فصلنامــه علمی پژوهشی دانش سرمایه‌گــذاری
سال دهم/ شماره چهل‌ام/ زمستان ۱۴۰۰

پیش‌بینی روند روزانه قیمت سهام با استفاده از متن‌کاوی احساسات کاربران شبکه اجتماعی و داده‌کاوی نماگرهای تکنیکال

کامل ابراهیمیان
دانشجوی دکترای مدیریت مالی،گروه مدیریت مالی، واحد قزوین، دانشگاه آزاد اسلامی، قزوین، ایران
Kamel61tb@gmail.com

ابراهیم عباسی
دانشیار و عضو هیئت علمی دانشگاه الزهرا، دانشکده علوم اجتماعی و اقتصادی، ده ونک ،تهران، ایران(نویسنده مسئول)
AbbasiEbrahim2000@alzahra.ac.ir

اکبر عالم تبریز
استاد، گروه مدیریت صنعتی، دانشکده مدیریت و حسابداری، دانشگاه شهید بهشتی، تهران، ایران
Tabriz-a@sbu.ac.ir

امیر محمدزاده
دانشیار، گروه مدیریت مالی، واحد قزوین، دانشگاه آزاد اسلامی، قزوین، ایران
A.Mohammadzadeh@qiau.ac.ir

تاریخ دریافت: ۱۴۰۰/۰۱/۲۸ تاریخ پذیرش: ۱۴۰۰/۰۲/۰۴

چکیده

این پژوهش به پیش‌بینی روند آتی قیمت سهام در بازه کوتاه مدت روزانه با استفاده از تحلیل نظرات سرمایه گذاران در شبکه‌ی اجتماعی سهامیاب می‌پردازد.قابلیت پیش‌بینی پذیری بازارهای سهام، به خاطر دارا بودن یک سیستم پیچیده، پویا و غیرخطی همواره از چالش‌های پژوهشگران بوده است در این تحقیق، برای اولین بار، با تحلیل احساسات نظرات کاربران و ترکیب آن با ۲۰ نماگر تکنیکال به کمک سه الگوریتم درخت تصمیم، بیز ساده و ماشین بردار پشتیبان ،مدلی با دقت ۷۲/۰۸ درصد برای پیش‌بینی جهت حرکت سهام توسعه‌یافت و به پیش‌بینی روند کوتاه مدت سهام پرداخته شد بر اساس نتایج، ماشین بردار پشتیبان، عملکردی بهتر از دو الگوریتم دیگر از خود نشان داد.همچنین مشخص شد حجم معاملات روز آتی و تعداد نظر ها دارای همبستگی معناداری است و نتایج آزمون علیت گرنجر نشان داد می‌توان برای پیش‌بینی قیمت سهام، از تجمیع احساسات روزانه کاربران نیز بهره جست.

واژه‌های کلیدی: پیش‌بینی قیمت سهام،الگوریتم‌های طبقه بندی، تحلیل احساسات، علیت گرنجر.

۱- مقدمه

در گذشته همیشه این سؤال مطرح بوده که تا چه میزان می‌توان از قیمت و رفتار گذشته سهام برای پیش‌بینی آن در آینده استفاده کرد؟(فاما ۱۹۶۵). هدف این پژوهش، پیش‌بینی روند حرکت سهام با استفاده از تحلیل احساسات و نماگرهای تکنیکال می‌باشد بررسی همبستگی و رابطه‌ی بین حجم نظرات با حجم معاملات روز بعد و بررسی علت گرنجری تجمیع احساسات روزانه بر نسبت تغییر قیمت روز بعدی و نیز تحلیل احساسات کاربران وب سایت سهام یاب از دیگر اهداف اصلی پژوهش می‌باشد بر همین اساس از سه الگوریتم درخت تصمیم[۱](DTREE) ، ماشین بردار پشتیبان[۲](SVM) و بیز ساده[۳](NBAYES) در این پژوهش استفاده می‌شود.

نتایج حاصل نشان دهنده دقت قابل قبول برای پیش‌بینی جهت حرکت سهام است و نوآوری ایـن پـژوهش در پیش‌بینی روند کوتاه مدت سهام با ترکیب تحلیل حساسیت نظر کاربران و استفاده از ۲۰ نماگر شاخص‌های علم تکنیکال است.

در بخش دوم مبانی نظری تحقیق ، روش ها و الگوریتم‌های به کار رفته در پژوهش توضیح داده می‌شوند. در بخش سوم روش شناسی پژوهش بیان شده و در بخش چهارم و پنجم به ترتیب نتایج حاصـل از مـدل سـازی و بحث و نتیجه گیری بیان خواهد شد.

۲- مبانی نظری و پیشینه‌ی تحقیق

۲-۱- داده‌کاوی

روش‌های داده‌کاوی را می‌توان به دو دسته‌ی روش‌های باناظر و بدون ناظر تقسیم کرد.روش‌های باناظر در حقیقت شامل الگوریتم‌هایی است که وظیفه‌ی پیش‌بینی را بر عهده دارند. این الگوریتم ها با یادگیری رابطه‌ی بین متغیرهای ورودی و متغیرهدف، الگوی پنهان بین آن ها را شناسایی کرده و از این طریق پیش‌بینی را انجام می‌دهند. الگوریتم‌های طبقه بندی و رگرسیون از روش‌های باناظر داده‌کاوی محسوب می‌شوند.در روش‌های بدون ناظر متغیرهدف یا متغیرقابل پیش‌بینی وجود ندارد. خوشه بندی و قوانین انجمنی از روش‌های بدون ناظر داده‌کاوی محسوب می‌شوند در روش‌های باناظر داده‌ها به دو مجموعه آموزشی (۷۰ درصد) و آزمایشی(۳۰ درصد) افراز می‌شوند که با مجموعه‌ی آموزشی، مدل آموزش داده شده وبا مجموعه‌ی آزمایشی صحت ودقت مدل را می سنجد.

۲-۲- متن‌کاوی

هدف از متن‌کاوی پیدا کردن دانش‌های نهفته شده در متون می‌باشد. فرایند متن‌کاوی دارای چهار مرحله اصلی می‌باشد.در ابتدا متون و نظرات افراد جمع‌آوری می‌شود. در مرحله دوم متون پاک سازی و عملیات پیش‌پردازش متون بصورت زیر انجام خواهد شد.

۱) تبدیل کلمات محاوره ای به رسمی

۲) حذف کاراکترهای اضافه مثل علامت سؤال، علامت تعجب،کاراکترها و کلمات غیرفارسی

۳) یکسان سازی حروفی که دارای چندین شکل هستند مثل حرف ی، ک، أ،إ،آ
۴) حذف کلماتstop word
۵) ریشه‌یابی

مرحله سوم تشکیل ماتریسی به عنوانTerm-Document –Matrix است که ستون‌های این ماتریس کلمات یا اصطلاح های(Term) استخراج شده هستند و سطرهای این ماتریس نظرات یا سندها(Document) هستند.به عبارت دیگر این ماتریس نشان‌دهنده فراوانی هر کلمه در هر سند (متن، نظر و ...) است.برای به‌دست آوردن نتایج از الگوریتم وزن دهی به نامTF-IDF [۴] استفاده می‌شود که میزان اهمیت یک اصطلاح نسبت به‌یک سند، را نشان می‌دهد فرض کنیدN تعداد کل اسناد، n_i تعداد اسنادی است که شامل اصطلاحi باشند، D_{ji} فراوانی اصطلاحi در سندjام و D_j تعداد کل اصطلاحات سندjام باشد.در این صورت اهمیت اصطلاحi در سندj از رابطه (۱) به دست می‌آید.

$$T_iF - ID_jF = \frac{D_{ji}}{D_j}\log\frac{N}{n_i} \quad (1)$$

در مرحله‌ی چهار، برای تعیین قطبیت نظرات در دو طبقه‌ی مثبت و منفی ، بخشی از نظرات را برچسب زده که نقش متغیرهدف را ایفا می‌کنند و سپس با اجرای گام‌های اول تا سوم، اصطلاحات، استخراج خواهند شد این اصطلاحات نقش متغیرهای ورودی را بازی خواهند کرد.

