مدیریت فناوری اطلاعات
Journal of Information Technology Management
دانشکدهٔ مدیریت دانشگاه تهران
دورهٔ ۸، شمارهٔ ۲
تابستان ۱۳۹۵
ص. ۴۳۴- ۴۱۵

# بهبود روش‌های متن‌کاوی در کاربرد پیش‌بینی بازار با استفاده از الگوریتم‌های انتخاب نمونهٔ اولیه

فرزاد نیکنام[۱]، علی‌اکبر نیک‌نفس[۲]

**چکیده:** امروزه محققان با حجم وسیعی از داده مواجه‌اند که بخش زیادی از آنها ساختار پردازش‌پذیری ندارند. دو مورد از چالش‌های اصلی در این زمینه بالا بودن ابعاد فضای ویژگی و حجیم بودن داده‌های در دسترس است. به‌منظور رفع این چالش‌ها، مقالهٔ پیش رو یک روش انتخاب ویژگی مبتنی بر ویژگی‌های هدف ارائه کرده است که در کاهش ابعاد فضای ویژگی تأثیر زیادی دارد و همچنین برای مقابله با حجم بسیار زیاد نمونه‌های آموزش، با استفاده از روش‌های انتخاب نمونهٔ اولیه، به ویرایش مجموعهٔ آموزش می‌پردازد. روش پیشنهادی در این مقاله در سه فاز اجرا شده است که هر فاز بهبودیافتهٔ فاز قبل است و علاوه‌بر دست‌یافتن به نتایج مناسب در هر فاز، در پایان فاز سوم روش پیشنهادی بیشترین کارایی را به‌دست آورد. برای ارزیابی کارایی روش پیشنهادی، این روش با یکی از الگوریتم‌های موفق در زمینهٔ پیش‌بینی بازار مقایسه شد که با وجود کاهش نمونه‌های آموزش توسط الگوریتم‌های انتخاب نمونهٔ اولیه، به نتایج بسیار بهتری نسبت به آن الگوریتم دست یافت.

***واژه‌های کلیدی: انتخاب نمونهٔ اولیه، پیش‌بینی بازار، طبقه‌بندی متن، متن‌کاوی.***

۱. دانشجوی کارشناسی ارشد مهندسی کامپیوتر، دانشکدهٔ فنی و مهندسی، دانشگاه شهید باهنر کرمان، کرمان، ایران
۲. استادیار بخش مهندسی کامپیوتر، دانشکدهٔ فنی و مهندسی دانشگاه شهید باهنر کرمان، کرمان، ایران

**تاریخ دریافت مقاله: ۱۳۹٤/۱۰/۲۲**
**تاریخ پذیرش نهایی مقاله: ۱۳۹۵/۰۲/۱۷**
**نویسندهٔ مسئول مقاله: فرزاد نیکنام**
**E-mail: fd.niknam@gmail.com**

## مقدمه

بازارهای هر کشور در حال توسعه، قلب تپندهٔ آن کشور به‌شمار می‌روند و اطلاع از نوسان‌های بازار برای هر کشور مهم و حیاتی است. به‌کمک علم پیش‌بینی بازار و یادگیری حرکات بازار، تاجران می‌توانند به سودهای شایان توجهی دست یابند یا حتی از ضررهای مالی فراوان دور بمانند. از این رو، بسیاری از تجار و اقتصاددانان به‌دنبال روش‌هایی برای پیش‌بینی نوسان‌های بازار هستند. اخبار اقتصادی و مالی یکی از منابع مهمی است که تاجران می‌توانند به‌کمک آنها از روند بازار و چگونگی حرکت آن آگاه شوند. تحقیقات نشان می‌دهد بین اخبار منتشرشده و آیندهٔ بازار، ارتباط بسیار محکمی وجود دارد (ایم و همکاران، ۲۰۱۴). بنابراین، استخراج عقاید و الگوهای پنهان از میان اخبار مشاهده‌شده به‌منظور پیش‌بینی آیندهٔ بازار، بسیار مفید است؛ اما با توجه به ساخت‌یافته‌نبودن اخبار منتشرشده، درک بشر از این‌گونه اطلاعات بسیار محدود است؛ به همین دلیل پژوهشگران کمتر به این موضوع پرداخته‌اند و معدود پژوهش‌های موجود در این زمینه نیز، دقت کم و نزدیک به ۵۰ درصد دارند. از جمله این کارها می‌توان به پژوهش علمی ایم و همکارانش (۲۰۱۴) اشاره کرد که دقت آن به ۷۱ درصد رسید یا میان پژوهش‌های موجود، پژوهش نصیرطوسی، آقابزرگی، وا و انگو (۲۰۱۴) در زمینهٔ پیش‌بینی بازار، به بهترین دقت دست یافت؛ به‌طوری که در برخی موارد دقت روش پیشنهادی آنها به ۸۳/۳۳ درصد رسید. بنابراین، ارائهٔ مدلی برای استفاده از اسناد متنی و خبرهای اقتصادی بسیار مهم است.

دانش متن‌کاوی ابزاری برای تجزیه و تحلیل اسناد متنی است و با تبدیل ساختار متون به ساختار پردازش‌پذیر، به استخراج دانش از میان اسناد متنی و غیر قابل پردازش می‌پردازد (ویس، ایندارخیا و ژانگ، ۲۰۱۰). پیش‌بینی بازار با استفاده از متون خبری با چالش‌های بزرگی از جمله بالا بودن ابعاد فضای ویژگی و فراوان بودن اسناد متنی در دسترس همراه است. زیاد بودن تعداد ویژگی‌ها در ماتریس سند ـ ویژگی، تأثیر بسیار زیادی بر افزایش پراکندگی ماتریس سند ـ ویژگی و کاهش کارایی الگوریتم‌های یادگیری ماشین در مراحل بعدی پردازش دارد (یانگ، لیو، ژو، لیو و ژانگ، ۲۰۱۲). همچنین تعداد اسناد خبری به‌منظور پردازش بسیار فراوان است؛ در حالیکه ممکن است حجم بسیار عظیمی از این اسناد بی‌اهمیت باشند و از کارایی فرایند پیش‌بینی بکاهند (پاسینی، لویزا، استفانز، فیگوردو و نیلسون، ۲۰۱۳). بنابراین چالش‌های بیان‌شده، انگیزهٔ این پژوهش شدند و راه‌حل‌های زیر به‌تفصیل مورد مطالعه قرار گرفتند.

۱. ویرایش مجموعهٔ آموزش و حذف اسناد نویزی از درون آن به‌کمک الگوریتم‌های انتخاب نمونه‌های اولیهٔ ویرایشی؛
۲. کاهش ابعاد فضای ویژگی با توجه به ویژگی‌های مجموعهٔ هدف (مجموعهٔ تست).

دو مورد بیان‌شده، از مهم‌ترین نوآوری‌های این مقاله است که با ترکیب آنها و استفاده از برخی روش‌های پیش‌پردازش همانند هیستوگرام و ریشه‌یابی، روش جدیدی به‌منظور افزایش کارایی پیش‌بینی بازار ارائه شده است. روش پیشنهادشدهٔ این مقاله، روی بازار ارز فارکس به اجرا درآمد و با استفاده از تیترهای خبری، روند تغییرات ارز یورو بر مبنای ارز دلار پیش‌بینی شد. این مسئله، مسئلهٔ کلاسیک طبقه‌بندی متن است که تیترهای خبری را در دو دستهٔ افزایش روند قیمت و کاهش روند قیمت ارز یورو دسته‌بندی می‌کند.

