پیش بینی حوزه اخبار

پوریا پارسه کارشناس کامپوتر
Purya.dataminig@gmail.com

ناصر پیکری کارشناسی ارشد جامع شناسی
safirghrn@gmail.com

علی هاشمی فرد کارشناس الکترونیک
Alihashemifard1393@gmail.com

حمیدرضا طاهری کارشناس کامپیوتر
Hamidtaheri1365@yahoo.com

چکیده

در دهه اخیر در دنیای اینترنت روزانه بالغ بر میلیون‌ها خبر، کامنت و دیگر انواع متن هر روز بر بستر اینترنت جاری می گردد. متن کاوی یکی از تکنیک‌هایی است که با بکارگیری الگوریتم‌های داده کاوی به کشف و استخراج خودکار اطلاعات از اسناد و محتوای متنی وب می پردازد. در واقع متن کاوی، فرآیند کشف اطلاعات و دانش ناشناخته و مفید از داده های وب می باشد. در این نوشتار با بکارگیری الگوریتم‌های درخت تصمیم، شبکه عصبی و ماشین برداری پشتیبان بر روی مجموعه‌ای از اخبار در چهار حوزه خبری؛اقتصادی، بین‌المللی، فرهنگی و ورزشی پردازش صورت گرفت که بهترین مدل بدست‌آمده با میزان ۹۷ درصد با مدل بردار پشتیبان (SVM) از داده های آزمایشی نتیجه مطلوبی حاصل گردید.

کلمات کلیدی

کلمنتاین، رپیدماینر، درخت تصمیم، شبکه عصبی، ماشین برداری پشتیبان

۱- مقدمه

بخش قابل توجهی از اطلاعات قابل دسترس در بستر اینترنت بصورت صفحات وب و پایگاه داده‌های متنی ذخیره شده‌اند. پایگاه داده‌های متنی به علت رشد فزاینده در حجم، سرعت و تنوع نیازمند به کارگیری ابزارهای نوین فناوری جهت طبقه بندی، کشف ساختار و معناهای ضمنی پنهان در متن می‌باشند. داده‌های ذخیره شده در این پایگاه، داده‌های ساختار نیافته هستند که هدف متن کاوی پردازش، رده‌بندی، استخراج اطلاعات و یافتن روابط در متون به کمک الگوریتم‌های یادگیری ماشین، آمار و زبان شناسی محاسباتی است. روش‌های متن کاوی شامل؛ پردازش مستندات و استخراج دانش از متن می‌باشد. در فاز اول خروجی به دو صورت: ۱)مبتنی بر سند ۲)مبتنی بر مفهوم است. در فرمت نمایش مبتنی بر سند آنچه که مهم است، نحوه نمایش بهتر برای مستندات است. در فرمت مبتنی بر مفهوم، مفاهیم و معانی موجود در سند و نیز ارتباط میان آنها و هر نوع اطلاعات مفهومی دیگر که قابل استخراج، از متن استخراج می‌شود.

در فاز دوم گروه بندی، طبقه بندی و تجسم سازی و نظایر آن بر روی مستندات اعمال می‌گردد. هدف این نوشتار معرفی مدلی است که با استفاده از نرم‌افزار Rapid Mainer و Clementine با ذخیره صفحات وب سایت

خبری در موضوعات اقتصادی، بین‌المللی، فرهنگی و ورزشی اقدام به یادگیری مدل و استفاده از مدل برای پردازش‌های بعدی می پردازیم.

۲- روش‌های متن کاوی

این حوزه تمام فعالیت‌هایی که به دنبال کسب دانش از متن هستند را شامل می‌گردد. در مراحل اولیه ضروریست اسناد پیش پردازش شوند و اطلاعات در یک ساختار داده‌ای مناسب برای پردازش‌های بعدی ذخیره شوند. در این زمینه چندین روش وجود دارد که سعی در بهره گیری از ساختار نحوی و معنایی متن دارند. بیشتر روش‌ها اسناد را به صورت مجموعه‌ای از کلمات نمایش می‌دهند. روش‌های متن کاوی الگوریتم‌های کاوش را روی برچسب‌های نسبت داده شده به هر سند اعمال می‌کنند. این برچسب‌ها ممکن است کلمات کلیدی استخراج شده از سند یا فقط لیستی از کلمات در سند مورد نظر باشند. برای نشان دادن اهمیت یک کلمه در سند معمولا از نمایش بردار استفاده می شود. برای هر کلمه یک مقدار اهمیت عددی ذخیره می‌گردد. روش‌های اصلی و مهم موجود که بر اساس این ایده هستند عباراتند از : مدل فضای بردار، مدل احتمالی و مدل منطقی.

