

## Summarization Techniques: Impact on Persian Texts Classification

F. Z. Arab-Ahmadi | S. Karbasi

Received: 14, Jan. 2019

Accepted: 12, May 2019

**Purpose:** To determine the impact of document summarization parameters on the evaluation metrics of classification algorithms for Persian texts.

**Methodology:** 1000 news texts were collected from yjc.ir news agency website based on the number of visits, with at least 100 and at most 350 words, out of which 250 were selected randomly. Titles, summaries, and the texts of the 250 docs were included in three groups. The number of documents were increased by 100 percent in two stages, to 500 and 1000. After text preprocessing and deleting stop words by programming code, TF-ISF summarization technique was implemented on them. 12 Excel files were created from the words of original texts. Then, Bayesian, Decision tree, SVM and Rule-based algorithms implemented by Rapid Miner software, which provided 120 Excel output files for verifying accuracy, precision, and recall. Finally, five comparisons between the results were considered including comparing of results with 100% increase in the number of documents, comparing the parameters of TF and ISF summarizer, comparison of Bayesian classification algorithms, decision tree, Rule and SVM, comparing the original text and summary and comparison of the documents labels.

**Findings:** The results indicated the superiority of evaluation criteria in classification of 1000 documents relative to those of 250 and 500, which in 84% of cases. Meanwhile, the ISF summarizer method compared to TF in 82% of comparison showed a greater impact on classification accuracy. In addition, the values of the accuracy in Bayesian classification and the SVM were better. The highest value obtained from the accuracy (96. 67%) in the SVM classification by 1000 documents of original text and ISF summarizer technique.

**Conclusion:** Appropriate parameters for summarization and efficient classification techniques can improve the accuracy of Persian text classification process, while the required time also decreases. The best results obtained in the evaluations show that ISF summarizer, Bayesian and SVM algorithms, 1000 documents, as well as the main text are more effective.

### Keywords:

Classification, Persian texts, TF-ISF Summarizer Classification algorithms, Classification metrics

DOI: 10.30484/našinfo.2019.2331

1. MA in Computer Science, Golestan University, Gorgan, f.arabahmadi7@gmail.com
2. Assistant Professor, Computer Science, Golestan University, Gorgan (Corresponding author), s.karbasi@gu.ac.ir

## تأثیر تکنیک‌های خلاصه‌سازی بر دسته‌بندی متون فارسی

فاطمه زهرا عرب احمدی<sup>۱</sup> | سهیلا کرباسی<sup>۲</sup>

**هدف:** استفاده از پارامترهای مناسب خلاصه‌سازی و تکنیک‌های دسته‌بندی کارآمد باعث بهبود صحت فرایند دسته‌بندی متون فارسی می‌شود. ضمن آنکه زمان لازم نیز کاهش می‌یابد. هدف پژوهش حاضر برآورد تأثیر پارامترهای خلاصه‌سازی اسناد بر معیارهای ارزیابی کیفیت الگوریتم‌های دسته‌بندی در متون فارسی است.

**روش‌شناسی:** ۱۰۰۰ سند در قالب متن خبری برحسب بیشترین بازدید و میانگین طول (حداقل ۱۰۰ و حداکثر ۳۵۰ واژه) از سایت خبری yjc.ir جمع‌آوری شد. ۲۵۰ سند از میان آنها به‌طور تصادفی انتخاب شد. عنوان، خلاصه، و متن آنها را سه دسته کردیم. در دو مرحله رشد ۱۰۰ درصدی، شمار آنها به ۵۰۰ و ۱۰۰۰ افزایش یافت. با پیش‌پردازش، ایست‌واژه را حذف، و تکنیک‌های خلاصه‌سازی TF-ISF را روی آنها اجرا کردیم. دوازده فایل اکسل خروجی از واژه‌های متن و خلاصه به‌دست آمد. الگوریتم‌های دسته‌بندی بیزین، درخت تصمیم، بردار پشتیبان، و قانون توسط نرم‌افزار Rapid Miner بر آنها اجرا شد. ۱۲۰ فایل اکسل خروجی از نتایج معیارهای استاندارد ارزیابی صحت، دقت، و فراخوان به‌دست آمد. درنهایت، پنج حالت مقایسه شامل مقایسه نتایج با رشد ۱۰۰ درصدی تعداد اسناد؛ مقایسه پارامترهای خلاصه‌ساز TF و ISF؛ مقایسه الگوریتم‌های دسته‌بندی بیزین، درخت تصمیم، قانون، و ماشین بردار پشتیبان؛ مقایسه متن اصلی و متن خلاصه؛ و مقایسه برچسب‌های اسناد بر نتایج به‌دست آمده را بررسی کردیم.

**یافته‌ها:** نتایج حاکی از برتری معیارهای ارزیابی دسته‌بندی در اسناد ۱۰۰۰ تایی نسبت به اسناد ۲۵۰ تایی و ۵۰۰ تایی بود؛ زیرا دسته‌بندی اسناد ۱۰۰۰ تایی در ۸۴٪ حالات نتایج بهتری داشت. همچنین، پارامتر خلاصه‌ساز ISF نسبت به TF در ۸۲٪ حالت مقایسه، تأثیر بیشتری بر دقت دسته‌بندی نشان داد. معیار صحت در روش‌های دسته‌بندی بیزین و بردار پشتیبان درباره متن اصلی برتر از روش‌های قانون و درخت تصمیم بود. بیشترین مقدار به‌دست آمده (۹۶/۶۷٪) از معیار صحت در دسته‌بندی SVM و تعداد اسناد ۱۰۰۰ تایی متن اصلی توسط پارامتر خلاصه‌ساز ISF بود.

**نتیجه‌گیری:** نتایج حاکی از تأثیر بیشتر پارامتر خلاصه‌ساز ISF، الگوریتم‌های دسته‌بندی بیزین و بردار پشتیبان، اسناد ۱۰۰۰ تایی و همچنین متون اصلی است.

