# بهبود دسته بندی متون فارسی با ترکیب روش دو مرحله ای انتخاب ویژگی و الگوریتم های یادگیری ماشین

**سید محسن هاشمی** [۱,*]

**۱- گروه کامپیوتر، موسسه آموزش عالی رهنما،اهواز، ایران**

**خلاصه**

امروزه با افزایش حجم داه ها و اسناد متنی نیاز به دسته بندی اسناد احساس می شود. از جمله کاربردهای متن کاوی می توان به دسته بندی متون، شاخص گذاری متون بر اساس یک لغتنامه، فیلتر کردن متون، تولید خودکار فرا داده، تولید کاتالوگ های سلسله مراتبی از منابع وبی و دیگر کاربرد های مشابه این زمینه نام برد. در این پژوهش به دسته بندی متون فارسی با ۲۳ دسته می پردازیم. پژوهش های پیشین اغلب بر تعداد کلاس های کمتر تمرکز داشته اند و لذا با افزایش تعداد کلاس ها دقت و کارایی به نسبت بسیار زیادی کاهش خواهد یافت. لذا به ضرورت موضوع در این مقاله، یک روش ترکیبی دو مرحله ای در قسمت انتخاب ویژگی با الگوریتم های IG و CFS پیشنهاد نموده ایم. بنابراین در مرحله یادگیری ابتدا با الگوریتم های یادگیری ماشین بصورت منفرد آزمایشاتی انجام شده است و سپس با الگوریتم های ترکیبی یادگیری ماشین آزمایش نهایی را بررسی نمودیم. نتایج نشان دهنده برتری ایده و روش پیشنهادی با درصد کارایی۹۵.۴۹٪می باشد.

**کلمات کلیدی:** دسته بندی، انتخاب ویژگی، دسته بندی متون، یادگیری ماشین، متن کاوی

## ۱- مقدمه

با افزایش اطلاعات، میزان پایگاه داده‌های متنی در شبکه جهانی وب به‌سرعت در حال رشد می‌باشند. این اطلاعات در فرم الکترونیکی، مانند نشریات الکترونیکی، انواع مختلف مدارک الکترونیکی، پست الکترونیکی می باشند. سیستم‌های زیادی برای پردازش اطلاعات وجود دارد که بطور عمده می‌توان از سیستم‌های اطلاعات مدیریتی، سیستم‌های مدیریت پایگاه داده ها، سیستم‌های تصمیم یار، سیستم‌های پرسش وپاسخ و سیستم‌های بازیابی اطلاعات نام برد. کاربران نیاز به ابزارهایی برای مقایسه اسناد مختلف، رتبه‌بندی اهمیت و ارتباط اسناد و یا پیدا کردن الگوها و روندها در اسناد متعدد هستند[۵]. بنابراین، متن‌کاوی به موضوعی محبوب و ضروری در داده‌کاوی تبدیل‌شده است.

* Email: Sayedmohsenhashemi@yahoo.com

## ۲- ضرورت دسته بندی

در سالهای اخیر به علت افزایش سریع اطلاعات و در دسترس قرار گرفتن این اطلاعات به اشکال مختلف دیجیتالی ( متن، فایلهای صوتی و تصویری و ...) نیازی مبرم به نوعی مدیریت جهت دسته بندی، پردازش و ارائه حس می‌گردد. دسته بندی متون به علت اهمیت بالای متن در ثبت اطلاعات این نیاز را دوچندان می‌کند. عمل برچسب گذاری موضوعی متون بر اساس یک مجموعه مشخص یکی از این موارد می‌باشد[۲]. دسته بندی متون به صورت دستی برای زمینه های تخصصی نیاز به دانش افراد خبره می‌باشد و از آنجا که برچسب گذاری دستی براساس تجربه و دانش فرد می‌باشد بسیار خطا پذیر است به همین علت تصمیم دو فرد خبره در برچسب گذاری می‌تواند متفاوت و حتی ناسازگار باشد.

