4th International Conference on Web Research
April 25-26, 2018; Tehran, IRAN
ICWR2018
کنفرانس بین المللی وب پژوهی
چهارمین
۵ و ۶ اردیبهشت ۱۳۹۷؛ تهران، ایران

# Improving Text Mining with Featured Word Selection

[1] M.Amin Abolghasemi, [2]Saeedeh Momtazi

[1] Master Student of Artificial Intelligence, Amirkabir University of Technology, Tehran, Iran
amin.a222@yahoo.com

[2] Assistant Professor of Artificial Intelligence, Amirkabir University of Technology, Tehran, Iran
momtazi@aut.ac.ir

## Abstract

Text mining is one of the main tasks in web research that aims at classification or clustering available texts in the web for different applications, such as news analysis and social network analysis. Since a very large amount of textual data is available on the Web, reducing the dimension of data using feature extraction techniques plays an important role in improving the efficiency and effectiveness of the text mining algorithms. Various techniques have been proposed in machine learning tasks that can also be applied in the text mining domain. In this paper we study the available techniques and compare their impact on improving Persian text classification performance. Our experimental results on Hamshahri corpus shows that using an appropriate feature selection technique can improve the classification f-measure from 88.12% to 93.07%.

**Keywords:** Web Mining, Text Mining, Text Classification, Feature Selection

4th International Conference on Web Research
April 25-26, 2018; Tehran, IRAN
ICWR2018
کنفرانس بین المللی وب پژوهی
چهارمین
۵ و ۶ اردیبهشت ۱۳۹۷؛ تهران، ایران

# بهبود متن‌کاوی با انتخاب کلمات ویژگی

محمدامین ابوالقاسمی[۱]، سعیده ممتازی[۲]

[۱] دانشجوی کارشناسی ارشد ، گروه هوش مصنوعی دانشکده مهندسی کامپیوتر و فناوری اطلاعات، دانشگاه صنعتی امیرکبیر، تهران، ایران
amin.a222@yahoo.com

[۲] استادیار، گروه هوش مصنوعی دانشکده مهندسی کامپیوتر و فناوری اطلاعات، دانشگاه صنعتی امیرکبیر، تهران، ایران
momtazi@aut.ac.ir

## چکیده

**متن‌کاوی یکی از فعالیت‌های اصلی در حوزه وب‌پژوهی محسوب می‌گردد که هدف آن دسته‌بندی یا خوشه‌بندی متون موجود در وب برای کاربردهای مختلف از جمله تحلیل خبر، تحلیل شبکه‌های اجتماعی و ... می‌باشد. با توجه به بالا بودن حجم دادگان موجود در وب برای پردازش‌های متنی، کاهش ابعاد دادگان با کمک روش‌های استخراج ویژگی نقش مهمی را در بهبود کیفیت متن‌کاوی و همین‌طور بهینه‌سازی زمان اجرا ایفا می‌نماید. روش‌های متنوعی برای استخراج ویژگی در الگوریتم‌های یادگیری ماشین ارائه شده است که قابلیت کاربردی‌سازی در حوزه متن‌کاوی را دارند. در مقاله حاضر به بررسی الگوریتم‌های موجود در این زمینه پرداخته می‌شود و نتایج حاصل از این الگوریتم‌ها در استخراج کلمات ویژگی متون فارسی مقایسه می‌گردد. همچنین تاثیر به‌کارگیری انتخاب ویژگی در دسته‌بندی متون فارسی مورد تحلیل قرار می‌گیرد. نتایج به‌دست آمده در آزمایش‌ها برروی پیکره همشهری فارسی نشان می‌دهد با کمک روش مناسب انتخاب ویژگی می‌توان نتایج دسته‌بندی متون فارسی را از ۸۸.۱۲٪ در معیار اف به ۹۳.۰۷٪ افزایش داد.**

## کلمات کلیدی

**وب‌کاوی، متن‌کاوی، دسته‌بندی متون، انتخاب ویژگی**

## ۱- مقدمه

بازیابی اطلاعات و متن‌کاوی (اعم از دسته‌بندی و یا خوشه‌بندی متون) جز اساسی‌ترین فعالیت‌ها در حوزه وب‌کاوی می‌باشند. هدف از بازیابی اطلاعات یافتن مستندات مرتبط با کوئری ارائه شده توسط کاربر می‌باشد. وظیفه دسته‌بندی متون تخصیص هر سند به یک کلاس مشخص و از پیش تعیین شده می‌باشد که با کمک الگوریتم‌های یادگیری بانظارت انجام می‌پذیرد[1,6]. خوشه‌بندی متون نیز سعی در تقسیم‌بندی مستندات دارد با این تفاوت که در آن نوع خوشه‌ها از پیش تعیین شده نیستند و این تقسیم‌بندی تنها براساس شباهت میان مستندات و با کمک الگوریتم‌های یادگیری بدون نظارت صورت می‌گیرد.