## پیشینهٔ پژوهش

روش‌های پیش‌بینی بازار در دو شاخهٔ کلی قرار می‌گیرند، شاخهٔ اول مربوط به پیش‌بینی بازار بر اساس تاریخچه بازار است و پژوهشگرانی که در این زمینه به مطالعه می‌پردازند، اعتقاد به تکرار تاریخچهٔ بازار دارند. این روش‌ها که به آنالیز فنی مشهورند، داده‌های مربوط به سال‌های گذشتهٔ کالایی را بر اساس روش‌های مختلف پیش‌بینی، مانند شبکه‌های عصبی و... ارزیابی می‌کنند و با محاسبات دقیق، به شناسایی الگوهای مخفی در آنها می‌پردازند و جهت حرکت قیمت یا قیمت نهایی کالا را در زمان مشخصی پیش‌بینی می‌کنند. شاخهٔ دوم که با عنوان آنالیز بنیادی شناخته می‌شود، نسبت به روش‌های فنی امیدوارکننده‌تر است (نصیرطوسی، آقابزرگی، وا و انگو، ۲۰۱۵) و به تحلیل داده‌های اساسی می‌پردازد. داده‌های این شاخه نیز از منابع مختلفی مانند، اطلاعات مالی شرکت، وضعیت جغرافیایی و آب‌وهوا مانند بلایای طبیعی و غیرطبیعی، موقعیت سیاسی، اطلاعات مالی دربارهٔ فعالیت‌های دولت و بانک‌ها و اخبار اقتصادی و مالی گوناگون جمع‌آوری می‌شود. یکی از مهم‌ترین چالش‌ها در این زمینه، ساخت‌یافته‌نبودن داده‌های جمع‌آوری‌شده است؛ به همین دلیل در حوزهٔ تحلیل بنیادین بازار، تحقیقات کمتری نسبت به آنالیز فنی انجام‌شده و آن معدود تحقیقات نیز کارایی کمی دارند.

در حوزهٔ تحلیل بنیادی، برخی مطالعات به ارائهٔ نمای کلی از سیستم‌های پیش‌بینی بر اساس خبرها اختصاص دارد (هگنو، لیبمن و نیومن، ۲۰۱۳؛ نیک‌فرجام، عمادزاده و موتایا، ۲۰۱۰ و نصیرطوسی و همکاران، ۲۰۱۴). این پژوهشگران در مطالعات خود اجزای سیستم‌های پیش‌بینی را بررسی کردند و به مرور مطالعات انجام‌شده در این زمینه پرداختند. همچنین نصیرطوسی و همکارانش (۲۰۱۵) به‌منظور تأکید بیشتر بر روش‌های متن‌کاوی و مقابله با برخی از جنبه‌های خاص از جمله مشکلات ابعاد زیاد و نادیده‌گرفتن احساسات و معناشناسی[1] در برخورد با زبان متنی، الگوریتم چندلایه‌ای ارائه کردند. الگوریتم آنها از نوعی روش انتخاب ویژگی بر مبنای

1. Sentiment

ویژگی‌های هدف بهره برده و به‌منظور پیش‌بینی هر نمونهٔ تست، مدل مجزایی برای هر نمونه ایجاد شده است. هگنو و همکارانش (۲۰۱۳) در پژوهشی بر روش‌های استخراج و انتخاب ویژگی تمرکز کردند. به همین منظور پس از ترکیب روش‌های پیشرفتهٔ انتخاب ویژگی، یعنی استفاده از ترکیب کلمات به‌عنوان ویژگی و ادغام این روش‌ها با بازخوردهای بازار، توانستند دقت طبقه‌بندی و آنالیزهای احساسات را افزایش دهند. برخی از پژوهشگران نیز به‌منظور تحلیل بازار با استفاده از داده‌های متنی، از روش‌های عقیده‌کاوی و تحلیل احساسات متون استفاده کردند. در این زمینه کیم، جنگ و غنی (۲۰۱۴) برای پیش‌بینی شاخص بورس کامپوزیت کُره، نوعی روش عقیده‌کاوی ارائه دادند. همچنین ایم و همکارانش (۲۰۱۴) نوعی سیستم آنالیز احساسات بر مبنای فرهنگ لغت طراحی کردند و با استفاده از فرهنگ لغت به هر کلمه یک امتیاز مثبت و منفی اختصاص دادند. آنها به‌منظور بررسی تأثیر تیترهای خبری بر بازار، آزمایش‌های خود را سه مرتبه تکرار کردند (فقط بر اساس تیتر، تیتر و محتوا، فقط بر اساس محتوا). در زمینهٔ تحلیل احساسات، هانگ، لیو، یانگ، چانگ و لو (۲۰۱۰) به‌منظور رسیدن به بار معنایی خبرها و تأثیر آنها بر بازار، ابتدا با استفاده از قوانین انجمنی وزن‌دار به شناسایی ویژگی‌های مهم پرداختند و به هر ویژگی متناسب با اهمیتش وزنی اختصاص دادند. سپس با بهره‌مندی از تکنیک‌های داده‌کاوی بار معنایی هر خبر را مشخص کردند. دفورتونیو، دسمیت، مارتینز و دلمانس (۲۰۱۴) به بحث و تحقیق دربارهٔ طراحی مدلی برای پیش‌بینی قیمت سهام بر مبنای تکنیک‌های متن‌کاوی پرداختند. مطالعهٔ آنها در سه بخش انجام گرفت؛ در بخش اول یک مدل پیش‌بینی قیمت بر مبنای تکنیک‌های متن‌کاوی طراحی کردند، در بخش دوم به مطالعهٔ معیارهای مناسب‌تر و دقیق‌تر برای ارزیابی مدل‌های پیش‌بینی قیمت بر اساس تکنیک‌های متن‌کاوی پرداختند و در بخش سوم به بحث در خصوص به‌دست آوردن بینشی جدید برای ارائهٔ مدل‌ها با دقت بیشتر پرداختند.

در بیشتر مطالعات پیشین در زمینهٔ تحلیل بنیادی بازار، به‌منظور استخراج ویژگی از روش کیسهٔ کلمات بهره‌برده شده که این روش با مشکل زیادبودن ابعاد فضای ویژگی همراه است (نصیرطوسی، آقابزرگی، وا و انگو، ۲۰۱۴). در برخی مطالعات نیز با وجود کاهش ابعاد فضای ویژگی، روش‌های پیشنهادی کارایی کمی دارند. از چالش‌های مهم دیگر مطالعات، تعداد زیاد نمونه‌ها و اسناد آموزش است که این موضوع سبب طولانی‌شدن محاسبات در فرایند آموزش مدل می‌شود؛ در حالیکه ممکن است بسیاری از اسناد آموزش بی‌اهمیت باشند و در فرایند آموزش مدل، سبب اختلال و یادگیری کم مدل شوند. به‌طور مثال امکان دارد لابه‌لای خبرها، برخی از اخبار از منابع نامعتبر منتشر شوند. این چالش در اغلب مطالعات دیده می‌شود، اما کمتر

به آن پرداخته شده است. مطالعهٔ پاسینی، لـویزا، اسـتفانز، فیگـوردو و نیلسـون (۲۰۱۳) از دسـته مطالعات طبقه‌بندی متن است که در آن از تکنیک‌های انتخاب نمونه‌های آموزش اسـتفاده شـده است؛ اما در حوزهٔ مطالعهٔ بازار، بهره‌مندی از این تکنیک‌ها و به‌خصـوص تکنیـک‌هـای انتخـاب نمونه‌های اولیه، مشاهده نشده است.

الگوریتم‌های انتخاب نمونه‌های اولیه، الگوریتم‌هایی هستند کـه پـس از ویـرایش مجموعـهٔ آموزش، از الگوریتم‌های داده‌کاوی مبتنی بر نمونه استفاده می‌کنند (گارسیا، لنگو و هررا، ۲۰۱۵ و مورتی و دو، ۲۰۱۱). اساس کار برخی از این الگوریتم‌هـا بـر مبنـای حـذف داده‌هـای نـویزی از مجموعهٔ آموزش است و استفاده از این الگوریتم‌ها علاوه‌بر حـذف نمونـه‌هـای نـویزی از میـان مجموعهٔ آموزش، به کاهش نمونه‌های آمـوزش و کـاهش میـزان محاسـبات در الگـوریتم‌هـای داده‌کاوی مبتنی بر نمونه، کمک فراوانی می‌کند. در این مطالعه تـأثیر ایـن روش‌هـا بـر مطالعـهٔ بنیادین بازار فارکس بررسی شده است.

## روش‌شناسی پژوهش

روش پیشنهادی این مقاله طی سه فاز متوالی و به‌هم‌پیوسته و آزمایش روش‌هـای متنـوع ارائـه شده است و جنبهٔ کاربردی دارد. در هر فاز تلاش شده است که با رفع نقص‌هـای فـاز قبـل، بـه بهبود الگوریتم پرداخته شود تا الگوریتم در فاز پایانی از نظر کارایی به نتایج مطلوبی دست یابد.