۲-۳- الگوریتم‌های داده‌کاوی

در این پژوهش‌از الگوریتم درخت تصمیم، بیز ساده و ماشین بردار پشتیبان(SVM)استفاده شده است

۲-۳-۱- الگوریتم درخت تصمیم

نحوه‌ی طبقه بندی درخت تصمیم ب، ین صورت است که به محض ورود یک نمونه‌ی ناشناخته (نمون، ی که طبقه‌یا برچسب آن مشخص نیست)، یک سری سؤالات به ترتیب در مورد آن نمونه پرسیده خواهد شد و با پاسخ گویی به هر سؤال مسیر حرکت در هر درخت مشخص می‌شود.

۲-۳-۲-الگوریتم بیز ساده

بیز ساده را می‌توان یک مدل بر مبنای احتمال شرطی [۵] در نظر گرفت.فرض کنید$X = \left(x_۱ ،x_۲ ، \dots ،x_n\right)$ برداری ازn ویژگی را بیان کند، که به صورت متغیرهای مستقل عمل می‌کنند.ب، ین ترتیب می‌توان احتمال رخدادC_k یعنی$P\left(C_k \middle| x_۱ ،x_۲ ، \dots ،x_n\right)$ به عنوان یکی از حالت‌های کلاس رخدادهای مختلف به ازایk های متفاوت از رابطه‌ی (۲) که همان قضیه‌ی بیز است نمایش داد.

$$P(C_k|X) = P\left(C_k \middle| x_۱ ،x_۲ ، \dots ،x_n\right) = \frac{P(C_k)P(X|C_k)}{P(X)} \quad (2)$$

۲-۳-۳- الگوریتم ماشین بردار پشتیبان(SVM)

الگوریتم ماشین بردار پشتیبان(SVM) توسط واپینگ و همکاران[۶](۱۹۹۵) معرفی شد بر این فرض استوار است که هدف مسأله ، طبقه بندی بین دو دسته است.SVM را می‌توان پدید آورنده‌ی خط یا ابرصفحه در بین مجموعه داده‌ها و به منظور طبقه بندی آن ها در نظر گرفت.

۲-۴- ارزیابی مدل

کارایی مدل با استفاده از ماتریس درهم ریختگی مطابق جدول (۱) سنجیده می‌شود.این ماتریس چگونگی عملکرد الگوریتم دسته بندی را با توجه به مجموعه داده‌ی ورودی و نحوه‌ی تفکیک آن در دسته‌های مناسب و نامناسب مورد تحلیل قرار داده و نمایش می‌دهد.در جدول ۱، TP نمونه‌های مثبتی است که به درستی، مثبت تشخیص داده شده‌اند..TN نمونه‌های منفی که به درستی، منفی تشخیص داده شده‌اند..FN نمونه‌های مثبتی که به صورت نادرست، منفی تشخیص داده شده‌اند وFP نمونه‌های منفی که به صورت نادرست، مثبت تشخیص داده شده‌اند.

جدول (۱): ماتریس درهم ریختگی

		پیش‌بینی ها	
		درستT	نادرستF
نمونه‌های واقعی	مثبتP	TP	FN
	منفیN	FP	TN

برای سنجش دقت مدل نیز از سنجه ی نسبت کل پیش‌بینی های درست استفاده می کنیم که نحوه محاسبه آن در جدول(۲) بیان شده است.

جدول (۲): سنجش دقت مدل

نسبت	شرح	نحوه محاسبه
TPR نرخ مثبت درست	نسبت نمونه هایی که به درستی مثبت تشخیص داده شده اند به تعداد کل نمونه های مثبت موجود	$\mathrm{TPR}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}$
TNR نرخ منفی درست	نسبت نمونه هایی که به درستی منفی تشخیص داده شده اند به تعداد کل نمونه های منفی موجود	$\mathrm{TNR}=\frac{\mathrm{TN}}{\mathrm{FP}+\mathrm{TN}}$
Accuracy دقت کلی الگوریتم	نسبت کل پیش‌بینی های درست	$\mathrm{Accuracy}=\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FN}+\mathrm{FP}}$

۲-۵- همبستگی

تحلیل همبستگی برای مشخص کردن نوع و شدت رابطه‌ی دو متغیرکمی‌از رابطه(۳)محاسبه می‌شود:

$$\rho = \frac{\sum_{i=1}^{n}\left((x_i - \bar{x})(y_i - \bar{y})\right)}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad (3)$$

مفهوم معنی داری همبستگی این است که همبستگی به دست آمده بین دو متغیررا می‌توان شانسی و تصادفی در نظر گرفت یا این که واقعاً بین دو متغیرهمبستگی وجود دارد آزمون فرضیه بصورت زیر:

$$\begin{cases} H_0 : \rho = 0 & \text{همبستگی وجود ندارد} \\ H_1 : \rho \neq 0 & \text{همبستگی وجود دارد} \end{cases}$$

و در سطح ۹۵ درصد سنجیده می شوداگر مقدارp-value کوچکتر از ۰/۰۵ باشد فرض صفر رد می‌شود یعنی همبستگی معنی داری بین دو متغیروجود خواهد داشت.

۲-۶- علیت گرنجری[۷]

گرنجر(۱۹۶۹) با استفاده از این واقعیت که آینده نمی‌تواند علت حال یا گذشته باشد بیان می‌کند اگر مقادیر جاری سری زمانیY_t با استفاده از مقادیر گذشته سریX_t با دقت بیشتری نسبت به حالتی که از اطلاعاتX_t استفاده نمی‌شود، پیش‌بینی شود، در این صورتX_t را علت گرنجریY_t گویند برای آزمودن فرضیه‌ی «X_t علت گرنجریY_t نیست»، مدل اتورگرسیو برداری[۸](VAR)از رابطه(۴)

$$Y_t = \sum_{i=1}^{k} \alpha_i Y_{t-1} + \sum_{i=1}^{k} \beta_i X_{t-1} + \mu_i \quad (4)$$

تعریف می‌شود.در این صورت اگر برای هر$\beta_i = 0$، $k, \ldots 2, 1 = i$ باشد،در نتیجهX_t علت گرنجریY_t نخواهد بود. جیوئیک[۹](۱۹۸۴) بیان می‌کند که اعتبار آزمون به مانایی[۱۰] هر دو سری زمانی بستگی دارد.یعنی هر دو سری زمانی می‌بایست مانا باشند.برای آزمون مانایی سری‌های زمانی می‌توان از آزمون دیکی فولر[۱۱] استفاده نمود.این آزمون، فرضیه زیر را مورد بررسی قرار می‌دهد.

$$\begin{cases} H_0 : & \text{سری زمانی مانا نمی‌باشد} \\ H_1 : & \text{سری زمانی مانا می‌باشد} \end{cases}$$

۲-۷- پیشینه تحقیق

احساسات می‌تواند نوسانات کوتاه مدت بازار را به دنبال داشته باشند شوماکر[۱۲] و چن[۱۳] (۲۰۰۹) تأثیر اخبارهای فوری بر قیمت سهام را پس از بیست دقیقه از انتشار آن ها ارزیابی کردند.بولن و همکاران (۲۰۱۱) از داده‌هـای توئیتر برای پیش‌بینی روند قیمت سهام استفاده کردنـد.آن هـا بـا اسـتفاده از الگـوریتم GPOMS[۱۴] بـه تحلیـل احساسات ده میلیون توئیت پرداخته و از این طریق قیمت بسته شدن شاخص صنعتی داوجونز[۱۵] را با دقت ۸۷/۶ ٪ پیش‌بینی کردند. میتال و گوئل[۱۶] (۲۰۱۲) اساس کار خود را بر مطالعه‌ی بولن وهمکاران (۲۰۱۱) قرار دادنـد اما با مجموعه‌ی بزرگتری از نظرات با اندازه‌ی ۴۰۰ میلیون و دقتی که آن ها به دست آوردند از ۸۷/۶ ٪به ۷۵ ٪ کاهش یافت. لی و همکاران (۲۰۱۴) روشی را برای تعیین اهمیت متن‌کاوی[۱۷] در پیش‌بینی سـهام ارائـه دادنـد. مشاری و همکاران(۱۳۹۹) در پژوهشی به بررسی قابلیت پیش‌بینی پذیری نقاط شروع(کف) و پایان(سـقف)روند کوتاه مدت قیمت سهام با استفاده از مدل نایوبیز پرداختند.وطن پرست و همکاران(۱۳۹۸) از یک شبکه عصبی LM-BP براساس سری‌های زمانی قیمتی روشی برای پیش‌بینی قیمت سهام معرفی کردند.

