دسته بندی متون بوسیله تکنیک های ترکیبی داده کاوی

عادل حیدری*[۱]، سید محسن هاشمی[۲]،عارف سیاحی[۳]، پیمان جلالی [۴]

۱- آموزشکده فنی و حرفه ای سما،دانشگاه آزاد اسلامی، واحد سوسنگرد، سوسنگرد، ایران

۲- آموزشکده فنی و حرفه ای سما،دانشگاه آزاد اسلامی، واحد سوسنگرد، سوسنگرد، ایران

۳- آموزشکده فنی و حرفه ای سما،دانشگاه آزاد اسلامی، واحد سوسنگرد، سوسنگرد، ایران

۴- آموزشکده فنی و حرفه ای سما،دانشگاه آزاد اسلامی، واحد سوسنگرد، سوسنگرد، ایران

خلاصه

با توجه به گستردگی حجم اطلاعات متنی الکترونیکی که به طور قابل توجهی از طریق اینترنت و سایر منابع قابل دسترسی می باشند، در صورت نبودن ایندکس گذاری و دسته بندی مناسب، کار بازیابی و پردازش اطلاعات متنی دسته بندی نشده با مشکلات زیادی مواجه می گردد .دسته بندی متون، کاربردهای زیادی از جمله پیگیری اسناد، مدیریت اسناد، گسترش اسناد و کاهش حجم اطلاعات دارد .روش های یادگیری زیادی در زمینه دسته بندی متن ها در سالهای اخیر بکار برده شده است. بررسی‌ها و مطالعات نشان می‌دهند که روش‌های دسته‌بندی متون مانند بیز ساده و ماشین بردار پشتیبان نمی توانند کیفیت دسته‌بندی را از یک حدی بیشتر افزایش دهند اما با بهره‌گیری از روش‌های ترکیبی، می‌توان کیفیت دسته‌بندی را ارتقا داد. در این پژوهش سعی خواهد شد تا یک مدل دسته‌بندی خودکار را با استفاده از ترکیبی از الگوریتم‌ها و تکنیک‌های متن‌کاوی باکار آیی و دقت بالا دسته‌بندی کرد.

کلمات کلیدی: متن، متن کاوی، روش ترکیبی، رای گیری

۱. مقدمه:

دسته بندی متن یک مسئله پردازش زبان طبیعی می باشد و می تواند به عنوان انتساب سندهای دسته بندی نشده به یک یا بیشتر دسته های از پیش تعریف شده، بر اساس محتوایشان دیده شود. دسته بندی متن خودکار در مصرف زمان و هزینه بسیار مفید می باشد و 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[۱۲].

با توجه به حجم بالای اطلاعات و منابع، نیاز به افزایش سرعت و دقت در مواردی همچون جستجوها، بایگانی اسناد و مدارک، دسته‌بندی مقالات جدید و دسته‌بندی صفحات وب، کاملاً محسوس است. پس مهم است که بتوانیم این منابع را به‌صورت دقیق و با سرعت‌بالا دسته‌بندی کنیم. در پژوهش‌های پیشین، طبقه‌بندی‌های مختلف و با نگرش‌های متفاوتی

[1] heidari.adel1@gmail.com

برای این کار مطرح‌شده است. نکته قابل‌توجه، امکان خودکارسازی این طبقه‌بندی‌ها به ازای متون جدید است. درزمینه پردازش زبان طبیعی و به‌خصوص پردازش متن، یکی از پایه‌ای‌ترین کارها دسته‌بندی یا طبقه‌بندی خودکار متون است. شناسایی رده، دسته یا طبقه یک متن می‌تواند اطلاعات مفیدی برای فرایندهای همچون ترجمه ماشینی، تبدیل نوشتار به گفتار و غیره فراهم کند.

۲. دسته بندی متون

دسته‌بندی می‌تواند به سه صورت دسته‌بندی دودویی[1]، دسته‌بندی چند دسته[2] و دسته‌بندی چند برچسبی[3] پیاده‌سازی گردد [۹].

- در دسته‌بندی دودویی سند متنی تنها به یکی از دودسته موجود متعلق است. بنابراین دسته‌بند باید سند متنی را به یکی از دسته‌ها نسبت دهد.
- در مورد دسته‌بندی چند دسته، چندین دسته وجود دارد که در این مورد، سند متنی تنها به یکی از دسته‌های از پیش تعریف‌شده موجود تعلق دارد.
- در مورد دسته‌بندی چند برچسبی همانند دسته‌بندی چند دسته، چندین دسته مختلف وجود دارد که یک سند متنی می‌تواند به یک یا تعدادی از دسته‌ها تعلق داشته باشد. به عبارتی می‌توان گفت که دسته‌ها ممکن است با یکدیگر تداخل داشته باشند.