### فاز اول

فاز اول مطالعه مربوط به ارائهٔ روش پیشنهادی در دو گام پیش‌پردازش و طبقه‌بندی اسناد اسـت که این فاز پایه و اساس فازهای دوم و سوم پژوهش محسوب می‌شـود. در فازهـای دوم و سـوم گام پیش‌پردازش، روش پیشنهادی نسبت به فاز اول تغییری نمی‌کند و عمدهٔ تمرکز این فازها بر بهبود فاز طبقه‌بندی روش پیشنهادی است.

مرحلهٔ اول روش پیشنهادی، مرحلهٔ پیش‌پردازش اسناد است. بـدین ترتیـب کـه ابتـدا اسـناد واردشده به این مرحله قطعه‌بندی[1] می‌شوند و سپس به مرحلهٔ حذف کلمات توقف وارد می‌شوند. در این مرحله، کلماتی مانند a, an, the و ... که به تعداد فراوان در هـر سـند تکـرار شـده‌انـد و هیچ‌گونه بار معنایی ندارند، از میان کلمات سند حذف خواهند شد (آگراوال و ژایی، ۲۰۱۲). مرحلهٔ بعد در گام پیش‌پردازش، مرحلهٔ ریشه‌یابی است. هدف از این مرحله یکسان‌سازی شـکل و فـرم کلمات موجود در اسناد است. به‌کمک روش‌های ریشه‌یابی، کلماتی که از نظر مفهوم مشابه‌انـد و

1. Tokenization

تنها در فرم ظاهری با یکدیگر تفاوت دارند، در یک گروه قرار می‌گیرند و به‌عنوان ویژگی در نظر گرفته می‌شوند. کاهش ابعاد فضای ویژگی از مزیت‌های ریشه‌یابی است و در روش پیشنهادی از ریشه‌یاب پورتر که یکی از مشهورترین ریشه‌یاب‌های زبان انگلیسی است، استفاده شده است (جیوانی، ۲۰۱۱). در مرحلهٔ بعد، حضور یا نبود دقیق کلمات هر سند در فرهنگ لغت وردنت[1] بررسی می‌شود؛ در صورتی که کلمه‌ای به‌طور دقیق در این فرهنگ لغت باشد، کلمه در بُردار ویژگی مربوط به هر سند حفظ می‌شود و در غیر این صورت از بردار ویژگی سندش حذف خواهد شد. گام بعد، یافتن معادل HYPERNYM برای کلمات موجود در هر سند است. معادل HYPERNYM هر کلمه برابر با طبقهٔ هر کلمه است. برای مثال معادل HYPERNYM برای سه کلمهٔ «آبی»، «قرمز» و «زرد» کلمهٔ «رنگ» است. چنانچه برای یک کلمه بیش از یک معادل یافت شود، اولین کلمهٔ یافت‌شده به‌جای کلمهٔ اصلی در بُردار ویژگی سند جایگزین می‌شود و در صورت نیافتن معادل برای یک کلمه، آن کلمه از بُردار ویژگی مربوط به هر سند حذف خواهد شد.

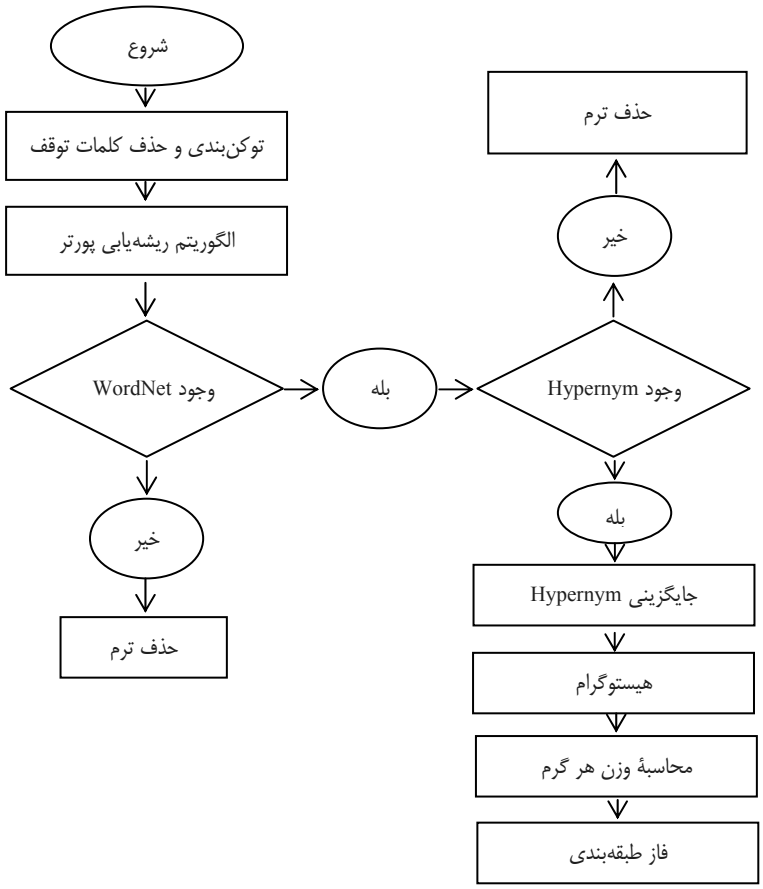
در مرحلهٔ بعد، هدف، یافتن وزن مناسب هر کلمه با توجه به روش‌های وزن‌دهی TF-IDF و Sum Score است. روش TF-IDF، روش وزن‌دهی متداول در مطالعات متن‌کاوی است که بر اساس تعداد تکرار هر کلمه در هر سند و تعداد تکرار آن در کل اسناد، به آن کلمه وزن مناسب اختصاص می‌دهد و از رابطهٔ ۱ محاسبه می‌شود. در این رابطه $t_k$ به کلمهٔ kام، $d_i$ به سند iام، N به تعداد کل اسناد موجود و $d_k$ به تعداد اسناد دارای ترم $t_k$ اشاره دارد (ویس و همکاران، ۲۰۱۰).

$$TFIDF(t_k, d_i) = TF(t_k, d_i) \times \log\left(\frac{N}{d_k}\right) \quad \text{رابطهٔ ۱)}$$

Sum Score نوعی معیار وزن‌دهی است که از فرهنگ لغت احساسی استخراج می‌شود. در این فرهنگ لغت، به هر کلمه با توجه به مفهومش سه امتیاز مثبت، منفی و امتیاز کل نسبت داده می‌شود و Sum Score برابر مجموع امتیازهای مثبت و منفی است (نصیرطوسی و همکاران، ۲۰۱۵). پس از محاسبهٔ TF-IDF و Sum Score برای هر کلمه در هر سند، از حاصل‌ضرب این دو معیار به‌عنوان وزن اختصاص داده‌شده به هر ترم استفاده می‌شود و تنها در مواردی که این حاصل‌ضرب برای همهٔ ترم‌های یک سند برابر صفر باشد، برای آن سند فقط از معیار TF-IDF استفاده می‌شود. شکل ۱ مراحل گام پیش‌پردازش روش پیشنهادی را نشان می‌دهد. مرحلهٔ آخر این گام فاز طبقه‌بندی نام دارد که بهبود این فاز از اهداف اصلی پژوهش است و این هدف در سه فاز متوالی اجرا شده است.

---

1. WordNet




**شکل ۱. فلوچارت روش پیشنهادی**

گام دوم روش پیشنهادی، طبقه‌بندی اسناد است که شکل ۲ مراحل آن را نمایش می‌دهد. ورودی این گام، ماتریس سند ویژگی به‌دست‌آمده از گام پیش‌پردازش است. در این فاز هدف طبقه‌بندی اسناد یا به بیانی، پیش‌بینی روند ارز یورو بر حسب دلار بعد از انتشار گروه‌های خبری است.