### کلیدواژه‌ها

دسته‌بندی متون فارسی، خلاصه‌ساز TF-ISF، الگوریتم‌های دسته‌بندی، معیارهای ارزیابی دسته‌بندی

دریافت: ۹۷/۱۰/۲۵ پذیرش: ۹۸/۰۲/۲۳

۱. کارشناس ارشد کامپیوتر، دانشگاه گلستان، گرگان  
f.arabahmadi7@gmail.com

۲. استادیار گروه کامپیوتر، دانشگاه گلستان، گرگان  
s.karbasi@gu.ac.ir (نویسنده مسئول)

مطالعات ملی کتابداری و سازماندهی اطلاعات،  
دوره سی، شماره سوم، پاییز ۱۳۹۸، ص ۲۳-۸.

DOI: 10.30484/našinfo.2019.2331

## مقدمه

رشد سریع تعداد اسناد الکترونیکی و مشکل کمبود زمان برای کاربران و نیاز اساسی به نگهداری، بازیابی، و پردازش اطلاعات منجر به استقبال پژوهشگران و دانش پژوهان از پژوهش‌های خلاصه‌سازی و دسته‌بندی خودکار متن شده است. رشد سریع شبکه جهانی وب و افزایش اسناد الکترونیکی، ارزش زمان برای مشتریان، طولانی‌بودن اسناد و اطلاعات و ممکن‌نبودن بررسی تمام اطلاعات، نیاز به خلاصه‌سازی و دسته‌بندی متن برای تشخیص سریع‌تر موضوع و محتوای متون را افزایش داده است. هدف متن‌کاوی، شناسایی و اکتشاف الگوها برای استخراج اطلاعات مفید از داده‌های متنی بدون ساختار است. متن‌کاوی از پایگاه داده، یادگیری ماشین<sup>۱</sup>، و مانند اینها بهره می‌برد. تکنیک‌های دسته‌بندی و خلاصه‌سازی متن بخشی از حوزه موسوم به متن‌کاوی است. (غضنفری، علیزاده، و تیمورپور، ۱۳۹۳). دسته‌بندی عبارت است از یادگیری به‌وسیله نمونه‌های موجود در دسته‌های مشخص که هر دسته دارای یک یا چند برجسب<sup>۲</sup> است برجسب‌ها مقادیر اسمی یا عددی هستند و مشخصه نمونه‌های هر دسته‌اند. دسته‌بندی دو مرحله ساخت مدل و استفاده از مدل و پیش‌بینی از طریق داده‌های قبلی است (Han & Kamber, 2012). الگوریتم‌های دسته‌بندی از قبیل «نزدیک‌ترین همسایه»<sup>۳</sup>، «ماشین بردار پشتیبان»<sup>۴</sup>، «بیزین»<sup>۵</sup>، «درخت تصمیم‌گیری»<sup>۶</sup>، و رگرسیون داده‌های متنی را به رده‌های تعریف‌شده، تقسیم می‌کنند (Brindha, Prabha, & Sukumaran, 2016).

همچنین حجم عظیم اطلاعات و محدودبودن زمان موجب می‌شود تا پژوهشگران به‌دنبال راهکارهایی برای انتخاب درست و سریع مطالب باشند. از این‌رو، خلاصه‌سازی متون برای دسته‌بندی اطلاعات اهمیت دارد خلاصه‌سازی، «کلان داده‌ها»<sup>۷</sup> را پردازش می‌کند و در عین حفظ نکات مهم و معنای کلی، طول و جزئیات آنها را می‌کاهد (شورای عالی اطلاع‌رسانی، ۱۳۸۸).

تاکنون تکنیک‌های خلاصه‌سازی و دسته‌بندی به‌طور ترکیبی بر متون فارسی انجام نشده است. خصوصیات و تفاوت خط فارسی با زبان‌های دیگر، نظیر چسبیدگی حروف و تنوع نگارش، تعریف‌نشده‌گی برخی نکات دستوری، و واژه‌های ترکیبی، پردازش خط فارسی را پیچیده می‌کند (شورای عالی اطلاع‌رسانی، ۱۳۸۸). تکنیک‌های خلاصه‌سازی و دسته‌بندی متون فارسی ضروری است. در این پژوهش با ادغام تکنیک‌های مختلف خلاصه‌سازی بر دسته‌بندی متون فارسی عملکرد آنها مقایسه و ارزیابی شده است:

– مقایسه تأثیر افزایش تعداد اسناد بر نتایج تکنیک‌های مختلف خلاصه‌سازی و دسته‌بندی

1. Machin learning
2. Label
3. K-Nearest-Neighbor
4. Suoport vector machine
5. Naive Bayesian
6. Decision Tree
7. Big data

- مقایسه پارامترهای خلاصه‌ساز<sup>۱</sup> TF و ISF<sup>۲</sup>
  - مقایسه الگوریتم‌های دسته‌بندی بیزین، درخت تصمیم، قانون، و بردار پشتیبان
  - مقایسه تأثیر متون خلاصه و متون اصلی بر نتایج تکنیک‌های دسته‌بندی
  - مقایسه بین برچسب اسناد دسته‌بندی شده
- برای بهبود تکنیک‌های دسته‌بندی و خلاصه‌سازی خودکار متن‌های فارسی دو تلاش انجام شده است (آهنگری، ۱۳۹۶؛ احمدی، ۱۳۹۰). در پژوهش حاضر از ترکیب چند روایت از هر دو تکنیک بهره گرفته‌ایم؛ نیز تأثیر پارامترهای خلاصه‌سازی اسناد را بر معیارهای ارزیابی کیفیت الگوریتم‌های دسته‌بندی در متون فارسی بررسی کرده‌ایم.