در تمامی فعالیت‌های فوق حجم بسیار بالایی از دادگان متنی مورد بازیابی، دسته‌بندی، و یا خوشه‌بندی قرار می‌گیرد که باعث بالا رفتن پیچیدگی محاسباتی این فعالیت‌ها می‌گردد در حالیکه در اکثریت کاربردهای آن‌ها در وب‌کاوی سرعت بالای پردازش نقش کلیدی در میزان رضایت کاربر دارد. برای رفع این مشکل ایده استخراج ویژگی از مستندات پیشنهاد شده‌است است که باعث کاهش حجم فضای برداری ایجاد شده برای هر سند می‌گردد. مهمترین مزیت انتخاب ویژگی ، کاهش ابعاد داده و متعاقبا افزایش سرعت اجرای الگوریتم‌ها می‌باشد. به‌علاوه این کار باعث می شود تا میزان استعداد بیش‌برازش الگوریتم‌های مورد استفاده در متن‌کاوی کاهش یابد . در این راستا در فعالیت‌های مختلف حوزه یادگیری ماشین به منظور استخراج ویژگی‌های مناسب و بهینه معیارهای گوناگونی ارائه شده است[3]. اگرچه تمام این روش‌ها می‌توانند نقش موثر در استخراج ویژگی‌های کلیدی داده داشته باشند اما بررسی عمیق‌تری مورد نیاز است تا بتوان از میان روش‌های متنوع موجود بهترین گزینه را برای کاربردهای مختلف انتخاب نمود.

در مقاله حاضر سه نوع از این معیارها که در حوزه‌های مختلف یادگیری ماشین مورد استفاده قرار گرفته است برای متن‌کاوی و پردازش زبان طبیعی مورد بررسی قرار می‌گیرد.الگوریتم‌های مدنظر مقاله حاضر اطلاعات متقابل[1][8]، بهره اطلاعات[2][7] و مربع کای[3][4,5] می‌باشند. در این راستا ابتدا نحوه کاربردی کردن آن‌ها در حوزه متن‌کاوی معرفی می‌شود. سپس تفاوت خروجی این معیارها در استخراج ویژگی از متون فارسی بررسی می‌گردد و در نهایت تاثیر استفاده از این معیارها در نتایج دسته‌بندی متون مطالعه می‌شود.

## ۲- معیارهای استخراج ویژگی

### ۲-۱- اطلاعات متقابل

یکی از متداول‌ترین معیارهای انتخاب ویژگی اطلاعات متقابل است. این معیار بررسی می کند که چه مقدار اطلاعات با وجود یا عدم وجود یک کلمه برای مدل یادگیری ما منجر به دسته‌بندی بهتر روی یک کلاس یا دسته می شود. رابطه (۱) روش محاسبه اطلاعات متقابل را نمایش می‌دهد.

$$MI(w,c_i)=\log\frac{P(w,c_i)}{P(w)P(c_i)} \qquad (1)$$

در رابطه فوق $w$ و $c_i$ به ترتیب بیان‌گر کلمه و کلاس iام می‌باشند. $MI(w,c_i)$ اطلاعات متقابل آن‌ها را ارائه می‌نماید. احتمالات مورد نیاز به صورت زیر محاسبه می‌گردد.

$$P(c_i)=\frac{N_i}{N} \quad P(w)=\frac{N_w}{N} \quad P(w,c_i)=\frac{N_{iw}}{N}.$$

که در آن $N$ برابر با تعداد کل اسناد، $N_i$ برابر با تعداد کل اسناد موجود در کلاس $c_i$، $N_w$ برابر با تعداد کل اسنادی که شامل کلمه‌ی $w$ هستند و $N_{iw}$ برابر تعداد اسناد کلاس $c_i$ که شامل کلمه $w$ می‌باشد است.
پس از به‌دست آوردن مقدار بالا به ازای تمامی کلمات و کلاس‌ها، آن‌ها را با استفاده از رابطه‌ی زیر رتبه‌بندی کرده و بهترین آن‌ها به عنوان ویژگی انتخاب می‌شوند.

$$MI(w)=\sum_{i=1}^{k}MI(w,c_i) \qquad (2)$$

### ۲-۲- بهره اطلاعات

این معیار اختلاف آنتروپی را در دو وضعیت وجود یک ویژگی و عدم وجود یک ویژگی بیان می کند. در واقع بهره اطلاعات این مسئله که با وجود یک ویژگی مقدار آنتروپی چه مقدار کاهش می یابد را مورد ارزیابی قرار می دهد و سپس از بین ویژگی ها، آن‌هایی که بهره اطلاعات آن‌ها بالاتر است انتخاب می‌شوند. رابطه (۳) روش محاسبه بهره اطلاعات را نمایش می‌دهد.

$$\begin{aligned}IG(w)&=-\sum_{i=1}^{K}P(c_i)\log P(c_i)\\&+P(w)\sum_{i=1}^{K}P(c_i|w)\log P(c_i|w)\\&+P(\overline{w})\sum_{i=1}^{K}P(c_i|\overline{w})\log P(c_i|\overline{w}) \qquad (3)\end{aligned}$$