۳. مروری بر کارهای پیشین در زمینه دسته بندی متون

جدول ۱: مروری بر روش‌ها ارائه‌شده پیشین

نام روش	ویژگی	مزایا	معایب
در مرجع[۳-۱] ترکیب SVM Rochio -K-NNM-K-NN	ترکیب دسته‌بندها	حالت بهتر با ترکیب SVM و K-NNM	عدم سازگاری برخی از الگوریتم‌های ترکیبی
در مرجع [4] استفاده روش Attribute Bagging	روش پوششی	ترکیب با هر الگوریتم یادگیری بهبود پایداری و دقت	زمان‌بر بودن
در مرجع[۵] روش SVM خطی و غیرخطی	دسته‌بندی خطی	افزایش کارایی در دسته‌بندی خطی	عدم کارایی مناسب درروش غیرخطی
در مرجع[۶] استفاده روش ژنتیک – تحلیل مؤلفه‌های اصلی و روش بهره اطلاعات	استفاده از دو روش کاهش ویژگی فیلتری و پوششی	کم نمودن ویژگی‌ها با روش بهره اطلاعات	روش تحلیل مؤلفه‌های اصلی ضعیف‌تر از روش بهره اطلاعات می‌باشد
در مرجع[۷] ترکیب Naïve Bayes- K-NN- Rochio	ترکیب دودویی آنان همراه بارأی گیری	کارایی نسبت به روش‌های ساده نسبتاً بهتر	استفاده از اسناد کم در مقاله از اطمینان کافی برخوردار نیست.
در مرجع[۸] روش Boosting - Bagging	روش‌های ترکیبی	افزایش کارایی	داشتن مجموعه آزمایشی زمان‌بر بودن ترتیبی بودن الگوریتم

[1] Binary Classification
[2] Multi-Class Classification
[3] Multi-Label Classification

۴. روش پیشنهادی:

استفاده از دسته‌بند ترکیبی باعث افزایش کارایی دسته‌بندی توسط ترکیب چند دسته‌بند منفرد می‌شود. روش‌های ترکیبی به دو صورت ترکیب ویژگی‌ها و ترکیب خروجی‌ها می‌باشند. روش پیشنهادی بر روی ترکیب خروجی دسته‌بندها عمل می‌کند که نسبت به ترکیب ویژگی‌ها از دو جنبه برتری دارد: اول اینکه درصورت افزایش بردار ویژگی موجب افزایش پیچیدگی روش ترکیب نمی‌شود و دوم اینکه به دلیل این که نیازی به دانستن ساختار دسته‌بندها و بردارهای ویژگی آن‌ها نمی‌باشد، دارای پیچیدگی کمتری می‌باشد. روش پیشنهادی از نوع همگن می‌باشد بدین معنی که از دسته‌بندهای یکسان استفاده می‌کند. روش پیشنهادی با استفاده از نمونه‌گیری‌های متفاوت همراه با جایگزینی[1] از مجموعه آموزشی، چندین مجموعه آموزشی جدید را به دست می‌آورد و بدین طریق از هر کدام یک دسته‌بند جداگانه را آموزش می‌دهد. این کار باعث افزایش کارایی در روش پیشنهادی می‌شود.

۴.۱. مجموعه داده:

اسناد ورودی یا مجموعه داده انتخابی، مجموعه داده اخبار رویترز-۲۱۵۷۸ می‌باشد[۱۲]. این مجموعه داده در سال ۱۹۸۷ گردآوری و توسط گروهی از کارکنان خبرگزاری رویترز تهیه و شاخص‌گذاری شد.. با توجه به اینکه در این تحقیق، روش تک برچسبی - چند دسته مورد توجه است، از زیرمجموعه R (8) استفاده می‌شود. این زیرمجموعه شامل ۸ دسته اصلی و دارای ۷۶۷۴ سند متنی می‌باشد. هر سند متنی تنها به یکی از دسته‌های موجود تعلق دارد. با توجه به روش جداسازی اسناد آموزشی از اسناد آزمایشی ModApte Split، مجموعه اسناد آموزشی شامل ۵۴۸۵ سند و مجموعه اسناد آزمایشی شامل ۲۱۸۹ سند می‌باشد. جزئیات زیرمجموعه R (8) اخبار رویترز در جدول (۱) نشان داده شده است.

