طبقه‌بندی خودکار متون فارسی

بهاره بینا[۱]،مسعود رهگذر[۲]، آذین ده‌موبد[۳]
(b.bina[۱], a.dehmoubed[۳])@ece.ut.ac.ir
قطب علمی کنترل و پردازش هوشمند، پردیس دانشکده‌های فنی، دانشکده برق و کامپیوتر،دانشگاه تهران
rahgozar[۲]@ut.ac.ir

چکیده

طبقه بندی خودکار متون از موارد کاربرد الگوریتم‌های یادگیری ماشینی در مبحث بازیابی اطلاعات می‌باشد. در این مقاله نیز نتایج طبقه بندی خودکار متون فارسی با استفاده از معیارهای شاخص‌گذاری ۳-gram, ۴-gram و کلمه ارائه شده‌است . در ضمن نتایج در دو حالت با حذف stop word و بدون حذف stop word های متون نیز مقایسه شده‌اند. به منظور دسته‌بندی متون از الگوریتم یادگیری ماشینی نزدیک‌ترین k همسایه(knn) استفاده شده است و در نهایت به منظور ارزیابی و مقایسه نتایج، دو معیار دقت و یادآوری برای هر روش شاخص‌گذاری نیز محاسبه شده‌اند. نتایج بدست آمده نشان داد که بهترین روش شاخص‌گذاری متون فارسی ۴-gram می‌باشد و حذف stop word ها نتایج را اندکی بهبود می‌بخشد.

کلمات کلیدی

طبقه بندی متن فارسی، شاخص گذاری n-gram ، الگوریتم یادگیری ماشینی، نزدیکترینk همسایه

۱- مقدمه

در دنیایی زندگی می کنیم که اطلاعات ارزش زیادی برای ما دارند. با افزایش حجم اطلاعات در دسترس روی اینترنت، نیاز فوق العاده به ابزارهایی که بتوانند در جستجو، فیلتر نمودن و مدیریت منابع کمک کنند کاملا محسوس است.

طبقه بندی متون ، فرآیندی است که در آن متن ها را به یک یا چند طبقه از قبل تعریف شده بر اساس محتوا یا زبان نگارش متن نسبت می‌دهیم[۹].

طبقه‌بندی ایمیل‌ها، تشخیص موضوع، فیلتر نمودن متون از جمله موارد کاربرد سیستم طبقه‌بندی خودکار متون می‌باشند[۶].

برخی از الگوریتم‌هایی که بر اساس خواص آماری متون و بر اساس الگوریتم‌های یادگیری ماشینی[۱] نیز در این مقوله استفاده شده‌اند عبارتند از : نزدیک ترین همسایه[۶] ، درخت تصمیم گیری[۶] ، طبقه بندی بیزی[۲] ، تشابه بر اساس بازخورد [۶]، شبکه های عصبی [۱] و غیره.

بیشتر سیستم‌های طبقه بندی خودکار متون[۲]، برای متون زبان انگلیسی طراحی شده‌اند و معمولاً قابل استفاده برای متون فارسی نیستند. توسعه سیستم طبقه بندی خودکار متون فارسی به دلیل ماهیت زبان فارسی و در دسترس نبودن مجموعه‌ای شامل ریشه کلمات و کلمات پر کاربرد زبان و مجموعه‌ای برای تست سیستم کاری نسبتاً دشوار می‌باشد .

برخی از تحقیقاتی که اخیرا بر روی متون فارسی در زمینه بازیابی اطلاعات انجام شده، عبارتند از: مقاله[۱۱] که در آن شش روش بازیابی اطلاعات با استفاده از پیکره متنی همشهری ارزیابی شده‌است. روش های پیاده سازی شده در این مقاله عبارتند از Vector Space(دو روش) و Language Modeling(چهار روش). مقاله [۱۲] که در آن مدل بازیابی N-gram Vector Space با دو روش وزن دهی موسوم به (atc.atc) و (Lnu.ltu) بررسی و با روش LCA بهبود داده شده است و مقاله [۱۳] که در آن ساخت یک ریشه یاب فارسی شرح داده شده است. در زمینه طبقه بندی خودکار متون فارسی تا به حال گزارش یا مقاله‌ای منتشر نگردیده، لذا در این تحقیق به بررسی این موضوع پرداخته‌ایم.