۳- سوالات پژوهش

- آیا تغییرات جهت حرکت سهام با کمک تکنیک‌های داده‌کاوی و متن‌کاوی امکان‌پذیر است؟
- الگوریتم ارائه شده نسبت به سایر روش‌ها چه مزیت‌هایی دارد ؟
- آیا قیمت‌های پیش‌بینی شده با دقت قابل قبولی پیش‌بینی پذیر هستند؟
- تاثیر شبکه‌های اجتماعی بر دقت پیش‌بینی نقاط ورود و خروج چقدر است؟

۴- روش شناسی و داده‌های پژوهش

هدف این پژوهش توسعه‌ی مدلی برای پیش‌بینی جهت حرکت سهام با رویکرد ترکیبی تحلیل احساسات و نماگرهای تکنیکالی است. بنابراین این پژوهش از نظر هدف، کاربردی و از نوع پژوهش‌های میدانی محسوب می‌شود.

۴-۱- داده‌های پژوهش

جامعه‌ی آماری اطلاعات جمع‌آوری شده‌ا ز ۱۴ شرکت پذیرفته شده در بورس تهران می‌باشد. داده‌های متنی پژوهش نظرات و توئیت‌های روزانه‌ی کاربران وب سایت سهام‌یاب است. لیست کاملی از خواص جمع‌آوری شده اولیه را می‌توان در جدول (۲) مشاهده نمود.

مدل سازی بر اساس ۲۰ نماگر تکنیکال قیمت‌های روزانه (جدول۳) و یک متغیرتحلیل احساسات صورت گرفته است از داده‌های سال‌های ۱۳۹۷ تا ۱۳۹۸ جهت مدل سازی استفاده شده است.

جدول (۲) لیستی از مشخصه‌ها برای مجموعه

ماهیت	مشخصه‌ها برای مجموعه
نظرات	نام، نام کاربری، محتوای نظرات، تاریخ انتشار، زمان انتشار، منبع پیام
کاربران	نام، نام کاربری، نظرات کاربر،

جدول(۳) : نماگر های تکنیکال

شرح متغیر	عبارت انگلیسی	منبع
نوسان ساز مبتنی بر حرکت	Know Sure Thing	پرینگ،۱۹۹۱
میانگین متحرک تطبیقی کافمن	Kaufman's Adaptive Moving Average	کافمن،۲۰۱۳
منحنی کوپک	Coppock Curve	کوپک ۱۹۶۲
شاخص ویلیام	Williams %R	ویلیامز ۲۰۱۴
شاخص ورتکس	Vortex Indicator	بوتس و داگلاس ۲۰۱۰
شاخص نوسان نگر لحظ، ی قیمت	Decision Point Price Momentum	مورفی ۲۰۰۹
شاخص نوسان ساز نهایی	Ultimate Oscillator	ویلیامز،۱۹۸۵
شاخص میانگین جهت دار	Average Directional Index	گوریب ایخلاص ۲۰۱۸
شاخص گردش پول	Money Flow Index	دنگ و ساکارای ۲۰۱۳
شاخص کانال کالا	Commodity Channel Index	باکر و همکاران ۱۹۸۳
شاخص قدرت واقعی	True strength index	بالو ۱۹۹۱
شاخص قدرت نسبی	relative strength index	مورفی،۲۰۰۹
شاخص حجم منفی	Negative Volume Index	دیزارت ۱۹۸۳
شاخص جریان پول چالکین	Chalkin Money Flow	نیمان ۲۰۰۹
شاخص توده	MASS Index	دروسی ۱۹۹۲
تغییر قیمت	Price Rate of Change	اسماعیل حسین،۲۰۱۸
برد واقعی میانگین	Average True Range	ویلدر،۱۹۷۸
اسیلاتور نوسان قیمت	percentage price oscillator	انتاکاریس،۲۰۱۹
استوکاستیک RSI	STOCHASTIC RSI	چانده ۱۹۹۴
استوکاستیک D٪	Stochastic oscillator	پرسون ۲۰۰۴

۴-۲- روش انجام پژوهش

۴-۲-۱- جمع‌آوری و پیش‌پردازش داده‌ها

نماد شرکت‌های استفاده شده همراه با تعداد نظرات جمع‌آوری شده در بازه‌ی زمانی۱۳۹۷/۱/۱ الی ۱۳۹۸/۱۲/۲۹ از وب سایت سهام یاب در جدول (۲) آورده شده‌ا ست.تعداد کل توئیتها این ۱۴ سهم ۵۵۰،۶۷۴

نظر بوده و با حذف نظرات تکراری به ۵۱۴،۷۰۵ نظر کاهش یافته است، برای پیش‌پردازش نظرات کاربران، مراحل زیر لحاظ شده‌ا ست:

۱) نظرات تکراری و نظرات آخر هفته‌ها و تعطیلات رسمی و زمان توقف نماد، حذف شدند.
۲) در نظرات علامت @ را با متن" اتساین" جایگزین شد.
۳) #نماد هر سهم در نظرات با کلمه" نمادسهم" جایگزین شد.

برای به دست آوردن نماگرها از نرم افزارTSE Client ، آمی‌بروکر و اکسل و از طریق برنامه نویسی استفاده شده است و به از زبان برنامه نویسیR برای داده‌کاوی و متن‌کاوی استفاده شده است.

۴-۲-۲- برچسب زنی نظرات کاربران

در این پژوهش به جای استفاده از واژه نامه‌های احساسی از تحلیل قطبیتی احساسات سرمایه گذاران استفاده می‌شود چون اگرچه در تعیین احساسات توییت، ذهنیت و عینیت در نظر گرفته می‌شود، اما در بسیاری از موارد، توییت‌های بیطرف (عینی) در شرایط نرمال باید به عنوان توییت‌های قطبی در این پژوهش تلقی شوند. به عنوان مثال، این نظر "#خودرو هیچ خبری نیست فقط وقت تلف کردنه"، به احتمال زیاد به عنوان جمله‌ای خنثی و بی طرف علامت گذاری خواهد شد و بنابراین از تجزیه و تحلیل قطبیتی خارج می‌شود، در حالی که در واقعیت نشان می‌دهد یک کاربر انتظار دارد قیمت سهام کاهش یابد. ثانیا"در حوزه زبان فارسی برای نظرکاوی بازارهای مالی و حتی سایر علوم فارسی، واژه‌نامه احساسی و نرم افزارهای خاص وجود ندارد برای غلبه بر این مشکل، از برچسب زنی بخشی از توییت‌ها به کمک خبرگان برای آموزش الگوریتم‌های طبقه بندی استفاده نمودیم.در جدول (۴) برخی از توییت ها و برچسب‌های نوعی آنها و همچنین استدلال‌های اینکه چرا آنها ب، ین ترتیب برچسب گذاری شده‌اند بیان شده‌ا ست.