با توجه به این که در این نوشتار از فضای مدل بردار استفاده گردیده این روش را توضیح می دهیم.

۲-۱- فضای بردار

این مدل قادر به آنالیز کارآمد مجموعه بزرگی از سندهاست. این روش در ابتدا برای بازیابی اطلاعات و ایندکس کردن معرفی شده بود اما همکنون در برخی از روش‌های متن کاوی نیز از آن استفاده می‌شود. در این مدل اسناد و *query* به عنوان بردارهایی در فضای m بعدی نمایش داده می‌شوند. که در این فضا هر بعد یک ترم است. منظور از ترم یک مفهوم پایه مثل کلمه یا عبارت است. عناصر بردار با وزن ترم متناظرند. سند d به صورت $d = (x_1, x_2, \ldots\ldots x_n)$ نمایش داده می‌شود که هر x_i اهمیت ترم i را توی سند d نشان می‌دهد. در اینجا شباهت را بر اساس فاصله بین بردارها(سند با سند یا سند با *query*) تعیین می‌کنیم. هر چه بردارها بهم نزدیکتر باشند شبیه‌تر در نظر گرفته می‌شوند. که این شباهت بر اساس زاویه دو بردار یا ضرب برداری یا شباهت کسینوسی تعریف می‌شود. برای نسبت دادن وزن به ترم ها از تخمین Tf * و IDF [] می‌توان استفاده کرد. در Tf یک ترم که تعداد تکرار در سند بیشتر باشد اهمیت آن بیشتر و وزنش بیشتر است. $tf(t,d) = f(t,d)$. اما مسلما هر چه سند طولانی‌تر باشد احتمال اینکه فرکانس ترمی در آن بیشتر باشد، بیشتر است. بنابراین باید tf را نرمال کرد و طول سند را نیز در نظر گرفت:

$$w(d,t) = \frac{\mathrm{tf}(d,t)\log(N/n_t)}{\sqrt{\sum_{j=1}^{m} \mathrm{tf}(d,t_j)^2 \ (\log(N/n_{tj}))^2}}$$

در مورد IDF هرچه یک ترم خاص‌تر باشد، وزنش بیشتر است.

$$IDF(t) = 1 + \log(n/k)$$

در این فرمول k برابر با تعداد سندهایی است که ترم t در آن‌ها تکرار شده است. شباهت بین دو سند را می‌توان به طرق زیر محاسبه نمود:

ضرب برداری:

$$S(d_1,d_2) = \sum_{k=1}^{m} w(d_1,t_k)\cdot w(d_2,t_k).$$

فاصله اقلیدسی

$$dist(d_1,d_2) = \sqrt[2]{\sum_{k=1}^{m} |w(d_1,t_k) - w(d_2,t_k)|^2}$$

شباهت کسینوسی

$$\cos\varphi = \frac{\vec{x}\vec{y}}{|\vec{x}|\cdot|\vec{y}|} = 1 - \frac{1}{2}d^2\left(\frac{\vec{x}}{|\vec{x}|}, \frac{\vec{y}}{|\vec{y}|}\right).$$

[1]

۲-۱-۱- درخت تصمیم

درخت تصمیم یک ابزار برای پشتیبانی از تصمیم است که از درختان برای مدل کردن استفاده می‌کند. درخت تصمیم به طور معمول در تحقیق درعملیات استفاده می‌شود، به طور خاص در آنالیز تصمیم، برای مشخص کردن استراتژی که با بیشترین احتمال به هدف برسد بکار، می‌رود. استفاده دیگر درختان تصمیم، توصیف محاسبات احتمال شرطی است.[2]

