تحلیل احساسات در شبکه اجتماعی توییتر با تکنیک متن کاوی

ناصر پیکری [۱]

سید علی اصغر یعقوبی [۲]

حمیدرضا طاهری[۳]

چکیده

با گسترش شگرف اینترنت، شبکه های اجتماعی و استفاده روزافزون از آن شاهد حجم انبوهی ازنظرات کاربران در ارتباط با موضوعات مختلف هستیم که مطالعه و تحلیل نظرات در حجم انبوه با مشکلات زیادی روبرو بوده و کاربرد تکنیک‌های علمی نوین ضرورتی اجتنابناپذیر می‌باشد. مقاله حاضر با کاربرد تکنیک متن‌کاوی و تحلیل محتوا پدیده فوت مرتضی پاشایی را در شبکه اجتماعی توییتر مورد مطالعه و بررسی قرار داده و تمام توییت‌های انتشاریافته شامل؛ ۱۷۷۱٤ توییت را در پنج مقوله؛ تبلیغ در مورد آلبوم پاشایی، انعکاس عیادت هنرمندان و بازیگران، بازتاب مراسم تشییع پاشایی، پیگیری خبر سلامتی وضعیت پاشایی و دعا برای وی و فکاهی کردن و لوث کردن مرگ پاشایی رده بندی و همبستگی بین رده‌ها را با ویژگی کاربران توصیف نموده است.

کلمات کلیدی

شبکه اجتماعی توییتر، توییت، متن‌کاوی.

احساسات همواره از دیرباز جنبه مرموز و ناشناخته انسان‌ها بوده و جایگاه مهمی در حیات اجتماعی افراد دارا می‌باشد. پرداختن به سوائق احساسی و عاطفی همچون؛ خشم، شادی، ترس، غم، کینه، بغض، عصبانیت، شرم، گناه و امثال آن به ضرورتی بنیادین در کنش‌های انسانی تبدیل شده و تحلیل رفتار آدمیان بدون در نظر گرفتن احساسات و عواطف ناقص بوده و ارزش چندانی بر آن مترتب نیست. بررسی درگذشت مرتضی پاشایی در فضای مجازی و تحلیل رفتار کاربران در این خصوص موضوع اصلی این مقاله می‌باشد. در این بررسی تمام توییت‌های شبکه اجتماعی توییتر مرتبط با پدیده فوت پاشایی در قالب جدول اکسل شامل افراد توییت کننده و محتوای توییت‌ها جمع‌آوری[۴] و سپس به کمک نرم‌افزار rapidmainer و الگوریتم های متن کاوی مورد تحلیل قرار گرفته است. هدف اصلی در متن‌کاوی، دسته‌بندی متون در قالب تعداد معینی از دسته‌های از پیش تعیین‌شده است. یک سند می‌تواند در یک یا چند دسته قرار بگیرد. این موضوع می‌تواند در قالب یک یادگیری خودکار بر روی تعدادی متن انجام و سپس در پردازش‌های بعدی بر روی اسناد مورد استفاده قرار گیرد.