## ۳- مروری بر کارهای پیشین

تا کنون اغلب کارهای انجام شده بر روی اسنادی به زبان انگلیسی بوده و اخیرا در مورد زبانهایی مثل چینی، عربی و ... فعالیت هایی صورت گرفته است. روش‌های معمول درزمینه دسته بندی متون، روش های یادگیری ماشین هستند در چند سال اخیرالگوریتمهای زیادی برای مسأله دسته بندی اسناد، پیشنهاد شده است. در زمینه پردازش زبان طبیعی و به خصوص پردازش متن، یکی از پایه‌ای‌ترین کارها رده‌بندی یا دسته بندی خودکار متون است[۲]. دسته بندی خودکار در کاهش زمان و هزینه لازم برای دسته بندی صدها یا هزاران سندی که در یک روز وجود دارد، بدون نیاز به افراد متخصص، بسیار مفید می باشد. روش ها و الگوریتم های متفاوتی برای دسته بندی متن بکار برده شده است، که این روش ها در دقت و محاسبات متفاوت می باشند.

دسته بندی متن شامل مراحل مختلفی می باشد و پروسه دسته بندی به الگوریتم های استفاده شده بستگی دارد برمبنای این روش، تعدادی از مستندات آموزشی طبقه بندی شده به صورت دستی، داده می‌شود[ ۶]. قوانین طبقه بندی متن باید به نحوی آموزش داده شوند که بتوانند مسأله یادگیری بانظارت را به راحتی حل کنند[ ۷].

در مرجع[ ۹] هاشمی و همکاران، پژوهشی به نام دسته بندی متون فارسی با استفاده از فیلتر کردن واژه ها بر اساس اندازه آنها انجام داده اند که در این مقاله با انجام مراحلی از پیش پردازش داده ها و استفاده از الگوریتم های یادگیری ماشین از خانواده بیز کارایی دسته بندی را با مدل اعتبار سنجی ۱۰ مرحله ای متقابل به ۸۴.۶۲ درصد با ۸ دسته رساندند.

در مرجع دیگری[۱۰] هاشمی و همکاران، پژوهشی به نام استفاده از تکنیک های متن کاوی برای دسته بندی متون فارسی با مجموعه داده همشهری انجام داده اند که با به کارگیری الگوریتم های یادگیری بیز کارایی دسته بندی را با مجموعه داده همشهری با ۷۰ درصد آموزش و ۳۰ درصد آزمایش مورد بررسی قرار داده ایم. نتایج کارایی به ۸۵.۴۸ درصد و خطای۱۴.۵۲ درصد با ۸ دسته بوده است.

در مرجع دیگری مهدی پور و همکاران [۱۵] سیستم خلاصه ساز خودکار متن فارسی با استفاده از الگوریتم ترکیبی SA-GA را ارائه کرده اند که این سیستم پس از ریشه یابی کلمات با استفاده ازترکیب روشهای مبتنی بر گراف و TF-IDF با ترکیب الگوریتم SA-GA عمل می‌کند. نتایج حاصل از ارزیابی متن خلاصه شده توسط سیستم ققنوس و مقایسه آن با سیستم پارسینا نشان می دهد که کیفیت متن خلاصه شده در سیستم ققنوس۶۴.۳۵٪ بوده که بیشتر از سیستم پارسینا با ۵۹.۸۶ ٪ است.

## ۴- مجموعه داده:

یک مجموعه‌های متنی ابزارهای مهمی برای پیشبرد تحقیقات در تعدادی از شاخه‌های علوم کامپیوتر مانند بازیابی اطلاعات[1]، زبانشناسی پیکره‌ای[2] و زبانشناسی محاسباتی[3] هستند. این مجموعه آزمایش متعلق به پیکره همشهری در سال ۲۰۰۷ می‌باشد[۸]. مجموعه آزمایش همشهری یکی از معتبرترین منابع در زبان فارسی است. از این مجموعه در همایش‌های معتبر بین‌المللی استفاده شده است. تعداد هر کدام از اسناد در جدول زیر قابل مشاهده می باشد. در این پژوهش از ۲۳ دسته که در جدول ۱ نشان داده داده ایم، برای فرآیند دسته بندی متون استفاده می‌نماییم.