در رابطه (۳) $w$ و $c_i$ به ترتیب بیان‌گر کلمه و کلاس iام می‌باشند. همچنین احتمالات موجود در این رابطه به صورت زیر محاسبه می‌شوند.

$$P(c_i)=\frac{N_i}{N} \quad P(w)=\frac{N_w}{N} \quad P(\overline{w})=\frac{N_{\overline{w}}}{N}$$
$$P(c_i|w)=\frac{N_{iw}}{N_w} \quad P(c_i|\overline{w})=\frac{N_{i\overline{w}}}{N_{\overline{w}}}$$

در روابط بالا $N$ برابر با تعداد کل اسناد، $N_i$ برابر با تعداد کل اسناد موجود در کلاس $c_i$، $N_w$ برابر با تعداد کل اسنادی که شامل کلمه‌ی $w$ هستند، $N_{\overline{w}}$ برابر با تعداد کل اسنادی که شامل کلمه‌ی $w$ نیستند، $N_{iw}$ برابر با تعداد کل اسنادی که در کلاس $c_i$ بوده و شامل کلمه‌ی $w$ نیز می‌باشند و در نهایت $N_{i\overline{w}}$ تعداد کل اسنادی که در کلاس $c_i$ بوده ولی شامل کلمه‌ی $w$ نمی‌باشند را بازنمایی می‌کند.

### ۲-۳- مربع کای

این معیار در مباحث آماری کاربرد بسیار متداولی دارد و هدف آن اندازه‌گیری میزان استقلال دو رویداد از هم می‌باشد. به‌طور خاص در حوزه متن‌کاوی این معیار کمک می‌کند تا بتوانیم اندازه بگیریم که به چه مقدار رخداد یک کلمه خاص و رخداد یک کلاس خاص از هم مستقل می‌باشند.
مسئله قابل ذکر در مورد این معیار نادقیق بودن آن به دلیل وجود یک درجه آزادی است. این نادقیق بودن باعث می شود تا برخی ویژگی‌های نامناسب نیز در نظر گرفته شوند که البته آزمایش‌ها نشان داده است که این میزان ویژگی نویزی تاثیر چندانی روی دقت کلی دسته‌بند ندارد. رابطه (۴) روش محاسبه مربع کای را نمایش می‌دهد.

$$\chi^2(w,c_i)=.\frac{N.(N_{iw}N_{\overline{i}\overline{w}}-N_{i\overline{w}}N_{\overline{i}w})^2}{(N_{iw}+N_{i\overline{w}}).(N_{\overline{i}w}+N_{\overline{i}\overline{w}}).(N_{iw}+N_{\overline{i}w}).(N_{i\overline{w}}+N_{\overline{i}\overline{w}})} \qquad (4)$$

پس از به‌دست آوردن مقدار بالا به ازای تمامی کلمات، آن‌ها را با استفاده از رابطه‌ی زیر رتبه‌بندی کرده و بهترین آن‌ها به عنوان ویژگی انتخاب می‌شوند.

$$\chi^2(w)=\sum_{i=1}^{K}P(c_i).\chi^2(w,c_i) \qquad (5)$$

## ۳- دسته‌بندی متون

برای دسته‌بندی متون از الگوریتم بیز ساده[۴] استفاده شده است. رویکرد بیز ساده با در نظر گرفتن استقلال ویژگی‌ها (مدل تک-نگاشتی[۵]) عمل می کند. به‌طوری‌که در محاسبه احتمالات تنها به وجود یا عدم وجود یک ویژگی (کلمه) توجه خواهد داشت. در این روش نحوه تعیین کلاس یک سند به صورت زیر می باشد:

$$c = argmax_{c_i} p(d|c_i)p(c_i) \qquad (6)$$

$\boldsymbol{p(c_i)}$ به عنوان احتمال اولیه[۶] و $\boldsymbol{p(d|c_i)}$ به عنوان احتمال شباهت[۷] معرفی می شود. درمورد احتمال اولیه می‌توان گفت که این احتمال در مواقعی که داده‌ها به صورت یکنواخت در مجموعه داده بین کلاس‌های مختلف توزیع شده‌اند، تاثیر خاصی روی عمل دسته‌بندی ندارند، اما هنگامی که مجموعه داده به صورت غیر یکنواخت است، تاثیر بسیار زیادی خواهند داشت و هر اندازه که میزان آن به ازای یک دسته بیشتر باشد احتمال انتخاب آن دسته افزایش خواهد یافت. احتمال شباهت نیز براساس رابطه (۷) محاسبه می‌شود.

$$P(d|c_i) = \prod_{w \in d} P(w|c_i) \qquad (7)$$

## ٤- نتایج

### ٤-١- انتخاب کلمات ویژگی از متون فارسی

برای بررسی تفاوت میان معیارهای مطرح شده دادگان پیکره همشهری مورد استفاده قرار گرفته‌است. در این پیکره در کنار هر سند موضوع آن سند نیز مشخص شده است[2].