در این مقاله، به منظور طبقه بندی خودکار متون فارسی از سه روش شاخص گذاری[۳] متون ۳-gram,۴-gram و کلمات استفاده شده‌است و الگوریتم یادگیری ماشینی موسوم به نزدیک‌ترین k همسایه[۴] بکار گرفته شده است. از سه معیار تشابه دایس و مانهاتن و ضرب داخلی نیز استفاده گردیده‌است. نتایج حاصله نشان داده‌است که بهترین روش شاخص گذاری متون ۴-gram با معیار تشابه ضرب داخلی می‌باشد.

برای طبقه بندی متون فارسی از مجموعه ۴۰۰۰ مقاله‌ی روزنامه همشهری استفاده شده‌است. این مقالات در مجموعه ای به نام پیکره همشهری گردآوری شده‌اند که در مرجع [۱۰] روش ساخت و اطلاعات

آماری پیکره متنی توضیح داده شده است. همچنین بعضی پارامتر های زبان فارسی مانند تبعیت با قانون Zipf و میزان آنتروپی آن نیز در این مقاله محاسبه شده‌اند.

از مجموعه مقالات منتخب از پیکره با توجه به محتوای آنها (هفت طبقه‌ی سیاسی، اجتماعی، خارجی، حوادث، ورزشی، اقتصادی و علم و فناوری)۷۰٪ به عنوان مجموعه یادگیری در نظر گرفته شد. در مجموعه یادگیری طبقه هر کدام از متن ها از قبل مشخص است و ۳۰٪ مقاله‌ها در مجموعه تست جای گرفتند که طبقه بندی خودکار روی آنها انجام می‌پذیرد.

طبقه بندی خودکار متون شامل دو فاز استخراج خصوصیات[۵] از متن و فاز یادگیری[۶] [۷] می‌باشد.

در ادامه، در قسمت دوم این مقاله به شرح استخراج خصوصیات پرداخته و در قسمت سوم نیز فاز یادگیری توضیح داده شده‌است . در قسمت چهارم مقاله نتایج، و درانتهانتیجه گیری آورده‌شده‌است.

۲- استخراج خصوصیات از متون فارسی

این مرحله شامل فازهای آماده سازی متون و شاخص گذاری مستندات و وزندهی شاخص‌ها می‌باشد.

۲-۱- فاز آماده سازی متون

در فاز آماده سازی متون، متن فارسی که شامل کاراکترهای پشت سر هم است به نمایشی که برای الگوریتم های یادگیری و طبقه بندی مناسب باشد تبدیل می‌شود.

این فرآیند در ابتدا معمولا شامل موارد زیر است[۳] :

- حذف tag های html یا xml
- کدگذاری[۷] متون به "utf-۸"
- حذف Stop word ها و علائم نگارشی
- بدست آوردن ریشه کلمات[۸] و حذف پیشوندها و پسوندها

Stop word ها کلماتی هستند که زیاد تکرار می شوند و حاوی اطلاعات نیز نمی‌باشند، مثلا حروف اضافه ، افعال ربطی ، کلمات وصل و...

در زیر اسامی ۱۶ Stop word پر استفاده آورده شده است :

و-در-به-از-که-را-آن-با-بودن-هر-برای-بر-داشتن-کردن-آنها-یا

۲-۲- شاخص گذاری مستندات

در فاز شاخص‌گذاری مستندات، یک متن d_j را به صورت برداری از وزن‌ها $\vec{d}_j = < w_{1j}, \ldots, w_{|T|j} >$ نمایش می‌دهند که در آن T مجموعه عبارات موجود در کل مستندات است و W_{kj} نشان دهنده اهمیت عبارت t_k در مستند d_j می‌باشد[۲]. روش‌های مختلف شاخص-گذاری در یکی از موارد زیر با هم تفاوت دارند :

- آنچه به عنوان عبارات شناسایی می‌شود و به آن وزن اختصاص داده می‌شود.
- روش وزن دادن به عبارات مختلف و پارامترهای موثر در این وزن‌دهی.

در این مقاله از روش شاخص گذاری متن بصورت کلمات ساده و روش N-gram (۳-gramو ۴-gram)استفاده شده است که در ادامه روش N-gram توضیح داده شده است.