جدول شماره ۱: مجموعه داده ۲۳ کلاسی همشهری

| نام فارسی | اسامی دسته ها | دسته ها |
|---|---|---|
| اقتصاد | Economy | ۱ |
| اقتصاد.بانک و بورس | Economy. Bank and Bourse | ۲ |
| ادب و هنر | Literature and Art | ۳ |
| ادب و هنر.هنر | Literature and Art. Art | ٤ |
| ادب و هنر.هنر.سینما | Literature and Art .Art. Cinema | ٥ |
| ادب و هنر.هنر.موزیک | Literature and Art .Art. Music | ٦ |
| ادب و هنر. هنر.تئاطر | Literature and Art. Art. Theater | ۷ |
| ادب و هنر.ادب | Literature and Art. Literature | ۸ |
| گوناگون | Miscellaneous | ۹ |
| گوناگون.حوادث | Miscellaneous. Happenings | ۱۰ |
| گوناگون.شهری | Miscellaneous. Urban | ۱۱ |
| گوناگون.خبر جهانی | Miscellaneous. World News | ۱۲ |
| سیاسی | Politics | ۱۳ |
| سیاسی. ایران سیاسی | Politics. Iran Politics | ۱٤ |
| علمی فرهنگی | Science and Culture | ۱٥ |
| علمی فرهنگی.علمی | Science and Culture. Science | ۱٦ |
| علمی فرهنگی.علمی.کتاب | Science and Culture. Science. Book | ۱۷ |
| علمی فرهنگی.علمی.ارتباطات و فناوری اطلاعات | Science and Culture. Science. Information and Communication Technology | ۱۸ |
| علمی فرهنگی.علمی.پزشکی و درمان | Science and Culture. Science. Medicine and Remedy | ۱۹ |
| اجتماعی | Social | ۲۰ |
| اجتماعی.جهان | Social. Religion | ۲۱ |
| ورزشی | Sport | ۲۲ |
| گردشگری | Tourism | ۲۳ |

[1] Information Retrieval
[2] Corpus Linguistics
[3] Computational Linguistics

## ۵- روش کار

### ۵-۱- مرحله پیش پردازش

در حقیقت پیش پردازش وظیفه نگاشت متن داده شده به یک نمای منطقی را بر عهده دارد. به عبارت دیگر استخراج ویژگی و وزن دهی و کاهش ابعاد در این قسمت انجام می‌گیرد. بسته به کاربرد استخراج ویژگی می تواند بسیارساده و یا بسیار مفصل باشد. تحلیل واژگانی شامل عملیات مربوط به یکسان سازی متن، قواعد مربوط به‌نشانه گذاری ها ومرزبندی بین کلمات می باشد. در این مرحله داده‌های ورودی باید به صورتی تبدیل شوند که قابل پردازش توسط مراحل بعد باشند. در مرحله پیش‌پردازش داده های فارسی معمولاً بر روی‌داده‌های ورودی عملیات زیر صورت می‌گیرد:

۱. جداسازی کلمات: در این مرحله تمام متن به صورت کلمات متمایز تبدیل می‌شود و همچنین علائم نگارشی، تگ‌ها و ... حذف می‌شوند.

۲. وزن دهی کلمات (ویژگی‌ها): پس از انجام مراحل فوق متن به صورت برداری از کلمات که ویژگی‌های متن هستند می‌تواند نمایش داده شود. حال می‌تواند این ویژگی‌ها را با توجه به میزان اهمیت آن‌ها نسبت به سند متنی و دسته آن وزن دهی کرد. یکی از روشهای وزن دهی TF-IDF می باشد. این وزن دهی بر مبنای تکرار کل کلمه ومعکوس تعداد متونی که این کلمه را در بر دارند، محاسبه می‌گردد . به عبارت دیگر، هرچه تعداد تکرار کلمه در متن بیش تر و تعداد متن هایی که این کلمه را در بر دارند کمتر باشد، وزن آن ویژگی بیشتر است.