مرحلهٔ اول گام طبقه‌بندی، مرحلهٔ انتخاب ویژگی است؛ در این مرحله تنها از ویژگی‌های غیرصفر نمونه‌های آزمایشی به‌عنوان ویژگی‌های برتر استفاده می‌شود. استفاده از ویژگی‌های غیرصفر نمونه‌های آزمایش، علاوه‌بر کاهش بسیار زیاد ابعاد فضای ویژگی، این امکان را می‌دهد که مدل بر اساس ویژگی‌های موجود در نمونه‌های آزمایش، آموزش ببیند. گام بعد به ویرایش

مجموعهٔ آموزش بر مبنای معیار فاصلهٔ اقلیدسی اختصاص دارد. بـه ایـن منظـور مـاتریس سـند ویژگی بر اساس حضور یا نبود ویژگی‌ها در اسناد ایجـاد مـی‌شـود و سـپس بـه ازای هـر نمونـهٔ آزمایشی، تعداد M نمونهٔ نزدیک به آن انتخاب‌شده و درون مجموعهٔ آموزش جدید قرار می‌گیرند. مجموعهٔ آموزش جدید ایجادشده به همراه فضای ویژگی کاهش داده‌شده، ورودی الگوریتم‌هـای یادگیری ماشین را شکل می‌دهد و از مدل آموزش‌داده‌شـده بـرای برچسـب‌گـذاری نمونـه‌هـای آزمایش استفاده می‌شود.

۱. انتخاب ویژگی‌های غیرصفر نمونه‌های آزمایش به‌عنوان ویژگی‌های برتر
۲. انتخاب m نمونهٔ نزدیک به هر دادهٔ آزمایش بر اساس فاصلهٔ اقلیدسی
۳. آموزش مدل بر اساس نمونه‌های انتخاب‌شده و فضای ویژگی جدید
۴. طبقه‌بندی نمونه‌های جدید

**شکل ۲. مراحل گام طبقه‌بندی روش پیشنهادی در فاز اول**

## فاز دوم

فاز دوم بر گام طبقه‌بندی فاز اول و بهبود کارایی این گام تمرکز دارد. بنابراین در این فـاز اسـناد مطابق شکل ۱ از گام پیش‌پردازش عبور داده می‌شوند و ماتریس سند ویژگی حاصل از خروجـی این دو الگوریتم، به‌عنوان ورودی به گام طبقه‌بندی وارد می‌شـود. گـام طبقـه‌بنـدی در فـاز دوم مطابق شکل ۳ است. همان‌طور که مشاهده می‌شود، در فاز دوم هدف اسـتفاده از الگـوریتم K-نزدیک‌ترین همسایه به‌منظور طبقه‌بندی اسناد است؛ اما این طبقه‌بند با مشکلاتی از جمله، میزان حافظهٔ مصرفی زیاد و حساس‌بودن به داده‌های نویز مواجه اسـت (گارسـیا، دراک، کـانو و هـررا، ۲۰۱۲). در روش پیشنهادی، اول به‌منظور کاهش میـزان حافظـهٔ مصـرفی و حـذف نمونـه‌هـای نویزی، از معیار سادهٔ شباهت اقلیدسی استفاده شد؛ اما به‌منظور رفع عیوب طبقه‌بند نزدیک‌تـرین همسایه، روش‌هایی برای انتخاب نمونه‌های مناسب از میان نمونه‌های آموزش ارائه شـده اسـت که این روش‌ها به دو دستهٔ انتخاب نمونه‌های اولیه[1] و تولید نمونه‌های اولیه[2] تقسـیم مـی‌شـوند (گارسیا و همکاران، ۲۰۱۲). روش‌های تولید نمونه‌های اولیه، علاوه‌بر انتخاب داده‌ها بـه اصـلاح داده‌های آموزش می‌پردازند و طبقه‌بند نزدیک‌ترین همسایه را بر اسـاس نمونـه‌هـای مصـنوعی ایجادشده آموزش می‌دهند؛ اما روش‌های انتخاب نمونه‌های اولیه، فقط بـه انتخـاب نمونـه‌هـای

1. Prototype Selection
2. Prototype Generation

آموزش بر اساس معیارهای مختلف می‌پردازند و طبقه‌بند نزدیک‌ترین همسایه را بر اساس نمونه‌های انتخاب‌شده آموزش می‌دهند. برخی از روش‌های انتخاب نمونه‌های اولیه از الگوریتم‌های ویرایشی محسوب می‌شوند و هدف این الگوریتم‌ها حذف نمونه‌های نویزی از میان نمونه‌های آموزش است. در اسناد آموزشی، ممکن است بسیاری از خبرها از منابع نامعتبر منتشر شوند یا حتی بسیاری از خبرها تأثیر زیادی در بازار نداشته باشند که وجود این خبرها در مجموعهٔ آموزش، از قدرت پیش‌بینی سیستم پیشنهادی می‌کاهد. بنابراین استفاده از برخی از روش‌های نمونه‌های آموزش، مانند الگوریتم ویرایش نزدیک‌ترین همسایه (گارسیا و همکاران، ۲۰۱۵) می‌تواند بسیاری از نمونه‌های بی‌اهمیت را از میان نمونه‌های آموزش حذف کند که این موضوع علاوه‌بر افزایش کارایی الگوریتم K- نزدیک‌ترین همسایه، دارای دو مزیت اصلی است: ۱. صرفه‌جویی در حافظهٔ مصرفی و ۲. کاهش حجم محاسبات.

۱. ویرایش مجموعهٔ آموزش با استفاده از الگوریتم‌های انتخاب نمونه‌های اولیه
۲. انتخاب ویژگی‌های غیرصفر نمونه‌های آزمایش به‌عنوان ویژگی‌های برتر
۳. استفاده از مدل K- نزدیک‌ترین همسایه برای پیش‌بینی نمونه‌های آزمایش

**شکل ۳. مراحل طبقه‌بندی روش پیشنهادی در فاز دوم**

## فاز سوم

روش پیشنهادی در فاز سوم، گام پیش‌پردازش مشابهی با روش‌های پیشنهادیِ فاز اول و دوم دارد (شکل ۱)؛ اما گام طبقه‌بندی آن مطابق شکل ۴، ترکیبی از گام‌های طبقه‌بندی فازهای اول و دوم است. در فاز طبقه‌بندی، ابتدا با استفاده از الگوریتم ویرایشی نزدیک‌ترین همسایه[1] به‌عنوان الگوریتمی کارا (نیکنام و نیک‌نفس، ۲۰۱۵)، به ویرایش مجموعهٔ آموزش و حذف نویز از آن پرداخته شده است. سپس همانند فاز اول، از معیار شباهت فاصلهٔ اقلیدسی به‌منظور ویرایش مجدد مجموعهٔ آموزش و انتخاب نمونه‌های نزدیک به نمونه‌های آزمایش استفاده شده است. برخلاف گام طبقه‌بندی روش پیشنهادی در فاز اول که ابتدا به انتخاب ویژگی پرداخته شده بود و ویژگی‌های صفر نمونه‌های آزمایش با حذف‌شدن از فضای ویژگی هیچ تأثیری در انتخاب نمونه‌های نزدیک به نمونه‌های آزمایش نداشتند، در روش پیشنهادی سوم این ویژگی‌ها حذف نشدند؛ بلکه ابتدا با حضور ویژگی‌های صفر، تعدادی از نمونه‌های آموزش نزدیک به نمونه‌های آزمایش انتخاب شدند تا تأثیر این ویژگی‌ها در فرایند انتخاب نمونه‌های آموزش اعمال شود. پس

1. Edited Nearest Neighbor

از آن، تأثیر انتخاب ویژگی‌های غیرصفر نمونه‌های آزمایش در طبقه‌بندی تیترهای خبری بررسی شد.

۱. ویرایش مجموعهٔ آموزش بر اساس روش نزدیک‌ترین همسایهٔ ویرایشی
۲. انتخاب M نمونهٔ نزدیک به هر دادهٔ آزمایش و ایجاد نمونهٔ آموزش
۳. انتخاب ویژگی‌های غیرصفر نمونه‌های آزمایش به‌عنوان ویژگی‌های برتر
۴. آموزش مدل بر اساس فضای ویژگی به‌دست‌آمده
۵. تخصیص برچسب مناسب به هر نمونهٔ آزمایش

**شکل ٤. مراحل گام طبقه‌بندی روش پیشنهادی در فاز سوم**

## یافته‌های پژوهش

### مجموعهٔ داده‌ها

مجموعهٔ دادهٔ استفاده‌شده در این پژوهش، شامل تیترهای خبری سال‌های ۲۰۰۸ تا ۲۰۱۲ است که ۶۹۰۴ رکورد و گروه خبری را دربردارد و این گروه‌های خبری بر اساس زمان انتشار مرتب شده‌اند. هر گروه خبری موجود در این مجموعهٔ داده، شامل خبرهای منتشرشده در بازه‌های دوساعته است و برای پیش‌بینی هر گروه خبری بر اساس زمان انتشار آن، از همهٔ گروه‌های خبری که قبل از خبر مدنظر منتشر شده‌اند، استفاده می‌شود (نصیرطوسی و همکاران، ۲۰۱۵) و این دیتاست در آدرس https://sites. google.com/ site / bigdatasetmining /Projects/ textmining (آقابزرگی، ۲۰۱۵) در دسترس است.