جدول (۴) برخی نظرهای کاربران و برچسب‌های آنها

نظر	استدل	برچسب
#خودرو فردا تا ساعت ۱۱ هر چقدر حقوقی عرضه کرد بخرید که بعدش حالا حالاها نمی‌توانید بخرید.	نطری کوتاه همراه با احساسات مثبت نسبت به نماد #خودرو که روند رو به رشدی را طی خواهد کرد	مثبت
#کیسون، درگذشته هم پروژه کم نداشته، پس چرا قبل از این هیچ تاثیری روی قیمت سهامش نگذاشته؟	کاربر بدون اینکه احساسی از خود بروز دهد ، در مورد تاثیر پروژه‌ها در شرایط حال و تفاوت آن نسبت به گذشته سوال می‌پرسد	خنثی
#شپلی ، این سهم ۱۰۰ تومانی با سفته بازی رفت بالا... سهام خود را تعویض کنید... ازدواج# نکنید	کاربر نظر خود را در مورد رشد کاذب سهم اعلام می‌نماید، و تعویض سهم را پیشنهاد می‌کند.	منفی

۴-۲-۳- مدل سازی

در بخش اول پژوهش، مدلی برای طبقه بندی نظرات در دو طبقه‌ی مثبت و منفی توسعه داده می‌شود و پس از آموزش مدل و مشخص شدن کلاس تمامی نظرات، متغیرتجمیع احساسات روزانه محاسبه می‌شود. این متغیراز جمع جبری احساسات روزانه افراد به دست آمده است و عبارت است از تفاضل تعداد نظرات منفی از تعداد نظرات مثبت سهام مربوطه. به عبارت دیگر اگرi نام یکی از سهم‌های موجود در جدول (5) باشد در این صورت تجمیع احساسات روزj ام این سهام از رابطه‌ی(۵) به دست می‌آید.

$$S_{ij} = S_{ij}^{+} - S_{ij}^{-} \quad (5)$$

که در آن S_{ij}^{+} تعداد نظرات مثبت سهامi در روز j ام، S_{ij}^{-} تعداد نظرات منفی همان سهام در روزj ام و S_{ij} تجمیع احساسات است.اگر $S_{ij} > ۰$ باشد به این معنی است که کاربران نظر مساعدی نسبت به آن سهم دارند و احتمال می‌رود در روز بعدی جهت حرکت سهام رو به افزایش باشد و اگر $S_{ij} < ۰$ باشد آن گاه عکس مطلب فوق می‌تواند برقرار باشد.

بخش دوم تحلیل مربوط می‌شود به رابطه همبستگی، علیت گرنجری و بخش سوم مربوط به مدل سازی با رویکرد ترکیبی براساس نماگرهای تکنیکالی و تحلیل احساسات برای پیش‌بینی جهت حرکت سهام است.بنابراین در این بخش ۲۱ متغیرورودی روزانه برای یک سهم وجود دارد. ۲۰متغیرمربوط به نماگرهای روزانه‌ی تکنیکالی و متغیردیگر، تجمیع احساسات روزانه آن سهام است.متغیرهدف در این بخش جهت حرکت سهام است که در دو طبقه‌ی افزایشی و کاهشی قرار می‌گیرد.اگر قیمت روز بعد سهام بیشتر از قیمت امروز باشد، در نتیجه امروز در طبقه افزایشی قرار می‌گیرد و بالعکس، اگر کمتر باشد، طبقه‌ی امروز کاهشی خواهد بود.

آمار تعداد نظرات منتشر شده توسط کاربران ، تعداد روزهای معاملاتی ، رتبه دنبال کننده در سایت سهامیاب و نیز متوسط بازدید روزانه آن در جدول (۵) توصیف شده است.

جدول (5): لیست سهام به همراه تعداد نظرات

نام سهام	تعداد نظرات	روز معاملاتی	رتبه دنبال کننده	متوسط بازدید روزانه	نام سهام	تعداد نظرات	روز معاملاتی	رتبه دنبال کننده	متوسط بازدید روزانه
اخابر	۴،۷۲۶	۴۳۵	۶۵	۲،۴۵۰	خودرو	۲۰،۸۴۴	۴۳۵	۶	۴۵،۵۸۱
حفاری	۱۰،۲۵۲	۳۴۹	۴۹	۳،۸۵۱	کیسون	۲۲،۴۸۴	۴۷۲	۱۱۷	۲،۷۳۰
فاراک	۱۱،۴۲۴	۴۳۶	۲۹	۴،۵۲۳	خساپا	۲۳،۷۹۷	۴۶۱	۵	۲۰،۶۲۴
شتران	۱۲،۱۳۹	۴۴۵	۱۳	۱۸،۱۶۶	وتجارت	۳۶،۰۹۱	۳۵۶	۲	۲۱،۸۵۵
وبصادر	۱۴،۳۳۷	۴۶۱	۱۲	۱۱،۲۹۵	ذوب	۳۸،۲۱۹	۴۶۳	۴	۷،۲۷۵
شبندر	۱۴،۳۶۸	۴۶۰	۱۰	۱۰،۴۶۱	تپکو	۴۲،۰۱۲	۴۵۸	۱۸	۳،۲۳۹
ستران	۱۷،۹۲۴	۴۵۳	۶۲	۲،۴۱۶	شپلی	۷۸،۰۱۵	۴۳۶	۲۳	۴،۷۷۸

۵- تجزیه و تحلیل

سه نوع تحلیل در این پژوهش مورد بررسی قرار گرفت اولین تحلیل مربوط به رویکرد پردازش زبان طبیعی و تحلیل احساسات، دومین تحلیل مربوط به همبستگی و علیت گرنجری و سومین تحلیل مربوط به رویکرد ترکیبی پیش‌بینی جهت حرکت سهام است.

۵-۱- پردازش زبان طبیعی و تحلیل احساسات

یک نمونه‌ی تصادفی به حجم ۲،۳۴۴ توئیت به عنوان نمونه‌های آموزشی و آزمایشی انتخاب گردید که به صورت دستی توسط خبرگان و کارشناسان بازار مالی، قطبیت آن ها در دو رده‌ی مثبت و منفی مشخص شد. از این تعداد ۱۶۴۰ توئیت (۷۰ درصد) به عنوان نمونه‌ی آموزشی و ۷۰۴ توئیت (۳۰درصد) به عنوان نمونه‌ی آزمایشی انتخاب شدند. تعداد نظرات مثبت و منفی در نمونه‌ی آموزشی به ترتیب برابر با ۸۲۵ و ۸۱۵ و همین اعداد در نظرات آزمایشی برابر با ۳۷۵ و ۳۲۹ است. قبل از بیان نتایج الگوریتم‌های طبقه بندی تشریح می‌گردد که برای ساخت ویژگی ها (متغیرها) ازunigram (تک کلمه ای ها) وbigram ها (دو کلمه ای ها) استفاده شده است.از الگوریتم‌های بیز ساده، درخت تصمیم و ماشین بردار پشتیبان برای ساخت مدل پیش‌بینی طبقه‌ی نظرات استفاده شده و نتایج ارزیابی آن ها برای نمونه‌های آموزشی و آزمایشی به ترتیب در جداول (۶) و (۷) آمده است. همان طور که از جدول (۶) مشخص است الگوریتم ماشین بردار پشتیبان در نمونه‌های آموزشی دارای بیشترین دقت با ۸۰/۳۷ درصد است و سپس الگوریتم بیز ساده دارای بیشترین دقت ٪۷۳/۶۶بوده است. باید توجه شود که الگوریتم درخت تصمیم عملکرد بسیار ضعیفی از خود نشان داده است بنابراین نتایج آن به هیچ عنوان قابل بررسی نمی‌باشد.