۳. **انتخاب ویژگی با روش بهره اطلاعات (IG)**: این روش یکی از پرکاربردترین و محبوب‌ترین روش‌ها می‌باشد که در بسیاری از پژوهش‌ها مورد استفاده قرارگرفته و نتایج مطلوبی داشته است. در سال‌های اخیر، سودمندی اطلاعات به عنوان معیاری برای میزان سودمندی واژه در زمینه یادگیری ماشین مطرح شده است. این معیار تعدادی قلم‌های اطلاعاتی را برای پیشگویی دسته، با کار بر روی وجود و عدم وجودیک حمله در یک دیتاست به دست می‌آورد. به عبارتی در این روش میزان سودمندی یک حمله در دیتاست تعداد حملاتی است که برای پیشگویی دسته، با توجه به وجود یا عدم وجود حمله دردیتا ست به دست می‌آید. رابطه (۱) مربوط به این روش می‌باشد [۱۴].

$$IG(t_k, c_i) = \sum_{c \in \{c_i, \bar{c}_i\}} \sum_{t \in \{t_k, \bar{t}_k\}} p(\bar{t}, c)\, log_2 \frac{P(t,c)}{P(\bar{t})P(c)} \qquad (1)$$

با داشتن یک مجموعه آموزشی، می‌توان بهره اطلاعاتی را برای هر واژه محاسبه نمود. حملاتی که بهره اطلاعاتی آن‌ها از یک حد آستانه کمتر است از فضای ویژگی حذف می‌شوند. طبق تعریف محاسبه شامل تخمینی از احتمال‌های شرطی از دسته تعیین‌کننده واژه و محاسبات آنتروپی است.

۴. **انتخاب ویژگی با روش CfsSubsetEval**: یک روش فیلتر است که زیر مجموعه های ویژگی ها را بر اساس قابلیت پیش بینی به همراه درجه افزونگی بین آنها ارزیابی می کند. در نتیجه ویژگی هایی که همبستگی بالایی با متغییر هدف داشته ودر عین حال همبستگی بین خود ویژگی ها پایین باشد ترجیح داده می شود[۱۴].

### ۵-۲- مرحله یادگیری

در مرحله یادگیری از الگوریتم های یادگیری ماشین زیر استفاده می نماییم.

#### ۵-۲-۱- الگوریتم Naïve Bayes

یکی دیگر از الگوریتم‌های مطرح‌شده در زمینه دسته‌بندی[۱]، بیز ساده است. این روش به دلایل متعددی اهمیت دارد. اینکه ساخت آن بسیار ساده است و نیازی به برنامه‌های تخمین پارامتر تکرارشونده پیچیده ندارد. بدین معنی که می‌توان از آن برای مجموعه داده‌های بسیار وسیع نیز استفاده نمود. این الگوریتم یکی از سریع‌ترین الگوریتم‌ها در زمینه دسته‌بندی می‌باشد.

#### ۵-۲-۲- الگوریتم Multinominal Naive Bayes

وظیفه کلاسه بندی متنی را می توان از یک دیدگاه یادگیری بیزین دانست که توزیع های واژه ای در اسناد از طریق مدل پارامتری خاصی بوجود آمده و پارامترها می توانند از طریق داده های یادگیری محاسبه گردند. با شناسایی(کاوش) مجموعه داده یادگیری ما می توانیم همه فراوانی های مورد نیاز را بدست آورده و سپس احتمال های شرطی متناظر را محاسبه نمائیم. یکی از مزایای مدل MNB این است که پیش بینی ها را با راندمان بالا انجام می دهد. بنابراین برای انجام پیش بینی ها MNB تنها نیازمند جستجو میان واژه های دارای شمارگان صفر می باشد. بطور متداول در کنار نمایش تکرار کم واژه های متنی بکار گرفته می شود. بطور عکس دیگر کلاسه بندها مانند شبکه ها بایستی نیازمند جستجو در میان همه واژگان در مجموعه یادگیری می باشند. از این رو که در کلاسه بندی متنی اندازه واژگان اغلب خیلی بزرگتر از طول سند می باشد. الگوریتم MNB می تواند پیش بینی ها را سریع تر از کلاسه بندهای دیگر به جز خانواده بیز انجام دهد [۳].

#### ۵-۲-۳- الگوریتم Discriminative Multinominal Naive Bayes

بیز ساده چند جمله ای متمایزگر (DMNB) یکی از روش یادگیری ماشین است. در رفرنس[۴] این الگوریتم را شرح داده شده است که از یک مدل بیز ساده چند جمله ای بهره می گیرد. هدف DMNB حفظ اطلاعات فراوانی در حالی که به ماهیت قابل تمایز کلاسه بندی توجه دارد و بنابراین ترکیبی از یادگیری عمومی و متمایز می باشد. الگوریتم DMNB از طریق اسناد یادگیری تولید می گردد. مشخص است که DMNB دارای همان خصوصیات محاسباتی MNB می باشد؛ این الگوریتم دقیقتر و نسبت به MNB سریعتر است و همچنین یادگیری آنلاین را پشتیبانی می نماید.