دادگان مورد استفاده برای این پژوهش تعداد ۲۵۸۰ سند است که با یکی از موضوعات ورزش (۵۹۹ سند)، اقتصادی (۵۶۹ سند)، سیاسی (۱۱۷ سند)، ادب و هنر (۵۷۹ سند) و علمی فرهنگی (۵۷۳ سند) و یا محیط زیست (۱۴۳ سند) برچسب‌دهی شده است.

هدف از آزمایش‌های این بخش این است که بر اساس هر یک از سه الگوریتم مذکور یک بردار ویژگی بهینه با حداکثر امتیاز تولید نماییم. جداول ۳-۱ کلمات استخراجی براساس هریک از معیارهای اطلاعات متقابل، بهره اطلاعات و مربع کای را نمایش می‌دهد. برای این‌کار ۲۰ کلمه اول این پیکره براساس هریک از معیارها مرتب شده‌اند. علاوه بر نمایش امتیاز هر کلمه، دامنه‌ای که هریک از کلمات بیشترین امتیاز را در آن کسب کرده مشخص شده‌است.

خروجی برای هر الگوریتم بصورت ۴ ستون زیر خواهد بود:

- $Word_i$ کلمه انتخاب شده برای بردار ویژگی است.
- $Score$ امتیاز کلی این کلمه در الگوریتم انتخاب ویژگی است.
- $domain$ اسم دامنه‌ای است که کلمه $Word_i$ بیشترین امتیاز را در آن بدست آورده است.
- $domain\text{-}score$ امتیاز دامنه‌ای است که کلمه $Word_i$ بیشترین امتیاز را در آن بدست آورده است.

شایان ذکر است با توجه به‌اینکه کلماتی که تعداد تکرار آن‌ها در کل پیکره بسیار کم است نقشی در کلمات کلیدی نخواهند داشت، می‌توان در یک مرحله پیش‌پردازش قبل از استخراج ویژگی‌ها کلماتی است که به ندرت در متن‌ها ظاهر شده اند را حذف نمود. این کار باعث می شود تا سرعت استخراج ویژگی در هنگامی که تعداد داده‌ها زیاد است افزایش

**جدول (۱): ویژگی‌های بدست آمده با روش اطلاعات متقابل**

| Word | score | domain | domain-score |
|---|---|---|---|
| رایانه | ۱۳.۴۴۵۶۰۷ | علمی فرهنگی | ۰.۳۱۹۸۸۹ |
| بورس | ۱۳.۴۰۶۷۵۳ | اقتصادی | ۰.۳۲۳۷۷۷ |
| قهرمانی | ۱۲.۸۰۰۷۵۰ | ورزش | ۰.۳۲۹۲۳۷ |
| گل | ۱۲.۶۹۸۶۲۹ | ورزش | ۰.۲۳۸۲۴۲ |
| جام | ۳.۱۴۴۸۲۴ | ورزش | ۰.۲۸۵۷۵۳ |
| فوتبال | ۳.۱۴۰۷۴۹ | ورزش | ۰.۳۲۴۰۰۹ |
| نرم | ۳.۱۱۷۴۳۹ | علمی فرهنگی | ۰.۲۹۸۲۴۳ |
| ورزشی | ۱.۰۷۱۸۲۶ | ورزش | ۰.۳۲۶۳۹۴ |
| تولید | ۱.۰۵۳۱۶۵ | اقتصادی | ۰.۱۲۶۰۳۱ |
| اینترنت | ۰.۹۶۹۱۹۷ | علمی فرهنگی | ۰.۳۱۱۶۵۴ |
| صنعتی | ۰.۹۵۰۱۹۹ | اقتصادی | ۰.۲۲۲۴۳۱ |
| صنایع | ۰.۹۴۳۸۱۳ | اقتصادی | ۰.۲۵۱۳۰۹ |
| عرضه | ۰.۹۲۷۵۳۰ | اقتصادی | ۰.۱۲۷۵۰۳ |
| مسابقه | ۰.۹۱۷۸۲۳ | ورزش | ۰.۲۵۸۸۳۲ |
| دیدار | ۰.۹۱۴۸۸۱ | ورزش | ۰.۲۹۲۸۱۷ |
| سهام | ۰.۹۰۸۹۴۷ | اقتصادی | ۰.۳۱۱۹۵۸ |
| صفحه | ۰.۸۶۳۰۴۶ | علمی فرهنگی | ۰.۲۱۸۷۵۸ |
| قیمت | ۰.۸۶۲۰۵۸ | اقتصادی | ۰.۲۰۹۱۸۵ |
| دولت | ۰.۸۵۷۴۰۸ | اقتصادی | ۰.۱۱۴۰۷۷ |
| میلیارد | ۰.۸۲۴۲۸۷ | اقتصادی | ۰.۲۴۶۱۵۹ |