---

* Tern frequency

[] Inverse document frequency

۲-۱-۲- شبکه عصبی

شبکه های عصبی مصنوعی (Artificial Neural Network) الگویی برای پردازش اطلاعات می باشند که با تقلید از شبکه های عصبی بیولوژیکی مثل مغز انسان ساخته شده اند.عنصر کلیدی این الگو ساختار جدید سیستم پردازش اطلاعات آن می باشد و از تعداد زیادی عناصر (نرون) با ارتباطات قوی داخلی که هماهنگ با هم برای حل مسائل مخصوص کار می کنند تشکیل شده اند. شبکه های عصبی مصنوعی با پردازش روی داده های تجربی، دانش یا قانون نهفته در ورای داده ها را به ساختار شبکه منتقل می کند که به این عمل یادگیری می گویند. اصولاً توانایی یادگیری مهمترین ویژگی یک سیستم هوشمند است. سیستمی که بتواند یاد بگیرد منعطف تر است وساده تر برنامه ریزی میشود، بنابراین بهتر می‌تواند در مورد مسایل و معادلات جدید پاسخگو باشد.[3]

۲-۱-۳- Feature Selection

در این روش با توجه به میزان تاثیر گزاری متغیرهای موجود، آن حوزه از متغیرها که در داده کاوی بی تاثیر یا کم تاثیر هستن حذف شده یا به عبارت دیگر انتخاب متغیرهای موثر صورت می گیرد.

که این کار باعث افزایش دقت و سرعت در مدل می شود.

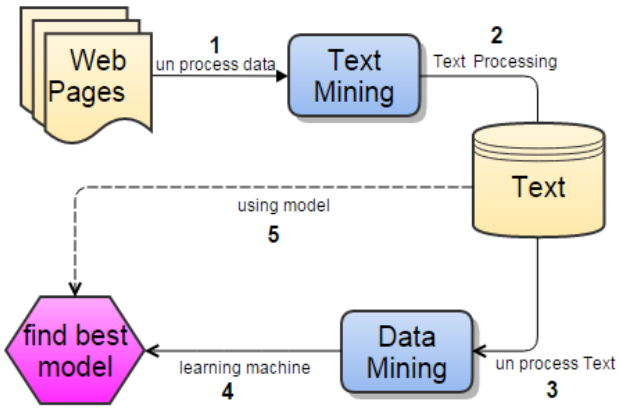
۲-۱-۴- ماشین برداری پشتیبان

یکی از ایده های جدید در شناسایی و دسته بندی الگوها، ماشین بردار پشتیبان یا SVMاست. ماشین بردار پشتیبان دارای خواص بسیار ارزشمندی است که آن را برای شناسایی الگو مناسب می سازد. از جمله اینکه SVM در آموزش خود مشکل بهینه های محلی را ندارد، دسته بندی کننده را با حداکثر تعمیم بنا می کند، ساختار و توپولوژی خود را بصورت بهینه تعیین می نماید و توابع تمایز غیر خطی را به راحتی و با محاسبات کم، با استفاده از مفهوم حاصلضرب داخلی در فضای هیلبرت، تشکیل می دهد.[4]

۲-۲- متدولوژی انجام کار

در ابتدا صفحات وب جمع آوری شده را مورد متن کاوی قرار داده سپس متن به دست آمده را با تکنیک های داده کاوی مورد آنالیز قرار داده تا به مدل مطلوب دست یابیم.

حال این فرایند که علاوه بر زمان گیر بودن و دارای هزینه بالا سیستمی و انسانی می باشد را می توان با داشتن مدل مطلوب بسیار آسان‌تر انجام داد به این ترتیب که متن های حاصل از متن کاوی را به صورت مستقیم درون مدل قرار می‌دهیم و بدون صرف زمان پردازش اولیه برای به دست آوردن مدل می توانیم خروجی را مشاهده کنیم.



شکل(2): فلوچارت تهیه مدل پیش بینی

۲-۲-۱- توصیف داده‌ها

در ابتدا چهارصد صفحه وب را به صورت HTML که مرتبط به بازه زمانی یک هفته‌ای یکی از سایت‌های خبری است را در چهار پوشه با دسته بندی به صورت حوزه های اخبار پیش گفته ذخیره می‌کنیم.