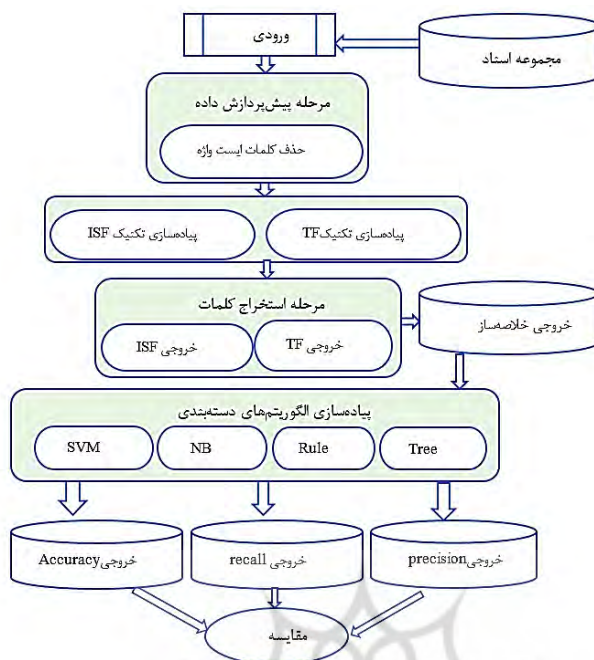
### روش‌شناسی

داده‌ها از سایت باشگاه خبرنگاران جوان به نشانی <https://www.yjc.ir> شامل هزار سند با طول حداقل ۱۰۰ و حداکثر ۳۵۰ واژه جمع‌آوری شد. سبب استفاده از سایت ذکرشده به‌روز بودن اخبار و سنجش کارایی روش پیشنهادی ما بر داده‌های واقعی و غیرآزمایشگاهی بود. عنوان، متن کامل، و خلاصه این هزار سند را در سه فایل متنی ذخیره کردیم. هر محتوای فایل را با برچسب‌های پزشکی، سیاسی، انتخابات، اقتصادی، فرهنگی، ورزشی، و بین‌الملل به ۷ گروه تقسیم کردیم. پیش‌پردازش ایست‌واژه‌ها از متون حذف شد. برای مقایسه تأثیر افزایش تعداد اسناد بر کارایی الگوریتم‌های دسته‌بندی، اسناد را به‌طور تصادفی ۲۵۰ تایی، ۵۰۰ تایی، و ۱۰۰۰ تایی تقسیم و در سه مرحله وارد پایگاه داده کردیم. سپس تکنیک‌های خلاصه‌سازی TF و ISF روی آنها پیاده‌سازی شدند و در نهایت، الگوریتم‌های دسته‌بندی اجرا شد و مقادیر معیارهای دقت، صحت و فراخوان ارزیابی شد.

این پژوهش در ۷ مرحله انجام شده است: (۱) جمع‌آوری داده‌ها؛ (۲) پیش‌پردازش؛ (۳) پیاده‌سازی خلاصه‌سازی TF و ISF؛ (۴) استخراج کلمات «متن کامل» و «متن خلاصه»؛ (۵) پیاده‌سازی الگوریتم‌های دسته‌بندی؛ (۶) استخراج نتایج پیاده‌سازی الگوریتم‌های دسته‌بندی؛ و (۷) مقایسه نتایج (شکل ۱).

1. Term frequency

2. Inverse sentence frequency



شکل ۱. چهارچوب گام‌های اجرایی در پژوهش

جدول ۱. تعداد برچسب‌ها در اسناد جمع‌آوری شده

تعداد برچسب در ۱۰۰۰ سند	تعداد برچسب در ۵۰۰ سند	تعداد برچسب در ۲۵۰ سند	برچسب
۴۲۱	۲۰۰	۲۱	ورزش
۷۹	۷۹	۲۱	بین‌الملل
۱۹۰	۲۱	۲۱	فرهنگی
۱۵	۱۵	۱۵	انتخابات
۴۹	۴۹	۴۹	اقتصاد
۵۴	۵۴	۵۴	سیاسی
۱۹۱	۸۲	۸۱	علمی-پزشکی

در این پژوهش به‌جای سیستم وزن دهی کلاسیک TF-IDF از معیار TF-ISF استفاده شد که معیار محاسبه وزن در واحد جمله است. پارامترهای TF و ISF به‌ترتیب نشان‌دهنده تعداد تکرار واژه در هر سند و معکوس تعداد جمله‌های حاوی کلمه در هر سند است و از آنها برای انتخاب جمله‌های مهم اسناد در تکنیک‌های خلاصه‌سازی استفاده می‌شود (Rahman & Borah, 2015). برای همه واژه‌های هر سند، فرکانس آنها برای پارامتر خلاصه‌ساز TF و معکوس فرکانس جمله‌هایی که این کلمات را دارند برای پارامتر ISF محاسبه شدند.

نتیجه پیاده‌سازی کدهای خلاصه‌ساز برای اسناد ۲۵۰ تایی، ۵۰۰ تایی، و ۱۰۰۰ تایی، در ۱۲ فایل اکسل ذخیره شد. این کار برای پیاده‌سازی الگوریتم‌های دسته‌بندی در مرحله بعد انجام شد و برای وضوح بیشتر طبق جدول ۲ نام‌گذاری شدند.

جدول ۲. اسامی فایل‌های اکسل خروجی حاوی مقادیر TF و ISF

	TF	ISF
متن اصلی	TF250(1)	ISF250(1)
	TF500(1)	ISF500(1)
	TF1000(1)	ISF1000(1)
خلاصه	TF250(2)	ISF250(2)
	TF500(2)	ISF500(2)
	TF1000(2)	ISF1000(2)

بعد از پیاده‌سازی کد خلاصه‌ساز TF و ISF بر متن کامل و خلاصه‌های پیش‌پردازش‌شده، خروجی‌هایی از تعداد تکرار کلمات و معکوس تعداد تکرار کلمات در جمله از متن کامل و متن خلاصه به‌دست آمد. چون ۱۰۰۰ سند جمع‌آوری شده به سه گروه تقسیم شده بودند، برای هر یک از روش‌های TF و ISF در متن کامل و از هر دسته ۳ و جمعاً ۱۲ فایل خروجی به‌دست آمد. برای پیاده‌سازی الگوریتم‌های دسته‌بندی از نرم‌افزار Rapid Miner<sup>۱</sup> استفاده شد. این نرم‌افزار موتور داده‌کاوی رایگان برای تجزیه و تحلیل داده‌هاست و در تحلیل و پیش‌بینی، یادگیری ماشین، امور تجاری و پژوهشی و آموزشی، و همچنین متن‌کاوی به‌کار می‌آید.

