ساخت واژگان به‌صورت خودکار برای تحلیل نظرات در حوزهٔ بورس

مرتضی آهنگری آهنگرکلائی[۱]، علی سبطی[*۲] مهدی یعقوبی[۳]

[۱و۲و۳] گروه مهندسی کامپیوتر، دانشکده مهندسی گرگان، دانشگاه گلستان، گرگان، ایران

چکیده

با رشد چشم‌گیر رسانه‌های اجتماعی، افراد و سازمان‌ها به‌طور فزاینده‌ای از افکار عمومی در این رسانه‌ها برای تصمیم‌گیری خود استفاده می‌کنند. هدف تحلیل احساسات، استخراج خودکار احساسات افراد از این شبکه‌های اجتماعی است. شبکه‌های اجتماعی مرتبط به بازارهای مالی، از جمله بازارهای سهام، به‌تازگی در مرکز توجّه بسیاری از افراد و سازمان‌ها بوده‌اند. افراد در این شبکه‌ها نظرات و عقاید خود را در مورد هر سهم در قالب یک پست یا توییت، به اشتراک می‌گذارند. درواقع تحلیل احساسات در این حوزه، سنجش نگرش افراد به هر سهم است. یکی از رویکردهای پایه‌ای و اصلی در تحلیل خودکار احساسات روش‌های مبتنی بر واژگان است. اغلب واژگان‌های مرسوم به‌صورت دستی استخراج شده‌اند که فرایندی بسیار دشوار و هزینه‌بر است. در این مقاله روشی جدید جهت استخراج واژگان به‌صورت خودکار در حوزهٔ شبکه‌های اجتماعی بورسی ارائه شده‌است. یک ویژگی خاص این شبکه‌ها، وجود اطلاعات قیمتی هر سهم در هر روز است. با در نظر گرفتن وضعیت قیمتی سهم در روز درج نظر برای آن سهم، واژگانی برای بهبود کیفیت عقیده‌کاوی در این شبکه‌ها استخراج شد. برای ارزیابی واژگان‌های تولیدشده با استفاده از روش پیشنهادی نیز، با نسخه فارسی واژگان SentiStrength که با هدف استفادهٔ عمومی طراحی شده‌است، مقایسه شد. نتایج آزمایش‌ها بیست درصد بهبود را در معیار صحت نسبت‌به استفاده از واژگان عمومی نشان می‌دهد.

واژگان کلیدی: تحلیل احساسات، عقیده کاوی، ساخت واژگان، واژگان فارسی.

Automatically generate sentiment lexicon for the Persian stock market

Morteza Ahangari Ahangarkolaei[1], Ali Sebti[2*], Mehdi Yaghoubi[3]

[1,2,3] Department of Computer Engineering, Faculty of Engineering, Golestan University, Gorgan, Iran

Abstract

With the significant growth of social media, individuals and organizations are increasingly using public opinion in these media to make their own decisions. The purpose of Sentiment Analysis is to automatically extract people's emotions from these social networks. Social networks related to financial markets, including stock markets, have recently attracted the attention of many individuals and organizations. People on these social networks share their opinions and ideas about each share in the form of a post or tweet. In fact, sentiment analysis in this area is measuring people's attitudes toward each share. One of the basic approaches in automatic analysis of emotions is lexicon-based methods. Most conventional lexicon is manually extracted, which is a very difficult and costly process. In this article, a new method for extracting a lexicon automatically in the field of stock social networks is proposed. A special feature of these networks is the availability of price information per share. Taking into account the price information of the share on the day of tweeting for that share, we extracted lexicon to improve the quality of opinion mining in these social networks. To evaluate the lexicon produced using the proposed method, we compared it with the Persian version of the SentiStrength

* نویسندهٔ عهده‌دار مکاتبات

* Corresponding author

• تاریخ ارسال مقاله: ۱۴۰۰/۳/۳۰ • تاریخ پذیرش: ۱۴۰۱/۶/۷ • تاریخ انتشار: ۱۴۰۲/۷/۳۰ • نوع مطالعه: پژوهشی

[DOI: 10.61186/jsdp.20.2.3]

lexicon, which is designed for general purpose. Experimental results show a 20% improvement in accuracy compared to the use of general lexicon.
Keywords: Sentiment Analysis, Opinion Mining, Lexicon Creation, Persian Lexicon.

۱- مقدمه

اینکه «دیگران به چه چیزی فکر می‌کنند»، همیشه در فرایند تصمیم‌گیری برای بیشتر افراد، جزء اطلاعات مهم است. مدت‌ها پیش که شبکهٔ جهانی وب[1] (www) گسترش نیافته‌بود، اغلب برای پیداکردن یک مکانیک خوب جهت تعمیر ماشین خود، از دوستان یا آشنایان کمک می‌گرفتیم، یا با مشورت از مصرف‌کنندگان دیگر تصمیم می‌گرفتیم کدام ماشین ظرفشویی را خریداری کنیم؛ اما اینترنت و شبکه اکنون (در میان امکانات دیگر) این امکان را فراهم آورده‌اند تا از نظرات و تجربیات طیف وسیعی از مردم استفاده کنیم که نه آشنای شخصی ما هستند و نه منتقدان حرفه‌ای شناخته‌شده؛ یعنی افرادی که ما هرگز چیزی از آنها نشنیده‌ایم [1].

با رشد چشم‌گیر رسانه‌های اجتماعی (به‌عنوان مثال، بررسی‌ها[2]، بحث‌های انجمن‌ها، وبلاگ‌ها و شبکه‌های اجتماعی) در وب، افراد و سازمان‌ها به‌طور فزاینده‌ای از افکار عمومی در این رسانه‌ها برای تصمیم‌گیری خود استفاده می‌کنند. با این حال، یافتن و نظارت بر نظرات در سایت‌های نظردهی در وب و بررسی اطلاعات موجود در آن‌ها به دلیل گسترش سایت‌های متنوع، همچنان یک کار دشوار است. هر سایت به‌طور معمول حاوی حجم عظیمی از نظرات است که همیشه در پست‌های طولانی در وبلاگ‌ها به‌راحتی قابل‌فهم نیست. یک خوانندهٔ معمولی در شناسایی تارنماهای موردنظر و جمع‌بندی دقیق اطلاعات و نظرات موجود در آن‌ها مشکل خواهدداشت[2].

امروزه مشتریان و صاحبان کسب‌وکار از این نظرات برای شناسایی نقاط قوت و ضعف محصولات استفاده می‌کنند، ولی باتوجه‌به افزایش حجم نظرات، مطالعهٔ موردی و نتیجه‌گیری نهایی از آن‌ها دیگر امکان‌پذیر نبوده و نیازمند سامانه‌ای جهت کسب دانش به‌صورت خودکار تحت عنوان تحلیل احساسات[3]، که به‌عنوان عقیده‌کاوی[4] نیز شناخته می‌شود، شد. تحلیل احساسات یا عقیده‌کاوی، در سیاست (انتخابات، پیش‌بینی تحولات سیاسی، میزان اتحاد مردم یا جامعه در یک مورد و ...)، علوم اجتماعی و روان‌شناسی (تحلیل مسائل اجتماعی و فرهنگی، تحلیل تأثیر اتفاقات مختلف در رفتار مردم و ...)، مدیریت و رهبری (کمک در تصمیم‌سازی و تصمیم‌گیری، آگاهی از میزان رضایت یا مطلوبیت مشتریان، مشترکان یا گروهی از مخاطبان، و ...) کاربردهای فراوانی دارد.

تاکنون پژوهش‌های زیادی در حوزهٔ عقیده‌کاوی و تجزیه و تحلیل احساسات در زبان‌های انگلیسی، چینی و روسی انجام شده‌است. باوجوداین‌که زبان فارسی، زبان اصلی ایران، افغانستان و تاجیکستان است و بیش از ۱۱۰ میلیون نفر در سرتاسر دنیا به این زبان سخن می‌گویند، ولی در متن‌های فارسی پژوهش‌های بسیار کمی در تجزیه و تحلیل احساسات انجام شده‌است و همچنان مشکلات و چالش‌های بسیاری در حوزهٔ تحلیل احساسات و عواطف در زبان فارسی وجود دارد.

امروزه از تجزیه و تحلیل احساسات می‌توان در کاربردهای مختلفی استفاده کرد. بخشی از آن شامل تشخیص احساسات نسبت‌به موضوعات خاص در حوزهٔ بررسی محصول، مدیریت ارتباط با مشتری، بازارهای مالی و چهره‌های سیاسی است [3].

بازارهای مالی، از جمله بازارهای سهام، بازارهایی هستند که به‌تازگی در کشور ایران، افراد بسیاری به آن توجه کرده‌اند. به همین دلیل، شبکه‌های اجتماعی که مرتبط با این بازارها هستند نیز، بسیار در مرکز توجّه هستند. در این شبکه‌های اجتماعی، کاربران نظرات خود را بر اساس وضعیت سهم‌ها و یا کلیت بازار، در قالب یک پست یا توییت منتشر می‌کنند. نظرات کاربران می‌تواند انعکاسی از عملکرد، اخبار یا ارزش قیمتی شرکت‌ها و بر اساس تحلیل‌هایی که از یک سهم ارائه می‌شود، باشد. یا حتی بدون هیچ تحلیل و پشتوانه، به‌طور کامل، احساسی و غیرفنی باشد. به همین دلیل می‌توان گفت که این نظرات با احساسات مثبت یا منفی همراه است.

تحلیل احساسات در این حوزه، سنجش نگرش افراد به هر سهم است، که می‌تواند برای مدیران سطح بالا در سازمان‌های مختلف، شرکت‌هایی که رضایت یا عدم رضایت سهام‌داران برایشان مهم است، یا برای پیش‌بینی روند آتی یک سهم استفاده شود. این نگرش‌ها به‌طور لزوم ممکن است درست نباشد، ولی تجمیع این نظرات می‌تواند یک موج در جهت صعود یا نزول یک سهم ایجاد کند.

[1] World Wide Web
[2] Reviews
[3] Sentiment Analysis
[4] Opinion Mining

[DOI: 10.61186/jsdp.20.2.3]

یک ویژگی منحصربه‌فرد در این شبکه‌ها این است که اطلاعات قیمتی هر سهم در روز درج نظر برای آن سهم موجود است. ازآنجاکه سهم‌ها، همه‌روزه در حال رشد یا ریزش هستند، عقاید یا نظراتی که سهامداران یا تحلیل‌گران، روزانه منتشر می‌کنند، از این رشد و ریزش‌های سهم تأثیر می‌پذیرند. ما در این پژوهش ثابت کردیم که احساس نهفته در نظرات یا توییت‌های درج‌شده در روزهایی که سهم در حالت ریزش یا رشد است، با اطلاعات قیمتی آن سهم هم‌بستگی دارند. در سامانهٔ پیشنهادی از این موضوع جهت استخراج واژگان به‌صورت خودکار برای بهبود کیفیت عقیده‌کاوی در این شبکه‌ها استفاده شد.

به‌طورکلی دو روش اصلی برای تحلیل احساسات وجود دارد. روش مبتنی بر واژگان[1] و روش مبتنی بر یادگیری ماشین[4]. روش‌های یادگیری ماشین، برای پیش‌بینی قطبیت احساسات بر اساس مجموعه داده‌های آموزشی و تست عمل می‌کنند. و می‌توانند از روش‌های یادگیری با ناظر (از داده‌های برچسب‌گذاری‌شده برای دسته‌بندی متن استفاده می‌کنند) و بدون ناظر (از داده‌های خام برای دسته‌بندی متن استفاده می‌کنند) استفاده کنند. درحالی‌که رویکرد مبتنی‌بر واژگان نیازی به آموزش اولیه برای استخراج داده‌ها ندارد. از یک فهرست کلمات که ازپیش‌تعریف‌شده است، استفاده می‌کند. به‌طوری‌که هر کلمه با یک احساس خاص همراه است. ما برای کار خود از روش مبتنی بر واژگان استفاده کردیم.