۲-۲-۱ روش N-gram

در این روش، شاخص‌گذاری به‌صورت ترتیبی از n حرف پشت سر هم می‌باشد. یک کلمه متن بصورت مجموعه ای از N-gram ها که با هم، همپوشانی دارند نشان داده می‌شود[۷]. بعنوان مثال کلمه "سلام " از N-gram های زیر تشکیل شده است :

tri-gram: -سل ، سلا،لام ، ام-

quad-gram: -سلا ، سلام ، لام-

"-" نمایش‌دهنده فاصله می‌باشد. مزیت N-gram با توجه به طبیعت آن می‌باشد . چون هر رشته از تعداد محدودی از کلمات تشکیل شده است ، خطاها منتشر نمی‌شوند و روی تعداد محدودی از رشته ها اثر می‌گذارند.

بر اساس قانون Zipf[۱۴] :

اگر f تعداد رخداد کلمه‌ای باشد و r رتبه آن کلمه در یک زبان باشد . آنگاه رابطه (۱) برقرار می‌باشد:

$$f = \frac{k}{r} \qquad (۱)$$

k نیز مقدار ثابتی می‌باشد.

از این رو قابل ذکر است که متونی که توزیع رخداد یکسانی از کلمات یا N-gram ها دارند متعلق به یک طبقه هستند و محتوای یکسانی نیز دارند.

۲-۳- وزن‌دهی به خصوصیات‌استخراجی

چندین روش برای پیدا کردن وزن یک عبارت متن وجود دارد که بیشتر این روش ها بر اساس دو قانون زیر می‌باشند[۱۵]:

- هر چه یک عبارت در متن بیشتر تکرار شود بیشتر با موضوع متن مرتبط است.
- هر چه یک عبارت بیشتر در همه متن ها تکرار شود اهمیت کمتری دارد و وزن کمتری می‌گیرد.

در این تحقیق از روش زیر برای وزندهی شاخص‌ها استفاده‌شده‌است :

$$w = \frac{\log(freq)}{(1+\log(avgTF))} \times \log\left(\frac{n}{filecount}+1\right) \qquad (۲)$$

freq: تعداد رخداد عبارت در متن مزبور

avgTF: میانگین رخداد عبارات

n: تعداد متون

filecount: تعداد رخداد عبارت در متنها

در این روش وزندهی، هر دو قانون بیان شده در فوق در نظر گرفته شده‌اند و log دوم بیانگر این امر است که هر چه عبارتی در متن‌های بیشتری تکرار شود وزن کمتری به خود اختصاص دهد.

۳- فاز یادگیری

این فاز شامل مراحل تعریف معیار شباهت، الگوریتم یادگیری و ارزیابی کارآیی می‌باشد.

۳-۱- معیار شباهت

معیارهای شباهت برای مقایسه بین متن‌های مجموعه تست و متن‌های مجموعه یادگیری در حین اجرای الگوریتم یادگیری بکار می‌روند. در اینجا از معیارهای شباهت ضرب نقطه‌ای[۹]، معیار تشابه دایس[۳][۱۰] و فاصله منهاتن[۳][۱۱]استفاده شده‌است. تعریف این معیارهای شباهت در زیر آمده‌است:

$$\text{Manhattan} = (P_i, P_j) = \sum_{h=1}^{k} |P_{ih} - P_{jh}| \quad (۳)$$

$$\text{Similarity Dice } (P_i, P_j) = \frac{|P_{ih} \cap P_{jh}|}{|P_i| + |P_j|} \quad (۴)$$

$$\text{Dot Product} = \sum_{h=1}^{k} |P_{ih} \times P_{jh}| \quad (۵)$$

۳-۲- الگوریتم طبقه‌بندی خودکار متون

در این مقاله از الگوریتم نزدیک‌ترین k همسایه (knn) استفاده شده است. در این روش k تا از نزدیکترین متون به متن d را با توجه به وزن آن متون از مجموعه یادگیری انتخاب می‌کنیم و با توجه به طبقات آن k متن، تصمیم‌گیری در مورد تعلق متن d به طبقات را انجام می‌دهیم.

قابل ذکر است که در مقاله‌هایی که الگوریتم‌های مختلف طبقه بندی متون انگلیسی را با هم مقایسه نموده‌اند، این الگوریتم نتایج بسیار خوبی داشته‌است[۲]. به همین دلیل در اینجا نیز از این الگوریتم استفاده شده است. در اینجا از چندین مقدار k استفاده شده است از جمله ۱،۱۰،۱۵،۲۰،۲۵ مقدار k نشان‌دهنده اینست که در لیستی از کلاس ها که به یک متن نسبت می دهیم از چند مقاله برای تصمیم گیری کلاس واقعی استفاده نماییم.