1. www.Rapidminer.com

برای ارزیابی روش پیشنهادی، از الگوریتم‌های دسته‌بندی بیزین، درخت تصمیم، قانون، و بردار پشتیبان که از مهم‌ترین الگوریتم‌های متن‌کاوی هستند استفاده شد (Brindha et al., 2016). از نتایج مرحله قبل، ۱۲ فایل اکسل از کلمات متن‌های کامل و خلاصه‌ها با تعداد تکرار آنها برای TF و معکوس تعداد جمله‌هایی که شامل آن واژه‌ها هستند برای ISF به دست آمد و به نرم‌افزار Rapid Miner وارد شد. در این مرحله به استفاده از عملگرهای پیش‌پردازش و آماده‌سازی نیاز نبود چون داده‌ها در قسمت قبل پیش‌پردازش شده بودند. بعد از پیاده‌سازی الگوریتم‌های دسته‌بندی بیزین، درخت تصمیم، و قانون، سه خروجی اکسل برای هر کدام از معیارهای دقت، صحت، و فراخوان حاصل شد. خروجی Role Model نیز بعد از پیاده‌سازی الگوریتم قانون با توجه به کلمات موجود، قانونی برای برچسب‌گذاری اسناد به دست آورد. همین‌طور برای الگوریتم بردار پشتیبان، یک جدول با محتوای معیار صحت برای کل سند و معیار فراخوان و دقت برای هر برچسب ایجاد شد. خروجی‌های این مرحله شامل ۳۶ فایل اکسل از معیار دقت، ۳۶ فایل اکسل از معیار فراخوان، و ۳۶ فایل اکسل از معیار صحت، و ۱۲ جدول از سه معیار صحت، فراخوان، و دقت برای هر کدام از اسناد ۲۵۰ تایی، ۵۰۰ تایی، و ۱۰۰۰ تایی برای متن کامل و متن خلاصه مجزاست. با اعمال الگوریتم‌های دسته‌بندی روی ۱۲ فایل خروجی از مرحله اعمال پارامترهای TF و ISF، ۴۸ مقدار برای معیار صحت، ۱۰۸ مقدار برای معیار دقت، و ۱۰۸ مقدار هم برای معیار فراخوان به دست آمد که جمعا ۲۶۴ مقدار برای مقایسه حاصل شد. مقادیر به دست آمده برای پارامترهای TF و ISF جداگانه مقایسه شد و در نهایت، بین این دو پارامتر مقایسه نهایی انجام شد.

#### یافته‌ها

نتایج ۱۲۰ خروجی اکسل با پیاده‌سازی الگوریتم‌های دسته‌بندی در جدول‌های ۳ الی ۱۰ خلاصه شد. جدول‌های ۳ تا ۶، نتایج الگوریتم‌های دسته‌بندی با پیاده‌سازی خلاصه‌ساز TF و ISF در اسناد ۲۵۰ تایی، ۵۰۰ تایی، و ۱۰۰۰ تایی از متون اصلی و خلاصه را نشان می‌دهند. چهار جدول ۷ الی ۱۰، مقایسه برچسب‌ها در روش‌های خلاصه‌ساز TF و ISF در اسناد ۲۵۰ تایی، ۵۰۰ تایی، و ۱۰۰۰ تایی از متون اصلی و متون خلاصه را توسط معیارهای دقت و فراخوان در الگوریتم دسته‌بندی ماشین بردار پشتیبان نشان می‌دهند.



جدول ۳. نمایش معیارهای دقت، صحت، و فراخوان در الگوریتم دسته‌بندی TREE با خلاصه‌ساز TF و ISF

الگوریتم دسته‌بندی TREE						اسناد
خلاصه‌ساز ISF			خلاصه‌ساز TF			
دقت %	فراخوان %	صحت %	دقت %	فراخوان %	صحت %	
۳۵/۹۱	۴۳/۱۲	۶۸	۲۱/۱۷	۳۰/۹۱	۵۴/۶۷	۲۵۰(۱)
۲۱/۱۸	۳۱/۴۸	۵۳/۳۳	۲۲/۱۴	۱۶/۲۱	۴۲/۶۷	۲۵۰(۲)
۴۶/۵۸	۴۳/۴۲	۷۱/۳۳	۲۴/۴۸	۲۸/۸۴	۵۴	۵۰۰(۱)
۲۵/۲۵	۲۰/۸۳	۴۴	۱۳/۲۲	۱۹/۰۵	۴۲/۶۷	۵۰۰(۲)
۳۷/۵۶	۴۲/۱۴	۷۹/۶۷	۲۶/۸۶	۳۳/۹۳	۶۳/۳۳	۱۰۰۰(۱)
۲۲/۵۴	۲۳/۳۳	۴۶/۳۳	۱۳/۳۳	۱۴/۱۸	۴۶/۳۳	۱۰۰۰(۲)

مقادیر جدول ۳ نشان می‌دهد الگوریتم دسته‌بندی TREE توسط پارامتر ISF در اسناد ۱۰۰۰ تایی متن کامل بهترین نتیجه را در معیار صحت با مقدار ۷۹/۶۷ درصد به دست آورد.