### نحوهٔ ارزیابی

به‌منظور ارزیابی نتایج به‌دست‌آمده از روش پیشنهادی، از معیارهای متداول ارزیابی روش‌های بازیابی اطلاعات مانند Precision، Recall و Accuracy بهره برده شده است که این معیارها با توجه به جدول ۱ به‌کمک رابطه‌های ۲، ۳ و ۴ محاسبه می‌شوند.

به‌طور مثال مقادیر جدول ۱ برای کلاس C عبارت‌اند از:

TP: تعداد نمونه‌های متعلق به کلاس C که بازیابی شده‌اند.

TN: تعداد نمونه‌هایی که متعلق به کلاس C نیستند و بازیابی نشده‌اند.

FP: تعداد نمونه‌هایی که متعلق به کلاس C نیستند و بازیابی شده‌اند.

FN: تعداد نمونه‌های متعلق به کلاس C که بازیابی نشده‌اند.

$$Accuracy = (TP + TN)/(TP + TN + FP + FN)$$ رابطهٔ ۲)

$$Precision(P) = TP/(TP + FP)$$ رابطهٔ ۳)

$$Recall(P) = TP/(TP + FN)$$ رابطهٔ ۴)

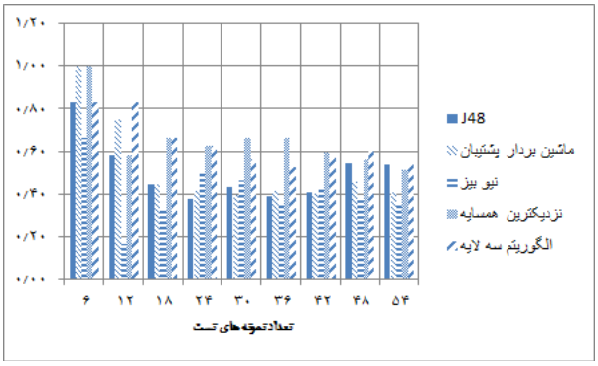
**جدول ۱. وضعیت نمونه‌های بازیابی‌شده**

| | Relevant | No relevant |
|---|---|---|
| **Retrieved** | true positives (TP) | false positives (FP) |
| **Not retrieved** | false negatives (FN) | true negatives (TN) |

منبع: مانینگ، رغوان و اسچوتز، ۲۰۰۸

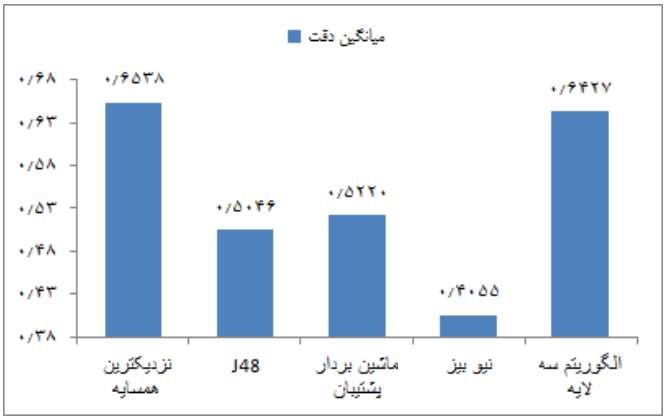
## نتایج فاز اول

روش پیشنهادی در فاز اول توسط چهار طبقه‌بند K- نزدیک‌ترین همسایه، نایو بیز، J48 و ماشین بردار پشتیبان (استفاده از مدل‌های موجود در نرم‌افزار R) ارزیابی می‌شود که در هریک از این ارزیابی‌ها، مقدار M برابر ۴۰ در نظر گرفته شده است؛ این مقدار طی آزمایش‌های متعدد با تغییر مقدار M از ۴۰ تا ۱۵۰ به‌دست آمد. مقدار K برای طبقه‌بند K-نزدیک‌ترین همسایه برابر یک در نظر گرفته شده است و کرنل استفاده‌شده برای ماشین بردار پشتیبان از نوع خطی است. شکل ۵ نتایج روش پیشنهادی در فاز اول را با استفاده از طبقه‌بندهای مختلف در مقایسه با الگوریتم سه‌لایه (نصیرطوسی و همکاران، ۲۰۱۵) نشان می‌دهد. همان‌گونه که مشاهده می‌شود، طبقه‌بند نزدیک‌ترین همسایه در اغلب موارد نسبت به سایر طبقه‌بندها و الگوریتم سه‌لایه برتر است.

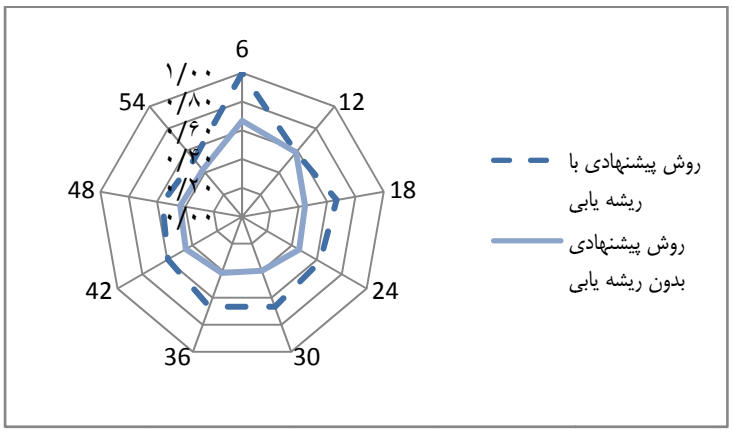



**شکل ۵. مقایسهٔ دقت طبقه‌بندهای مختلف در فاز اول**

شکل ۶ مقایسهٔ روش پیشنهادی در فاز اول را از نظر میانگین با الگوریتم سه‌لایه نشان می‌دهد. همان‌گونه که در این شکل مشاهده می‌شود، کارایی روش پیشنهادی با طبقه‌بند نزدیک‌ترین همسایه، از الگوریتم سه‌لایه به‌طور میانگین بیشتر است. شکل ۷ تأثیر استفاده از الگوریتم ریشه‌یابی بر کارایی روش پیشنهادی در فاز اول را به نمایش گذاشته است. در این شکل نیز کارایی روش پیشنهادی در حالتی که از الگوریتم ریشه‌یابی استفاده‌شده در همهٔ حالات نسبت به حالتی که از الگوریتم ریشه‌یابی استفاده نشده، برتر است.



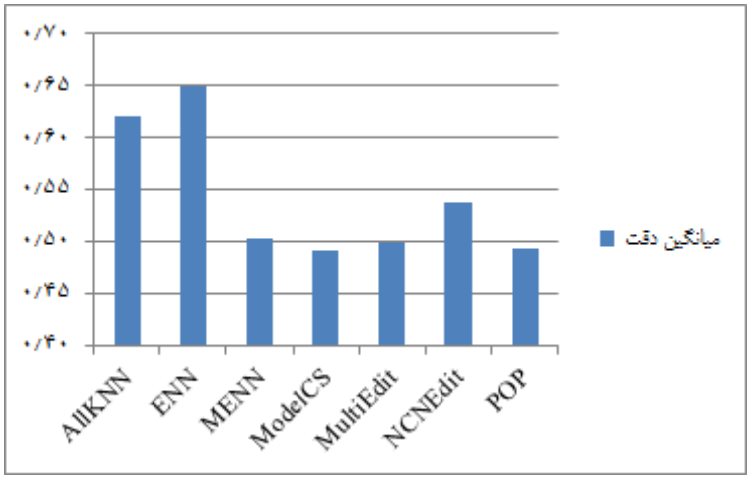

**شکل ٦. مقایسهٔ میانگین دقت روش پیشنهادی در فاز اول با الگوریتم سه‌لایه و طبقه‌بندهای متفاوت**




**شکل ۷. مقایسهٔ دقت روش پیشنهادی در حالت‌های استفاده و عدم استفاده از ریشه‌یابی بر اساس تعداد نمونه‌ها**

## نتایج فاز دوم

شکل ۸ نتایج کارایی روش پیشنهادی در فاز دوم را به‌صورت میانگین برای روش‌های مختلف انتخاب نمونه‌های اولیه نشان می‌دهد. در این تحقیق از روش‌های انتخاب نمونهٔ اولیه موجود در نرم‌افزار KEEL استفاده شده است. همان‌طور که مشاهده می‌شود، با استفاده از الگوریتم نزدیک‌ترین همسایهٔ ویرایشی برای ویرایش مجموعهٔ آموزش، کارایی الگوریتم نسبت به سایر روش‌ها بهتر می‌شود.