#### ۵-۲-۴- الگوریتم Complement Naive Bayes

دسته بندی بیز ساده مکمل(CNB) ضعف طبقه بندی بیز ساده را با پارامترهایی از داده ها در تمام دسته‌های حساس بهبود می بخشد[۳].

### ۵-۲-۵- الگوریتم ترکیبی Bagging

الگوریتم Bagging تنها یک روش یادگیری دارد. با نمونه‌گیری‌های متفاوت همراه با جایگزینی از مجموعه آموزشی، چندین مجموعه آموزشی جدید با همان اندازه اولیه به دست می‌آید و بدین طریق از هر کدام می‌توان یک دسته‌بند جداگانه را آموزش داد. در عمل نشان داده شده است که Bagging نتایج بسیار خوبی به ویژه وقتی که اندازه مجموعه داده‌ها کوچک باشد به دست می‌دهد [۱۲].

### ۵-۲-۶- الگوریتم ترکیبی Boosting

در روش Boosting الگوریتم دسته‌بندی که به آن الگوریتم یادگیری ضعیف می‌گویند[۱۳]. چندین بار و هر بار با داده‌های آموزشی متفاوت که با توجه به اجرای قبلی انتخاب می‌شوند اجراشده و در نهایت آن جوابی که بیشتر تکرار شده انتخاب می‌شود. اگر چه این روش وقت‌گیر است ولی جواب‌های آن مطمئن خواهد بود. در این روش به هر نمونه از مجموعه داده‌های آموزشی یک وزن انتساب می‌شود. در ابتدا وزن همه نمونه‌ها برابر است و در تکرارهای بعدی با توجه به کارایی دسته‌بند، وزن هر نمونه اصلاح می‌شود، به طوری که نمونه‌هایی که اشتباه دسته‌بندی شده‌اند وزنشان افزایش و نمونه‌هایی که درست دسته‌بندی شده‌اند وزنشان کاهش می‌یابد. AdaBoost* مشهورترین روش Boosting می‌باشد[۱۱].

## ۶- نتایج

**جدول ۲: مقایسه کارایی وخطای دسته بندی با 1Gram و روش وزن دهی TF-IDF**

| Algorithms | 1Gram+ TF-IDF | | |
|---|---|---|---|
| | Accuracy | Error classification | Time(s) |
| NB | ۹۲.۹۲ | ۷.۰۸ | ۹۹.۶۹ |
| MNB | ۹۳.۲۶ | ۶.۷۴ | ۰.۵۶ |
| **DMNB** | **۹۳.۴۳** | **۶.۵۷** | **۱۰.۲۵** |
| CNB | ۹۳.۳۵ | ۶.۶۵ | ۰.۶۶ |

درجدول ۲ با روش ۱-gram تعداد ۸۰۰۷ ویژگی تولید شده است که با چهار الگوریتم آزمایش هایی را انجام دادیم. نتایج نشان می دهدکه الگوریتم DMNB با کارایی ٪ ۹۳.۴۳ بالاترین کارایی و کمترین درصد خطا دسته بندی را دارد.

**جدول ۳: مقایسه کارایی وخطای دسته بندی با 2Gram و روش وزن دهی TF-IDF**

| Algorithms | 2Gram+ TF-IDF | | |
|---|---|---|---|
| | Accuracy | Error classification | Time(s) |
| NB | ۹۰.۱۵ | ۹.۸۵ | ۲۵۴.۹۳ |
| MNB | ۹۱.۰۹ | ۸.۹۱ | ۰.۶۹ |
| **DMNB** | **۹۳.۳۵** | **۶.۶۵** | **۱۲.۶۶** |
| CNB | ۹۳.۰۹ | ۶.۹۱ | ۰.۸۱ |

* Adaptive Boosting

درجدول ۳ با روش ۲-gram تعداد ۱۸۸۹۷ ویژگی تولید شده است که چهار الگوریتم بالا را آزمایش نمودیم. نتایج نشان می دهدکه الگوریتم DMNB با کارایی ٪ ۹۳.۳۵ بالاترین کارایی و کمترین درصد خطا دسته بندی را دارد. ولی در مقایسه با ۱-gram کارایی کمتری را دارد.