دو نوع واژگان احساسی برای تحلیل احساسات وجود دارد؛ واژگان احساسی عمومی که هر کلمه در آن با یک قطبیت همراه است و واژگان احساسی برای حوزه خاص[2]، که قطبیت هر کلمه باتوجه‌به آن حوزه مشخص می‌شود. برای مثال کلمهٔ «مشخص» در حوزه فیلم قطبیت منفی دارد، اما در حوزهٔ سیاست قطبیت مثبت خواهدداشت.

برای ساخت واژگان احساسی، سه روش کلی وجود دارد. روش دستی، که هر کلمه توسط یک خبره باتوجه‌به قطبیت آن کلمه برچسب‌گذاری می‌شود، و کاری زمان‌بر و پرهزینه است. ساخت واژگان احساسی بر اساس فرهنگ واژگان، مانند WordNet [5] که مترادف و متضاد کلمات در آن قرار دارد. و ساخت واژگان احساسی با استفاده از پیکره[3]. اغلب برای ایجاد واژگان احساسی برای حوزه‌های خاص از این روش استفاده می‌شود. ساخت واژگان احساسی در این روش را می‌توان یا با استفاده از وقوع کلمات با یکدیگر ایجاد کرد؛ به‌عنوان مثال، اگر یک کلمه همراه کلمه‌ای با قطبیت مثبت بیاید، می‌توان گفت آن کلمه نیز قطبیت مثبت دارد؛ یا می‌توان با استفاده از پیکره در یک حوزهٔ خاص، قطبیت کلمات یک فرهنگ واژگان را که از قبل وجود دارد، با آن پیکره تطبیق داد.

ما در کار خود از روش مبتنی بر پیکره برای ساخت واژگان پیشنهادی خود استفاده می‌کنیم. پیکرهٔ مورداستفاده در این روش در حوزهٔ بورس است که ما آن‌ها را از پایگاه سهام‌یاب[4] استخراج کردیم و شامل ۱.۱۰۰.۰۰۰ نظر از سهامداران بورس است. که لغات و قطبیت آن‌ها به‌صورت خودکار استخراج می‌شود. آزمایش‌های ما نشان می‌دهد که استفاده از این روش برای ساخت واژگان، در مقایسه با واژگانی که به‌صورت عمومی برای تعیین احساسات در سطح جملات استفاده می‌شود، دقت را در حدود ۲۰ درصد افزایش می‌دهد.

در ادامهٔ این مقاله به موضوعات زیر پرداخته می‌شود. بخش ۲، به برخی از مهم‌ترین پژوهش‌های مربوط به توسعهٔ واژگان احساسی برای زبان فارسی پرداخته می‌شود. در بخش ۳، فرایند توسعهٔ واژگان احساسی که ما تولید کردیم، به‌طور مفصل شرح داده می‌شود. ارزیابی و تست واژگان تولیدشده در بخش ۴ انجام گرفت. و در بخش ۵، نتیجه‌گیری و پیشنهادهایمان مطرح شد.

۲- پیشینهٔ پژوهش

مهم‌ترین و اصلی‌ترین مرحله در فرایند جلب رضایت مشتری، شناسایی توقعات، انتظارات و گاهی الزامات طرح‌شده از طرف مصرف‌کننده است. باتوجه‌به‌این‌که تمام مشتریان به‌صورت حضوری در دسترس نیستند، یا افرادی که نظرات مهمی در مورد محصول دارند، آن را با تولیدکنندگان در میان نمی‌گذارند، می‌توان با استفاده از شبکه‌های اجتماعی و با درنظرگرفتن این نکته که افراد در این شبکه‌ها نظرات خود را با دیگران به اشتراک می‌گذارند، اطلاعات موردنیاز خود را بدون نیاز به حضور مشتری به‌دست‌آورد.

کاربردهای تحلیل احساسات در سه سطح سند[5]، جمله و وجه[6] بررسی می‌شود[6]. تمرکز کار ما بر روی

[1] Lexicon based
[2] Domain Specific
[3] Corpus
[4] https://www.sahamyab.com/
[5] Document Level
[6] Aspect Level

[DOI: 10.61186/jsdp.20.2.3]

پیش‌بینی احساسات در سطح سند است؛ که ما در آن به هر سند یک برچسب مثبت یا منفی اختصاص می‌دهیم.

به‌طور کلی اطلاعات موجود در اسناد متنی را می‌توان به دو دسته تقسیم کرد: ۱) عینی[1] ۲) ذهنی[2]. منظور از عینی، واقعیت‌ها، دستورهای واقعی و قابل‌مشاهده دربارهٔ موجودیت‌های مستقل و اتفاقاتی است که در جهان می‌افتد. اما اطلاعات ذهنی، بازتاب عواطف انسانی یا مشاهداتی است که مردم نسبت‌به دنیای خارج و اتفاقات آن دارند[7]. نظرات کاربران جزء اسناد متنیِ ذهنی هستند که به‌سرعت در دنیای مجازی در حال تولید هستند؛ بنابراین، حجم زیادی دارند و امکان بررسی نظرات به‌صورت دستی نیست. از این‌رو، نیازمند به‌کارگیری و انتخاب روش‌های مناسب جهت بررسی نظرات به‌صورت خودکار هستیم. به‌صورت کلی روش‌های موجود برای تحلیل احساسات به دو دسته تقسیم می‌شوند؛ روش‌های بر اساس یادگیری ماشین و روش‌های بر اساس واژگان. تاکنون بیشتر کارهایی که در حوزهٔ تحلیل احساسات صورت گرفته، از روش‌های یادگیری ماشین بهره برده‌اند[8, 9]. بیشتر این کارها از الگوریتم ماشین بردار پشتیبان[3] و نایو بیز[4] برای کار خود استفاده کرده‌اند. آنها می‌توانند از روش‌های یادگیری با ناظر[5] یا بدون ناظر استفاده کنند. روش‌های با ناظر، برای طبقه‌بندی متن از داده‌های دارای برچسب استفاده می‌کنند، درحالی‌که روش‌های بدون ناظر فقط از داده‌های خام استفاده می‌کنند[10]. روش‌های مبتنی بر واژگان برای تشخیص احساسات یک متن، از یک واژگان استفاده و با محاسبات آماری بر روی کلمات مثبت و منفی، قطبیت متن داده‌شده را محاسبه می‌کنند. از مهم‌ترین مزایای این روش، سرعت بالا و عدم نیاز به داده‌های آموزش است. و عیب اصلی این روش نیز عدم مقیاس‌پذیری آن است[11]. در ادامهٔ این بخش به بررسی برخی از روش‌های ارائه‌شده مبتنی بر یادگیری ماشین و مبتنی بر واژگان خواهیم پرداخت.

۱-۲- روش‌های مبتنی بر یادگیری ماشین

ترنی[6] [12]، یک الگوریتم یادگیری ساده بدون نظارت برای دسته‌بندی بررسی‌ها به‌عنوان توصیه‌شده (انگشت شست بالا[7]) یا توصیه نشده (انگشت شست پایین[8]) ارائه داده است. که در آن دسته‌بندی یک بررسی با میانگین گرایش معنایی عبارات موجود در بررسی، که حاوی صفت یا قید است، پیش‌بینی می‌شود. الگوریتم پیشنهادی او هنگام ارزیابی در ۴۱۰ بررسی از پایگاه Epinions، نمونه‌برداری از چهار دامنهٔ مختلف (بررسی اتومبیل، بانک، فیلم، و مقصد سفر) به میانگین صحت ۷۴٪ می‌رسد.

رانی[9] و *کومار*[10] [13]، تحلیل احساسات را بر روی داده‌های بررسی فیلم، که به زبان هندی است، با استفاده از تنظیمات مختلف پیکره‌بندی شبکه عصبی پیچیده[11] انجام دادند. آن‌ها نتایج به‌دست‌آمده را توسط الگوی شبکهٔ عصبی پیچیده خود با پیشرفته‌ترین نتایج الگوریتم‌های سنتی یادگیری ماشین مقایسه کردند. نتایج کار آن‌ها نشان داد که الگوی پیشنهادی آن‌ها قادر به دستیابی به عملکرد بهتری نسبت‌به رویکردهای سنتی یادگیری ماشین است و به دقت ۹۵٪ رسید.

برای غلبه بر کاستی‌ها در روش‌های تحلیل احساسات فعلی، یک روش تحلیل احساسات بر اساس شبکه‌های عصبی بازگشتی[12] به نام حافظهٔ طولانی کوتاه‌مدت دوجهته[13] ارائه شده‌است[14]. آن‌ها داده‌های خود را که شامل ۱۵۰۰۰ متن است، بعد از تعبیهٔ کلمات[14] به‌وسیلهٔ الگوی Word2vec و محاسبهٔ وزن کلمات به‌وسیلهٔ الگوریتم TF_IDF، به برداری وزن‌دار تبدیل و سپس به شبکهٔ حافظهٔ طولانی کوتاه‌مدت دوجهته اعمال کردند. آزمایش‌های آن‌ها نشان می‌دهد که الگوی پیشنهادیشان دقت[15] و F1-Score بالاتری نسبت‌به شبکه عصبی پیچیده، شبکهٔ عصبی بازگشتی، حافظهٔ طولانی کوتاه‌مدت و الگوریتم نایو بیز دارد. دقت و F1-Score به‌دست‌آمده در کار آن‌ها به‌ترتیب در حدود ۹۱ و ۹۲ درصد است. در کاری مشابه [15]، برای تعبیهٔ کلمات، به‌جای استفاده از Word2vec از الگوی Doc2vec استفاده کردند؛ همچنین برای استخراج هرچه بهتر ویژگی‌ها از شبکهٔ عصبی پیچیده نیز در کنار شبکه حافظهٔ طولانی کوتاه‌مدت استفاده کردند. آن‌ها با آزمایش‌ها نشان دادند که روش پیشنهادی آن‌ها در مقایسه با شبکه عصبی پیچیده، حافظهٔ طولانی کوتاه‌مدت، حافظهٔ طولانی

[1] Objective
[2] Subjective
[3] SVM
[4] Naïve Bayes
[5] Supervised
[6] Turney
[7] Thumbs Up
[8] Thumbs Down
[9] Rani
[10] Kumar
[11] Convolutional Neural Network (CNN)
[12] Recurrent Neural Network (RNN)
[13] Bidirectional Long Short-Term Memory (BiLSTM)
[14] Word Embedding
[15] Precision

[DOI: 10.61186/jsdp.20.2.3]

کوتاه‌مدت دوجهته و CNN-LSTM صحت[1] بالاتری دارد که صحت به‌دست‌آمده برای روش پیشنهادی آن‌ها، بر روی مجموعه‌دادگان خود که از مقالات فرانسوی به‌دست‌آمده از روزنامه‌های ملی و بین‌المللی است، در حدود ۹۱ درصد است.

تاکنون کارهای زیادی با استفاده از الگوریتم ماشین بردار پشتیبان و نایو بیز برای تحلیل احساسات انجام شده‌است[16-20]. اولین کاری که برای مسئلهٔ دسته‌بندی در تحلیل احساسات به زبان فارسی صورت گرفت[21]، از دو روش استاندارد ماشین بردار پشتیبان و نایو بیز، در حوزهٔ بررسی فیلم استفاده شده‌است. همچنین شش ویژگی حضور[2] و تکرار[3] Unigramها، Bigramها و Trigramها برای نمایش اسناد، با هم مقایسه شدند. باتوجه‌به ارزیابی‌هایی که انجام دادند، متوجه شدند که در کار آن‌ها الگوریتم ماشین بردار پشتیبان نسبت‌به نایو بیز عملکرد بهتری داشته‌است و همچنین، استفاده از ویژگی Unigram در مقایسه با Bigram و Trigram کارایی دسته‌بندی کننده را بهبود می‌بخشد و همچنین، در نظر گرفتن فقط حضور یک ویژگی نتیجهٔ بهتری از تکرار آن دارد. در کاری مشابه [22]، با بررسی چهار معیار اطلاعاتی مختلف، شامل فراوانی اسناد[4]، واریانس تکرار اصطلاح[5]، اطلاعات متقابل[6] و اطلاعات متقابل اصلاح‌شده[7] که توسط آن‌ها پیشنهاد شد، متوجه شدند که روش پیشنهادی عملکرد به‌نسبت‌بهتری از رویکردهای فراوانی اسناد، واریانس تکرار اصطلاح و اطلاعات متقابل دارد. اطلاعات متقابل اصلاح‌شده به‌طور کلی می‌تواند به ۸۵٪ از معیار F-Score برسد.