جدول شماره (۱) نمونه‌ای از داده‌های استخراجی از شبکه اجتماعی توییتر

---

[۱]. کارشناس ارشد جامعه شناسی

[۲]. کارشناس مخابرات

[۳]. کارشناس الکترونیک

[٤]. شکل یک نمونه‌ای از داده‌های استخراجی از شبکه اجتماعی توییتر

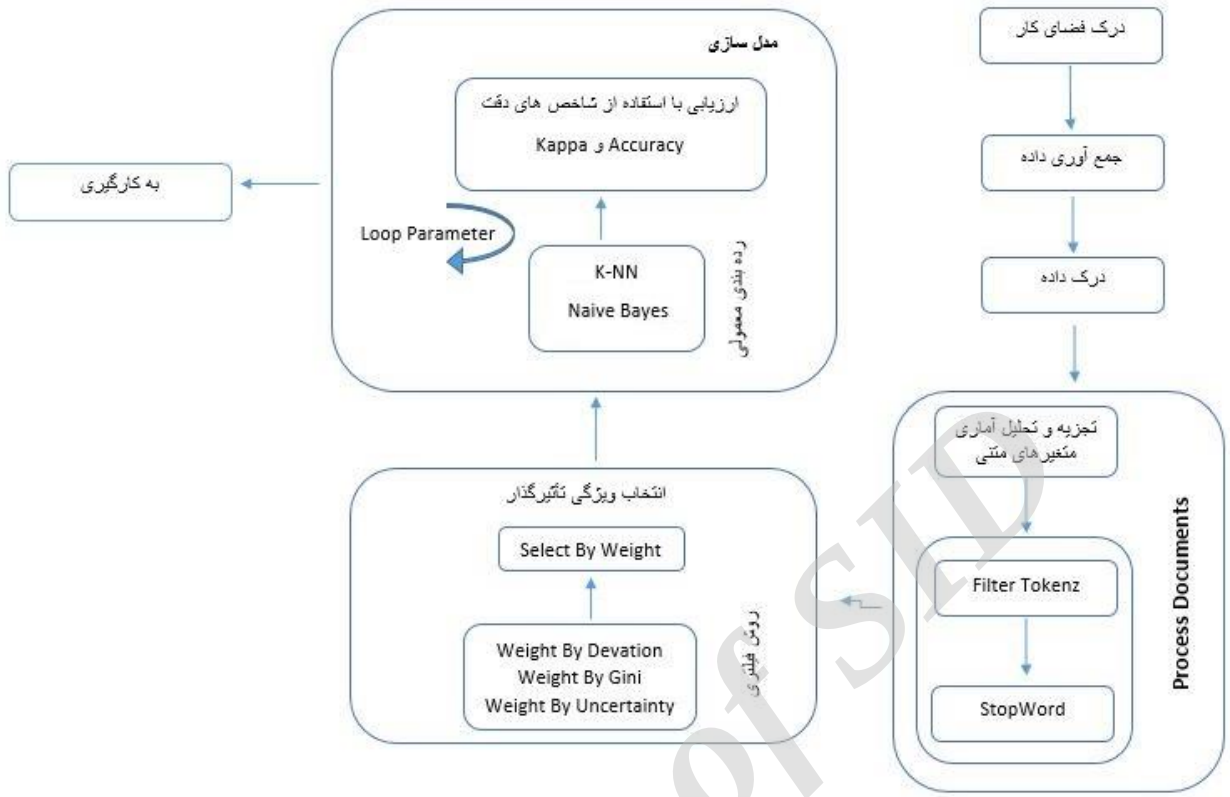
t u favourites	t u statuses	t u followers	rt u favourites	rt u followers	rt u friends	t text
8660	2129	209	0	0	0	بعد از اون اپ های مرتضی پاشایی باید منتظر نذری فایندرهم می بودیم
35966	18161	500	0	0	0	این کلیپی که واسه آهنگ گلزار و پاشایی ساختن رو درک نمیکنم.دختره حامله میشه میره گل میده به ماشینا بعد ماشین میزنه بهش کسیر میشه.
285	11649	813	0	0	0	#جام_جم : درخواست رضا صادقی: برای مرتضی پاشایی دعا کنید http://t.co/R25hbg8h1f http://t.co/gBoFFbMK0T
16	31	11	24007	3046	153	RT @chaghzz: ر شوهرخاله م آهنگ مرتضی پاشایی پیشوازش بود. وضع کسشر زیر گوشمون داره پیشروی میکنه و ما تو خواب غفلتیم
5446	495	233	59987	994	377	RT @sepeehr: طبیعیش این بود که هرکس کاور این آهنگ رو میدید اصلن آهنگ رو دانلود نمی‌کرد و مرتضی پاشایی کلا معروف نمیشد
5446	496	233	0	0	0	@sepeehr ن بار که میخواستم آهنگ مرتضی پاشایی و دانلود کنم از رادیو جوان ، هی به کاورش نگاه میکردم ، هی به لایکا و پلی هاش
129	803	112	0	0	0	G+: http://t.co/lYk5k38V پیش درآمدی بر آنچه خواهیم شنید . دیوار موسیقی : بهروز پاشایی تصویر برداری و ویدئو : آرش بلوری
54	87362	456	0	0	0	http://t.co/pzl2eTtlxj سرگرمی :: درخواست رضا صادقی از مردم: برای مرتضی پاشایی دعا کنید: به گزارش تی وی پلاس، رضا...
2	26	1	0	0	0	دانلود آهنگ مرتضی پاشایی – بی تو... http://t.co/D4Klw5oktq# #آهنگایرانی #آهنگغمگین #دانلودآهنگ
2	26	1	0	0	0	دانلود آهنگ اطلسی – مرتضی پاشایی:... http://t.co/wERl2NmtFj #آهنگایرانی #آهنگغمگین #دانلودآهنگ #t.co/aZkzxsTcDc
2	26	1	0	0	0	دانلود آهنگ مرتضی پاشایی به نام... http://t.co/6BnvP1aqcE #آهنگایرانی #آهنگغمگین #دانلودآهنگ #p://t.co/MiES60Oqo6