جدول (۶): نتایج ارزیابی الگوریتم‌های طبقه بندی نمونه‌های آموزشی (تحلیل احساسات)

الگوریتم	دقت کلی ٪	فراخوان توئیت‌های مثبت ٪	فراخوان توئیت‌های منفی ٪	ماتریس درهم ریختگی		
					منفی	مثبت
SVM	۸۰/۳۷	۷۷/۳۳	۸۳/۴۴	مثبت	۱۸۷	۶۳۸
				منفی	۶۸۰	۱۳۵
					منفی	مثبت
DTREE	۶۲/۲۰	۳۵/۵۲	۸۹/۲۱	مثبت	۵۳۲	۲۹۳
				منفی	۷۲۷	۸۸
					منفی	مثبت
NBAYES	۷۳/۶۶	۶۹/۸۲	۷۷/۵۵	مثبت	۲۴۹	۵۷۶
				منفی	۶۳۲	۱۸۳

نتایج مشابهی نیز در ارزیابی نمونه‌های آزمایشی با این سه الگوریتم در جدول (۷) وجود دارد.با توجه به این جدول الگوریتم ماشین بردار پشتیبان دارای دقت ۷۰/۲۰ درصد است.لذا از این الگوریتم برای پیش‌بینی طبقه‌ی سایر نظرات استفاده خواهد شد.

جدول (۷): نتایج ارزیابی الگوریتم‌های طبقه بندی نمونه‌های آزمایشی (تحلیل احساسات)

الگوریتم	دقت کلی ٪	فراخوان توئیت‌های مثبت ٪	فراخوان توئیت‌های منفی ٪	ماتریس درهم ریختگی
SVM	۷۰/۲۰	۷۲/۵۳	۶۷/۴۷	منفی ۱۰۳ مثبت ۲۷۲ مثبت ۱۰۷ منفی ۲۲۲
DTREE	۵۷/۳۹	۳۲/۲۷	۸۶/۰۲	منفی ۲۵۴ مثبت ۱۲۱ مثبت ۴۶ منفی ۲۸۳
NBAYES	۶۵/۲۰	۶۰/۵۳	۷۰/۵۲	منفی ۱۴۸ مثبت ۲۲۷ مثبت ۹۷ منفی ۲۳۲

پس از برچسب گذاری تمامی‌نظرات متغیرهای حجم روزانه‌ی نظرات ، تعداد نظرات مثبت روزانه، تعداد نظرات منفی روزانه و تجمیع احساسات روزانه برای هر سهم محاسبه گردید.

۵-۲- همبستگی ها و علت گرنجری حجم توئیت ها

سؤال اساسی مهم این است که آیا حجم توئیت‌های نماد x و حجم معاملات آن نماد در روز بعد رابطه و وجود دارد؟ و می‌توان انتظار داشت که بین حجم توئیت‌های هر نماد با حجم معاملات روز بعد آن همبستگی معناداری وجود داشته باشد؟ مقدار ضریب همبستگی وp-value مربوط به آزمون فرض همبستگی در جدول (۸) بیانگر این است که در همه نمادها بجز خساپا و کیسون، بین حجم کل توئیت ها و حجم معاملات روز بعد، همبستگی مثبت و معناداری وجود دارد.اما همبستگی بین حجم توئیت‌های مثبت و حجم معاملات روز بعد، فقط در ۹ نماد معنادار شده‌ا ست و در آخر بین حجم توئیت‌های منفی و حجم معاملات روز بعد، همبستگی معناداری مشاهده نمی‌شود.

با توجه به همبستگی بین حجم توئیت ها و حجم معاملات روز بعد، می‌توان یک گام فراتر رفته و این سؤال را مطرح نمود ک، یا تجمیع احساسات روزانه دارای اطلاعاتی برای پیش‌بینی نسبت تغییر قیمت سهام(PRC) می‌باشد؟ برای پاسخ ب، ین سؤال می‌بایست از آزمون علیت گرنجر استفاده نمود اما قبل از آن، آزمون دیکی فولر برای بررسی مانایی هر دو سری زمانی الزامی می‌باشد. ذکر این نکته ضروری است که هر دو سری حداکثر باید در تأخیر یک مانا شوند. برای توضیح و بررسی‌های بیشتر، فرض کنید که علیت گرنجری تجمیع احساسات امروز (S_i) برای پیش‌بینی تغییر قیمت فردا(PRC_{i+1}) توسط مدل اتورگرسیو برداری (۶) تعریف شود.

$$PRC_{i+1} = \alpha_0 + \alpha_1 PRC_i + \beta_0 + \beta_1 S_i \quad (6)$$

بنابراین هر دو سری زمانی می‌بایست حداکثر در تأخیر یک مانا گردند و اگر نمادهایی نتوانند در هـر یـک از دو آزمون دیکی فولر و علیت گرنجر معنادار شوند، از تحلیل حذف خواهند شـد.بـا توجـه بـه جـدول (۹) مشـخص می‌شود که در تمامی‌نمادها آزمون دیکی فولر معنادار شـده‌ا سـت و لـذا هـر دو متغیـرPRC وS دارای شـرایط مانایی در تأخیر یک هستند.

جدول (۸) همبستگی بین حجم توئیت ها و حجم معاملات روز بعد

نماد	کل توئیت ها		توئیت‌های مثبت		توئیت‌های منفی	
	همبستگی	P-value	همبستگی	P-value	همبستگی	P-value
اخابر	۰/۲۵۹۰	۰/۰۰۰۰	۰/۱۹۵۰	۰/۰۰۰۰	۰/۰۵۵۰	۰/۲۵۲۸
تپکو	۰/۲۱۴۰	۰/۰۰۰۱	۰/۱۲۲۰	۰/۰۲۳۳	۰/۰۷۳۶	۰/۱۷۰۸
تجارت	۰/۱۹۴۰	۰/۰۰۰۰	۰/۱۴۰۰	۰/۰۰۳۵	۰/۰۳۷۲	۰/۴۳۹۶
حفاری	۰/۲۰۰۰	۰/۰۰۰۰	۰/۱۹۳۰	۰/۰۰۰۰	-۰/۰۱۰۷	۰/۸۲۱۳
خساپا	۰/۰۸۹۹	۰/۰۵۴۱	۰/۰۲۷۲	۰/۵۶۰۲	۰/۰۶۲۹	۰/۱۷۸۱
خودرو	۰/۱۷۳۰	۰/۰۰۰۲	۰/۰۸۲۲	۰/۰۷۸۷	۰/۰۸۱۶	۰/۰۸۰۶
ذوب	۰/۱۰۲۰	۰/۰۳۰۴	۰/۰۳۲۴	۰/۴۹۱۸	۰/۰۷۶۴	۰/۱۰۵۰
ستران	۰/۱۶۵۰	۰/۰۰۰۶	۰/۰۷۸۰	۰/۱۰۴۹	۰/۰۷۸۸	۰/۱۰۱۱
شتران	۰/۲۰۹۰	۰/۰۰۰۰	۰/۱۷۶۰	۰/۰۰۰۱	۰/۰۰۲۲۶	۰/۹۶۱۰
شبندر	۰/۱۹۶۰	۰/۰۰۰۰	۰/۱۲۷۰	۰/۰۰۶۳	۰/۰۵۸۹	۰/۲۰۷۶
شپلی	۰/۱۹۲۰	۰/۰۰۰۳	۰/۱۵۲۰	۰/۰۰۴۲	۰/۰۱۱۳	۰/۸۳۱۶
فاراک	۰/۱۷۷۰	۰/۰۰۰۱	۰/۱۹۹۰	۰/۰۰۰۰	-۰/۰۵۷۰	۰/۲۲۱۱
کیسون	۰/۰۷۳۰	۰/۱۱۸۹	۰/۰۶۱۳	۰/۱۹۰۹	۰/۰۱۲۷	۰/۷۸۶۸
وبصادر	۰/۱۷۳۰	۰/۰۰۰۳	۰/۱۳۵۰	۰/۰۰۴۸	۰/۰۳۰۱	۰/۵۳۱۲