نسخهٔ دیگری از ماشین بردار پشتیبان که با الگوریتم بهینه‌سازی ازدحام ذرات برای داده‌های بررسی فیلم در توییتر استفاده شده[23]، نشان می‌دهد که استفاده از الگوریتم بهینه‌سازی ازدحام ذرات برای تعیین شاخص‌های ماشین بردار پشتیبان، و همچنین استفاده از ویژگی‌های n-grams و به‌ویژه Unigrams می‌تواند صحت را تا ۴ درصد بهبود ببخشد. این بهبود با پاک‌سازی داده‌ها، می‌تواند تا حدود ۲ درصد بیشتر نیز افزایش یابد.

نخستین مجموعه‌دادگان به نام Pars-ABSA که به‌طور کامل مبتنی‌بر وجه در زبان فارسی است، در [24] ارائه شد. برای آزمایش مجموعه‌دادگان، از شش الگویی که به‌تازگی برای تحلیل احساسات بر اساس وجه، روی حوزه‌های مختلف به زبان انگلیسی، کار شده‌است و تمرکزشان بر روی روش‌های یادگیری عمیق است، استفاده شد. در میان نتایج تمامی الگوها روی کار پیشنهادی آن‌ها، الگوی TD-LSTM نتایج شگفت‌انگیزی داشته‌است، زیرا این الگو نسبت‌به الگوهای دیگر در مجموعه‌دادگان‌های انگلیسی نتیجهٔ ضعیف‌تری داشت، اما در کار آن‌ها نتیجهٔ بهتری نسبت‌به سایر الگوها دارد.

در کاری دیگر[25]، از دو الگوی یادگیری عمیق (رمزگذارهای خودکار[8] و شبکه‌های عصبی پیچیده) در مجموعه داده‌های بررسی فیلم فارسی استفاده شد. نتایج به‌دست‌آمده از این دو الگو با پرسپترون چند لایه[9] مقایسه شد، نتایج نشان می‌دهد که رمزکننده‌های خودکار از پرسپترون چندلایه، دقت بالاتری دارند و الگوی شبکهٔ عصبی پیچیده پیشنهادی نیز از رمزکننده‌های خودکار با دقت ۸۲/۶ درصد، عملکرد بهتری دارد.

۲-۲- روش‌های مبتنی بر واژگان

بیشتر افرادی که در حوزهٔ تحلیل احساسات کار می‌کنند، تمرکز خود را بر روی روش‌های مبتنی بر یادگیری ماشین قرار داده‌اند و تعداد کمی از آن‌ها توجه خود را به روش‌های مبتنی بر واژگان معطوف کردند.

تا به حال فهرست واژگان زیاد و خوبی برای کارهای انگلیسی‌زبان تولید شده‌است [26-29]. اما به تولید واژگان برای تحلیل احساسات به زبان فارسی زیاد توجه نشده‌است.

یکی از واژگان‌های قطبیت شناخته‌شده برای انگلیسی، SentiWordNet است [30]. در SentiWordNet، سه امتیاز به هریک از مجموعه‌های هم‌معنی WordNet اختصاص داده شده‌است که نشان می‌دهد این مجموعه‌ها چقدر مثبت، منفی و خنثی هستند. این منبع شامل بیش از ۱۱۷۰۰۰ مجموعه هم‌معنی است. پیشنهاد اصلی ساخت SentiWordNet طبقه‌بندی مجموعه‌های هم‌معنی WordNet با استفاده از واژه‌نامه‌های مجموعه‌های هم‌معنی است. نکتهٔ مهمی که باید در مورد امتیازات اختصاص‌یافته به هر مجموعهٔ هم‌معنی درنظرگرفته‌شود، این است که این امتیازات قدرت قطبیت را نشان نمی‌دهد. آن‌ها فقط نشان می‌دهند که یک مجموعهٔ هم‌معنی چقدر مثبت، منفی

[1] Accuracy
[2] Presence
[3] Frequency
[4] Document Frequency (DF)
[5] Term Frequency Variance (TFV)
[6] Mutual Information (MI)
[7] Modified Mutual Information (MMI)
[8] Autoencoders
[9] Multilayer Perceptron (MLP)

[DOI: 10.61186/jsdp.20.2.3]

یا خنثی است. در نتیجه، جمع این سه نمره برای هر مجموعه هم‌معنی برابر با یک است. یکی دیگر از واژگان قطبیت عمومی شناخته‌شده برای زبان انگلیسی، واژگان NRC است [31]. این واژگان، به هر کلمه برچسب‌های احساسی، عصبانیت، ترس، انتظار، انزجار، ... و همچنین برچسب‌های مثبت و منفی اختصاص می‌دهد که حدود ۱۵۰۰۰ کلمه است و به‌صورت دستی از طریق Amazon's Mechanical Turk ایجاد شده‌است.

نخستین رویکرد تحلیل احساسات به زبان فارسی [32] که مبتنی بر واژگان است، با ارائهٔ یک چارچوب برای تحلیل احساسات انجام شد. آن‌ها دو منبع برای تحلیل احساسات در زبان فارسی معرفی کردند: ۱- یک واژگان فارسی که مرتبط است با کلمات احساسی فارسی به همراه قطبیت آن. ۲- یک مجموعه‌دادگان که به‌صورت دستی جمع‌آوری و توسط یک شخص خبره برچسب‌گذاری شده‌است. آن‌ها از نظریهٔ *دمپستر شافر*[1] برای تعیین قطبیت هر سند استفاده کردند. نتایج کار آن‌ها نشان می‌دهد که روش پیشنهادی آن‌ها نسبت‌به روش‌های مبتنی بر یادگیری ماشین، رتبهٔ F-Score بالاتری در حدود ۹۰ درصد کسب می‌کند. آن‌ها در کاری دیگر [33]، برای کمک به رفع مشکل کمبود منابع برای تحلیل احساسات در زبان فارسی، دو منبع جدید به نام‌های SPerSent و CNRC را برای تحلیل احساسات در زبان فارسی را ارائه می‌دهند. SPerSent، یک مجموعه‌دادگان در سطح جمله است که هر جمله با دو برچسب همراه است، یک برچسب دودویی برای تعیین قطبیت و یک برچسب رتبه‌بندی پنج ستاره. CNRC، یک واژگان فارسی است که برای ایجاد آن از لغت‌نامهٔ NRC [34] همراه با سه مرحله پردازش استفاده شده‌است. آن‌ها برای ارزیابی واژگان CNRC، آن را با نسخهٔ فارسی واژگان‌های NRC و Senti_Str [35] بر روی مجموعه‌دادگان SPerSent توسط الگوریتم یادگیری ماشین نایو بیز مقایسه کردند، نتایج نشان داد که واژگان CNRC کارایی و دقت بالاتری را نسبت‌به دو واژگان دیگر دارد. آن‌ها در کاری مشابه با این روش [36]، برای اثبات این که ترجمهٔ مستقیم واژگان انگلیسی به فارسی در کار تحلیل احساسات کیفیت مناسب را ندارد، چهار واژگان را با هم مقایسه کردند که شامل نسخهٔ فارسی واژگان‌های NRC، SentiStrength، CNRC و Adjectives است. نتایج نشان می‌دهد که ترجمهٔ مستقیم استفاده‌شده در NRC، ضعیف‌ترین عملکرد را به همراه دارد؛ درحالی‌که پیش‌پردازش و پالایش واژگان مورد استفاده در SentiStrength و CNRC باعث بهبود عملکرد می‌شود. همچنین، نتایج نشان می‌دهد که استفاده از فقط صفت در مقایسه با استفاده از NRC، منجر به نتایج بهتری می‌شود.

ثابتی و همکاران [37]، یک روش جدید مبتنی‌بر گراف، برای انتخاب و گسترش مجموعهٔ اولیه[2] برای تولید واژگان قطبیت عمومی، به نام LexiPers پیشنهاد داده‌اند، که این واژگان قطبیت شامل بیش از ۶۰۰۰ کلمه است. آن‌ها از رویکردهای مبتنی‌بر واژگان و همچنین، پیکره بهره‌مند شدند.

امیری و همکارانش روش دیگری را برای تحلیل احساسات مبتنی بر واژگان، ارائه دادند [38]. آنها واژگانی فارسی را که متشکل از صفات، واژه‌ها و اصطلاحات نظریافته در دو دستهٔ رسمی و غیررسمی است، جمع‌آوری کردند. همچنین واژگان جمع‌شده با فارسی استاندارد یا فارسی منسوخ‌شده که توسط عدهٔ خاصی از بومی‌زبانان استفاده می‌شود، مطابقت دارد. آن‌ها همچنین، یک رابط شبکه ایجاد کردند که افراد بومی را قادر می‌سازد به‌صورت دستی یک امتیاز به کلمات واژگان اختصاص دهند. به‌دنبال ایجاد یک واژگان احساسات فارسی حاشیه‌نویسی‌شده، آن‌ها یک پایپ‌لاین[3] زبانی بر اساس چارچوب GATE طراحی و توسعه دادند. که اجزای آن شامل یک توکن‌ساز[4] فارسی، تقسیم‌کنندهٔ جمله[5]، برچسب گذار POS[6] و فرهنگ لغت[7] بود. در نتیجه، آن‌ها صحت حدود ۶۵٪ را گزارش کردند که در مقایسه با رویکردهای مبتنی‌بر واژگان مشابه، به‌عنوان یک پیشرفت در نظر گرفته‌شد.

آقای *دهخرقانی* [39]، یک روش مبتنی‌بر ترجمهٔ جدید برای ایجاد واژگان قطبیت در زبان‌هایی که منابع واژگانی کمی برای تحلیل احساسات وجود دارد، پیشنهاد می‌دهد و آن را برای زبان فارسی به‌کارمی‌برد. برای ساخت این واژگان قطبی که SentiFars نام دارد، از چندین منبع واژگان قطبیت انگلیسی به نام‌های NRC، SentiWordNet، SenticNet و واژگان قطبیت Liu's کمک گرفت. روش پیشنهادی آن‌ها در چهار مرحله انجام می‌شود. نخست، کلمات به فارسی ترجمه می‌شوند. سپس کلمات ترجمه‌شده به‌صورت دستی برچسب مثبت، منفی

[1] Dempster-Shafer

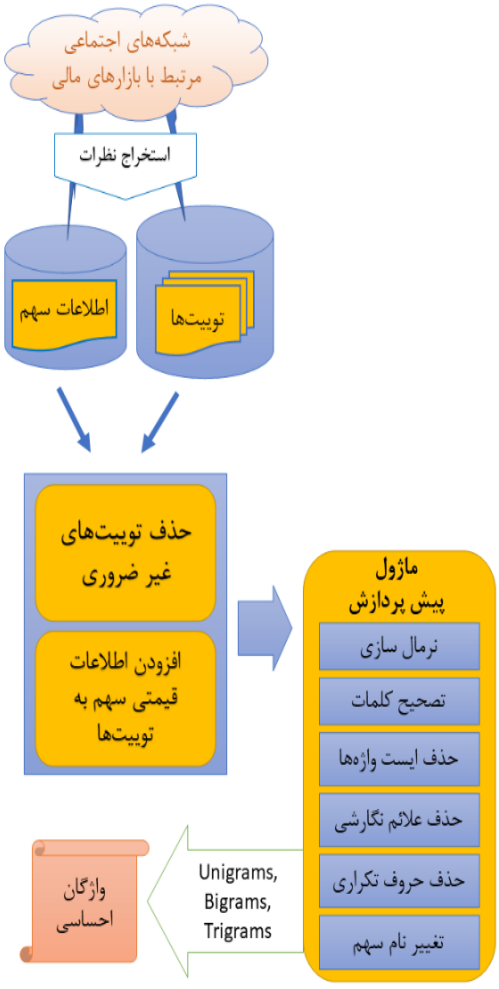
[2] Seed
[3] Pipeline
[4] Tokenizer
[5] Sentence Splitter
[6] Part of Speech Tags
[7] Gazetteer

[DOI: 10.61186/jsdp.20.2.3]

و خنثی دریافت می‌کنند. مرحلهٔ بعد استخراج ویژگی توسط واژگان قطبیت انگلیسی است. سپس کار دسته‌بندی توسط دسته‌بندی‌کنندهٔ منطقی[1] صورت می‌گیرد که این کار توسط یادگیری نگاشت بین ورودی‌های توصیف‌شده توسط ویژگی‌های استخراج‌شده، و سه برچسب ردهٔ (مثبت، منفی و خنثی) انجام می‌شود. در نهایت بعد از آزمایش‌های انجام شده توانست با در نظر گرفتن هر چهار واژگان انگلیسی یادشده برای استخراج ویژگی، به صحت ۹۵/۹۲ درصد برسد.