جدول (۱): جزئیات زیرمجموعه اخبار رویترز

دسته	تعداد اسناد آموزشی	تعداد اسناد آزمایشی	تعداد کل اسناد دسته
Acq	1596	696	2292
Crude	253	121	374
Earn	2840	1083	3923
Grain	41	10	51
Interest	190	81	271
Money-fx	206	87	293
Ship	108	36	144
Trade	251	75	326

۴.۲. پیش پردازش:

برای دسته‌بندی متون پس از فراهم کردن مجموعه داده، نوبت به مرحله پیش‌پردازش می‌رسد. در این مرحله اسناد متنی خام باید به فرمی که قابل استفاده در مرحله انتخاب ویژگی و الگوریتم یادگیری باشند، تبدیل شوند. در مرحله پیش‌پردازش روش پیشنهادی، عملیات Transform Case، Tokenization، Filter StopWords، Stemming و Generate n-gram انجام می‌شود. عملکرد هر کدام در زیر بیان شده است.

[1] Sampling with Replacing

Transform Case: در این مرحله تمامی کاراکترهای موجود در متن به فرم یکسان تبدیل می‌شوند. در این مرحله تمامی کاراکترها به حروف کوچک تبدیل می‌شوند.

Tokenization: در این مرحله کل متن به صورت کلمات متوالی جدا از هم تقسیم می‌شود.

Filter StopWords: در این مرحله کلمات زائد در زبان انگلیسی حذف خواهند شد.

Stemming: در این مرحله ریشه کلمات در زبان انگلیسی با استفاده از الگوریتم ریشه‌یابی پورتر برای حذف پسوندها و پیشوندهای کلمات در جهت کاهش طول کلمات و تا زمانی که به فرم ریشه‌شان تبدیل شوند، استفاده شده است.

Generate n-gram: در این مرحله برای شاخص‌بندی و کاهش ابعاد متن از n-gram استفاده شده است. با استفاده از n-gram می‌توان متن را به صورت یک سری از کلمات متوالی به طول n نشان داد. این مدل ابتدا برای مسائل پردازش گفتار مطرح شد. ولی اکنون نگارش‌های مختلفی از این مدل برای مسائل پردازش زبان طبیعی و دسته‌بندی متون مطرح شده است[۱۶و۱۵]. با توجه به آزمایش‌های انجام‌شده بر روی مقادیر مختلف n و برای جلوگیری از افزایش پیچیدگی، از n-gram با میزان n=2 استفاده شد.پس از مراحل فوق، وزن‌دهی ویژگی‌ها انجام می‌شود. در این تحقیق از روش وزن‌دهی TF-IDF استفاده شده است که عملکرد دسته‌بندها با استفاده از این روش وزن‌دهی مورد ارزیابی قرار می‌گیرند.

۴.۳.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 [۱۲].

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

۴.۴.مرحله یادگیری

پس از انتخاب ویژگی‌های مناسب، با استفاده از نمونه‌گیری‌های متفاوت همراه با جایگزینی از مجموعه آموزشی، چندین مجموعه آموزشی جدید به هرکدام از دسته‌بندها داده می‌شود. سپس خروجی دسته‌بندها با یکدیگر ترکیب می‌شود. در مرحله ترکیب از رأی‌گیری استفاده‌شده است. سپس ارزیابی صورت گرفته و کارایی دسته‌بندی بر اساس معیارهای مختلف موردبررسی قرار می‌گیرد.

۵. نتایج

جدول (۲): مقایسه روشهای ترکیبی

	Proposed SVM	Proposed NB	Proposed J48
Average Precision	99.39	81.47	86.24
Average Recall	99.20	91.26	88.58
Average F1	99.46	85.58	86.20
Accuracy	99.88	94.89	97.13
Error	0.12	5.11	2.87

در معیار میانگین دقت، بهترین کارایی مربوط به مدل پیشنهادی با استفاده از روش یادگیری ماشین بردار پشتیبان با میزان ٪۹۹.۳۹ است و ضعیف‌ترین عملکرد را مدل Naive Bayes با میزان ٪۸۱.۴۷ خود نشان می‌دهد. در معیار میانگین فراخوانی بهترین کارایی مربوط به مدل پیشنهادی با استفاده از روش یادگیری ماشین بردار پشتیبان با میانگین ٪۹۹.۲۰ است. ضعیف‌ترین عملکرد را نیز مدل J48 با میزان ٪۸۸.۵۸ از خود نشان می‌دهد. در معیار میانگین F1 نیز بهترین کارایی مربوط به مدل پیشنهادی با استفاده از روش یادگیری ماشین بردار پشتیبان با میانگین ٪۹۹.۴۶ است. ضعیف‌ترین عملکرد را مدل Naive Bayes با میزان ٪۸۵.۵۸ از خود نشان می‌دهد. درستی با میزان ٪۹۹.۸۸ و خطای دسته‌بندی با روش یادگیری ماشین بردار پشتیبان و معیار Classification ERROR نیز بهترین عملکرد متعلق به مدل پیشنهادی با استفاده از روش یادگیری ماشین بردار پشتیبان با میزان ٪۰.۱۲ است. با توجه به ارزیابی‌های انجام‌شده، مدل پیشنهادی با استفاده از دسته‌بند ماشین بردار پشتیبان بهترین عملکرد را نسبت به سایر مدل‌ها دارد.

