Characteristics [↑](#footnote-ref-92)
93. Big Data Quality [↑](#footnote-ref-93)
94. Big Data Analytics [↑](#footnote-ref-94)
95. Volume [↑](#footnote-ref-95)
96. Velocity [↑](#footnote-ref-96)
97. Value [↑](#footnote-ref-97)
98. Variety [↑](#footnote-ref-98)
99. Veracity [↑](#footnote-ref-99)
100. Valence [↑](#footnote-ref-100)
101. Hadoop Eco Systems [↑](#footnote-ref-101)
102. Batch processing [↑](#footnote-ref-102)
103. Frame Based Ontology [↑](#footnote-ref-103)
104. aggrigation [↑](#footnote-ref-104)
105. Scalar Value [↑](#footnote-ref-105)
106. Compositional Data [↑](#footnote-ref-106)
107. Functional Data [↑](#footnote-ref-107)
108. tensor [↑](#footnote-ref-108)
109. Valence [↑](#footnote-ref-109)
110. Betweenness centrality [↑](#footnote-ref-110)
111. [U.S. Securities and Exchange Commission](https://www.sec.gov/news/press-release/2017-79)  [↑](#footnote-ref-111)
112. Istanbul Stock Exchange [↑](#footnote-ref-112)
113. Agent based Automated trading [↑](#footnote-ref-113)
114. Soft real time [↑](#footnote-ref-114)
115. با توجه به شکل 4- نمودار پراکندگی روش­ها بر اساس سال انتشار را نشان می­دهد تعداد مقالات منتشر شده در دو سال 2018 و 2017 نسبت به سال­های گذشته رو به رشد بوده است. [↑](#footnote-ref-115)
116. به شکل 8- روش سازماندهی متن بر اساس سال انتشار منبع مورد استفاده مراجعه شود. [↑](#footnote-ref-116)
117. Linguistics [↑](#footnote-ref-117)
118. Diffusion pattern [↑](#footnote-ref-118)
119. Co-movements [↑](#footnote-ref-119)