۳- سامانهٔ پیشنهادی

همان‌طورکه پیشتر گفته‌شد، به‌دلیل اهمیت بازارهای مالی، ما به‌صورت خودکار یک واژگان برای تحلیل احساسات در حوزه بورس ایجاد کردیم. این کار در چند مرحله صورت گرفت که در شکل (۱) آمده‌است.



(شکل-۱): روش پیشنهادی برای ایجاد واژگان
(Figure-1): Proposed method for lexicon creation

[1] Logistic Classifier

داده‌ها یا توییت‌هایی که برای این کار در نظرگرفته‌شده‌است، و همچنین، برای ارزیابی و آزمون استفاده می‌شود، از پایگاه سهام‌یاب استخراج و در پایگاه داده خود ذخیره کردیم. همچنین، ارزش قیمتی هر سهم را در زمان درج آن، از سایت رسمی بورس ایران[2] استخراج و در یک جدول جداگانه از پایگاه داده ذخیره کردیم. جدول (۱) اطلاعات سهم‌های ذخیره‌شده را در پایگاه داده نمایش می‌دهد.

باتوجه‌به سامانهٔ پیشنهادی در شکل (۱)، بعد از جمع‌آوری داده‌ها، پیش‌پردازش‌هایی بر روی آن‌ها صورت می‌گیرد. در ادامه، به بررسی و توضیح هریک از این پیش‌پردازش‌ها می‌پردازیم.

نام سهم	تعداد	سال
دی	85461	1392-1399
فملی	43054	1392-1399
فولاد	35978	1392-1399
حفاری	8917	1392-1399
حکشتی	30907	1392-1399
خودرو	162371	1392-1399
خساپا	157308	1392-1399
ستران	31383	1392-1399
شبندر	147451	1392-1399
شپنا	88118	1391-1399
شتران	22605	1392-1399
تاپیکو	51276	1392-1399
وتجارت	50341	1392-1399
وبملت	77265	1392-1399
وبصادر	107565	1392-1399

(جدول-۱): اطلاعات سهم‌های انتخاب‌شده
(Table-1): Information of selected shares

۳-۱- حذف توییت‌های غیرضروری:

متأسفانه اطلاعات قیمتی روزانه سهم‌ها، برای همهٔ روزهای درج نظر، در اختیار نیست. ازآنجاکه ما برای ساخت خودکار واژگان موردنظر به این اطلاعات قیمتی نیاز داشتیم، ناچار به حذف این دسته از توییت‌ها شدیم. تعداد این توییت‌ها بعد از این مرحله به ۹۰۰۰۰۰ توییت کاهش یافت. چند نمونه از این اطلاعات قیمتی به همراه زمان و نام سهم، در جدول (۲) آورده شده‌است.

[2] http://www.tsetmc.com/

[DOI: 10.61186/jsdp.20.2.3]

(جدول-۲): اطلاعات قیمتی سهم
(Table-2): Share price information

نام سهم	تاریخ	بیشترین قیمت	قیمت پایانی	قیمت دیروز
دی	1399/12/25	19941	19486	18813
خودرو	1399/02/27	1174	1174	587
حکشتی	1397/11/29	4116	4072	4121
حفاری	1393/07/30	5687	5605	5607
فولاد	1388/03/04	1900	1865	1884

۲-۳- افزودن اطلاعات قیمتی سهم به توییت‌ها:

بعد از حذف توییت‌های غیرضروری، اطلاعات قیمتی برای هر توییت به آن اضافه شد. این اطلاعات که قسمتی از آن در جدول (۲) آورده‌شده، جهت تعیین وزن و قطبیت کلمات واژگانِ نهایی استفاده شده‌است. جدول (۳)، بخشی از توییت‌ها را به همراه اطلاعات قیمتی آن نمایش می‌دهد.

(جدول-۳): توییت‌ها به همراه اطلاعات قیمتی
(Table-3): Tweets with price inforamtion

توییت	اطلاعات قیمتی
#خودرو #حکشتی امروز هر دو صف می‌شوند. خوشبین	1392/08/26, 'تاریخ': ,'خودرو' :'نماد سهم' 'کمترین قیمت': 3150, 'بیشترین قیمت': آخرین ' ,3090 :'قیمت پایانی' ,3070 قیمت ' ,3111 :'اولین قیمت' ,3075 :'معامله 25003468 :'حجم' ,3075 :'دیروز
#شبندر سال ۹۴ تا سال ۹۶ شبندر بازم بهتر می‌شود.	1393/01/26, 'تاریخ': ,'شبندر' :'نماد سهم' 'کمتری قیمت': 11800, 'بیشترین قیمت': آخرین ' ,11185 :'قیمت پایانی' ,11104 ,11104 :'اولین قیمت' ,11190 :'معامله 17345720 :'حجم' ,11566 :'قیمت دیروز'
#فولاد سلام خدمت دوستان فولادی باتوجه‌به رشد خوب فولاد که نوش همه سهامدارن آن باشد، متاسفانه واگرایی منفی را نشان می‌دهد. احتیاط بیشتر پیشنهاد می‌شود. الهی همه پرسود باشند.	1396/09/27, 'تاریخ': ,'فولاد' :نماد سهم' 'کمترین قیمت': 2900, 'بیشترین قیمت': آخرین ' ,2871 :'قیمت پایانی' ,2854 قیمت ' ,2860 :'اولین قیمت' ,2860 :'معامله ,2850 :'دیروز' 32281473 :'حجم'

۳-۳- پیش‌پردازش:

یکی از مراحل مهم در تحلیل احساسات، پیش‌پردازش داده‌ها است. انتخاب روش‌های پیش‌پردازش مناسب، می‌تواند باعث بهبود طبقه‌بندی صحیح داده‌ها شود[40]. به همین دلیل ما پیش‌پردازش داده‌ها را در شش مرحله انجام می‌دهیم.

۱-۳-۳ -عادی‌سازی:

برای عادی‌سازی داده‌ها ما از ابزار هضم[1] که یک کتابخانه برای زبان پایتون[2] است، استفاده کردیم. این ابزار از ماژول‌های متفاوتی تشکیل شده‌است. که ما از ماژول عادی‌سازی آن برای کار خود استفاده کردیم. بعضی کلمات در فارسی از چند بخش تشکیل شده‌اند. این کلمات به‌طور معمول، با نیم‌فاصله از هم جدا می‌شوند. اما اغلب، افراد در توییت‌هایی که در شبکه‌های اجتماعی قرار می‌دهند، این نکته را رعایت نمی‌کنند. برای مثال، کلمه «آن‌ها» به‌اشتباه «آن ها» نیز نوشته می‌شود. یکی از کارهایی که این ماژول انجام می‌دهد، تبدیل این نوع فاصله‌ها به نیم‌فاصله است. همچنین برخی از حروف در فارسی دارای یونیکد متفاوت هستند. به‌طور خاص، کلمه «ی» و «ک» در بعضی مواقع به‌ترتیب به شکل «ي» و «ك» نیز نوشته می‌شود. این حروف نیز توسط این ماژول به شکل استانداردِ خود در می‌آیند. در حقیقت، در این قسمت ما کلمات غیراستاندارد را به شکل استاندارد در آوردیم.

۲-۳-۳- تصحیح کلمات:

در بیشتر متون غیررسمی صفحات وب فارسی، کاربران کلمات را همان‌گونه‌که در مکالمات روزانه استفاده می‌کنند، می‌نویسند. به همین دلیل این متون حاوی تعداد زیادی از کلمات با املای غیراستاندارد است. بنابراین، بررسی املای کلمات در زبان فارسی چالش برانگیزتر از زبان انگلیسی است [32].

برای حل این مسئله، فهرستی از کلمات به‌همراه تعداد رخداد آن‌ها از توییت‌های ذخیره‌شده در پایگاه داده استخراج شد. سپس اندازهٔ این فهرست در دو مرحله کاهش یافت. در مرحله نخست، از پیکرهٔ شناخته‌شدهٔ همشهری [41]، به‌منظور تطبیق هر کلمه از فهرست به‌دست‌آمده با آن استفاده کردیم. با این فرض که اگر آن کلمه در پیکرهٔ همشهری وجود داشته‌باشد، املای آن استاندارد است. در مرحلهٔ دوم با فرض اینکه کلماتی که تعداد رخداد آن در کل داده‌ها کمتر از ده بار باشد، تأثیر ناچیزی روی تخمین قطبیت احساسی جملات

[1] https://www.sobhe.ir/hazm/
[2] Python

[DOI: 10.61186/jsdp.20.2.3]

خواهندگذاشت، از فهرست حذف شدند؛ و در نهایت، تعداد ۲۷۸۲ کلمه از فهرست باقی ماند. در جدول (۴) تعدادی از کلمات این فهرست به همراه معادل صحیح آن آورده شده‌است. در پایان، معادل درست هریک از این کلمات به‌صورت دستی و بر روی کل دیتابیس اعمال شد.

(جدول-۴): بخشی از فهرست کلمات فارسی غیراستاندارد به همراه معادل استاندارد آن

(Table-4): Part of the list of informal Persian words with its formal equivalent

معادل صحیح آن کلمات	کلمات با املای ناصحیح
منفی است	منفیه
قیمت‌ها	قیمتا
می‌ریزد	میریزه
می‌گردد	میگرده
نشوند	نشن

۳-۳-۳-حذف ایست‌واژه‌ها:

برای کاهش اندازه داده‌ها و بهبود دقت تحلیل احساسات، ایست‌واژه‌هایی را که تأثیر قابل‌توجه‌ای در تشخیص احساسات ندارند و پیوسته در مجموعه‌دادگان تکرار شده‌اند، حذف کردیم. برای این کار برخلاف روش‌های دیگر که از فهرست ایست‌واژه‌های شناخته‌شده برای تشخیص استفاده می‌کنند، ما با ایجاد یک Unigrams از کلمات درون مجموعه‌دادگان، به‌همراه تکرار هرکدام از آن‌ها برای شناسایی ایست‌واژه‌ها استفاده کردیم. با این فرض که تکرار کلمات ایست‌واژه در مجموعه‌دادگان، اختلاف چشم‌گیری با دیگر کلمات خواهندداشت. این روش باعث می‌شود کلمات تأثیرگذار در مجموعه‌دادگان که ممکن است توسط فهرست‌های ایست‌واژهٔ موجود، به‌عنوان یک ایست‌واژه شناخته شوند، در مجموعه‌دادگان باقی بمانند.