جدول ۴. نمایش معیارهای دقت، صحت، و فراخوان در الگوریتم دسته‌بندی Bayesian با خلاصه‌ساز TF و ISF

الگوریتم دسته‌بندی Bayesian						اسناد
خلاصه‌ساز ISF			خلاصه‌ساز TF			
دقت %	فراخوان %	صحت %	دقت %	فراخوان %	صحت %	
۵۴/۸۳	۴۵/۳۵	۶۹/۳۳	۵۶/۴۸	۵۶/۱۹	۸۰	۲۵۰(۱)
۶۴/۳۳	۵۵/۱۳	۷۰/۶۷	۴۷/۴	۴۲/۳۶	۶۰/۸۱	۲۵۰(۲)
۶۱/۷	۶۳/۱۶	۸۷/۳۳	۵۵/۹۷	۶۰/۸۷	۸۹/۴۱	۵۰۰(۱)
۷۳/۹۸	۶۰/۳	۸۳/۳۳	۷۶	۶۶/۰۸	۸۸/۴۴	۵۰۰(۲)
۷۶/۸	۶۴/۴۴	۸۸/۵۹	۸۱/۰۶	۶۶/۱	۹۲/۷۲	۱۰۰۰(۱)
۷۴/۰۴	۶۴/۲۵	۸۷	۶۴/۷۳	۵۶/۵۳	۸۲/۶۷	۱۰۰۰(۲)

مقادیر جدول نشان می‌دهد الگوریتم دسته‌بندی Bayesian توسط پارامتر TF در اسناد ۱۰۰۰ تایی متن اصلی بهترین نتیجه را در معیار صحت با مقدار ۹۲/۷۲ درصد به دست آورد.



جدول ۵. نمایش معیارهای دقت، صحت، و فراخوان در الگوریتم دسته‌بندی Rule با خلاصه‌ساز TF و ISF

الگوریتم دسته‌بندی Rule						اسناد
خلاصه‌ساز ISF			خلاصه‌ساز TF			
دقت %	فراخوان %	صحت %	دقت %	فراخوان %	صحت %	
۶۹/۸۵	۶۶/۷۱	۸۵/۳۳	۵۵/۳	۵۴/۶۸	۷۰/۶۷	۲۵۰(۱)
۴۴/۰۸	۳۳/۹	۵۰/۶۷	۳۵/۵۸	۲۷/۶۴	۴۲/۶۷	۲۵۰(۲)
۷۱/۵۹	۷۵/۲۹	۸۸/۶۷	۵۸/۱۲	۵۷/۱۳	۸۴	۵۰۰(۱)
۶۶/۳۱	۴۶/۹۶	۶۶	۳۷/۲	۳۸/۷۵	۵۸/۶۷	۵۰۰(۲)
۸۷/۸۲	۹۰/۷۲	۹۵	۵۸/۹۸	۵۸/۸۱	۸۴	۱۰۰۰(۱)
۶۶/۵۷	۵۸/۲۳	۶۹/۶۷	۵۵/۷۸	۴۹/۱۶	۶۶/۳۳	۱۰۰۰(۲)

مقادیر جدول ۵ نشان می‌دهد الگوریتم دسته‌بندی Rule توسط پارامتر ISF در اسناد ۱۰۰۰ تایی متن کامل بهترین نتیجه را در معیار صحت با مقدار ۹۵ درصد به دست آورد.

جدول ۶. نمایش معیار صحت در الگوریتم دسته‌بندی SVM برای خلاصه‌ساز TF و ISF

معیار Accuracy (%) در الگوریتم دسته‌بندی SVM		اسناد
خلاصه‌ساز ISF	خلاصه‌ساز TF	
۸۰	۸۰	۲۵۰(۱)
۵۶	۵۳/۳۳	۲۵۰(۲)
۹۲/۶۷	۸۸	۵۰۰(۱)
۸۷/۶۷	۸۷/۳۳	۵۰۰(۲)
۹۶/۶۷	۸۸/۳۳	۱۰۰۰(۱)
۸۲/۳۳	۸۴/۶۷	۱۰۰۰(۲)

مقادیر جدول ۶ نشان می‌دهد الگوریتم دسته‌بندی SVM توسط پارامتر ISF در اسناد ۱۰۰۰ تایی متن کامل بهترین نتیجه را در معیار صحت با مقدار ۹۶/۶۷ درصد به دست آورد.

جدول ۷. نمایش معیار Recall در الگوریتم دسته‌بندی SVM با خلاصه‌ساز TF

معیار Recall (%) در الگوریتم دسته‌بندی SVM برای خلاصه‌ساز TF						برچسب‌ها
۱۰۰۰(۲)	۱۰۰۰(۱)	۵۰۰(۲)	۵۰۰(۱)	۲۵۰(۲)	۲۵۰(۱)	
۸۵/۷۱	۸۵/۷۱	۹۳/۱۰	۹۳/۱	۷۲	۹۲	علمی-پزشکی
۵۸/۳۳	۵۰	۸۶/۶۷	۸۶/۶۷	۴۷/۳۷	۶۸/۲۲	سیاسی
.	.	.	.	.	.	انتخابات
۱۵	۲۵	۶۲/۵	۶۲/۵	۶۴/۷۱	۹۴/۱۲	اقتصادی
۹۳/۶۲	۹۷/۸۷	.	.	۲۵	۷۵	فرهنگی
۵۶/۵۲	۵۶/۵۲	۱۰۰	۱۰۰	۱۰	۵۰	بین‌المللی
۹۹/۲۹	۹۸/۵۷	۹۶/۸۳	۹۸/۴۱	-	-	ورزشی

جدول ۸. نمایش معیار Precision در الگوریتم دسته‌بندی SVM با خلاصه‌ساز TF

معیار Precision (%) در الگوریتم دسته‌بندی SVM برای خلاصه‌ساز TF						برچسب‌ها
۱۰۰۰(۲)	۱۰۰۰(۱)	۵۰۰(۲)	۵۰۰(۱)	۲۵۰(۲)	۲۵۰(۱)	
۸۲/۲۱	۸۸/۸۹	۸۴/۳۸	۸۴/۳۸	۹۴/۴۴	۸۵/۹۹	علمی-پزشکی
۶۳/۶۴	۶۰	۶۵	۶۵	۹۲/۵۹	۷۲/۲۲	سیاسی
.	.	.	.	۱۰۰	.	انتخابات
۷۵	۱۰۰	۹۰/۹۱	۹۰/۹۱	۷۶/۹۲	۸۴/۲۱	اقتصادی
۷۰/۹۷	۶۸/۶۶	.	.	.	۷۵	فرهنگی
۱۰۰	۸۶/۶۷	۸۰	۸۰	.	۱۰۰	بین‌المللی
۹۰/۸۵	۹۲/۶۲	۹۸/۳۹	۹۸/۴۱	-	-	ورزشی