جدول (۹): نتایج آزمون دیکی فولر

نماد	نسبت تغییر قیمت		تجمیع احساسات		نماد	نسبت تغییر قیمت		تجمیع احساسات	
	آماره دیکی فولر	P-value	آماره دیکی فولر	P-value		آماره دیکی فولر	P-value	آماره دیکی فولر	P-value
اخابر	-۱۲/۸۸	۰/۰۱*	-۱۴/۷۲	۰/۰۱*	ستران	-۱۰/۵۵	۰/۰۱*	-۱۴/۴۶	۰/۰۱*
تپکو	-۱۲/۹۳	۰/۰۱*	-۱۱/۲۲	۰/۰۱*	شتران	-۱۴/۱۲	۰/۰۱*	-۱۷/۱۸	۰/۰۱*
تجارت	-۱۴/۵۱	۰/۰۱*	-۱۵/۱۹	۰/۰۱*	شبندر	-۱۵/۲۵	۰/۰۱*	-۱۵/۵۹	۰/۰۱*
حفاری	-۱۲/۵۰	۰/۰۱*	-۱۵/۴۹	۰/۰۱*	شپلی	-۱۲/۹۰	۰/۰۱*	-۱۳/۰۸	۰/۰۱*
خساپا	-۱۲/۴۹	۰/۰۱*	-۱۴/۵۲	۰/۰۱*	فاراک	-۱۳/۹۴	۰/۰۱*	-۱۷/۴۱	۰/۰۱*
خودرو	-۱۳/۰۶	۰/۰۱*	-۱۵/۱۰	۰/۰۱*	کیسون	-۱۴/۵۱	۰/۰۱*	-۱۵/۰۱	۰/۰۱*
ذوب	-۱۴/۳۴	۰/۰۱*	-۱۵/۵۴	۰/۰۱*	وبصادر	-۱۴/۸۰	۰/۰۱*	-۱۵/۰۷	۰/۰۱*

* کوچکتر از ۰/۰۱

با توجه به جدول (۱۰) آزمون علت گرنجری در تمامی‌نمادها بجز تپکو و تجارت معنادارشده‌ا ست بنابراین این دو نماد از ادامه تحلیل حذف خواهند شد.

جدول (۱۰): نتایج آزمون علیت گرنجری تجمیع احساسات برای پیش‌بینی نسبت تغییر قیمت

نماد	آماره F	P-value	نماد	آماره F	P-value	نماد	آماره F	P-value	نماد	آماره F	P-value
اخابر	۲۷/۵۹	۰/۰۰۰۰	ستران	۳۴/۳۵	۰/۰۰۰۰	خساپا	۱۹/۰۲	۰/۰۰۰۰	فاراک	۶۴/۱۱	۰/۰۰۰۰
تپکو	۰/۸۴	۰/۳۵۸۷	شتران	۴۱/۶۰	۰/۰۰۰۰	خودرو	۵۰/۱۲	۰/۰۰۰۰	کیسون	۲۶/۶۵	۰/۰۰۰۰
تجارت	۰/۳۹	۰/۵۲۹۲	شبندر	۸/۱۱	۰/۰۰۴۶	ذوب	۲۱/۳۹	۰/۰۰۰۰	وبصادر	۷/۴۵	۰/۰۰۶۶
حفاری	۲۵/۵۴	۰/۰۰۰۰	شپلی	۳۳/۷۴	۰/۰۰۰۰						

۵-۳- رویکرد ترکیبی پیش‌بینی جهت حرکت سهام

بین احساسات کاربران و حجم معاملات و قیمت سهام در روز بعد رابطه و همبستگی وجود دارد.آمار توصیفی متغیرها (با حذف دو نماد تپکو و تجارت) و همچنین نقش آن ها در مدل سازی‌های این پژوهش در جدول (۱۱) نشان داده شده‌ا ست.

جدول (۱۱): شاخص‌های آمار توصیفی متغیرهای مدل

نام متغیر	نقش	مینیمم	ماکسیمم	میانگین	انحراف معیار	چولگی	کشیدگی
VOLUME	ورودی	۲۰۰۰	۲۸۵۸۰۸۶۰۰۰	۶۳۶۰۸۱۷۰	۱۲۱۷۲۵۶۰۰	۷/۸۱۹۹	۹۹/۴۴۱۳
PRC	ورودی	-۱/۷۲۲۵	۰/۲۳۸۹	۰/۰۰۲۴	۰/۰۴۲۸	-۳۸/۱۶۶۸	۱۹۱۵/۲۱۴۳
ADX	ورودی	۳/۴۸۷۲	۹۶/۳۴۶۶	۲۹/۳۵۵۴	۱۷/۱۲۱۰	۰/۹۶۲۰	۰/۵۷۷۹
CCI	ورودی	-۶۴۷/۲۵۳۰	۴۵۹/۰۷۳۷	۲۹/۷۹۵۹	۱۱۳/۹۶۵۷	-۰/۱۸۹۲	۰/۵۶۵۴
CPPC	ورودی	-۱۲۲/۱۰۴۴	۱۹۱/۸۶۶۰	۱۰/۹۵۴۵	۳۱/۲۰۶۰	۱/۵۸۰۴	۹/۹۸۸۴
ATR	ورودی	۲/۵۷۱۴	۲۲۵۶	۱۹۹/۷۹۱۶	۲۳۰/۳۷۲۹	۳/۶۱۸۲	۲۱/۰۲۲۷
KAMMA	ورودی	۴۰۸/۱۰۸۷	۴۰۴۵۰/۵۴۰۰	۴۴۰۸/۳۲۶۰	۴۵۲۶/۴۳۸۰	۲/۹۰۵۳	۱۳/۶۶۸۸
KST	ورودی	-۵۹۳/۲۶۵۷	۱۴۹۳/۷۸۳۰	۹۶/۲۹۵۲	۲۱۶/۸۰۳۹	۲/۰۳۱۹	۹/۸۱۳۳
MASS	ورودی	۱۶/۹۹۶۵	۳۳/۷۱۷۴	۲۵/۳۵۶۹	۱/۵۹۲۶	-۰/۲۳۳۹	۳/۸۳۸۷
MFI	ورودی	۳/۷۳۷۹	۱۰۰	۵۹/۰۹۵۴	۱۹/۰۱۵۳	-۰/۲۶۷۲	-۰/۴۶۴۲
NVI	ورودی	۶۱۴/۳۷۰۹	۱۱۱۷/۵۶۰۰	۷۸۱/۳۸۰۳	۱۰۳/۰۰۵۸	-۰/۹۵۹۹	۱/۰۳۳۲
PPO	ورودی	-۲۶/۱۱۰۱	۲۳/۶۵۰۹	۱/۸۸۶۱	۵/۱۷۰۵	-۱/۴۷۴۶	۱۱/۱۷۷۲
RSI	ورودی	۱۳/۰۰۸۸	۹۸/۷۱۳۱	۵۵/۳۴۳۱	۱۴/۷۳۴۸	-۰/۱۲۶۰	-۰/۱۳۲۶
STOCHASTIC	ورودی	۰/۰۱۰۰	۱۰۰/۰۱۰۰	۵۴/۸۳۶۷	۳۰/۵۳۵۵	-۰/۱۶۷۲	-۱/۲۷۸۴
STOCHASTIC-۱۶	ورودی	۰/۰۰۰۰	۰/۹۹۹۹	۰/۵۰۲۴	۰/۳۷۱۶	-۰/۰۴۱۳	-۱/۵۳۱۱
TSI	ورودی	-۶۲/۷۶۸۹	۹۴/۴۵۵۲	۱۰/۶۶۰۲	۲۵/۹۶۷۸	۰/۰۱۵۳	۰/۰۸۶۰