۴-۳-۳-حذف علائم نگارشی[1]:

تمامی علائم نگارشی بجز «#» و «.» از مجموعه‌دادگان حذف شد.

۵-۳-۳-حذف حروف تکراری:

در بعضی مواقع در میان توییت‌ها، افراد برای تأکید بر یک موضوع بعضی کلمات را با تکرار بر یک حرف آن می‌نویسند. مثلا برای تأکید بر کلمه «نه»، از «نهههههه» استفاده می‌کنند. در این ماژول، تکرار این حروف حذف شد.

[1] Punctuations

۶-۳-۳-تغییر نام سهم:

ازآنجاکه نام سهم تأثیری بر روند تشخیص احساسات ندارد و باعث افزونگی در واژگان ما می‌شود، در این قسمت نام سهم‌ها را به یک اسم خاص (#سهم) تغییر دادیم. که هم از افزونگی در مجموعه‌دادگان جلوگیری کند و هم روند تطبیق کلمات بهبود یابد.

۴-۳- ساخت واژگان:

همان‌طورکه در بخش ۱ نیز ذکر شد، سه روش کلی برای ساخت واژگان وجود دارد؛ که کار ما ساخت واژگان بر اساس پیکره است. رویکرد سامانهٔ پیشنهادی برای ساخت واژگان، استفاده از ویژگی‌های n-grams (به‌طور ویژه Unigram، Bigram و Trigram) در داده‌های توییت است. همچنین، اطلاعات قیمتی متناسب با هر توییت برای محاسبهٔ نرخ رشد هر سهم در زمان درج توییت نیز محاسبه می‌شود. دلیل محاسبهٔ نرخ رشد، این است که ما فرض می‌کنیم زمانی‌که نرخ رشد یک سهم مثبت است، افرادی که توییت‌های خود را در این شبکه‌های اجتماعی برای آن سهم ارسال می‌کنند، به دلیل حس رضایت، کلمات مثبت بیشتری را در نظرات خود به‌کارمی‌برند. همچنین، اگر نرخ رشد سهم منفی باشد، در توییت‌هایی که این افراد منتشر می‌کنند، به‌دلیل حس ناراحتی، کلمات منفی بیشتری به‌کاربرده‌می‌شود. با در نظر گرفتن این فرض، برای هر توییت نرخ رشد متناسب با زمان ارسال آن و این‌که این نظر برای چه سهمی ارسال شده، محاسبه شد. برای محاسبهٔ نرخ رشد، ما از رابطه (۱)- (نسخه اول)، و رابطه (۲)- (نسخه دوم) استفاده کردیم:

Growth Rate = نرخ رشد

P_{Close} = قیمت پایانی سهم

$P_{PrevClose}$ = قیمت پایانی سهم در روز قبل

P_{Min} = کمترین قیمت سهم در طول روز

P_{Max} = بیشترین قیمت سهم در طول روز

$$\textbf{Growth Rate} = \begin{cases} \frac{P_{Close}-min(P_{Min},P_{PrevClose})}{min(P_{Min},P_{PrevClose})}, P_{Close}-P_{PrevClose}>0 \\ \frac{P_{Close}-\max(P_{Max},P_{PrevClose})}{\max(P_{Max},P_{PrevClose})}, P_{Close}-P_{PrevClose}<0 \\ 0 \qquad\qquad\qquad\quad , P_{Close}-P_{PrevClose}=0 \end{cases} \quad (۱)$$

$$\textbf{Growth Rate} = \begin{cases} \frac{P_{Close}-P_{Min}}{P_{Min}}, P_{Close}-P_{PrevClose}>0 \\ \frac{P_{Close}-P_{Max}}{P_{Max}}, P_{Close}-P_{PrevClose}<0 \\ 0 \qquad\quad , P_{Close}-P_{PrevClose}=0 \end{cases} \quad (۲)$$

[DOI: 10.61186/jsdp.20.2.3]

در رابطهٔ پیشنهادی هنگامی‌که نرخ رشد برابر با صفر شود، نشان‌دهندهٔ آن است که این کلمه فاقد قطبیت بالاست و تأثیر آن در روند تحلیل احساسات در کار ما بسیار پایین است. در داده‌های اطلاعاتی ما، «قیمت پایانی» برای هر سهم در هر روز موجود است. قیمت پایانی عبارت است از میانگین قیمت‌های معامله‌شده در طول ساعات معاملاتی آن روز برای یک سهم. ما برای بازتاب بهتر تغییرات قیمتی در طول روز، از «بیشترین قیمت» و «کمترین قیمت» به‌منظور محاسبهٔ نرخ رشد استفاده کردیم. دلیل این کار این است که هیجانات هم‌راستای قیمت پایانی سهم را در نظر بگیریم. برای مثال، اگر در آغاز شروع ساعات معاملاتی، یک سهم با ارزش بالایی شروع شود و در انتهای ساعات معاملاتی با کاهش چشم‌گیری مواجه شود، نظرات یا توییت‌هایی که کاربران برای آن سهم منتشر می‌کنند، از نظر احساسات بسیار مهم و قابل‌توجه است. همچنین، برعکس این موضوع نیز صادق است. یعنی اگر یک سهم در آغاز ساعات معاملاتی بازار با ارزش کم شروع به معامله شود و این ارزش در ادامهٔ ساعات معاملاتی افزایش یابد، نظراتی که کاربران منتشر می‌کنند، اغلب با احساسات مثبت همراه است. ازاین‌رو، استفاده از کمترین قیمت و بیشترین قیمت، باعث تمایز بهتری برای محاسبهٔ نرخ رشد می‌شود. منظور از تمایز بهتر این است که اگر شرایط ذکرشده اتفاق بیفتد، نرخ رشد برای آن توییت‌ها مثبت‌تر، یا منفی‌تر خواهدشد.

در شرایطی که سهمی از آغاز، در صف خرید باشد، کمترین قیمت سهم (P_{Min}) و قیمت پایانی سهم (P_{Close}) در طول روز یکی می‌شود، همچنین برعکس این موضوع نیز صادق است. به‌این‌صورت‌که اگر سهمی از آغاز در صف فروش باشد، بیشترین قیمت سهم (P_{Max}) و قیمت پایانی سهم (P_{Close}) نیز در طول روز یکی می‌شود، ازاین‌رو، رابطهٔ (۱)- (نسخه اول) که نسخهٔ پایه برای روش پیشنهادی است، در این حالت به‌جای استفاده از کمترین قیمت و بیشترین قیمت از قیمت پایانی سهم در روز قبل ($P_{PrevClose}$) برای محاسبه نرخ رشد استفاده می‌کند. این در حالی است که ضابطهٔ اول و دوم رابطه (۲)- (نسخه دوم) در این حالت برابر با صفر می‌شود. رابطهٔ پیشنهادی (۲)، قطبیت بیشتر را به کلمات یا عباراتی اطلاق می‌کند که در روزهای پرنوسان دیده‌شوند و همچنین، در روزهایی که صف فروش یا خرید قفل است، روش پیشنهادی مقداری نزدیک به صفر می‌دهد. در واقعیت نیز روزهایی که قیمت‌ها دچار نوسان هستند (چه در جهت مثبت و چه در جهت منفی)، بیشتر شاهد توییت کاربران خواهیم‌بود. یا برعکس، در روزهایی که نوسانات قیمتی کم است، کاربران فعالیت کمتری دارند و نظرات آن‌ها دارای قطبیت قابل‌توجه کمتری است. به عبارت دیگر، فعالیت کاربران با احساسات قوی را بیشتر در روزهای نوسانی بالا شاهد هستیم. رابطهٔ پیشنهادی (۲)- (نسخه دوم) بیانگر این واقعیت است. همچنین، باید به این نکته توجه داشت که محور محاسبهٔ نرخ رشد در روابط پیشنهادی، بر اساس میانگین سهم است و نوساناتی که بسیار لحظه‌ای هستند و یا حجم معاملات در آن پایین است، در قیمت پایانی سهم تأثیر مهمی نمی‌گذارند، یا به عبارت دیگر، این نوسانات معیار احساسات و قضاوت اهالی بازار نیستند. ما در کار خود هر دوی این روش‌ها را استفاده کرده و نتایج به‌دست‌آمده این موضوع را نیز تأیید می‌کند. نتایج در بخش ۳-۴ (ارزیابی نهایی) قابل‌مشاهده است. ازاین‌رو، در ادامه بیشتر به روش پیشنهادی (۲)- (نسخهٔ دوم) پرداخته خواهدشد.

به منظور محاسبهٔ نرخ رشد، نخست، بررسی می‌کنیم که قیمت پایانی سهم نسبت‌به روز گذشته افزایش داشته یا کاهش. اگر قیمت پایانی افزایش داشته‌باشد، از ضابطهٔ اول رابطهٔ (۲)، و اگر کاهش داشته باشد، از ضابطه دوم رابطه (۲)، نرخ رشد را محاسبه می‌کنیم.

همان‌طورکه ذکر شد، استخراج کلمات واژگان، توسط ویژگی n-gramها انجام شد و ما نرخ رشد و رخداد این کلمات را در کنار آن وارد کردیم. محاسبهٔ نرخ رشد هریک از این کلمات واژگان به این صورت است که نخست، زمانی‌که این کلمات توسط ویژگی n-gram از توییت‌ها انتخاب می‌شوند نرخ رشد هر یک از توییت‌ها که آن کلمه در آن قرار دارد، در یک فهرست، برای آن کلمه در نظر گرفته می‌شود. پس از پیمایش تمامی توییت‌ها میانگین این نرخ رشدها برای هر کلمه به‌عنوان نرخ رشد هر کلمه در نظر گرفته می‌شود. برای مثال اگر کلمه «فروش» در ۱۰۰۰ توییت وجود داشته‌باشد، میانگین نرخ رشد این توییت‌ها به‌عنوان نرخ رشد این کلمه در نظر گرفته‌می‌شود.

برای درک عمیق‌تر موضوع، رابطهٔ پیشنهادی در قالب یک مثال ساده از چند توییت به‌همراه نحوهٔ محاسبهٔ نرخ رشد و عملکرد روش پیشنهادی برای حالات مختلف آورده شده‌است. برای این کار در جدول (۵) اطلاعات

[DOI: 10.61186/jsdp.20.2.3]

قیمتی دو سهم به‌همراه توییتی که برای هرکدام گذاشته شده‌بود، آورده شده‌است. در ادامه روند محاسبهٔ نرخ رشد در رابطهٔ پیشنهادی با استفاده از این اطلاعات قیمتی شرح داده می‌شود.

نمونهٔ اول:

نخست، نیاز است که نرخ رشد سهم نسبت‌به روز پیش محاسبه شود، تا از این طریق بفهمیم از کدام ضابطه باید استفاده کرد.

$P_{Close} = ۱۰۹۸۱, P_{PrevClose} = ۱۰۲۲۲$

$P_{Min} = ۱۰۱۶۵, P_{Max} = ۱۱۲۱۰$

$P_{Close} - P_{PrevClose} = ۱۰۹۸۱ - ۱۰۲۲۲ > ۰$

$\Longrightarrow$ Growth Rate $= \frac{P_{Close}-P_{Min}}{P_{Min}} = \frac{۱۰۹۸۱ - ۱۰۱۶۵}{۱۰۱۶۵} \approx ۸\%$

نمونه دوم:

$P_{Close} = ۲۴۲۵۶, P_{PrevClose} = ۲۵۸۰۴$

$P_{Min} = ۲۳۲۸۹, P_{Max} = ۲۶۵۷۸$

$P_{Close} - P_{PrevClose} = ۲۴۲۵۶ - ۲۵۸۰۴ < ۰$

$\Longrightarrow$ Growth Rate $= \frac{P_{Close}-P_{Max}}{P_{Max}} = \frac{۲۴۲۵۶ - ۲۶۵۷۸}{۲۶۵۷۸} \approx -۸.۷$ %

مقادیر به‌دست‌آمده برای نرخ رشد برای هر دو روش پیشنهادی (نسخهٔ اول و دوم) یکسان است. به این ترتیب نرخ رشد برای تمامی توییت‌ها محاسبه می‌شود، و سپس میانگین نرخ رشد هر n-gram بر اساس رخداد آن محاسبه می‌شود.