**جدول (۲): ویژگی‌های بدست آمده با روش مربع کای**

| Word | score | domain | domain-score |
|---|---|---|---|
| ورزشی | ۶۳۳.۸۸۳۹۴۲ | ورزش | ۵۰۳.۴۸۱۰۹۷ |
| تیم | ۵۳۷.۲۶۲۵۷۶ | ورزش | ۴۲۷.۹۲۳۲۱۸ |
| فوتبال | ۳۴۷.۵۶۲۰۳۲ | ورزش | ۲۷۷.۰۸۵۵۵۶ |
| رایانه | ۲۸۸.۲۷۶۲۳۳ | علمی فرهنگی | ۲۲۷.۷۰۱۳۳۶ |
| گروه | ۲۷۹.۱۲۵۷۲۲ | ورزش | ۲۰۱.۱۸۱۸۰۲ |
| بورس | ۲۵۲.۱۳۴۹۷۵ | اقتصادی | ۱۹۹.۶۶۸۵۴۵ |
| جام | ۲۴۰.۵۴۱۰۴۳ | ورزش | ۱۹۲.۳۷۸۵۰۷ |
| اینترنت | ۲۳۸.۹۰۹۶۷۱ | علمی فرهنگی | ۱۸۷.۵۱۰۸۶۳ |
| قهرمانی | ۲۳۰.۵۸۲۴۹۲ | ورزش | ۱۸۳.۵۷۱۶۳۴ |
| هنری | ۲۲۹.۰۷۶۶۱۹ | ادب و هنر | ۱۸۰.۸۷۸۲۳۲ |
| سرمایه | ۲۰۷.۲۸۵۷۲۱ | اقتصادی | ۱۵۴.۴۸۹۷۶۲ |
| فیلم | ۱۹۷.۳۷۰۹۰۶ | ادب و هنر | ۱۵۲.۵۱۲۲۰۸ |
| بازیکن | ۱۸۷.۰۴۳۹۶۳ | ورزش | ۱۴۹.۲۴۷۵۱۸ |
| بازی | ۱۸۴.۴۶۸۲۰۰ | ورزش | ۱۳۸.۷۳۴۰۲۵ |
| بانک | ۱۸۲.۲۷۸۲۰۵ | اقتصادی | ۱۴۱.۸۱۱۰۱۱ |
| دیدار | ۱۸۲.۰۱۳۳۴۹ | ورزش | ۱۳۹.۴۱۰۷۷۰ |
| سهام | ۱۸۰.۹۶۱۲۷۷ | اقتصادی | ۱۴۲.۶۸۴۲۷۳ |
| سایت | ۱۷۹.۴۷۹۷۱۰ | علمی فرهنگی | ۱۴۱.۷۵۵۴۵۳ |
| گذاری | ۱۷۰.۹۴۸۱۹۵ | اقتصادی | ۱۲۷.۹۰۰۴۳۳ |
| نرم | ۱۷۰.۴۶۸۷۹۴ | علمی فرهنگی | ۱۳۴.۹۹۸۸۳۸ |

**جدول (۳): ویژگی‌های بدست آمده با روش بهره اطلاعات**

| Word | score | domain | domain-score |
|---|---|---|---|
| ورزشی | ۰.۴۱۵۰۳۳ | ورزش | ۰.۲۶۱۲۰۸ |
| تیم | ۰.۳۳۶۶۹۱ | ورزش | ۰.۲۱۳۰۱۲ |
| فوتبال | ۰.۲۰۴۷۷۰ | ورزش | ۰.۱۱۶۸۳۴ |
| گروه | ۰.۲۰۱۷۳۷ | ورزش | ۰.۱۳۸۵۲۰ |
| رایانه | ۰.۱۷۰۶۵۲ | علمی فرهنگی | ۰.۰۹۶۷۴۹ |
| بورس | ۰.۱۵۱۰۵۹ | اقتصادی | ۰.۰۸۲۶۴۴ |
| جام | ۰.۱۴۲۵۷۳ | ورزش | ۰.۰۸۶۱۰۰ |
| اینترنت | ۰.۱۴۰۱۸۲ | علمی فرهنگی | ۰.۰۷۹۵۹۹ |
| قهرمانی | ۰.۱۳۵۹۷۷ | ورزش | ۰.۰۷۱۸۴۱ |
| هنری | ۰.۱۳۱۲۱۶ | ادب و هنر | ۰.۰۷۶۵۰۹ |
| زیست | ۰.۱۳۱۱۷۳ | محیط زیست | ۰.۱۰۲۵۷۹ |
| سرمایه | ۰.۱۲۶۴۶۹ | اقتصادی | ۰.۰۷۸۵۴۳ |
| فیلم | ۰.۱۱۹۱۵۸ | ادب و هنر | ۰.۰۶۶۷۴۸ |
| بازی | ۰.۱۱۴۹۷۵ | ورزش | ۰.۰۶۸۰۱۲ |
| بانک | ۰.۱۰۹۳۶۳ | اقتصادی | ۰.۰۶۳۷۳۱ |
| دیدار | ۰.۱۰۹۲۱۴ | ورزش | ۰.۰۵۸۶۵۸ |
| بازیکن | ۰.۱۰۸۱۱۴ | ورزش | ۰.۰۵۶۹۲۹ |
| سایت | ۰.۱۰۵۸۷۷ | علمی فرهنگی | ۰.۰۵۵۸۳۴ |
| سهام | ۰.۱۰۵۳۱۰ | اقتصادی | ۰.۰۵۹۰۴۵ |
| گذاری | ۰.۱۰۴۷۶۹ | اقتصادی | ۰.۰۶۳۲۵۲ |

یابد و همچنین از لحاظ مصرف حافظه نیز بهبود خواهیم داشت.