با توجه به جدول های ۲ و ۳ نتایج نشان دهنده این مساله است که ۱-gram ویژگی های بهتری را در انتخاب و دسته بندی دارد و لذا در این مرحله از ۱-gram استفاده خواهد شد. در مرحله بعد از الگوریتم IG برای شناسایی ارزش ویژگی های مفید استفاده گردید که به شناسایی ۵۰۰ ویژگی مفید منجر شد. این ویژگی ها از ۸۰۰۷ مورد به ۵۰۰ ویژگی تقلیل داده شد ولی ایده پژوهش حاضر بر کاهش مجدد ویژگی می باشد. ویژگی های کاهش یافته از خروجی روش IG را ذخیره و با روش انتخاب ویژگی CFS به تشخیص ویژگی های مفید پرداختیم. نتایج ویژگی های مفید به ۱۰۷ مورد رسیده است که کاهش بسیار مطلوبی در این روش داشته ایم.

**جدول ۴: مقایسه کارایی روش انتخاب ویژگی پیشنهادی با ۱۰۷ ویژگی موثر**

| Algorithms | 1Gram+ TF-IDF+(IG+CFS) | | |
|---|---|---|---|
| | Accuracy | Error classification | Time(s) |
| NB | ۸۳.۳۰ | ۱۶.۷۰ | ۱.۵۶ |
| MNB | ۹۱.۰۱ | ۸.۹۹ | ۰.۱۶ |
| **DMNB** | **۹۳.۳۱** | **۶.۶۹** | **۱.۲۵** |
| CNB | ۸۷.۸۲ | ۱۲.۱۸ | ۰.۳۱ |

در جدول ۴ بالا توانستم با ۱۰۷ ویژگی با انتخاب و کاهش آن ویژگی های به کارایی ۹۳.۳۱ ٪ برسیم که نتایج نشان دهنده کاهش در زمان تمامی الگوریتم ها می باشیم. مسئله کاهش ویژگی و انتخاب موثر آنها به زمان اجرای آزمایش الگوریتم های ما کمک بسزایی نموده است.

**جدول ۵: مقایسه روش ترکیبی انتخاب ویژگی با روش ترکیبی یادگیری Bagging**

| Algorithms | Bagging | | |
|---|---|---|---|
| | Accuracy | Error classification | Time(s) |
| NB | ۹۲.۹۳ | ۷.۰۷ | ۸.۴۹ |
| MNB | ۹۲.۷۰ | ۷.۳۰ | ۰.۵۴ |
| **DMNB** | **۹۳.۳۲** | **۶.۶۸** | **۹.۱۴** |
| CNB | ۷۸.۷۸ | ۲۱.۲۲ | ۰.۲۳ |

در جدول ۵ روش ترکیبی انتخاب ویژگی با الگوریتم ترکیبی Bagging بهمراه ۴ الگوریتم بیز را آزمایش نموده و نتایج در این جدول بهبود بعضی روش ها را شامل شده است.

**جدول ۶: مقایسه روش ترکیبی انتخاب ویژگی با روش ترکیبی یادگیری Adaboost**

| Algorithms | Adaboost | | |
|---|---|---|---|
| | Accuracy | Error classification | Time(s) |
| NB | ۹۳.۳۱ | ۶.۶۹ | ۶.۶۹ |
| MNB | ۹۳.۳۱ | ۶.۶۹ | ۲.۴۶ |
| DMNB | ۹۳.۴۷ | ۶.۵۳ | ۱.۳۴ |
| CNB | **۹۵.۴۹** | **۴.۵۱** | **۱.۵۵** |

در جدول ۶ روش ترکیبی انتخاب ویژگی با الگوریتم ترکیبی Adaboost به همراه ۴ الگوریتم بیز را آزمایش نموده و به بالاترین درصد با زمان نسبتاً مناسبی دست یافته ایم. کارایی نشان می دهد که الگوریتم بیز ساده مکمل با ۹۵.۴۹٪ و خطای دسته بندی ۴.۵۱٪ بدست آمده است.