نام متغیر	نقش	مینیمم	ماکسیمم	میانگین	انحراف معیار	چولگی	کشیدگی
UO	ورودی	۴/۸۱۴۴	۹۸/۹۰۰۳	۴۸/۷۰۱۳	۱۴/۵۷۵۷	۰/۵۲۶۶	۰/۶۰۷۴
VI	ورودی	۰/۵۴۸۹	۱/۵۲۱۸	۱/۰۱۳۰	۰/۱۵۴۲	۰/۰۵۲۹	-۰/۲۶۷۲
CMF	ورودی	-۰/۶۰۰۹	۰/۷۰۰۰	۰/۰۵۲۱-	۰/۲۰۸۶	۰/۱۹۴۷	-۰/۱۱۰۲
PMO	ورودی	-۱۷/۰۲۰۲	۳۳/۰۵۹۶	۳/۶۱۰۹	۶/۳۲۶۳	۰/۷۹۲۳	۳/۳۳۵۷
WILLIAMSR	ورودی	-۱۰۰	۰	۴۵/۱۷۳۴-	۳۰/۵۳۵۵	-۰/۱۶۷۲	-۱/۲۷۸۴
MACD	ورودی	-۲۲۷۶/۱۳۱۱	۲۹۶۶/۱۰۰۰	۵۹/۸۵۹۳	۲۴۵/۰۷۱۲	۳/۳۳۰۲	۴۱/۴۴۸۳
S	ورودی	-۹۵	۱۰۵	۶/۳۷۸۸	۴۰/۷۴۹۵	-۰/۰۳۳۵	-۱/۱۷۹۷
(جهت حرکت سهام)Y	هدف	طبقه افزایشی: ۲۷۹۰		طبقه کاهشی: ۲۵۴۵			

متغیرهدف، یک متغیررسته‌ای با دو دسته‌ی افزایشی و کاهشی است و در کل اطلاعات این ۱۲ سهم ، ۵۳۳۵ روز را شامل می‌شوند که از این تعداد ۲۷۹۰ روز منجر به حرکتی افزایشی و ۲۵۴۵ روز نیز منجر به حرکت کاهشی شده‌اند..

یک نمونه‌ی تصادفی به اندازه‌ی ۳۷۳۴ روز برای آموزش مدل‌های طبقه بندی در نظر گرفته شده است.نتایج آموزش و آزمایش مدل‌های طبقه بندی در جداول(۱۲) و(۱۳) گزارش شده است. با توجه به این دو جدول مشخص می‌شود که در نمونه‌های آموزشی و آزمایشی الگوریتمSVM دارای دقت بیشتری نسبت به دو الگوریتم دیگر می‌باشد و برخلاف مدل سازی تحلیل احساسات، الگوریتم بیز ساده عملکرد ضعیفی از خود نشان داده است. لذا از الگوریتم ماشین بردار پشتیبان برای رویکرد ترکیبی پیش‌بینی جهت حرکت سهام استفاده می‌شود و بر این اساس نتایج مدل سازی براساس این الگوریتم به صورت جداگانه برای هر یک از نمادها در جدول (۱۴) آورده شده‌ا ست.

جدول (۱۲): نتایج ارزیابی الگوریتم‌های طبقه بندی نمونه‌های آموزشی (رویکرد ترکیبی)

الگوریتم	دقت کلی ٪	فراخوان مثبت ٪	فراخوان منفی ٪	ماتریس درهم ریختگی		
				خطا	منفی	مثبت
SVM	۷۵/۲۵	۷۶/۳۳	۷۴/۰۵	۲۳/۶۷	۴۶۸	۱۵۰۹ مثبت
				۲۵/۹۵	۱۳۰۱	۴۵۶ منفی
DTREE	۶۹/۶۶	۶۹/۴۵	۶۴/۲۰	۳۰/۵۵	۶۰۴	۱۳۷۳ مثبت
				۳۵/۸۰	۱۱۲۸	۶۲۹ منفی
NBAYES	۶۱/۴۶	۶۲/۳۷	۶۰/۴۴	۳۷/۶۳	۷۴۴	۱۲۳۳ مثبت
				۳۹/۵۶	۱۰۶۲	۶۹۵ منفی

جدول (۱۳): نتایج ارزیابی الگوریتم‌های طبقه بندی نمونه‌های آزمایشی (رویکرد ترکیبی)

الگوریتم	دقت کلی ٪	فراخوان مثبت ٪	فراخوان منفی٪	ماتریس درهم ریختگی		
					کاهشی	افزایش
SVM	۷۲/۰۸	۶۸/۶۳	۷۵/۶۳	خطا	۲۵۵	۵۵۸
				افزایشی	۳۱/۳۷	۱۹۲
				کاهشی	۲۴/۳۷	۵۹۶
DTREE	۶۴/۲۷	۵۸/۷۹	۶۹/۹۲	خطا	۳۳۵	۴۷۸
				افزایشی	۴۱/۲۱	۲۳۷
				کاهشی	۳۰/۰۸	۵۵۱
NBAYES	۴۹/۴۱	۵۱/۹۱	۴۶/۸۳	خطا	۳۹۱	۴۲۲
				افزایشی	۴۸/۰۹	۳۶۹
				کاهشی	۵۳/۱۷	۴۱۹

جدول (۱۴): نتایج ارزیابی الگوریتم SVM نمونه‌های آموزشی (رویکرد ترکیبی)

نماد	٪دقت کلی	٪فراخوان نظرمثبت	٪فراخوان نظرمنفی	ماتریس درهم ریختگی		
					کاهشی	افزایش
اخابر	۷۲/۴۱	۶۸/۷۸	۷۵/۶۵	خطا	۶۴	۱۴۱
				افزایشی	۳۱/۲۲	۵۶
				کاهشی	۲۴/۳۵	۱۷۴
حفاری	۷۲/۱۳	۷۰/۵۶	۷۳/۸۳	خطا	۶۸	۱۶۳
				افزایشی	۲۹/۴۴	۵۶
				کاهشی	۲۶/۱۷	۱۵۸
خساپا	۷۳/۷۵	۷۲/۰۰	۷۵/۸۳	خطا	۷۰	۱۸۰
				افزایشی	۲۸/۰۰	۵۱
				کاهشی	۲۴/۱۷	۱۶۰
خودرو	۷۰/۶۵	۶۷/۹۲	۷۲/۹۸	خطا	۶۸	۱۴۴
				افزایشی	۳۲/۰۸	۶۷
				کاهشی	۲۷/۰۲	۱۸۱
ذوب	۷۳/۷۳	۷۲/۵۳	۷۵/۰۰	خطا	۶۴	۱۶۹
				افزایشی	۲۷/۴۷	۵۵
				کاهشی	۲۵/۰۰	۱۶۵
ستران	۷۲/۶۴	۷۳/۴۴	۷۱/۶۵	خطا	۶۴	۱۷۷
				افزایشی	۲۶/۵۶	۵۵
				کاهشی	۲۸/۳۵	۱۳۹