همانطور که در نتایج ارائه شده مشاهده می‌شود تمامی کلمات استخراج شده توسط سه روش کلمات مطرح و کلیدی در دامنه خود هستند که می‌توانند به خوبی موضوع آن دامنه را در ذهن خواننده متن تداعی نمایند. مقایسه سه روش نشان می‌دهد هر سه روش در استخراج کلمات موفق بودند و تفاوت زیادی در خروجی آن‌ها مشاهده نمی‌شود.

## ٤-٢- دسته‌بندی متون فارسی با انتخاب ویژگی

در این بخش به بررسی تاثیر انتخاب کلمات ویژگی در کیفیت دسته‌بندی متون فارسی پرداخته می‌شود.

برای این منظور از معیارهای صحت[۸]، فراخوانی[۹] و معیار اف[۱۰] استفاده شده‌است که فرمول‌های آن‌ها در روابط (۸-۱۰) بیان شده‌اند. در مسئله دسته‌بندی مستندات، فراخوانی بیانگر درصد سندهای کلاس i است که درست دسته‌بندی شده‌اند و صحت بیانگر تعداد اسناد درست دست بندی شده کلاس i به نسبت کل سندهایی است که در دسته i قرار گرفته‌اند. از سوی دیگر به دلیل وجود یک مصالحه بین معیارهای صحت و فراخوانی از معیار اف که میانگین وزن‌دار همگن صحت و فراخوانی می باشد استفاده می‌شود.

$$\text{Precision}_i \quad \frac{\#True\ Positives}{\#True\ Positives + \#False\ Positives} \qquad (8)$$

$$\text{Recall}_i \quad \frac{\#True\ Positives}{\#True\ Positives + \#False\ Negatives} \qquad (9)$$

$$\text{F1measure} \quad 2 * \frac{Recall * Precision}{Recall + Precision} \qquad (10)$$

دادگان آموزشی مورد استفاده در این بخش شامل یک پیکره‌ی ۲۵۸۰ سندی و دادگان آزمایشی شامل ۸۶۲ سند می‌باشد. بر این اساس ابتدا الگوریتم بیز ساده بدون استفاده از الگوریتم‌های انتخاب ویژگی روی پیکره‌ی آموزشی پیاده شده‌است. برای این منظور ۲۰۰۰ کلمه پرکاربرد متن را به‌عنوان ویژگی مدنظر قرار داده‌ایم. نتایج این آزمایش در قالب ماتریس آشفتگی[۱۱] در جدول ۴ نمایش داده شده‌است. در این جدول عنوان هر سطر مشخص کننده برچسب صحیح و عنوان ستون مشخص کننده برچسب انتخاب شده توسط الگوریتم بیز ساده است. از طریق این جدول می توانیم تشخیص دهیم که به ازای هر زوج کلاس $c_1$ و $c_2$ ، چه تعداد سند از کلاس $c_1$ به طور اشتباه به عنوان $c_2$ برچسب خورده‌اند.

**جدول(٤): ماتریس آشفتگی خروجی الگوریتم بیز ساده بدون انتخاب ویژگی**

| | | Predicted | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | علمی فرهنگی | اقتصادی | محیط زیست | ادب و هنر | سیاسی | ورزش | Recall |
| Desired | علمی فرهنگی | ۱۵۹ | ۲ | ۵ | ۱۴ | ۵ | ۶ | ۰.۸۳ |
| | اقتصادی | ۵ | ۱۶۹ | ۳ | ۲ | ۹ | ۲ | ۰.۸۹ |
| | محیط زیست | ۰ | ۰ | ۴۲ | ۵ | ۱ | ۰ | ۰.۸۷ |
| | ادب و هنر | ۲ | ۲ | ۱ | ۱۸۶ | ۱ | ۲ | ۰.۹۵ |
| | سیاسی | ۰ | ۰ | ۰ | ۲ | ۳۷ | ۰ | ۰.۹۴ |
| | ورزش | ۳ | ۰ | ۱ | ۶ | ۳ | ۱۸۷ | ۰.۹۳ |
| | precision | ۰.۹۴ | ۰.۹۷ | ۰.۸۰ | ۰.۸۶ | ۰.۶۶ | ۰.۹۴ | |