**شکل ۸. مقایسهٔ میانگین نتایج روش پیشنهادی در فاز دوم با استفاده از الگوریتم‌های انتخاب نمونهٔ اولیهٔ متفاوت**

## نتایج فاز سوم

در روش پیشنهادشدهٔ فاز اول پژوهش، به‌منظور تشکیل مجموعهٔ آموزش، از M دادهٔ نزدیک به هر نمونهٔ آزمایش بر اساس فاصلهٔ اقلیدسی استفاده شده است؛ اما در این روش، نمونه‌های نویزی از مجموعهٔ آموزش حذف نشده بودند و ممکن بود بین M نمونهٔ جمع‌آوری‌شده، دادهٔ نویز وجود داشته باشد. به همین دلیل در گام طبقه‌بندی، پیش از انتخاب M نمونهٔ نزدیک به هر نمونهٔ آزمایش، ابتدا مجموعهٔ آموزش به‌کمک الگوریتم ویرایشی نزدیک‌ترین همسایه ویرایش شد. ویرایش مجموعهٔ آموزش احتمال حضور نمونه‌های نویزی یا بی‌اهمیت را در M نمونهٔ نزدیک به هر نمونهٔ آزمایشی کاهش می‌دهد. بنابراین هدف از فاز سوم ارائهٔ روش ترکیبی از

فازهای اول و دوم است تا کارایی الگوریتم در همهٔ آزمایش‌ها به‌طور مناسبی افزایش یابد. در جدول‌های ۲ تا ۶ کارایی روش پیشنهادی در فاز سوم درج شده است و شکل ۹ میانگین نتایج به‌دست‌آمده از روش پیشنهادیِ فاز سوم را به ازای مقادیر متفاوت M نشان می‌دهد که با انتخاب M برابر ۱۰۰، نتایج روش پیشنهادی بهبود یافته و نسبت به سایر حالت‌ها میانگین بیشتری دارد.

**جدول ۲. نتایج روش پیشنهادی در فاز سوم به ازای مقدار M برابر ۵۰**

| تعداد آزمایش | دقت | Precision(p) | Precision(N) | Recall(p) | Recall(N) |
|---|---|---|---|---|---|
| ۶ | ۰/۸۳۳۳ | ۰ | ۱ | ۰ | ۰/۸۳۳۳ |
| ۱۲ | ۰/۶۶۶۷ | ۰/۳۳۳۳ | ۰/۷۷۷۸ | ۰/۳۳۳۳ | ۰/۷۷۷۸ |
| ۱۸ | ۰/۶۱۱۱ | ۰/۶۶۶۷ | ۰/۵۸۳۳ | ۰/۴۴۴۴ | ۰/۷۷۷۸ |
| ۲۴ | ۰/۶۶۶۷ | ۰/۸ | ۰/۵۷۱۴ | ۰/۵۷۱۴ | ۰/۸ |
| ۳۰ | ۰/۶۶۶۷ | ۰/۷۸۵۷ | ۰/۵۶۲۵ | ۰/۶۱۱۱ | ۰/۷۵ |
| ۳۶ | ۰/۶۶۶۷ | ۰/۷۶۴۷ | ۰/۵۷۸۹ | ۰/۶۱۹ | ۰/۷۳۳۳ |
| ۴۲ | ۰/۶۴۲۹ | ۰/۷۶۱۹ | ۰/۵۲۳۸ | ۰/۶۱۵۴ | ۰/۶۸۷۵ |
| ۴۸ | ۰/۶۰۴۲ | ۰/۶۵۲۲ | ۰/۵۶ | ۰/۵۷۶۹ | ۰/۶۳۶۴ |
| ۵۴ | ۰/۶۲۹۶ | ۰/۶۹۲۳ | ۰/۵۷۱۴ | ۰/۶ | ۰/۶۶۶۷ |

**جدول ۳. نتایج روش پیشنهادی در فاز سوم به ازای مقدار M برابر ۱۰۰**

| تعداد آزمایش | دقت | Precision(p) | Precision(N) | Recall(p) | Recall(N) |
|---|---|---|---|---|---|
| ۶ | ۰/۸۳۳۳ | ۰ | ۱ | ۰ | ۰/۸۳۳۳ |
| ۱۲ | ۰/۶۶۶۷ | ۰/۳۳۳۳ | ۰/۷۷۷۸ | ۰/۳۳۳۳ | ۰/۷۷۷۸ |
| ۱۸ | ۰/۷۲۲۲ | ۰/۷۵ | ۰/۷ | ۰/۶۶۶۷ | ۰/۷۷۷۸ |
| ۲۴ | ۰/۷۰۸۳ | ۰/۸۱۸۲ | ۰/۶۱۵۴ | ۰/۶۴۲۹ | ۰/۸ |
| ۳۰ | ۰/۷۳۳۳ | ۰/۸۱۲۵ | ۰/۶۴۲۹ | ۰/۷۲۲۲ | ۰/۷۵ |
| ۳۶ | ۰/۶۹۴۴ | ۰/۷۷۷۸ | ۰/۶۱۱۱ | ۰/۶۶۶۷ | ۰/۷۳۳۳ |
| ۴۲ | ۰/۶۶۶۷ | ۰/۷۷۲۷ | ۰/۵۵ | ۰/۶۵۳۸ | ۰/۶۸۷۵ |
| ۴۸ | ۰/۶۲۵ | ۰/۶۶۶۷ | ۰/۵۸۳۳ | ۰/۶۱۵۴ | ۰/۶۳۶۴ |
| ۵۴ | ۰/۶۲۹۶ | ۰/۶۷۸۶ | ۰/۵۷۶۹ | ۰/۶۳۳۳ | ۰/۶۲۵ |