جدول ۹. نمایش معیار Recall در الگوریتم دسته‌بندی SVM با خلاصه‌ساز ISF

معیار Recall (%) در الگوریتم دسته‌بندی SVM برای خلاصه‌ساز ISF (%)						برچسب‌ها
۱۰۰۰(۲)	۱۰۰۰(۱)	۵۰۰(۲)	۵۰۰(۱)	۲۵۰(۲)	۲۵۰(۱)	
۷۶/۷۹	۱۰۰	۷۹/۳۱	۱۰۰	۸۴	۹۵/۸۲	علمی-پزشکی
۳۳/۳۳	۵۸/۳۳	۴۰	۸۶/۶۷	۴۷/۳۷	۸۲/۳۵	سیاسی
.	.	۳۳/۳۳	.	.	۱۰۰	انتخابات
۵۰	۹۵	۵۶/۵۲	۹۳/۷۵	۵۲/۹۴	۸۸/۸۹	اقتصادی
۹۳/۶۲	۱۰۰	.	.	.	.	فرهنگی
۴۳/۴۸	۹۵	۹۰	۹۵	۳۰	۵۴/۵۵	بین‌المللی
۹۷/۱۴	۹۹/۱۹	۹۶/۸۳	۱۰۰	-	-	ورزشی

جدول ۱۰. نمایش معیار Precision در الگوریتم دسته‌بندی SVM با خلاصه‌ساز ISF

معیار Precision در الگوریتم دسته‌بندی SVM برای خلاصه‌ساز ISF						برچسب‌ها
۱۰۰۰(۲)	۱۰۰۰(۱)	۵۰۰(۲)	۵۰۰(۱)	۲۵۰(۲)	۲۵۰(۱)	
۸۱/۱۳	۱۰۰	۷۹/۳۱	۱۰۰	۴۵/۶۵	۷۱/۹۸	علمی-پزشکی
۶۶/۶۷	۷۷/۷۸	۶۶/۶۷	۸۶/۶۷	۶۹/۲۳	۷۳/۶۸	سیاسی
.	.	۳۳/۳۲	.	.	۱۰۰	انتخابات
۱۰۰	۱۰۰	۱۰۰	۱۰۰	۹۰	۱۰۰	اقتصادی
۶۱/۱۱	۹۵/۹۲	.	.	.	.	فرهنگی
۷۱/۴۳	۹۵/۹۲	۷۵	۱۰۰	۱۰۰	۱۰۰	بین‌الملل
۹۳/۷۹	۹۶/۵۳	۸۰/۲۶	۸۷/۵۰	-	-	ورزشی

#### • مقایسه نتایج با رشد ۱۰۰ درصدی تعداد اسناد

در این قسمت، ابتدا مقایسه رشد ۱۰۰ درصدی تعداد اسناد در دو مرحله از ۲۵۰ به ۵۰۰ و از ۵۰۰ به ۱۰۰۰ انجام شد. به‌عنوان نمونه، برای معیار صحت در دسته‌بندی SVM در جدول ۶ برای خلاصه‌ساز TF و ISF شاهد افزایش متناسب با افزایش تعداد اسناد هستیم. میزان صحت ۸۰ درصد در اسناد ۲۵۰ تایی متن کامل به ۸۸ درصد در اسناد ۵۰۰ تایی متن کامل و همین‌طور به ۸۸/۳۳ درصد در اسناد ۱۰۰۰ تایی متن کامل در خلاصه‌ساز TF می‌رسد. این روند برای متن خلاصه نیز تکرار شده است و با توجه به مقادیر دقت، صحت، و فراخوان موجود در جدول‌های ۳ الی ۶ به این نتیجه می‌رسیم که در ۸۴ درصد حالات، اسناد ۱۰۰۰ تایی نسبت به اسناد ۵۰۰ تایی و همین‌طور اسناد ۵۰۰ تایی نسبت به اسناد ۲۵۰ تایی و در نتیجه اسناد ۱۰۰۰ تایی نسبت به اسناد ۲۵۰ تایی دقت، صحت، و فراخوان بیشتری دارند. بنابراین، می‌توان نتیجه گرفت که هرچه شمار اسناد بیشتر باشد، معیارهای ارزیابی نیز بهبود پیدا می‌کند. این را یک پژوهش دیگر (Brindha et al., 2016) نیز تأیید کرده است. نکته این است که انتخاب مناسب تعداد نمونه‌های آموزشی برای یادگیری الگوریتم‌های دسته‌بندی لزوماً با توجه به مقادیر مورد انتظار پارامترهای ارزیابی مشخص می‌شود.

#### • مقایسه پارامترهای خلاصه‌ساز TF و ISF

برای مقایسه پارامترهای خلاصه‌ساز TF و ISF، با توجه به مقادیر جدول‌های ۳ تا

۶ در ۸۲ درصد حالات، برتری ارقام به‌دست‌آمده از پیاده‌سازی پارامتر ISF نسبت به روش TF مشهود است. به‌طور مثال، در جدول ۳ معیار صحت ۵۴/۶۷ درصد با روش خلاصه‌ساز TF و معیار صحت ۶۸ درصد با روش خلاصه‌ساز ISF برای اسناد ۲۵۰ تایی متن اصلی با پیاده‌سازی الگوریتم دسته‌بندی Tree به‌دست آمده است. کمترین مقدار صحت توسط پارامتر TF و الگوریتم‌های Rule و Tree با مقدار ۴۲/۶۷ درصد و بیشترین مقدار صحت نیز توسط پارامتر ISF و الگوریتم SVM با مقدار ۹۶/۶۷ درصد برآورد شد.