۳-۳-ارزیابی کارآیی الگوریتم

به منظور ارزیابی روش استفاده شده و مقایسه شاخص‌گذاری های مختلف استفاده شده از دو معیار دقت[۱۲] و یادآوری[۱۳] [۱] استفاده شده‌است . در زیر تعریف این دو معیار آمده‌است :

$$recall = \frac{a}{a+c} \quad (۶)$$

$$precision = \frac{a}{a+b} \quad (۷)$$

a – تعداد متونی که درست به یک طبقه منسوب شده‌اند.

b - تعداد متونی که نادرست به یک طبقه منسوب شده‌اند.

c - تعداد متونی که نادرست از یک طبقه رد شده‌اند.

و در نهایت برای ارزیابی کارایی روی تمام طبقات از روش میانگین گیری کلان[۱۴] [۱] استفاده شده‌است. در میانگین گیری کلان میانگین مقادیر دقت و یادآوری تمام طبقات محاسبه می‌شود. در این روش به همه طبقات وزن مساوی داده می‌شود .

۴- نتایج

نتایج استفاده از روش شاخص‌گذاری ۴-gram در متون فارسی بسیار بهتر از نتایج ۳-gram می باشد و نتایج هر دو نیز بسیار بهتر از نتایج شاخص‌گذاری کلمه است . شاید دلیل خوب نتیجه ندادن شاخص‌گذاری کلمه عدم استفاده از الگوریتم ریشه‌یابی کلمات در اینجا باشد. دراینجا فقط پیشوندهاو پسوندهای کلمات حذف گردیده‌است.

بهترین معیار شباهت در متون فارسی نیز معیار ضرب نقطه‌ای بود و بهترین مقدار k نیز در الگوریتم نزدیک‌ترین k همسایه ۲۰ به دست آمد.

همانطور که قبلا اشاره شد، در این مقاله به مقایسه طبقه‌بندی متون فارسی در دو حالت با حذف stop word ها و بدون حذف stop word ها نیز پرداخته شده‌است. نتایج این دو حالت فرق قابل ملاحظه ای با هم نداشتند. دلیل این امر ماهیت n-gram می باشد زیرا بعنوان مثال دو کلمه "می‌باشد" و "باشد" در ۴-gram "باشد" مشترک هستند و حذف "می" اثر چندانی در کالکشن n-gram ها ندارد. و یا حتی حذف کلمات پر استفاده و کم ارزش از نظر اطلاعاتی نظیر "است" به دلیل روش وزندهی استفاده شده که درآن وزن عباراتی که در متون زیادی تکرار می‌شوند کم در نظر گرفته می‌شود، اثری چندانی ندارد. با این حال شاید اثر کم دیکشنری کلمات پر استفاده زبان فارسی مربوط به تعداد لغات کم آن که حدوداً یکصد لغت می‌باشد است و اگر دیکشنری stop word ها گسترده تر شود اثر محسوس تری در نتایج داشته باشد . در جداول زیر مقادیر دقت و یادآوری در مورد شاخص-گذاری‌های مختلف با k =۲۰ و با استفاده از معیار شباهت ضرب داخلی آورده شده است.

جدول (۱): *نتایج نهایی*

	دقت	یادآوری
۳-gram	۰،۳۶	۰،۵۶
٤-gram بدون‌حذف stop word	۰،۶۱	۰،۷۱
٤-gram با حذف stop word	**۰،۶۸**	**۰،۷۸**
Word	۰،۳۸	۰،۵۷

جدول (۲): *جزئیات نتایج در هر طبقه*

طبقه	تعداد مقالات	Recall	Precision
سیاسی	۸۸۰	۰،۵۲	۰،۸۸
اجتماعی	۱۹۷	۰،۹۵	۰،۱۳
اقتصادی	۶۷۲	۰،۵۴	۰،۹۶
ورزشی	۶۹۰	۰،۸۸	۰،۹۴
تکنولوژی	۲۹۵	۰،۳۸	۰،۵۱
خارجی	۸۷۳	۰،۷۴	۰،۹۴
حوادث	۳۹۳	۰،۳۶	۰،۷۸

۵- نتیجه‌گیری

این مقاله نتایج سیستم طبقه بندی خودکار متون فارسی که روی مجموعه۴۰۰۰ مقاله‌ی روزنامه همشهری که در هفت کلاس بودند را نشان می‌دهد.