۶. نتیجه گیری:

در سال‌های اخیر استفاده از تکنیک‌های متن و الگوریتم‌های هوشمند فراگیر شده است. بسیاری از کارهایی که درگذشته با صرف هزینه‌های زمانی و مالی فراوان حاصل‌شده‌اند، می‌تواند توسط این تکنیک‌ها و الگوریتم‌ها انجام گردد. از سویی دیگر، بسیاری از منابع ما منابع متنی هستند که می‌توان آن‌ها را موضوع‌بندی و طبقه‌بندی کرد. گسترش روزافزون اطلاعاتی که بشر در اختیار دارد، مساله سازماندهی خودکار این اطلاعات اهمیت ویژه ای می یابد. در این میان کار دسته بندی اسناد متنی در گروه های جداگانه بعنوان یک مساله مرکزی باید مورد بحث و بررسی قرار گیرد. دسته بندی بعنوان یک روش مهم در آنالیز داده ها مطرح است و روش های متعددی درهوش مصنوعی و شناسایی آماری الگو برای این کار پیشنهاد شده است اما استفاده مستقیم از این روشها در کار دسته بندی متن امکان پذیر نمی‌باشد چرا که در این مساله با تعداد زیادی مشخصه روبرو خواهیم بود. در این پژوهش یک روش ترکیبی که از روش فیلتری IG برای انتخاب ویژگی بهره اطلاعات استفاده شده است.

مراجع:

1) Azam, N., Yao, J. T., 2012. Comparison of Term Frequency and Document Frequency Based Feature Selection Metrics in Text Categorization, Expert Systems with Applications, 39 (5), pp. 4760-4768.
2) Baoli, L., Shiwen, Y., Qin, L., 2003. An improved k-nearest neighbor algorithm for text categorization, In Proceedings of the 20th International Conference on Computer Processing of Oriental Languages, China.
3) Bell, D. A., Guan, J. W., Bi, Y. X., 2005. An evidential approach to classification combination for text categorization, Studies in Fuzziness and Soft Computing, 185, pp. 13-22.
4) Bryll, R., Osuna, R. G., Quek, F., 2003. Attribute Bagging: improving accuracy of classifier ensembles by using random feature subsets, Pattern Recognition, 36, pp. 1291-1302.
5) Klinkenberg, R., Joachims, T., 2000. Detecting concept drift with support vector machines, In Proceedings of the 17th International Conference on Machine Learning, Stanford, USA, pp. 487-494.
6) Uguz, H., 2011. A Two-Stage Feature Selection Method for Text Categorization by Using Information Gain, Principal Component Analysis and Genetic Algorithm, Knowledge-Based Systems, 24, pp. 1024-1032.

7) Larkey, L. S., Croft, W. B., 1996. Combining classifiers in text categorization, In Proceedings of the 19th ACM International Conference on Research and Development in Information Retrieval (Switzerland), pp. 289-297.
8) Sebastiani, F., 2002. Machine Learning in Automated Text Categorization, ACM Computing Surveys, 34 (1), pp. 1–47.
9) Arturo, M. R., 2006. Automatic Text Categorization of documents in the High Energy Physics domain, Ph. D. Thesis, University of Granada.
10) Peng, F., Schuurmans, D., Wang, S., 2003. Language and Task Idependent Text Categorization with Simple Language Models, In Proceedings of the HLT-NAACL.
11) Lan, M., Tan, C. L., 2007. Supervised and Traditional Term Weighting Methods for Automatic Text Categorization, Journal of IEEE Pami, 10 (10), pp. 1-36.
12) Yang, Y., & Liu, X. (1999, August). A re-examination of text categorization methods. In Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval (pp. 42-49). ACM.
13) Dataset, available in: http://archive.ics.uci.edu/ml/datasets/Reuters-21578+Text+Categorization+Collection, Retrived on 28 July 2013.
14) Peng, F., Schuurmans, D., Wang, S., 2003. Language and Task Idependent Text Categorization with Simple Language Models, In Proceedings of the HLT-NAACL.
15) Wei, Z., Miao, D., Hugues, J., Zhao, R., Li, W., 2009. N-grams based feature selection and text representation for Chinese Text Classification, International Journal of Computational Intelligence Systems, 2 (4), pp. 365-374.