## ۷- نتیجه گیری

با گسترس اطلاعات و فراگیر شدن آنها در شبکه هایی مانند شبکه‌های اجتماعی، سایت ها و دیگر رسانه های ارتباطی جهت کاوش متون نیازی احساس گردید که اغلب حجم انبوه داده ها و گاهاً برخی از آنها تکراری و بی ارزش هستند که ممکن است که موتور های جستجو نتوانند به درستی یک متن را کاوش نمایند. حذف ویژگی های تکراری و انتخاب ویژگی های مناسب دراین مقاله بیشترین اهمیت را داشته است. در این مقاله یک روش ترکیبی انتخاب ویژگی ارئه گردیدکه باعث کاهش زمان یادگیری الگوریتم و تسریع در امر پاسخگویی می باشد. در روش های ترکیبی همیشه زمان بیشتری را صرف یادگیری می کنند ولی با ارائه روش انتخاب ویژگی، زمان یادگیری زمان بر را کاهش داده ایم همچنین نتایج ارائه شده برای ۲۳ دسته با ۹۵.۴۹٪ بالاترین کارایی در این پژوهش و کمترین خطای دسته بندی را دارد.

## مراجع:

[1] Ganiz, M. C., George, C., & Pottenger, W. M. (2011). Higher order Naive Bayes: A novel non-IID approach to text classification. Knowledge and Data Engineering, IEEE Transactions on, 23(7), 1022-1034.
[2] Sebastiani, F. (2002). Machine learning in automated text categorization. ACM computing surveys (CSUR), 34(1), 1-47.
[3] Rennie, J. D., Shih, L., Teevan, J., & Karger, D. R. (2003, August). Tackling the poor assumptions of naive bayes text classifiers. In ICML (Vol. 3, pp. 616-623).
[4] Su, J., Zhang, H., Ling, C. X., & Matwin, S. (2008, July). Discriminative parameter learning for Bayesian networks. In Proceedings of the 25th international conference on Machine learning (pp. 1016-1023). ACM.
[5] Berry, M. W. (2004). Survey of text mining. Computing Reviews, 45(9), 548.
[6] Dumais, S., Platt, J., Heckerman, D., Sahami, M. (1998). Inductive learning algorithms and representations for text categorization. In Proceedings of the 7th International Conference on Information and Knowledge Management.
[7] Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. In Proceedings of the European Conference on Machine Learning (ECML).
[8] http://ece.ut.ac.ir/dbrg/hamshahri/faindex.html
[9] Hashemi, S. M., & Dorfeshan, z., &Javadi, S., &Dezfuli, M.M. (2014). Using text mining techniques to categorize Persian texts dataset Hamshahri, International Conference on Engineering, Arts and the Environment.

[10] Hashemi, S. M., & Dorfeshan, z., &Javadi, S., & Dezfuli, M.M. (2014). Categories of Persian texts Using filter words by size, International Conference on Engineering, Arts and the Environment.

[11] Shipp, C. A., & Kuncheva, L. I. (2002, January). An investigation into how adaboost affects classifier diversity. In Proceedings of 9th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (pp. 203-208).

[12] Polikar, R. (2006). Ensemble based systems in decision making. Circuits and Systems Magazine, IEEE, 6(3), 21-45.

[13] Freund, Y., & Schapire, R. E. (1996, July). Experiments with a new boosting algorithm. In ICML (Vol. 96, pp. 148-156).

[14] Jing, L. P., Huang, H. K., & Shi, H. B. (2002). Improved feature selection approach TFIDF in text mining. In Machine Learning and Cybernetics, 2002. Proceedings. 2002 International Conference on (Vol. 2, pp. 944-946). IEEE.

[۱۵] الهام مهدی پور، معصومه باقری قرقروک، افسانه رضایی، ۱۳۹۲ . سیستم خلاصه ساز خودکار متن فارسی با استفاده از الگوریتم ترکیبیSA-GA، هشتمین سمپوزیوم بین المللی پیشرفتهای علوم و تکنولوژی. مشهد، موسسه آموزش عالی خاوران.