(جدول-۵): نمونه‌ای از توییت‌ها به همراه اطلاعات قیمتی

(Table-5): Sample tweets with price inforamtion

توییت	اطلاعات قیمتی
#سهم این سهم پایین نداره کلی خبر خوب در شرکت نهفته است تغییر کاربری زمین و افزایش سرمایه و فروش زیر مجموعه و صادراتی بودن و افزایش نرخ سیمان و خرید حتی در صف خرید هم لذت خودش رو داره در این سهم چرا که فرداش هم صف خرید می‌شود و فردای فرداش هم همینطور	نماد سهم: ستران، بیشترین قیمت: 11210 (+9.66٪)، کمترین قیمت: 10165 (-0.55٪)، پایانی: 10981 (+7.42٪)، آخرین معامله: 11070 (+8.29٪)، اولین قیمت: 10699 (+4.66٪)، قیمت پایانی دیروز: 10222، حجم: 13432085
#سهم بایداز حقوقی سهم به وزارت اطلاعات و قوه محترم قضایی شکایت بشه	نماد سهم: حکشتی، حداکثر قیمت: 26578 (+3%)، حداقل قیمت: 23289 (-9.74%)، قیمت پایانی: 24256 (-6%)، آخرین معامله: 23431 (-9.2%)، اولین قیمت: 24800 (-3.9%)، قیمت پایانی دیروز: 25804،حجم: 37087720

همان‌طورکه ذکر شد، ما از ویژگی‌های Unigram، Bigram و Trigram به‌عنوان ویژگی‌های n-gram استفاده کردیم. در ادامه، مشخصات هریک از واژگان‌های به‌دست‌آمده ذکر می‌شود.

۱-۱-۱- Unigram:

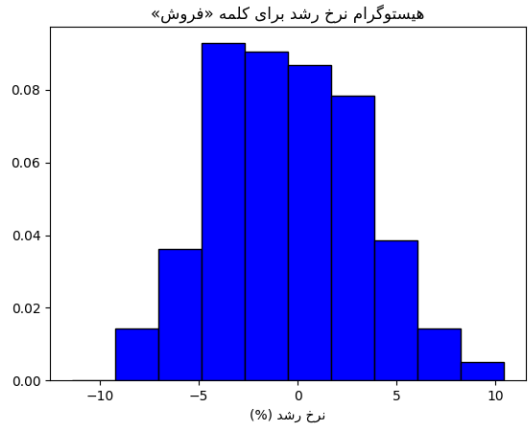
واژگانی که با استفاده از ویژگی Unigram به‌دست‌آمد، به نسبت واژگان‌هایی که با استفاده از ویژگی‌های Bigram و Trigram به‌دست‌آمد، در کار تحلیل احساسات از کیفیت کمتری برخوردار است. جدول (۶)، تعدادی از کلمات موجود در واژگان unigram را نمایش می‌دهد.

(جدول-۶): قسمتی از واژگان به‌دست‌آمده با Unigram

(Table-6): Part of the Lexicon obtained with Unigram feature

لغت	رخداد در واژگان	میانگین نرخ رشد (٪)
فروش	89179	-0.533
خرید	128991	0.051
خوشبین	161192	0.266
حمایت	30432	-0.604

همان‌طورکه پیشتر ذکر شد، واژگان‌های ما به‌صورت خودکار استخراج شده‌است. برای بررسی بیشتر و مشاهده پراکندگی کلمات به‌دست‌آمده با Unigram، نسبت‌به رشد قیمتی سهم در روز درج نظر، در قالب نمودار هیستوگرام چند نمونه را بررسی کردیم که یکی از آن‌ها در شکل (۲) آورده شده‌است.



(شکل-۲): نمودار هیستوگرام برای یک نمونه کلمهٔ به‌دست‌آمده با Unigram

(Figure-2): Histogram for a sample word obtained with unigram

باتوجه‌به شکل (۲)، مشاهده می‌کنیم که توزیع آماری کلماتی که با استفاده از ویژگی Unigram استخراج شده‌اند، اغلب مختلط است. در ادامه، دلیل این مسئله را بررسی می‌کنیم. برای مثال، توییت «امروز روز خوبی برای سرمایه‌گذاری است و برای این سهم از فردا صف خرید تشکیل می‌شود.»، زمانی‌که ارزش آن سهم نسبت‌به روز قبلش در حالت صعودی بود، منتشر شد. همچنین، توییت «این هفته، سهم‌ها وضعیت خوبی ندارند، و اصلاً توصیه به

[DOI: 10.61186/jsdp.20.2.3]

سرمایه گذاری نمی‌کنم.» در زمانی منتشر شد که سهم نسبت‌به روز قبلش در وضعیت نزولی قرار داشت. اگر به این دو نظر، کمی دقت کنیم، متوجه می‌شویم که کلمه «خوبی»، هم در روزهای منفی سهم و هم در روزهای مثبت سهم رؤیت می‌شود. درصورتی‌که اگر کلمات بعد یا قبل را نیز در نظر می‌گرفتیم این دو، متفاوت از یکدیگر می‌شدند. بنابراین، می‌توان نتیجه گرفت که استفاده از واژگان مبتنی بر Unigram، برای تحلیل احساسات در این حوزه از کارآمدی مناسبی برخوردار نباشد.

۱-۳-۴- Bigram:

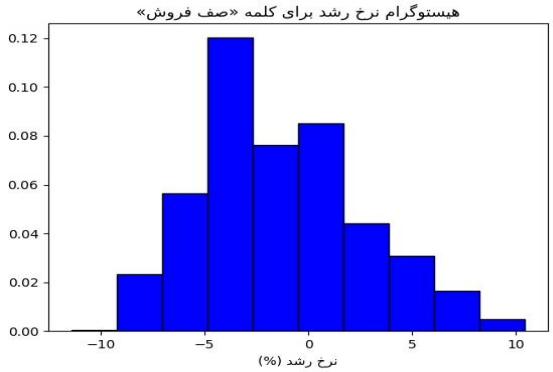
جدول (۷)، بخشی از واژگان به‌دست‌آمده توسط Bigram را نمایش می‌دهد.

(جدول-۷): قسمتی از واژگان به‌دست‌آمده با Bigram

(Table-7): Part of the lexicon obtained with Bigram feature

لغت	رخداد در واژگان	میانگین نرخ رشد (٪)
صف فروش	29904	-1.374
صف خرید	40820	0.588
با ضرر	2047	-0.866
خرید خوشبین	2159	0.508

یکی از مزایای استفاده از ویژگی Bigram برای انتخاب کلمات واژگان این است که احتمال این‌که دو کلمهٔ متوالی که احساس منفی یا مثبتی را انعکاس می‌دهند، هم در شرایط منفی بازار و هم در شرایط مثبت بازار رؤیت شوند، بسیار کمتر از کلمات واحد (که توسط Unigram انتخاب شده‌اند)، است. به عبارت دیگر، می‌توان گفت که توزیع آماری کلماتی که با استفاده از Bigram استخراج می‌شوند، بیشتر به سمت مثبت یا منفی تجمع می‌یابند. ولی توزیع کلماتی که با استفاده از ویژگی Unigram استخراج می‌شوند، بیشتر نزدیک به صفر است. به همین دلیل استفاده از واژگان استخراج شده با ویژگی Bigram، نسبت‌به واژگان استخراج‌شده با Unigram، کیفیت تحلیل احساسات را روی کار ما بهبود بخشید. در شکل (۳) توزیع آماری کلمه «صف فروش» برای واژگان به‌دست‌آمده با Bigram نمایش داده شده‌است. نتایج استفاده از این واژگان در کار تحلیل احساسات بر روی توییت‌های بورسی، در بخش (۴) آورده شده‌است.



(شکل-۳): نمودار هیستوگرام برای یک نمونهٔ به‌دست‌آمده با Bigram

(Figure-3): Histogram for a sample obtained with bigram

۲-۳-۴-Trigram:

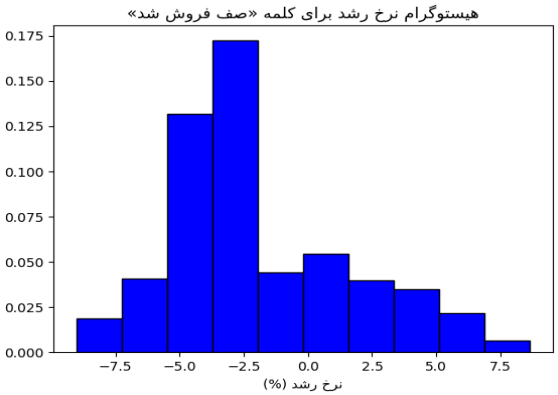
جدول (۸)، بخشی از واژگان به‌دست‌آمده با Trigram را نمایش می‌دهد.

(جدول-۸): قسمتی از واژگان به‌دست‌آمده با Trigram

(Table-8): Part of the lexicon obtained with Trigram feature

لغت	رخداد در واژگان	میانگین نرخ رشد (٪)
صف فروش شد	705	-1.952
صف خرید مبارک	324	2.432
پر سود باشید	2574	-0.385
بازار منفی است	489	-1.064

انتخاب کلمات واژگان با استفاده از ویژگی Trigram، به نسبت واژگان به‌دست‌آمده با Bigram، باعث بهبود در کیفیت تحلیل احساسات بر روی توییت‌های بورسی شد. دلیل این بهبود همان‌طورکه پیشتر ذکر شد، احتمال کمتر دیده‌شدن ترکیبات سه‌تایی کلمات، هم در شرایط مثبت بازار و هم در شرایط منفی بازار است. شکل (۴)، نمودار هیستوگرام کلمه «صف فروش شد» را نمایش می‌دهد.



(شکل-۴): نمودار هیستوگرام برای یک نمونهٔ به‌دست‌آمده با Trigram

(Figure-4): Histogram for a sample obtained with trigram

[DOI: 10.61186/jsdp.20.2.3]

۴- ارزیابی و آزمایش‌ها

به‌طور کلی، موضوعات تحلیل زبان طبیعی برای هر زبان مستقلند و قابل قیاس با زبان‌های دیگر نیستند. همچنین، در بازار بورس ایران اصطلاحات یا کلمات خاصی به کار می‌رود که نمونه مشابه آن را نمی‌توان در بازارهای مالی دیگر یا در زبان‌های دیگر مشاهده کرد. به همین دلیل ما برای ارزیابی یک‌سری آزمایش‌ها و همچنین، یک مقایسه، بر روی واژگان تولیدشده به روش پیشنهادی و واژگان عمومی که در کار آقای بصیری و همکارانش [32] استفاده شده‌است، انجام دادیم. شکل (۵) روند ارزیابی را نمایش می‌دهد. به‌منظور مقایسهٔ تعداد ۱۰۰۰ توییت به‌صورت تصادفی از دیتابیس خود انتخاب و قطبیت هریک از آن‌ها را به‌صورت دستی مشخص کردیم. به‌منظور انتخاب قطبیت ما از برچسب‌گذاری دودویی (مثبت و منفی) استفاده کردیم. جدول (۹)، مشخصات مجموعه‌دادگان ما را نشان می‌دهد.

(شکل-۵): مراحل ارزیابی
(Figure-5): Evaluation steps

۴-۱- واژگان SentiStrength:

یک کتابخانهٔ در دسترس است که برای تشخیص قطبیت و قدرت متون اجتماعی کوتاه و غیررسمی استفاده می‌شود [42]. آقای بصیری و همکارانش [32]، از این کتابخانه برای تحلیل احساسات روی کار خود استفاده کردند. از آنجاکه این نرم‌افزار برای زبان انگلیسی طراحی و ایجاد شده‌است، بصیری و همکارانش نخست، فهرست اصلی واژه‌های آن را به‌صورت دستی به زبان فارسی ترجمه، و بعد از حذف کلمات تکراری در کار خود استفاده کردند.