مشاهده می شود که به طور کلی الگوریتم بیز با استفاده از ۲۰۰۰ کلمه پرکاربرد عملکرد خوبی داشته‌است، به‌جز دسته سیاسی و محیط زیست که به نسبت صحت پایین‌تری دارند. یکی از علت‌های این موضوع می‌تواند تفاوت توزیع دسته‌های مختلف و در واقع کم بودن تعداد داده‌های آموزشی این دو دسته باشد. مسئله قابل ذکر دیگر میزان پیش‌بینی‌های اشتباه به ازای یک دسته است. در ماتریس آشفتگی می‌بینیم که بیشترین میزان پیش‌بینی اشتباه و در واقع کمترین میزان فراخوانی مربوط به دسته علمی فرهنگی است. این موضوع می تواند به این دلیل باشد که کلمات این دسته تا حدی با سایر دسته‌ها همپوشانی دارد و در نتیجه میزان اطلاعاتی که این کلمات منتقل می‌کند در مقایسه با سایر دسته‌ها به میزان کمتری متمایزکننده می‌باشد.

در بخش بعدی آزمایش‌ها سه روش مختلف انتخاب ویژگی مورد استفاده قرار گرفته‌اند و در هریک از آزمایش‌ها ۲۰۰ کلمه اول استخراج شده توسط هر یک از روش‌های بهره اطلاعات، اطلاعات متقابل و مربع کای به‌عنوان ویژگی مورد استفاده قرار گرفته‌اند و سپس همانند بخش اول الگوریتم بیز ساده بر روی دادگان اجرا شده است.

نتایج بدست آمده در قالب ماتریس آشفتگی در جداول ۵-۷ آمده است. در این جداول نیز عنوان هر سطر مشخص کننده برچسب صحیح و عنوان ستون مشخص کننده برچسب انتخاب شده توسط الگوریتم بیز ساده است. همچنین

خلاصه‌ای از نتایج کلی هر ۴ حالت آزمایش شده در جدول ۸ نشان داده‌شده‌است.

همانطور که در نتایج مشاهده می‌شود هر سه روش پیشنهادی برای استخراج ویژگی باعث بهبود دقت سیستم دسته‌بندی متون شده‌اند. مقایسه سه روش با یکدیگر نشان می‌دهد روش بهره اطلاعات بیشترین بهبود را در نتایج به‌دست آورده و توانسته در مجموع ۵٪ معیار اف را افزایش دهد. این در حالی است که در اکثر روش‌های متن‌کاوی روش اطلاعات متقابل به‌عنوان روش استخراج ویژگی مورد توجه قرار گرفته است.

بررسی معیار فراخوانی دسته علمی فرهنگی در هریک از نتایج فوق نشان می‌دهد که نقص این مورد با روش‌های بهره اطلاعات و مربع کای از بین رفته است و شاهد بالا آمدن فراخوانی این دسته هستیم. این امر نشان می‌دهد که این روش‌ها موجب شده‌اند تا ویژگی‌های مناسب‌تری به ازای دسته علمی فرهنگی انتخاب شوند.

## ۵- نتیجه

در این مقاله به بررسی سه روش مختلف انتخاب ویژگی به‌نام‌های اطلاعات متقابل، بهره اطلاعات و مربع کای در متن‌کاوی پرداخته شد. نتایج این بررسی نشان داد اگرچه مقایسه کلمات انتخاب شده توسط هر سه روش کیفیت یکسانی را نشان می‌دهد و تفاوت مشهودی در خروجی سه روش مشاهده نمی‌شود، استفاده از هریک از این روش‌ها در دسته‌بندی متون فارسی به نتایج متفاوتی دست‌یافته‌است. در نهایت با استفاده از روش مربع کای برای انتخاب ویژگی دقت دسته بندی ۵٪ در معیار اف افزایش داده شد.

**جدول (۵): ماتریس آشفتگی خروجی الگوریتم بیز ساده با انتخاب ویژگی به روش بهره اطلاعات**

| | | predicted | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | علمی فرهنگی | اقتصادی | محیط زیست | ادب و هنر | سیاسی | ورزش | recall |
| Desired | علمی فرهنگی | ۱۷۸ | ۵ | ۳ | ۲ | ۲ | ۱ | ۰.۹۳ |
| | اقتصادی | ۶ | ۱۷۷ | ۳ | ۰ | ۴ | ۰ | ۰.۹۳ |
| | محیط زیست | ۰ | ۰ | ۴۶ | ۲ | ۰ | ۰ | ۰.۹۵ |
| | ادب و هنر | ۳ | ۰ | ۰ | ۱۸۲ | ۸ | ۱ | ۰.۹۳ |
| | سیاسی | ۰ | ۰ | ۰ | ۱ | ۳۸ | ۰ | ۰.۹۷ |
| | ورزش | ۰ | ۰ | ۲ | ۰ | ۱ | ۱۹۷ | ۰.۹۸ |
| | precision | ۰.۹۵۱ | ۰.۹۷ | ۰.۸۵ | ۰.۹۷ | ۰.۷۱ | ۰.۹۸ | |