۲-۲-۲- مدل سازی

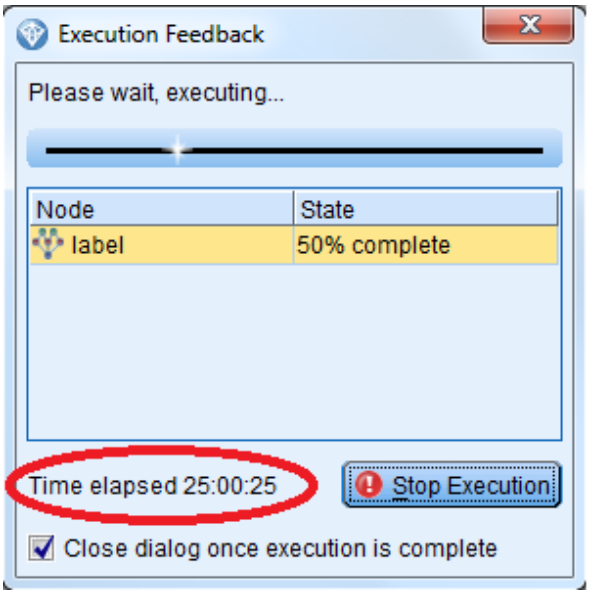
برای ساخت مدلی که متن های وب ذخیره شده را مورد متن کاوی قرار دهد به سراغ نرم افزار Rapid Miner می رویم ابتدا از یک process document استفاده می‌کنیم که در قسمت text directories فولدرهای ساخته شده را به ترتیب محتوا که در اینجا چهار حوزه اقتصادی، بین‌المللی، فرهنگی و ورزشی انتخاب شده‌اند به نرم افزار معرفی می‌کنیم که در هر حوزه صد صفحه وب وجود دارد، سپس گزینه Vector Creation را بر روی TF-IDF می گزاریم.

حال به درون خود process document می‌رویم و گره Tokenize را برای مشخص کردن کلمات قرار می دهیم سپس FilterTokes را قرار می دهیم به این منظور که کلمات بالای ۲۰ حرف یا زیر ۴ حرف را نشان ندهد و گره بعدی Filter Stopword می‌باشد که با ساخت یک فایل متنی که در آن کلماتی را که نمی خواهیم برای ما مورد پردازش قرار گیرد یادداشت می‌کنیم مانند کلماتی که در بدنه خود سایت به زبان html است که در این مدل بالغ بر ۹۰۰ کلمه می باشند، این

فایل متنی را در قسمت file که در خود Filter Stopword موجود است به نرم افزار معرفی می کنیم.
در نهایت با قرار دادن Stem کلماتی که بار معنایی یکسان دارند ولی به طور مختلف نوشته می‌شوند را در یک فایل متنی نوشته و در قسمت File به نرم افزار معرفی می‌کنیم.
مانند:

آشنائی:آشنایی.*

سپس با قرار دادن یک گره Write CSV تمام کلمات بدست آمده را در یک فایل متنی ذخیره می‌کنیم، لازم به ذکر است که در تمام روند انجام کار باید Encoding برای متن فارسی بر روی UTF-8 باشد.
حال کار با نرم افزار Rapid Miner به اتمام می‌رسد چون که کارهای متن کاوی به اتمام رسیده است و اینک با باز کردن نرم افزار modeler Clementine شروع به داده کاوی این فایل متنی می‌کنیم در نرم افزار ابتدا یک گره Var.File از قسمت Sources می آوریم و فایل متنی به دست آمده در مرحله قبل را در قسمت File به نرم افزار معرفی می‌کنیم، سپس یک گره Type گذاشته برای خواندن صحیح داده ها و همچنین مشخص کردن Target که همان فیلد label می‌باشد سپس از سربرگ Modeling یک Feature Selection برای انتخاب تاثیر گزارترین متغیرها مورد استفاده قرار می گیرد را انتخاب و Type را به آن متصل می‌کنیم تادقت مدل افزایش یابد، پس از اجرای گره، الماس Feature Selection به دست آمده را باز کرده و در بالای آن گزینه Generate را انتخاب کرده و گزینه Filter را می زنیم و در مرحله بعد رادیو باتم Important را میزنیم تا یک گره Filter بر روی صفحه نرم‌افزار ایجاد کند حال Type که در دو مرحله قبل ایجاد کرده بودیم را به این Filter نیز وصل می‌کنیم، سپس یک گره Partition از سر برگ Field Ops انتخاب و فیلتر را به آن متصل می‌کنیم، که در تنظیمات باید قسمت Training را برابر ۷۰ و Testing را برابر ۳۰ قرار داد به این معنی که ۷۰ درصد داده ها برای آموزش مورد استفاده قرار گیرد و ۳۰ درصد ما بقی برای تست این که مدل صحیح است یا خیر استفاده می‌شود.
حال آماده هستیم که مدل‌های مختلف را امتحان کنیم تا ببینیم کدام مدل بهترین مدل است که مدلهای شبکه عصبی ،انواع درخت تصمیم و ماشین برداری پشتیبان استفاده شد که به علت بالا بودن بسیار زیاده داده‌ها هر کدام از مدل‌ها بالغ بر ۲۴ ساعت زمان می برد تا به نتیجه برسد.