(جدول-۹): مشخصات مجموعه‌دادگان
(Table-9): Dataset specifications

برچسب	تعداد
مثبت	668
منفی	332

۴-۲- تجمیع[1]:

به‌منظور محاسبهٔ قطبیت هر توییت در مجموعه‌دادگان، ما از روش‌های آماری در کار خود استفاده و برای تعیین قطبیت هر توییت از دو روش معروف تجمیع استفاده کردیم؛ روش جمع امتیازها و روش دمپستر شافر.

۴-۲-۱- جمع امتیازها:

این روش امتیازهای مثبت و منفی هر جمله را با هم جمع و عدد حاصل، قطبیت جمله را مشخص می‌کند. محاسبهٔ قطبیت هر توییت به این صورت است که هر کلمه از توییت یک امتیاز در واژگان دارد. به‌عنوان مثال، در واژگان تولیدشده با Unigram، برای هر کلمه یک میانگین نرخ رشد در نظر گرفته شده که همان امتیاز قطبیت کلمه است. بنابراین برای هر توییت، امتیاز هریک از کلمات آن را از واژگان موردنظر محاسبه و با یکدیگر جمع می‌کنیم و عدد حاصل را به‌عنوان قطبیت نهایی آن توییت در نظر می‌گیریم. از آنجاکه حد آستانه را در این کار صفر در نظر گرفتیم، اگر قطبیت به‌دست‌آمده بیشتر از صفر شود، آن توییت از نظر احساسی مثبت، و اگر کمتر از صفر شود آن توییت از نظر احساسی منفی در نظر گرفته‌می‌شود. همچنین، به‌منظور مقایسه با کار آقای بصیری ما از روش آن‌ها (دمپستر شافر) نیز استفاده و نتایج را با هم مقایسه کردیم.

۴-۲-۲-دمپستر شافر:

این راهبرد مبتنی‌بر نظریهٔ شواهد دمپستر شافر است [43]. این روش برای اولین بار در کار تحلیل احساسات به زبان فارسی توسط بصیری و همکارانش [32]، پیشنهاد شد. آن‌ها در کار خود نمرهٔ هر جمله از یک بررسی را به‌عنوان یک رویداد برای نمرهٔ کلی آن بررسی در نظر گرفتند. همچنین، بعد از تعریف رویداد در کار خود برای محاسبه تابع جرم[2]، از معادله (۳) استفاده کردند.

[1] Aggregation
[2] Mass Function

[DOI: 10.61186/jsdp.20.2.3]

$$\text{تابع جرم (جمله)} = \frac{\text{کمترین نمره} - \text{نمره جمله}}{\text{کمترین نمره} - \text{بیشترین نمره}} \quad (۳)$$

در معادلهٔ ۳ «نمره جمله»، خروجی واژگان SentiStrength برای آن جمله است. «کمترین نمره» و «بیشترین نمره» به‌ترتیب کوچک‌ترین نمرهٔ منفی و بزرگ‌ترین نمرهٔ مثبت در جملات منفی و مثبت است. و در نهایت، برای محاسبهٔ نمره نهایی نمرات رویدادهای به‌دست‌آمده برای هر بررسی را از طریق معادلهٔ (۴) تجمیع می‌کنند.

$$m(A) = \frac{\sum_{X \cap Y=A} m_n(X) m_o(Y)}{1 - \sum_{X \cap Y=\phi} m_n(X) m_o(Y)} \quad (۴)$$

مقدار تجمیع کلی یک عدد حقیقی در بازهٔ [۰، ۱] است و گرد می‌شود تا به‌عنوان یک مقدار دودویی (مثبت/منفی) استفاده شود. ما نیز به‌جهت مقایسه با کار آن‌ها، از این روش برای تجمیع امتیازات استفاده و نتایج را با هم مقایسه کردیم.

۳-۴- ارزیابی نهایی:

برای ارزیابی از چهار معیار ارزیابی معروف استفاده شد. این معیارها شامل صحت، دقت، بازیابی و F-Score است. جدول (۱۰)، نتایج ارزیابی با استفاده از واژگان‌های به‌دست‌آمده با روش پیشنهادی (۱)- (نسخه اول) و همچنین جدول (۱۱)، نتایج ارزیابی با استفاده از واژگان‌های به‌دست‌آمده با روش پیشنهادی (۲)- (نسخه دوم) و واژگان SentiStrength را با کمک تجمیع جمع امتیازها، نمایش می‌دهد.

(جدول-۱۰): نتایج ارزیابی واژگان‌های به‌دست‌آمده با روش پیشنهادی (۱)- (نسخه اول) با استفاده از جمع نمرات

(Table-10): Results of the evaluation of the lexicons based on the proposed method 1 and SentiStrength using sum of scores

واژگان	صحت	دقت	بازیابی	F-Score
واژگان استخراج شده با Unigram	0.66	0.62	0.64	0.63
واژگان استخراج شده با Bigram	0.64	0.67	0.57	0.62
واژگان استخراج شده با Trigram	0.61	0.74	0.36	0.48

(جدول-۱۱): نتایج ارزیابی واژگان‌های به‌دست‌آمده با روش پیشنهادی (۲)- (نسخه دوم) و SentiStrength با استفاده از جمع نمرات

(Table-11): Results of the evaluation of the lexicons based on the proposed method 2 and SentiStrength using sum of scores

واژگان	صحت	دقت	بازیابی	F-Score
واژگان استخراج‌شده با Unigram	0.66	0.9	0.41	0.56
واژگان استخراج‌شده با Bigram	0.66	0.85	0.51	0.64
واژگان استخراج‌شده با Trigram	0.66	0.83	0.55	0.66
واژگان SentiStrength	0.46	0.52	0.08	0.13

همان‌طورکه در جدول (۱۱) مشاهده می‌شود، کار تشخیص قطبیت با استفاده از واژگان SentiStrength، که با هدف استفاده عمومی ایجاد شده‌است، در داده‌های بورسی کمترین دقت را دارد. دلیل این دقتِ پایین این است که کلماتی که در داده‌های حوزهٔ بورس وجود دارد، مختص بورس است و در حوزه‌های دیگر کاربرد زیادی ندارد. ازاین‌رو واژگان SentiStrength در میان واژگان‌های دیگر، کمترین دقت را به‌دست‌آورد. همچنین بیشترین رتبه F-Score نیز مربوط به واژگان به‌دست‌آمده با ویژگی Trigram است. دلیل برتری واژگان به‌دست‌آمده با Trigram به نسبت واژگان به‌دست‌آمده با Unigram و Bigram، همان‌طورکه در بخش ۳-۴-۳ نیز ذکر شد، همبستگی بیشتر کلمات به‌دست‌آمده با Trigram در کل مجموعه‌دادگان نسبت‌به اطلاعات قیمتی است. در واقع کلمات به‌دست‌آمده با این ویژگی، امکان این‌که هم در روزهای مثبت و هم در روزهای منفی بازار رؤیت شوند، به‌مراتب کمتر از کلمات به‌دست‌آمده با Unigram و Bigram دارند.

جدول (۱۲) نتایج ارزیابی با استفاده از واژگان‌های پیشنهادی و واژگان SentiStrength را با کمک راهبرد دمپستر شافر، نمایش می‌دهد.

همان‌طورکه در جدول (۱۲) نیز مشاهده می‌شود، تشخیص قطبیت با واژگان SentiStrength و تجمیع نمرات با استفاده از نظریهٔ دمپستر شافر کمترین صحت را دارد و بیشترین صحت را نیز واژگان به‌دست‌آمده با ویژگی Bigram دارد. البته گفتنی است که استفاده از این روش برای تجمیع نمرات، نسبت‌به تجمیع نمرات با استفاده از جمع ساده، باعث بهبود تشخیص قطبیت با استفاده از واژگان SentiStrength شد. اما همان‌طورکه مشاهده

[DOI: 10.61186/jsdp.20.2.3]

می‌شود، تجمیع نمرات با استفاده از تئوری دمپستر شافر به نسبت تجمیع نمرات با استفاده از روش جمع ساده، دقت تشخیص احساسات را برای کار ما علاوه‌براینکه افزایش نمی‌دهد، بدر حدود ۷٪ کاهش نیز می‌دهد. ازاین‌رو نتیجه می‌گیریم که استفاده از تئوری دمپستر شافر و همچنین، استفاده از واژگان‌های ساخته‌شده با اهداف استفادهٔ عمومی، علاوه‌بر اینکه ممکن است در بعضی از حوزه‌ها کاربرد خوبی نداشته‌باشند، ممکن است نتایج بسیار ضعیفی ایجاد نمایند.

(جدول-۱۲): ارزیابی واژگان‌های پیشنهادی و SentiStrength با استفاده از دمپستر شافر

(Table-12): Result of the evaluation of the proposed lexicons and SentiStrength using Dempster-Shafer theory

واژگان	صحت	دقت	بازیابی	F-Score
واژگان استخراج شده با Unigram	0.49	0.61	0.05	0.1
واژگان استخراج شده با Bigram	0.59	0.77	0.47	0.58
واژگان استخراج شده با Trigram	0.59	0.82	0.31	0.45
واژگان SentiStrength	0.51	0.68	0.26	0.38

۵- نتیجه‌گیری و پیشنهاد

هدف از تحلیل احساسات استخراج خودکار عقاید افراد نسبت‌به موضوعات مختلف در سطح شبکه است. شبکه‌های اجتماعی محیطی است که همه‌روزه افراد به بحث و تبادل نظر می‌پردازند. یکی از این شبکه‌ها که به‌تازگی، بسیار به آن توجه شده، شبکه‌های اجتماعی مرتبط با بازارهای مالی و یکی ازآن‌ها بازار بورس است که افراد خرید و فروش سهم را در آن انجام می‌دهند. در این شبکه‌ها افراد عقاید خود را نسبت‌به سهم‌ها بیان می‌کنند. ما در کار خود عقیده‌کاوی را بر روی توییت‌هایی که همه‌روزه در بازار بورس منتشر می‌شوند، انجام دادیم. ویژگی منحصربه‌فرد این شبکه‌ها، وجود اطلاعات قیمتی هر سهم در هر روز است. ما از این موضوع جهت استخراج واژگانی به‌صورت خودکار از بین ۹۰۰۰۰۰ توییت که در این شبکه‌ها موجود است، استفاده کردیم. درواقع، از این ویژگی‌ها برای محاسبهٔ نرخ رشد هر سهم در زمان درج نظر استفاده و امتیازات کلمات به‌دست‌آمده را برای واژگان به کمک این نرخ رشدها محاسبه کردیم. برای ارزیابی روش پیشنهادی برای تولید واژگان، واژگان‌های تولیدشده را با نسخهٔ فارسی واژگان SentiStrength که آخرین پروژه‌های زبان فارسی آن را استفاده می‌کنند، مقایسه کردیم. برای مقایسهٔ بهتر، روشی که آن‌ها در کار خود به‌کاربردند، نیز استفاده و نتایج ارزیابی شد. نتایج نشان داد که واژگان استفاده‌شدهٔ آن‌ها، در حوزهٔ بورس کارایی کمتری نسبت‌به واژگان تولیدشده به روش پیشنهادی دارد. همچنین، صحت به‌دست‌آمده با روش پیشنهادی، برابر با ۶۶ درصد است که ۲۰ درصد بیشتر از واژگان SentiStrength روی مجموعه‌دادگان ما است. علاوه‌بر این، پس از بررسی‌های انجام‌شده در انتخاب ویژگی مناسب برای استخراج کلمات واژگان، متوجه شدیم که استفاده از ویژگی Trigram نسبت‌به ویژگی‌های Unigram و Bigram، باعث بهبود کیفیت تحلیل احساسات در حوزه بازارهای مالی(بورس) می‌شود. زیرا کلماتی که با استفاده از ویژگی Trigram استخراج می‌شوند، از نظر توزیع آماری غیرشایع‌تر هستند و امکان مشاهدهٔ آن‌ها در روزهای مثبت و منفی بازار به‌صورت توأمان، کمتر است. در روش ما جملاتی که نمرهٔ نهایی آن‌ها نزدیک به صفر ارزیابی می‌شوند، جملات چالش‌برانگیزی هستند. تحلیل‌های قیمت با استفاده از تحلیل‌های همان روز سهم است. اما روزهای آتی سهم می‌تواند به ما در تشخیص احساسات کمک کند. زیرا بسیاری از نظرات درج شدهٔ کاربران، انعکاس وضعیت سهم از روزهای آیندهٔ سهم است. ما قصد داریم در کارهای آینده با کمک اطلاعات قیمتی روزهای آینده سهم، کیفیت واژگان استخراج‌شده و امتیازهای تخصیص‌یافته به کلمات آن را بهبود بخشیم.