**جدول(٦): ماتریس آشفتگی خروجی الگوریتم بیز ساده با انتخاب ویژگی به روش اطلاعات متقابل**

| | | predicted | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | علمی فرهنگی | اقتصادی | محیط زیست | ادب و هنر | سیاسی | ورزش | recall |
| Desired | علمی فرهنگی | ۱۵۵ | ۷ | ۴ | ۱۲ | ۱۱ | ۲ | ۰.۸۱ |
| | اقتصادی | ۳ | ۱۸۱ | ۳ | ۰ | ۳ | ۰ | ۰.۹۵ |
| | محیط زیست | ۱ | ۱ | ۴۲ | ۳ | ۱ | ۰ | ۰.۸۷ |
| | ادب و هنر | ۵ | ۰ | ۱ | ۱۸۳ | ۴ | ۱ | ۰.۹۴ |
| | سیاسی | ۰ | ۰ | ۰ | ۱ | ۳۸ | ۰ | ۰.۹۷ |
| | ورزش | ۰ | ۰ | ۲ | ۱ | ۲ | ۱۹۵ | ۰.۹۷ |
| | Precision | ۰.۹۴۵ | ۰.۹۵ | ۰.۸۰ | ۰.۹۱ | ۰.۶۴ | ۰.۹۸ | |

**جدول(۷): ماتریس آشفتگی خروجی الگوریتم بیز ساده با انتخاب ویژگی به روش مربع کای**

| | | predicted | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | علمی فرهنگی | اقتصادی | محیط زیست | ادب و هنر | سیاسی | ورزش | recall |
| Desired | علمی فرهنگی | ۱۷۶ | ۳ | ۴ | ۴ | ۳ | ۱ | ۰.۹۲ |
| | اقتصادی | ۴ | ۱۷۶ | ۵ | ۰ | ۵ | ۰ | ۰.۹۲ |
| | محیط زیست | ۰ | ۱ | ۴۳ | ۰ | ۴ | ۰ | ۰.۸۹ |
| | ادب و هنر | ۲ | ۰ | ۰ | ۱۸۵ | ۶ | ۱ | ۰.۹۵ |
| | سیاسی | ۰ | ۰ | ۲ | ۳ | ۳۴ | ۰ | ۰.۸۷ |
| | ورزش | ۰ | ۰ | ۲ | ۰ | ۰ | ۱۹۸ | ۰.۹۹ |
| | precision | ۰.۹۶ | ۰.۹۷ | ۰.۷۶ | ۰.۹۶ | ۰.۶۵ | ۰.۹۹ | |

**جدول (۸): خلاصه نتایج رویکردهای مختلف انتخاب ویژگی**

| | بدون ویژگی | بهره اطلاعات | اطلاعات متقابل | مربع کای |
|---|---|---|---|---|
| Precision | ۰.۸۶۱۶ | ۰.۹۰۹۳ | ۰.۸۷۵۶ | ۰.۸۸۶۶ |
| Recall | ۰.۹۰۱۶۷ | ۰.۹۵۳۱۳ | ۰.۹۲۱۹ | ۰.۹۲۶۴ |
| F-measure | ۰.۸۸۱۲ | ۰.۹۳۰۷ | ۰.۸۹۸۲ | ۰.۹۰۶۱ |

## مراجع


[1] Deepak Agnihotri, Kesari Verma, Priyanka Tripathi, *Variable Global Feature Selection Scheme for automatic classification of text documents*, Expert Systems with Applications, Volume 81, 2017.

[2] Abolfazl AleAhmad, Hadi Amiri, Ehsan Darrudi, Masoud Rahgozar, and Farhad Oroumchian. *Hamshahri: A Standard Persian Text Collection*. Knowledge Based Systems, 2:382–387, 2009.

[3] Huosong, X., & Jian, L. (2011). *The Research of Feature Selection of Text Classification Based on Integrated Learning Algorithm*. 10th International Symposium on Distributed Computing and Applications to Business, Engineering and Science, 20-22. 2011

[4] Zhou Qian, Zhao Mingsheng, Hu Min. *Study of Feature Selection in Chinese Text Categorization*. Journal of Chinese Information Processing, 2004, 18(3):17-23.

[5] Chen Tao, Xie Yangqun. *Literature Review of Feature Dimension Reduction in Text Categorization*. Journal of The China Society For Scientific and Technical Information, 24(6):690-695, 2005.

[6] Alper Kursat Uysal, *An improved global feature selection scheme for text classification*, Expert Systems with Applications, Volume 43, 2016.

[7] Shang Wenqian, Huang Houkuan, Zhu Haibin, et al. *A novel feature selection algorithm for text categorization*. Expert Systems with Application, 2007, 33:1-5.

[8] Y Yang, J. Q Pedersen. *Feature selection in statistical learning of text categorization*. In the 14th Int Conf on Machine Learning, 1997:412-420.


## زیرنویس‌ها

[1] Mutual Information

[2] Information Gain
[3] Chi-Square
[4] Naïve Bayes
[5] Unigram
[6] Prior probability
[7] Likelihood probability
[8] Precision
[9] Recall
[10] F-measure
[11] Confusion Matrix