شکل(2): زمان صرف شده برای رسم ۵۰ درصد از یک مدل

پس از استفاده از چهار مدل پیش بینی به درصد مورد نظر و قابل قبول ۹۷,۸۷ درصد با مدل ماشین برداری پشتیبان (SVM) می رسیم.
پس از به دست آمدن الماس این مدل حال به جای صرف ساعتهای بسیار برای به دست آوردن مدل و یا خواندن تک تک اخبار به راحتی فقط باصرف کمتر از یک ساعت وقت برای اجرای الماس می‌توان داده هایی به میزان بسیار زیادی را به نمودار داد تا به سرعت نتیجه پیش بینی را به ما بدهد، به عبارت دیگر پس از ایجاد خروجی متنی توسط Rapid miner فقط کافیست کلمات حاصل را درون نرم افزار Clementine وارد کنیم و آنهارا به الماس مذکور وصل کنیم تا پیش بینی این که خبر در کدام حوزه است صورت پذیرد.

Results for output field label
Comparing $S-label with label

'Partition'	1_Training		2_Testing	
Correct	206	100%	92	97.87%
Wrong	0	0%	2	2.13%
Total	206		94	

Coincidence Matrix for $S-label (rows show actuals)

'Partition' = 1_Training	eghtesad	farhangi	varzeshi
eghtesad	64	0	0
farhangi	0	73	0
varzeshi	0	0	69
'Partition' = 2_Testing	eghtesad	farhangi	varzeshi
eghtesad	34	2	0
farhangi	0	27	0
varzeshi	0	0	31

شکل(۳): بالاترین درصد پیش بینی

۲-۳- مقایسه مدل های پیش بینی

در جدول پایین درصد صحیح بودن هر یک از مدل‌های پیش بینی را ملاحضه می‌کنید ،مقادیر آزمایشی بر اساس همان مقدار

70% که در پارتیشن برای آموزش مدل تعیین کرده بودیم و مقادیر تست نیز **30%** است که با آنها صحیح بودن مدل را چک می‌کنیم و معیار بهتر بودن مدل صحیح بودن مقادیر تست می باشد.

نام‌مدل	مقادیر آزمایشی	مقادیر تست
Neural Net	92.72%	78.72%
C&R Tree	95.15%	88.3%
CHAID Tree	88.83%	97.79%
SVM	100%	97.89%

جدول(۱): مقایسه مدل های

۳- *نتیجه گیری*

در اینجا با ادغام روش‌های متن‌کاوی و داده‌کاوی فرایند وب کاوی را بر روی ۴۰۰ صفحه وب انجام داده‌ایم و نتیجه بدست آمده برای پیش بینی موضوعات خبری با ضریب خطای کمتر از ۳ درصد حاصل گردید. به کمک این فرایند می توان به راحتی صفحات وب را مورد بررسی قرار داد و طبق آن موضوع‌بندی که در ابتدا انجام داده‌ایم می‌توان پیش بینی‌های مورد نظر را به بهترین نحو ممکن انجام داد . کاربرد این مدل باعث صرفه جویی در زمان، سرمایه انسانی و سازمانی را بوجود و به جای صرف چندین روز وقت می‌توان کمتر از یک ساعت به نتیجه مطلوب دست یافت.

مراجع

[۱] مصباح، سارا؛ ۱۳۸۸، " **متن کاوی** "، استاد راهنما: دکتر مسعود رهگذر، پایان نامه،

[۲] درخت تصمیم
http://fa.wikipedia.org/wiki/%D8%AF%D8%B1%D8%AE%D8%AA_%D8%AA%D8%B5%D9%85%DB%8C%D9%85

[۳] شبکه های عصبی
/http://www.filedc.com/Products/150

[۴] کبودیان، جهانشاه؛ محمد رحمتی و محمدمهدی همایون پور، ۱۳۸۲، استفاده از ماشین بردار پشتیبان (SVM) در سه مساله شناسایی الگو، اولین کنفرانس بین المللی فناوری اطلاعات و دانش، تهران، دانشگاه صنعتی امیرکبیر، http://www.civilica.com/Paper-ICIKT01-ICIKT01_034.html