#### • مقایسه الگوریتم‌های دسته‌بندی بیزین، درخت تصمیم، قانون، و ماشین بردار پشتیبان

خروجی الگوریتم دسته‌بندی SVM که معیار فراخوان و دقت را برای هر برجسب به‌طور جداگانه محاسبه کرده است نشان می‌دهد نمی‌توان الگوریتم‌های دسته‌بندی را با این معیارها مقایسه کرد. در جدول‌های ۳ الی ۶، الگوریتم‌های دسته‌بندی بیزین، بردار پشتیبان، قانون، و درخت تصمیم در معیار صحت مقایسه شده است. از ۱۲ حالت مقایسه ۵۰ درصد برتری الگوریتم بیزین و ۴۱ درصد برتری الگوریتم بردار پشتیبان و ۹ درصد برتری الگوریتم قانون را نشان می‌دهد. پس می‌توان گفت الگوریتم بیزین و بردار پشتیبان تقریباً یکسانند و همان‌طور که در پژوهش‌هایی که از این دو الگوریتم استفاده شده، چنین بوده است.

#### • مقایسه متن اصلی و متن خلاصه

پیشتر گفتیم در دسته‌بندی و خلاصه‌سازی متون چینی (Jiang, Fan, & Chenn, 2007) و عربی (Thwain, 2014) که پژوهشگران از دسته‌بندی SVM استفاده کردند معیارهای ارزیابی نشان داد این شیوه درباره متن خلاصه بهتر از متن کامل عمل می‌کند. برای متون انگلیسی نیز در پژوهش (Jeong, Ko, & Seo, 2016) از ترکیب دو تکنیک خلاصه‌سازی و دسته‌بندی استفاده شد. از خلاصه‌سازی متن با استفاده از اطلاعات دسته‌بندی و از دسته‌بندی متن با استفاده از اطلاعات خلاصه‌سازی بهره گرفته شد. روش پیشنهادی آنها در مقایسه با شش روش خلاصه‌ساز استخراجی، عملکرد بهتری داشت که تأییدی بر تأثیر متقابل تکنیک‌های دسته‌بندی و خلاصه‌سازی در متون انگلیسی بود. در پژوهش (Ferreira, Simske, & Riss, 2015) مقایسه میان متن کامل و متن خلاصه‌شده با ۱۵ روش خلاصه‌ساز انجام و بعد از پیاده‌سازی الگوریتم دسته‌بندی

بیزین روی آنها، نتایج با معیار ارزیابی صحت سنجیده شد. بالاترین معیار ابتدا برای متن کامل و سپس متن خلاصه شده با روش خلاصه‌ساز اسم خاص و با تفاوت اندکی روش خلاصه‌سازی مبتنی بر فرکانس جمله و حروف بزرگ حاصل شد. در پژوهش حاضر با توجه به مقادیر جدول‌ها، در معیار صحت با روش خلاصه‌ساز ISF برای الگوریتم دسته‌بندی SVM روی اسناد ۱۰۰۰ تایی از متن کامل بهترین نتیجه با مقدار ۹۶/۶۷ درصد به دست آمد (جدول ۶). با مقایسه میان مقادیر به دست آمده در متن کامل و متن خلاصه، به طور مثال، برای اسناد ۲۵۰ تایی از متن کامل مقدار صحت ۸۰ درصد و برای اسناد ۲۵۰ تایی از متن خلاصه مقدار صحت ۵۳/۳۳ درصد به دست آمد. همین‌طور میان متن خلاصه و متن کامل از اسناد ۵۰۰ تایی و ۱۰۰۰ تایی و دیگر جدول‌ها، در ۸۸ درصد حالات، متن کامل دقت، صحت، و فراخوان بالاتری دارد. نتایج مقایسه در رشد ۱۰۰ درصدی اسناد، هرچه متن کامل‌تر باشد، احتمال دریافت برچسب صحیح‌تر بیشتر می‌شود. مهم‌تر اینکه با افزایش تعداد اسناد بررسی شده، معیارهای صحت، دقت، و فراخوان به‌طور چشمگیر افزایش یافته است. این امر لزوم استفاده از متون خلاصه را نشان می‌دهد در زمانی که تعداد اسناد دسته‌بندی شده بسیار زیاد است.

#### • مقایسه برچسب اسناد دسته‌بندی شده

چهار جدول ۷ تا ۱۰، برچسب «ورزشی» با بیش از ۸۰ درصد بهترین نتیجه، و برچسب «انتخابات» با ۲۰ مقدار صفر بدترین نتیجه را به دست دادند. جدول ۱، برچسب «ورزشی» بیشترین تعداد (۴۲۱) و برچسب «انتخابات» کمترین تعداد (۱۵) سند را داشتند که البته مقایسه نتایج بقیه برچسب‌ها به ما اجازه نتیجه‌گیری نمی‌دهد؛ زیرا برچسب «فرهنگی» با فراوانی ۱۹۰، کمتر از برچسب «اقتصاد» با فراوانی ۴۹ نتیجه دارد. در پژوهش دیگری نیز برچسب «ورزشی» بهترین و برچسب «جهان»، به سبب بیش از حد کلی بودنش، کمترین نتیجه را نشان داده است (Ferreira et al, 2015). بنابراین می‌توان گفت نتایج با محتوای برچسب نیز ارتباط مستقیم دارند. برچسب «انتخابات» به سبب نزدیکی محتوا با برچسب «سیاسی» و برچسب «فرهنگی» با داشتن تعداد ۱۳ صفر، به دلیل کلی بودنش، کمترین مقادیر را در جدول دارند.