جدول ٤. نتایج روش پیشنهادی در فاز سوم به ازای مقدار M برابر ۱۵۰

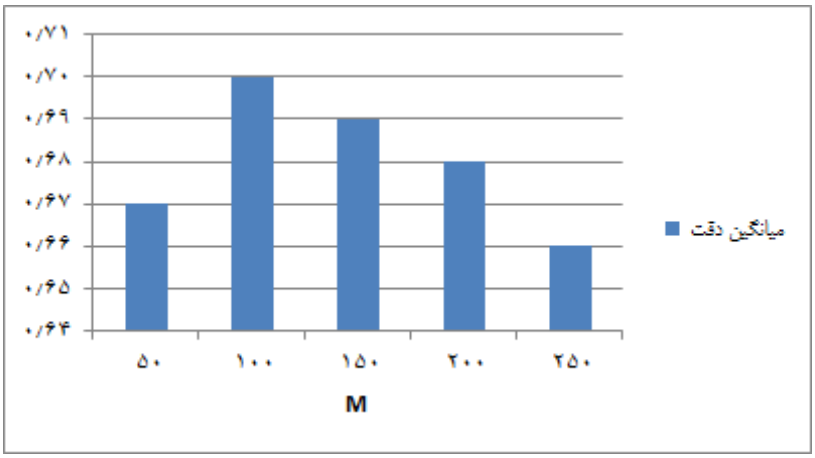
| تعداد آزمایش | دقت | Precision(p) | Precision(N) | Recall(p) | Recall(N) |
|---|---|---|---|---|---|
| ۶ | ۰/۸۳۳۳ | ۰ | ۱ | ۰ | ۰/۸۳۳۳ |
| ۱۲ | ۰/۷۵ | ۰/۵ | ۰/۸۷۵ | ۰/۶۶۶۷ | ۰/۷۷۷۸ |
| ۱۸ | ۰/۷۲۲۲ | ۰/۷۵ | ۰/۷ | ۰/۶۶۶۷ | ۰/۷۷۷۸ |
| ۲۴ | ۰/۶۶۶۷ | ۰/۸ | ۰/۵۷۱۴ | ۰/۵۷۱۴ | ۰/۸ |
| ۳۰ | ۰/۷ | ۰/۸ | ۰/۶ | ۰/۶۶۶۷ | ۰/۷۵ |
| ۳۶ | ۰/۶۶۶۷ | ۰/۷۶۴۷ | ۰/۵۷۸۹ | ۰/۶۱۹ | ۰/۷۳۳۳ |
| ۴۲ | ۰/۶۴۲۹ | ۰/۷۶۱۹ | ۰/۵۲۳۸ | ۰/۶۱۵۴ | ۰/۶۸۷۵ |
| ۴۸ | ۰/۶۲۵ | ۰/۶۶۶۷ | ۰/۵۸۳۳ | ۰/۶۱۵۴ | ۰/۶۳۶۴ |
| ۵۴ | ۰/۶۱۱۱ | ۰/۶۶۶۷ | ۰/۵۵۵۶ | ۰/۶ | ۰/۶۲۵ |

جدول ۵. نتایج روش پیشنهادی در فاز سوم به ازای مقدار M برابر ۲۰۰

| تعداد آزمایش | دقت | Precision(p) | Precision(N) | Recall(p) | Recall(N) |
|---|---|---|---|---|---|
| ۶ | ۰/۸۳۳۳ | ۰ | ۱ | ۰ | ۰/۸۳۳۳ |
| ۱۲ | ۰/۷۵ | ۰/۵ | ۰/۸۷۵ | ۰/۶۶۶۷ | ۰/۷۷۷۸ |
| ۱۸ | ۰/۶۶۶۷ | ۰/۷۱۴۳ | ۰/۶۳۶۴ | ۰/۵۵۵۶ | ۰/۷۷۷۸ |
| ۲۴ | ۰/۷۰۸۳ | ۰/۸۱۸۲ | ۰/۶۱۵۴ | ۰/۶۴۲۹ | ۰/۸ |
| ۳۰ | ۰/۷ | ۰/۸ | ۰/۶ | ۰/۶۶۶۷ | ۰/۷۵ |
| ۳۶ | ۰/۶۱۱۱ | ۰/۷۳۳۳ | ۰/۵۲۳۸ | ۰/۵۲۳۸ | ۰/۷۳۳۳ |
| ۴۲ | ۰/۶۴۲۹ | ۰/۷۶۱۹ | ۰/۵۲۳۸ | ۰/۶۱۵۴ | ۰/۶۸۷۵ |
| ۴۸ | ۰/۶۰۴۲ | ۰/۶۵۲۲ | ۰/۵۶ | ۰/۵۷۶۹ | ۰/۶۳۶۴ |
| ۵۴ | ۰/۵۹۲۶ | ۰/۶۵۳۸ | ۰/۵۳۵۷ | ۰/۵۶۶۷ | ۰/۶۲۵ |

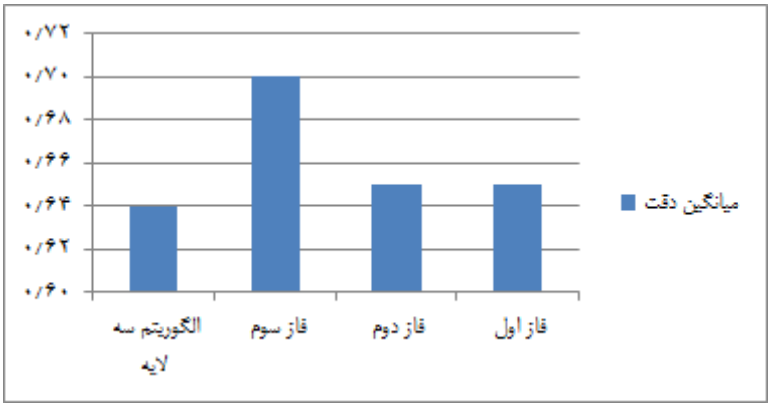
جدول ٦. نتایج روش پیشنهادی در فاز سوم به ازای مقدار M برابر ۲۵۰

| تعداد آزمایش | دقت | Precision(p) | Precision(N) | Recall(p) | Recall(N) |
|---|---|---|---|---|---|
| ۶ | ۰/۸۳۳۳ | ۰ | ۱ | ۰ | ۰/۸۳۳۳ |
| ۱۲ | ۰/۷۵ | ۰/۵ | ۰/۸۷۵ | ۰/۶۶۶۷ | ۰/۷۷۷۸ |
| ۱۸ | ۰/۶۶۶۷ | ۰/۷۱۴۳ | ۰/۶۳۶۴ | ۰/۵۵۵۶ | ۰/۷۷۷۸ |
| ۲۴ | ۰/۶۲۵ | ۰/۷۷۷۸ | ۰/۵۳۳۳ | ۰/۵ | ۰/۸ |
| ۳۰ | ۰/۶۶۶۷ | ۰/۷۸۵۷ | ۰/۵۶۲۵ | ۰/۶۱۱۱ | ۰/۷۵ |
| ۳۶ | ۰/۶۱۱۱ | ۰/۷۳۳۳ | ۰/۵۲۳۸ | ۰/۵۲۳۸ | ۰/۷۳۳۳ |
| ۴۲ | ۰/۶۱۹ | ۰/۷۵ | ۰/۵ | ۰/۵۷۶۹ | ۰/۶۸۷۵ |
| ۴۸ | ۰/۵۸۳۳ | ۰/۶۳۶۴ | ۰/۵۳۸۵ | ۰/۵۳۸۵ | ۰/۶۳۶۴ |
| ۵۴ | ۰/۵۵۵۶ | ۰/۶۱۵۴ | ۰/۵ | ۰/۵۳۳۳ | ۰/۵۸۳۳ |

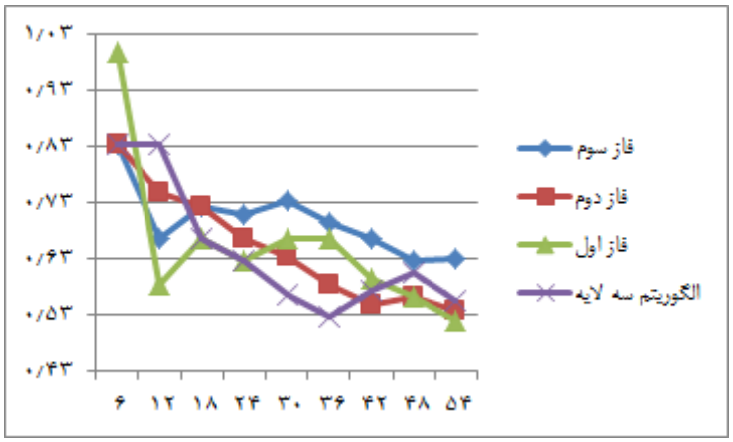



**شکل ۹. میانگین نتایج روش پیشنهادی در فاز سوم به ازای مقادیر متفاوت M**

با توجه به شکل ۹، مشاهده می‌شود میانگین نتایج روش در فاز سوم که از ترکیب دو فاز اول و دوم به‌دست آمده است با داشتن کارایی این دو فاز به‌صورت همزمان، در مقایسه با الگوریتم سه‌لایه و فازهای اول و دوم به برتری دست یافته است؛ شکل ۱۰ نشان‌دهندهٔ برتری روش پیشنهادی در فاز سوم است. همچنین در شکل ۱۰ می‌توان برتری فازهای اول و دوم را نسبت به الگوریتم سه‌لایه مشاهده کرد؛ به‌طوری که روش پیشنهادی در فاز سوم به حداکثر کارایی خود رسیده است. شکل ۱۱ نتایج آزمایش‌های روش پیشنهادی در سه فاز را در مقایسه با الگوریتم سه‌لایه نشان می‌دهد.