با استفاده از تست های مختلف انجام شده بهترین روش شاخص گذاری روش **۴-gram** بدست آمدو بهترین مقدار N در الگوریتم knn عدد ۲۰ بدست آمد. در ضمن با مقایسه طبقه بندی بین دو حالت حذف stop wordها و بدون حذف stop wordها تفاوت چندانی حاصل نشد . مقدار یادآوری با شاخص گذاری ۴-gram ۰،۶۸ و مقدار دقت نیز ۰،۷۸ به دست آمد که نسبتا نتیجه خوبی می‌باشد.

برای بهتر نمودن الگوریتم می‌توان در اینده از الگوریتم‌های ریشه-یابی فارسی نیز استفاده نمود.

مراجع

[۱] Fabriz Sebastiani, *Machine Learning in Text Categorization*, ACM Computing Surveys, Vol.۳٤, No. ۱, pp.۱-٤۷, March ۲۰۰۲

[۲] Kjersti Aas, Line Eikvil, *Text Categorization: A Survey* , June ۱۹۹۹

[۳] Laila Khreisat, *Arabic Text Classification Using N-Gram Frequency Statistics*, Tech. report Fairleigh Dickinson University, ۲۰۰٤

[۴] Teresa Goncalves, Paulo Quaresma, *Evaluating preprocessing techniques in a text classififcation problem*

[۵] Yang, Y., Pedersen J.P. *A Comparative Study on Feature Selection in Text Categorization*.Proceedings of the Fourteenth International Conference on Machine Learning (ICML'۹۷), ۱۹۹۷

[۶] Peter Nather, *N-gram Based Text Categorization*, Diploma thesis, ۲۰۰۵

[۷] W. B. Cavnar and J. M. Trenkle, *N-gram-based text categorization*. Proceedings of SDAIR-۹٤, ۳rd Annual Symposium on Document Analysis and Information Retrieval, ۱۹۹٤.

[۸] Andras Kornai, J.Micheal Richards, *Linear Discriminant Text Classification in High Dimension*

[۹] Helmut Berger, Dieter Merkl, *A Comparision of Text-Categorization Methods applied to N-Gram Frequency Statistics*, Report of University of Technology,NSW

[۱۰] Darrudi, E., Hejazi, M. R, Oroumchian, F. *Assessment of a Modern Farsi Corpus*. In Proceedings of the ۲nd Workshop on Information Technology & its Disciplines (WITID) ۲۰۰٤, ITRC, Kish Island, Iran.

[۱۱] Hadi Amiri, Abolfazl AleAhmad, Farhad Oroumchian, Caro Lucas, Masoud Rahgozar, *Using OWA Fuzzy Operator to Merge Retrieval System Results*, The Second Workshop on Computational Approaches to Arabic Script-based Languages, LSA ۲۰۰۷ Linguistic Institute, Stanford University, USA, ۲۰۰۷

[۱۲] Abolfazl Aleahmad, Parsia Hakimian, Farzad Mahdikhani and Farhad Oroumchian. ۲۰۰۷. *N-Gram and Local Context Analysis For Persian Text Retrieval*. International Sympo-sium on Signal Processing and Its Applications, Sharjah U.A.E.

[۱۳] Alireza Mokhtaripour, Saber Jahanpour "*Introduction to a new Farsi stemmer*," Proceedings of the ۱۰th ACM international conference on Information and Knowledge Management, Pages: ۸۲٦ - ۸۲۷, ۲۰۰٦, ISBN:۱-۵۹۵۹۳-٤۳

[۱۴] G. K. Zipf, "*Human Behavior and the Principle of Least Effort, an Introduction to Human Ecology*," Addison-Wesley, Reading, Mass., ۱۹۴۹.

[۱۵] Ciya Liao, Shamim Alpha, Paul Dixon, *Feature Preparation in Text Categorization*, Oracle Corporation

[۱۶] Thorsten Joachims, *Text Categorization with Support Vector Machines*

زیرنویس‌ها

۱ Machine Learning Algorithm
۲ Automatic Text Classification
۳ Feature Selection
٤ K Nearest Neighbor
۵ Feature Extraction
٦ Learning phase
۷ Encoding
۸ Word Stemming
۹ Dot Product
۱۰ Dice Similarity
۱۱ Manhattan distance
۱۲ Precision
۱۳ Recall
۱٤ Macro averaging