#### نتیجه‌گیری

در این پژوهش مقایسه‌ای میان ترکیب تعدادی تکنیک خلاصه‌سازی و دسته‌بندی بر

متون فارسی انجام شد. با رشد ۱۰۰ درصدی تعداد اسناد، چهار تکنیک دسته‌بندی بیزین، درخت تصمیم، قانون، و بردار پشتیبان و دو تکنیک خلاصه‌سازی برمبنای پارامترهای TF و ISF و معیارهای ارزیابی دقت، صحت، و فراخوان روی متون اصلی و خلاصه اجرا شد. مقایسه نتایج نشان داد در ۸۴ درصد حالات، اسناد ۱۰۰۰ تایی از اسناد ۵۰۰ تایی و ۲۵۰ تایی، دقت، صحت، و فراخوان بیشتری دارند. بدیهی است با اسناد بیشتر برای مقایسه، میزان دقت در مقایسه نیز افزایش می‌یابد. مقایسه بین پارامترهای خلاصه‌ساز TF و ISF نیز نشان داد در ۸۲ درصد حالات، ارقام به دست آمده از پیاده‌سازی پارامتر ISF نسبت به روش TF برتر است. مقایسه میان معیار صحت در الگوریتم‌های دسته‌بندی، ۵۰ درصد برتری الگوریتم بیزین، ۴۱ درصد برتری الگوریتم بردار پشتیبان، و ۹ درصد برتری الگوریتم قانون را نشان داد. همچنین در ۸۸ درصد حالات، متن کامل دقت، صحت، و فراخوان بالاتری نسبت به متن خلاصه دارد و با افزایش تعداد اسناد معیارهای صحت، دقت، و فراخوان نیز به‌طور چشمگیر افزایش می‌یابد که این امر لزوم استفاده از مزایای متون خلاصه را در حجم بالا نشان می‌دهد.

در مقایسه نتایج دو پارامتر دقت و فراخوان در دسته‌بندی برچسب‌ها، برچسب «ورزشی» بهترین نتیجه و برچسب «انتخابات» بدترین نتیجه را نشان دادند. در مقایسه بین متن کامل و متن خلاصه بیشترین مقدار مربوط به الگوریتم دسته‌بندی بردار پشتیبان برای اسناد متن کامل ۱۰۰۰ تایی با مقدار ۹۶/۶۷ درصد در معیار صحت حاصل شد.

نتایج به دست آمده از این پژوهش، حاکی از تأثیر مثبت استفاده از تکنیک‌های خلاصه‌سازی در کارایی الگوریتم‌های دسته‌بندی متون فارسی است (حداکثر مقدار ۸۸/۴۴ درصد در معیار صحت الگوریتم دسته‌بندی بیزین توسط پارامتر TF در اسناد ۵۰۰ تایی متن خلاصه).

یافته‌های این پژوهش می‌تواند در مطالعات آینده برای یکپارچه‌کردن تکنیک‌های خلاصه‌سازی و دسته‌بندی نقش مؤثری داشته باشد و زمان لازم را برای انجام دسته‌بندی متون فارسی کاهش چشمگیر دهد. ریشه‌یابی و استفاده از ویژگی‌های معنایی و ادراکی در مرحله پیش‌پردازش داده‌ها می‌تواند بهبود روش پیشنهادی ما را بهبود بخشد. همچنین، برای داشتن افزایش تعداد مقایسه‌های بیشتر می‌توان پارامترهای خلاصه‌سازی دیگر مانند عنوان، موقعیت، شبه ربط، و مقدار اطلاعات متنی را نیز در بررسی منظور کرد و پیاده‌سازی آنها را بر متون چندسندی ارزیابید.

## مآخذ

- آهنگری، فاطمه (۱۳۹۶). معرفی خلاصه‌ساز خودکار متون فارسی مبتنی بر الگوریتم‌های فراابتکاری. پایان‌نامه کارشناسی ارشد، دانشگاه گلستان، گرگان.
- احمدی، سیدمحمدحسین (۱۳۹۰). *دسته‌بندی موضوعی متون فارسی براساس روش قواعد انجمنی*. پایان‌نامه کارشناسی ارشد، دانشگاه پیام نور، تهران.
- شورای عالی اطلاع‌رسانی (۱۳۸۸). بررسی مستندات ابزارهای خودکار خلاصه‌سازی زبان‌های دنیا برای به‌کارگیری در خلاصه‌سازی متون زبان فارسی، طرح جامع ایجاد پیکره زبان فارسی با موضوع ایجاد پیکره متنی زبان فارسی (ویرایش ۱). بازیابی ۲ آبان ۱۳۹۸، از [http://www.prosody.ir/attachments/059\\_26-Summerization.pdf](http://www.prosody.ir/attachments/059_26-Summerization.pdf)
- غضنفری، مهدی؛ علیزاده، سمیه؛ و تیمورپور، بابک (۱۳۹۳). *داده کاوی و کشف دانش*. تهران: دانشگاه علم و صنعت ایران.
- Brindha, S., Prabha, K., & Sukumaran, S. (2016). A survey on classification techniques for text mining. In *3rd International Conference on Advanced Computing and Communication Systems (ICACCS)*, January 22-23. Retrieved October 9, 2019, from <https://ieeexplore.ieee.org/document/7586371>
- Ferreira, R., Simske, S., & Riss, M. (2015). *Automatic document classification using summarization strategies*. In DocEng'15, September 8-11, (pp. 69-72). New York, N.Y.: ACM.
- Han, J., & Kamber, M. (2012). *Data minin: Concepts and techniques* (3<sup>rd</sup> ed.). Waltham: Morgan Kaufmann Publisher,.
- Jeong, H., Ko, Y., & Seo, J. (2016). How to improve text summarization and classification by cooperation on an integrated framework. *Expert Systems with Applications*, 60 (C), 222-233.
- Jiang, X., Fan, X., & Chen, K. (2007). Chinese text classification based on summarization technique. In *Third International Conference on Semantics, Knowledge and Grid, October 29-31*, (pp. 362-365). Retrieved October 20, 2019, from <https://ieeexplore.ieee.org/document/4438570>
- Rahman, N., & Borah, B. (2015). A survey on existing extractive techniques for query-based text summarization. In *International Symposiwn on*



*Advanced Computing and Communication (ISACC)*, September 14-15, (pp. 98-102). Retrieved October 20, 2019, from <https://ieeexplore.ieee.org/document/7377323>

Thwaib, E. (2014). Text summarization as Feature Selection for Arabic Text Classification. *World of Computer Science and Information Technology Journal*, 4 (7), 101-104.

### استناد به این مقاله:

عرب‌احمدی، فاطمه زهرا؛ کرباسی، سهیلا (۱۳۹۸). تأثیر تکنیک‌های خلاصه‌سازی بر دسته‌بندی متون فارسی. *مطالعات ملی کتابداری و سازماندهی اطلاعات*، ۳۰ (۳)، ۲۳-۸.