**شکل ۱۰. مقایسهٔ میانگین نتایج روش پیشنهادی در سه فاز و الگوریتم سه‌لایه**




**شکل ۱۱. مقایسهٔ روش پیشنهادی در سه فاز مختلف با الگوریتم سه‌لایه**

## نتیجه‌گیری و پیشنهادها

امروزه حجم کثیری از داده‌های متنی برای پردازش موجود است و هرساله بر میزان این داده‌ها افزوده می‌شود. از چالش‌های مهم در پردازش داده‌های متنی بالا بودن ابعاد فضای ویژگی و حجیم‌بودن داده‌های آموزش است. بنابراین در این پژوهش با اعمال روش انتخاب ویژگی مناسب بر اساس ویژگی‌های نمونه‌های هدف و ویرایش مجموعهٔ آموزش از طریق روش‌های انتخاب نمونهٔ اولیه، راه‌حلی به‌منظور استفاده از تیترهای خبری در پیش‌بینی روند بازار ارز فارکس (مطالعهٔ موردی: ارز یورو) ارائه شده است. به همین منظور، روش پیشنهادی در این مقاله طی سه فاز طراحی شد و هدف هر فاز، بهبود فاز پیش‌بینی مرحلهٔ قبل بود تا درنهایت بتوان کارایی روش پیشنهادی را به وضعیت مطلوب رساند. جدول ۷ مراحل هر سه فاز از الگوریتم و دستاوردهای هر فاز را به‌طور خلاصه نشان می‌دهد. در فاز اول، بهره‌گیری از روش‌های ریشه‌یابی و ویرایش مجموعهٔ آموزش بر مبنای فاصلهٔ اقلیدسی، کارایی مدل پیشنهادی را نسبت به الگوریتم سه‌لایه بهبود داد. در فاز دوم، به‌منظور حذف نویز در مجموعهٔ آموزش، از روش‌های انتخاب نمونهٔ اولیه استفاده شد که کارایی مدل را در نمونه‌های تست کم به‌طور مناسبی بهبود داد و برتری روش‌های انتخاب نمونه‌های آموزش بر فاصلهٔ اقلیدسی در ویرایش مجموعهٔ آموزش را نشان داد. در فاز سوم با ترکیب فازهای اول و دوم و همچنین اثر دادن ویژگی‌های صفر در ویرایش مجموعهٔ آموزش، کارایی روش پیشنهادی نسبت به الگوریتم سه‌لایه و فازهای اول و دوم بهبود شایان توجه‌ای داشت.

با در نظر گرفتن داده‌های بنیادین در روش پیشنهادی، می‌توان این روش را برای پیش‌بینی شاخص بورس در ایران به‌کار برد. همچنین می‌توان با استفاده از الگوریتم ازدحام ذرات، نوعی روش انتخاب ویژگی مناسب طراحی کرد و کارایی روش پیشنهادی را افزایش داد.

**جدول ۷. خلاصهٔ روش پیشنهادشده و نتایج آن**

| فاز تحقیق | ایده‌های اصلی | روش‌های استفاده‌شده | نتایج |
|---|---|---|---|
| فاز اول | کاهش ابعاد فضای ویژگی، کاهش محاسبات، کاهش نمونه‌های آموزش | الگوریتم ریشه‌یابی پورتر، استفاده از ویژگی‌های غیرصفر نمونه‌های آزمایش به‌عنوان ویژگی‌های برتر، حذف نمونه‌های دور از نمونه‌های آزمایش بر اساس فاصله اقلیدسی | افزایش دقت پیش‌بینی، وابسته‌نبودن مدل آموزش به یک داده آزمایشی، کاهش محاسبات، کاهش نمونه‌ها موجب کمترشدن حافظهٔ مصرفی شده است. |
| فاز دوم | کاهش نمونه‌های آموزش | بهره‌بردن از روش‌های انتخاب نمونه‌های اولیه برای حذف نمونه‌های نویزی | افزایش دقت پیش‌بینی نسبت به فاز اول |
| فاز سوم | ترکیب فازهای اول و دوم | ویرایش مجموعهٔ آموزش با الگوریتم ENN و ایجاد فضای آموزش با انتخاب نمونه‌های نزدیک به نمونه‌های آزمایش | افزایش کارایی الگوریتم نسبت به فازهای اول و دوم و الگوریتم سه‌لایه |

## *References*


Aggarwal, C. C. & Zhai, C. (2012). Mining text data. *Springer Science & Business Media*. ISBN: 978-1-4614-3222-7 (Print) 978-1-4614-3223-4. (Online)

Aghabozorgi, S. (2016). *Big Data Mining*. Retrieved January 09, 2016, from https://sites.google.com/site/bigdatasetmining/Projects/textmining.

de Fortuny, E. J., De Smedt, T., Martens, D. & Daelemans, W. (2014). Evaluating and understanding text-based stock price prediction models. *Information Processing & Management,* 50(2): 426-441.

Garcia, S., Derrac, J., Cano, J. R., & Herrera, F. (2012). Prototype selection for nearest neighbor classification: Taxonomy and empirical study. *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* 34(3), 417-435.

García, S., Luengo, J. & Herrera, F. (2015). *Data preprocessing in data mining*. Switzerland: Springer.

Hagenau, M., Liebmann ,M., & Neumann, D. (2013). Automated news reading: Stock price prediction based on financial news using context-capturing features. *Decision Support Systems,* 55(3): 685-697.

Huang, C.-J., Liao, J.-J., Yang, D.-X., Chang, T.-Y., & Luo, Y.-C. (2010). Realization of a news dissemination agent based on weighted association rules and text mining techniques. *Expert Systems with Applications,* 37(9): 6409-6413.

Im, T. L., San, P. W., On, C. K., Alfred, R., & Anthony, P. (2014). Impact of Financial News Headline and Content to Market Sentiment. *International Journal of Machine Learning and Computing,* 4(3): 237-242.

Jivani, A. G. (2011). A comparative study of stemming algorithms. *Int. J. Computer Technology and Applications, 2*(6): 1930-1938.

Kim, Y., Jeong, S. R., & Ghani, I. (2014). Text opinion mining to analyze news for stock market prediction. *Int. J. Advances in Soft Computing and its Applications,* 6(1): 1-13.

Manning, C. D., Raghavan, P. & Schütze, H. (2008). *Introduction to information retrieval* (Vol. 1): Cambridge university press Cambridge.

Murty, M. N. & Devi, V. S. (2011). *Pattern recognition: An algorithmic approach*. India : Springer Science & Business Media.

Nassirtoussi, A. K., Aghabozorgi, S., Wah, T. Y. & Ngo, D. C. L. (2014). Text mining for market prediction: A systematic review. *Expert Systems with Applications,* 41(16): 7653-7670.

Nassirtoussi, A. K., Aghabozorgi, S., Wah, T. Y. & Ngo, D. C. L. (2015). Text mining of news-headlines for FOREX market prediction: A Multi-layer Dimension Reduction Algorithm with semantics and sentiment. *Expert Systems with Applications,* 42(1): 306-324.

Nikfarjam, A., Emadzadeh, E. & Muthaiyah, S. (2010). Text mining approaches for stock market prediction. Paper presented at the Computer and Automation Engineering (ICCAE). *2010 The 2nd International Conference on*. 26-28 Feb. 2010: Singapore

Niknam, F. & Niknafs, A. (2015). *Using Training Set Selection Methods to Improve Text Mining on Market Prediction via News Headlines*. Paper presented at the The International Congress on Technology, Communication and Knowledge, Mashhad, Iran. (*in Persian*)

Passini ,C., Luiza, M., Estébanez, K., Figueredo, G., Ebecken, F. & Nelson, F. (2013). A strategy for training set selection in text classification problems. *International Journal of Advanced Computer Science & Applications,* 4(6): 54-60.

Weiss, S. M., Indurkhya, N. & Zhang, T. (2010). *Fundamentals of predictive text mining*: Springer Science & Business Media.

Yang, J., Liu, Y., Zhu, X., Liu, Z., & Zhang, X. (2012). A new feature selection based on comprehensive measurement both in inter-category and intra-category for text categorization. *Information Processing & Management,* 48(4): 741-754.