۶- مراجع 6- Refrence

[1] B. Pang and L. Lee, "Opinion mining and sentiment analysis," Foundations and Trends® in Information Retrieval, vol. 2, no. 1–2, pp. 1-135, 2008.

[2] B. Liu and L. Zhang, "A survey of opinion mining and sentiment analysis," in Mining text data: Springer, 2012, pp. 415-463.

[3] I. Chaturvedi, E. Cambria, R. E. Welsch, and F. Herrera, "Distinguishing between facts and opinions for sentiment analysis: Survey and challenges," Information Fusion, vol. 44, pp. 65-77, 2018.

[4] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, "Lexicon-based methods for sentiment analysis," Computational linguistics, vol. 37, no. 2, pp. 267-307, 2011.

[5] G. A. Miller, WordNet: An electronic lexical database. MIT press, 1998.

[6] B. Liu, "Sentiment analysis and opinion mining," Synthesis lectures on human language technologies, vol. 5, no. 1, pp. 1-167, 2012.

[DOI: 10.61186/jsdp.20.2.3]

[7] S. Li, "Sentiment classification using subjective and objective views," International Journal of Computer Applications, vol. 80, no. 7, 2013.
[8] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? Sentiment classification using machine learning techniques," arXiv preprint cs/0205070, 2002.
[9] Y. Jo and A. H. Oh, "Aspect and sentiment unification model for online review analysis," in Proceedings of the fourth ACM international conference on Web search and data mining, 2011, pp. 815-824.
[10] D. Maynard and A. Funk, "Automatic detection of political opinions in tweets," in Extended Semantic Web Conference, 2011: Springer, pp. 88-99.
[11] K. Dashtipour, A. Hussain, Q. Zhou, A. Gelbukh, A. Y. Hawalah, and E. Cambria, "PerSent: A freely available Persian sentiment lexicon," in International Conference on Brain Inspired Cognitive Systems, 2016: Springer, pp. 310-320.
[12] P. D. Turney, "Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews," arXiv preprint cs/0212032, 2002.
[13] S. Rani and P. Kumar, "Deep learning based sentiment analysis using convolution neural network," Arabian Journal for Science and Engineering, vol. 44, no. 4, pp. 3305-3314, 2019.
[14] G. Xu, Y. Meng, X. Qiu, Z. Yu, and X. Wu, "Sentiment analysis of comment texts based on BiLSTM," Ieee Access, vol. 7, pp. 51522-51532, 2019.
[15] M. Rhanoui, M. Mikram, S. Yousfi, and S. Barzali, "A CNN-BiLSTM model for document-level sentiment analysis," Machine Learning and Knowledge Extraction, vol. 1, no. 3, pp. 832-847, 2019.
[16] M. Ahmad, S. Aftab, and I. Ali, "Sentiment analysis of tweets using svm," Int. J. Comput. Appl, vol. 177, no. 5, pp. 25-29, 2017.
[17] M. Ahmad, S. Aftab, M. S. Bashir, N. Hameed, I. Ali, and Z. Nawaz, "SVM optimization for sentiment analysis," Int. J. Adv. Comput. Sci. Appl, vol. 9, no. 4, pp. 393-398, 2018.
[18] K. Korovkinas, P. Danėnas, and G. Garšva, "SVM and k-means hybrid method for textual data sentiment analysis," Baltic Journal of Modern Computing, vol. 7, no. 1, pp. 47-60, 2019.
[19] L. Dey, S. Chakraborty, A. Biswas, B. Bose, and S. Tiwari, "Sentiment analysis of review datasets using naive bayes and k-nn classifier," arXiv preprint arXiv:1610.09982, 2016.
[20] V. Narayanan, I. Arora, and A. Bhatia, "Fast and accurate sentiment classification using an enhanced Naive Bayes model," in International Conference on Intelligent Data Engineering and Automated Learning, 2013: Springer, pp. 194-201.
[21] M. S. Hajmohammadi and R. Ibrahim, "A SVM-based method for sentiment analysis in Persian language," International Conference on Graphic and Image Processing (ICGIP 2012), vol. 8768, p. 876838, 2013.
[22] M. Saraee and A. Bagheri, "Feature selection methods in Persian sentiment analysis," International Conference on Application of Natural Language to Information Systems, pp. 303-308, 2013.
[23] A. S. H. Basari, B. Hussin, I. G. P. Ananta, and J. Zeniarja, "Opinion mining of movie review using hybrid method of support vector machine and particle swarm optimization," Procedia Engineering, vol. 53, no. 7, pp. 453-462, 2013.
[24] T. S. Ataei, K. Darvishi, S. Javdan, B. Minaei-Bidgoli, and S. Eetemadi, "Pars-ABSA: an Aspect-based Sentiment Analysis dataset for Persian," arXiv preprint arXiv:1908.01815, 2019.
[25] K. Dashtipour, M. Gogate, A. Adeel, C. Ieracitano, H. Larijani, and A. Hussain, "Exploiting deep learning for persian sentiment analysis," International Conference on Brain Inspired Cognitive Systems, pp. 597-604, 2018.
[26] E. Cambria, D. Olsher, and D. Rajagopal, "SenticNet 3: a common and common-sense knowledge base for cognition-driven sentiment analysis," in Proceedings of the AAAI Conference on Artificial Intelligence, 2014, vol. 28, no. 1.
[27] M. Hu and B. Liu, "Mining and summarizing customer reviews," in Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, 2004, pp. 168-177.
[28] L. Deng and J. Wiebe, "Mpqa 3.0: An entity/event-level sentiment corpus," in Proceedings of the 2015 conference of the North American chapter of the association for computational linguistics: human language technologies, 2015, pp. 1323-1328.
[29] A. Neviarouskaya, H. Prendinger, and M. Ishizuka, "SentiFul: A lexicon for sentiment analysis," IEEE Transactions on Affective Computing, vol. 2, no. 1, pp. 22-36, 2011.
[30] A. Esuli and F. Sebastiani, "Sentiwordnet: A publicly available lexical resource for opinion mining," in LREC, 2006, vol. 6: Citeseer, pp. 417-422.
[31] S. M. Mohammad and P. D. Turney, "Crowdsourcing a word–emotion association lexicon," Computational intelligence, vol. 29, no. 3, pp. 436-465, 2013.
[32] M. E. Basiri, A. R. Naghsh-Nilchi, and N. Ghassem-Aghaee, "A framework for sentiment analysis in persian," Open transactions on information processing, vol. 1, no. 3, pp. 1-14, 2014.
[33] M. E. Basiri and A. Kabiri, "Sentence-level sentiment analysis in Persian," in 2017 3rd International Conference on Pattern Recognition

[DOI: 10.61186/jsdp.20.2.3]

and Image Analysis (IPRIA), 2017: IEEE, pp. 84-89.
[34] S. M. Mohammad, S. Kiritchenko, and X. Zhu, "NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets," arXiv preprint arXiv:1308.6242, 2013.
[35] M. Thelwall, K. Buckley, and G. Paltoglou, "Sentiment strength detection for the social web," Journal of the American Society for Information Science and Technology, vol. 63, no. 1, pp. 163-173, 2012.
[36] M. E. Basiri and A. Kabiri, "Translation is not enough: comparing lexicon-based methods for sentiment analysis in Persian," in 2017 International Symposium on Computer Science and Software Engineering Conference (CSSE), 2017: IEEE, pp. 36-41.
[37] B. Sabeti, P. Hosseini, G. Ghassem-Sani, and S. A. Mirroshandel, "LexiPers: An ontology based sentiment lexicon for Persian," arXiv preprint arXiv:1911.05263, 2019.
[38] F. Amiri, S. Scerri, and M. Khodashahi, "Lexicon-based sentiment analysis for Persian text," in Proceedings of the International Conference Recent Advances in Natural Language Processing, 2015, pp. 9-16.
[39] R. Dehkharghani, "Sentifars: A persian polarity lexicon for sentiment analysis," ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), vol. 19, no. 2, pp. 1-12, 2019.
[40] E. Haddi, X. Liu, and Y. Shi, "The role of text pre-processing in sentiment analysis," Procedia Computer Science, vol. 17, pp. 26-32, 2013.
[41] A. AleAhmad, H. Amiri, E. Darrudi, M. Rahgozar, and F. Oroumchian, "Hamshahri: A standard Persian text collection," Knowledge-Based Systems, vol. 22, no. 5, pp. 382-387, 2009.
[42] M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, and A. Kappas, "Sentiment strength detection in short informal text," Journal of the American society for information science and technology, vol. 61, no. 12, pp. 2544-2558, 2010.
[43] G. Shafer, A mathematical theory of evidence. Princeton university press, 1976.

علی سبطی مدرک کارشناسی خود را در رشته مهندسی کامپیوتر گرایش نرم‌افزار در سال ۱۳۸۵ از دانشگاه یزد و مدرک کارشناسی ارشد خود را در سال ۱۳۸۸ در رشته مهندسی کامپیوتر گرایش هوش مصنوعی از دانشگاه صنعتی امیرکبیر اخذ کرد و دکترای خود را در رشته‌ی مهندسی کامپیوتر گرایش هوش مصنوعی در دانشگاه صنعتی شاهرود در سال ۱۳۹۶ به پایان رساند. ایشان در حال حاضر استادیار گروه مهندسی کامپیوتر دانشگاه گلستان است. زمینه‌های پژوهشی موردعلاقهٔ ایشان پردازش زبان طبیعی، پردازش تصویر و الگوریتم‌های تکاملی است.
نشانی رایانامهٔ ایشان عبارت است از:
a.sebti@gu.ac.ir

مرتضی آهنگری آهنگرکلائی کارشناس ارشد مهندسی نرم‌افزار از دانشگاه گلستان است. زمینه‌های پژوهشی وی تحلیل احساسات و پردازش زبان طبیعی است.
نشانی رایانامهٔ ایشان عبارت است از:
m.ahangari98@stu.gu.ac.ir

مهدی یعقوبی مدرک کارشناسی خود را در رشته مهندسی کامپیوتر گرایش نرم‌افزار در سال ۱۳۸۰ از دانشگاه صنعت نفت تهران و مدرک کارشناسی ارشد خود را در سال ۱۳۸۳ در رشته مهندسی کامپیوتر گرایش نرم‌افزار از دانشگاه صنعتی امیرکبیر دریافت کرد و دکترای خود را در رشته‌ی مهندسی کامپیوتر گرایش هوش مصنوعی در دانشگاه صنعتی شاهرود به پایان رساند. ایشان در حال حاضر استادیار گروه مهندسی کامپیوتر دانشگاه گلستان است. زمینه‌های پژوهشی موردعلاقهٔ ایشان داده‌کاوی، فرآیندکاوی و سامانه‌های اطلاعاتی مبتنی بر فرآیند است.
نشانی رایانامهٔ ایشان عبارت است از:
m.yaghoubi@gu.ac.ir

[DOI: 10.61186/jsdp.20.2.3]