نماد	٪دقت کلی	٪فراخوان نظرمثبت	٪فراخوان نظرمنفی	ماتریس درهم ریختگی			
					خطا	کاهشی	افزایش
شتران	۷۵/۲۱	۷۳/۱۲	۷۶/۶۳	افزایشی	۲۶/۸۸	۶۸	۱۸۵
				کاهشی	۲۳/۳۷	۱۷۰	۴۹
					خطا	کاهشی	افزایش
شبندر	۷۳/۷۵	۷۱/۶۰	۷۶/۱۵	افزایشی	۲۸/۴۰	۶۹	۱۷۴
				کاهشی	۲۳/۸۵	۱۶۶	۵۲
					خطا	کاهشی	افزایش
شپلی	۶۹/۶۶	۶۷/۲۰	۷۲/۳۵	افزایشی	۳۲/۸۰	۶۱	۱۲۵
				کاهشی	۲۷/۶۵	۱۲۳	۴۷
					خطا	کاهشی	افزایش
فاراک	۷۶/۴۶	۷۴/۲۳	۷۹/۳۱	افزایشی	۲۵/۷۷	۶۷	۱۹۳
				کاهشی	۲۰/۶۹	۱۶۱	۴۲
					خطا	کاهشی	افزایش
کیسون	۷۵/۵۵	۷۲/۶۹	۷۸/۱۰	افزایشی	۲۷/۳۱	۵۹	۱۵۷
				کاهشی	۲۱/۹۰	۱۸۹	۵۳
					خطا	کاهشی	افزایش
وبصادر	۶۸/۵۸	۷۰/۲۰	۷۹/۸۳	افزایشی	۲۹/۸۰	۵۹	۱۳۹
				کاهشی	۲۰/۱۷	۱۹۰	۴۸

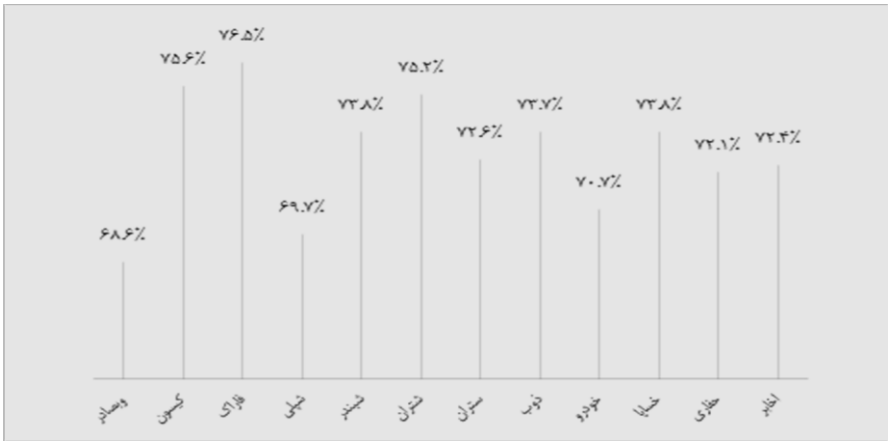
دقت استفاده از الگوریتم ماشین بردار پشتیبان برای رویکرد ترکیبی پیش‌بینی جهت حرکت سهام در نمودار (۱) بیانگر افزایش دقت در فراخوان نطرات منفی است می توان نتیجه گرفت که هر وقتع احساسات منفی باشد بیشترین حرکت سهام با احساسات منفی رخ می دهدو در نتیجه قدرت پیش بینی حرکات منفی سهام بهتر می شود.

نمودار(۱) دقت فراخوان الگوریتم طبقه بندی SVM

۶- بحث و نتیجه گیری

در این پژوهش با استفاده از یک رویکرد ترکیبی تحلیل احساسات و نماگرهای تکنیکالی مدلی برای پیش‌بینی جهت حرکت سهام با سه نوع الگوریتم داده کاوی توسعه داده شد نتایج به دست آمده از الگوریتم ماشین بردار پشتیبان نشان داد دقت مدل در بخش تحلیل احساسات ۷۰/۲۰٪ ودر بخش پیش‌بینی جهت حرکت سهام ۸۰/۳۹ ٪ بوده است. در پاسخ به سوالات پژوهش می‌توان گفت:

- با کمک احساسات و نماگرهای تکنیکالی می‌توان تا ۸۰٪ جهت حرکت سهام را پیش بینی نمود.
- طیف وسیعی از نظرات و احساسات معامله گران آنلاین و اثر آن بر جهت حرکت سهام بررسی شد در حالی که در پژوهش‌های گذشته فقط از اخبار و یا شایعات برای پیش‌بینی استفاده می‌شد
- بررسی رفتار معامله گران خرد در پیش‌بینی قیمت مهم است ولی اثرات نامتقارن سهامداران عمده می‌تواند در نتیجه مدل موثر باشد.در این پژوهش فقط رفتار معامله گران حقیقی و فعال در سهام یاب مورد بررسی قرار گرفت و فقدان بررسی رفتارهای حقوقی ها، سهامداران عمده و بازیگردان‌های سهم از نقاط ضعف این پژوهش به شمار می‌رود.
- با توجه به نمودار (۲) نتایج به دست آمده از الگوریتمSVM دقت پیش‌بینی جهت حرکت سهام را حدوداً ۸۰ درصد نشان می‌دهد که میتوان برای توسعه مدل های آتی از آن استفاده نمود.



نمودار(۲) دقت الگوریتم SVM در پیش بینی جهت حرکت سهام

برای پژوهش‌های آتی می‌توان متغیرهای بنیادی موثر بر ارزش سهام، متغییرهای کلان اقتصادی نظیر نرخ بهره، بودجه و تغییر قیمتهای جهانی و اثرآنها بر نظرات کاربران را نیز به عنوان متغیرهای ورودی به مدل اضافه کرد، استفاده از تحلیل گرافها و شبکه‌های پیوسته بین کاربران و و سطح دانش آنها در سرمایه گذاری و نیز توسعه‌یک واژه نامه مخصوص بازار سرمایه در زمینه تحلیل احساسات ، مورد تحقیق و پژوهش‌های آتی قرار گیرد.

فهرست منابع

* Fama, E. F. (1965). The behavior of stock-market prices. The journal of Business, Vol.38, no. 1, pp. 34-105.
* Geweke, J. (1984). Inference and causality in economic time series models. Handbook of econometrics, Vol.2, pp.1101-1144.
* Granger, C. W. (1969). Investigating causal relations by econometric models and cross-spectral methods. Econometrica: journal of the Econometric Society, pp.424-438.
* J.Bollen، H. Mao، and X. Zeng. (2011). Twitter mood predicts the stock market. Journal of Computational Science.
* Lee, H., Surdeanu, M., MacCartney, B., & Jurafsky, D. (2014, May). On the Importance of Text Analysis for Stock Price Prediction. In LREC (Vol. 2014, pp. 1170-1175).
* Mittal، Anshul، and Arpit Goel. (2012). Stock Prediction Using Twitter Sentiment Analysis. Stanford University، CS229. Available online: http://cs229.stanford.edu/proj2011/GoelMittalStockMarketPredictionUsingTwitterSenti
* Moshari.M، Didehkhani.H،Dameghani.KK، Abbasi.E. (2020).” Investigating the Predictability of Starting Point and Ending Short-Term Trend of Stock Price Using the Bayesian Likelihood Network”، Journal of Financial Management Strategy، Vol. 8، No. 28 Spring

* Schumaker، R. P.، & Chen، H. (2009). Textual analysis of stock market prediction using breaking financial news: The AZF in text system. ACM Transactions on Information Systems، 27(2)
* Vapnik.V. (1995). "The Nature of Statistical Learning Theory"، New York، NY: Springer.
* Vatanparast Mohammadreza، Asadi masoud، Mohammadi Shaban، Babaei abbas .2019.Stock price prediction based on LM-BP neural network and over-point estimation by counting time intervals: Evidence from the Stock Exchange. FINANCIAL summer، Volume 10 ، Number 39; Page(s) 193 to 218.

یادداشت‌ها

1 Decision Tree
2 Support Vector Machine
3 Naive Bayes
4 Term Frequency - Inverse Document Frequency
5 Conditional probability
6 Vapnik et al
7 Granger Causality
8 Vector autoregressive
9 Geweke
10 Stationary
11 Dickey Fuller
12 Schumaker
13 Chen
14 Google-Profile of Mood States
15 Dow Jones Industrial Average
16 Mittal and Goel
17 Text mining