۱. مفهوم متن کاوی

متن کاوی را می‌توان به عنوان متدها و الگوریتم‌هایی از فیلدهای یادگیری ماشین و آماری برای متن‌ها با هدف پیدا کردن الگوهای مفید در نظر گرفت. برای این هدف پیش پردازش کردن متون ضروری است. در بسیاری از روش‌ها، متدهای استخراج اطلاعات، پردازش کردن زبان طبیعی یا برخی پیش پردازش‌های ساده برای استخراج داده از متون استفاده می‌شود. سپس می‌توان الگوریتم‌های داده کاوی را بر روی داده های استخراج شده اعمال کرد. در این مقاله ما بیشتر متن کاوی را به عنوان کشف داده متنی در نظر می‌گیریم [۵] و بیشتر تمرکز بر روش‌های استخراج الگوهای مفید از متن شامل دسته‌بندی مجموعه‌های متنی یا استخراج اطلاعات مفید است.

جدول شماره (۲) فرایند متن کاوی

۵. مدل فرایند متن‌کاوی



۲. پیش پردازش متن

برای کاوش کردن مجموعه بزرگی از اسناد ضروریست که اسناد پیش پردازش شوند و اطلاعات در یک ساختار داده‌ای مناسب برای پردازش‌های بعدی ذخیره شوند. در این زمینه چندین روش وجود دارند که سعی در بهره‌گیری از ساختار نحوی و معنایی متن دارند. در بیشتر روش‌ها، اسناد به صورت مجموعه‌ای از کلمات نمایش داده می‌شوند. بیشتر روش‌های متن کاوی، الگوریتم‌های کاوش را روی برچسب‌های نسبت داده شده به هر سند اعمال می‌کنند. این برچسب‌ها ممکنه کلمات کلیدی استخراج شده از سند یا فقط لیستی از کلمات در سند مورد نظر باشند. برای نشان دادن کمترین اهمیت یک کلمه در یک سند معمولا از نمایش بردار استفاده می‌شود، برای هر کلمه یک مقدار اهمیت عددی ذخیره می‌گردد. روش‌های اصلی و مهم موجود که بر اساس این ایده هستند عبارتند از: مدل فضای بردار، مدل احتمالی و مدل منطقی. چون برخی از روش‌های متن کاوی که بیان می‌شوند از مدل فضای برداری استفاده می‌کنند این روش را مختصرا توضیح می‌دهیم.

۱.۲ 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 $d =(x_1, x_2,.....x_n)$ 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[۶] و IDF[۷] 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:

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

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

۳. روش‌های متن کاوی

دلیل اصلی به کار بردن روش‌های داده کاوی برای اسناد متنی، ساختار بندی کردن آنهاست. ساختارهای دیتابیس معرف عبارتند از: کاتالوگ‌های کتابخانه یا ایندکس‌های کتاب مشکل ایندکس‌های طراحی شده به صورت دستی، زمان مورد نیاز برای نگه‌داری آن‌ها است. بنابراین برای منابع اطلاعاتی که خیلی تغییر می‌کنند مثل وب مناسب نیستند. متدهای موجود برای ساختاربندی کردن مجموعه‌ها عبارتند از :روش‌های رده بندی و روش‌های خوشه‌بندی. ترکیب این روش‌ها با روش‌های ساختاربندی (خوشه-بندی و رده بندی)ابزارهای قدرتمندی برای کاوش الگوهای مفید در مجموعه‌های متنی فراهم می‌کنند.

---

[۶] Tern frequency
[۷] Inverse document frequency

٤. فازهای اصلی فرآیند متن کاوی

متن کاوی فرآیندی است که شامل فیلدهای تکنولوژیکی فراوانی است. بازیابی اطلاعات، داده کاوی و هوش مصنوعی و زبانشناسی محاسباتی همه فیلدهایی هستند که در این زمینه، نقشی را دارا هستند. اما به طور کلی دو فاز اصلی در فرآیند متن کاوی می‌توان در نظر گرفت. اولین فاز پیش پردازش مستندات است. خروجی این فاز می تواند دو شکل مختلف داشته باشد:۱) مبتنی برسند. ۲) مبتنی بر مفهوم. در فرمت نمایش مبتنی بر سند، آنچه که مهم است، نحوه ی نمایش بهتر برای مستندات است. مثلا تبدیل اسناد به یک فرمت میانی و نیمه ساخت یافته، یا بکار بردن یک ایندکس بر روی آن‌ها یا هر نوع نمایش دیگری که کار کردن با اسناد را کاراتر می کند. هر موجودیت در این نمایش در نهایت باز هم یک سند خواهد بود. در نوع دوم نمایش اسناد بهبود بخشیده می‌شود، مفاهیم و معانی موجود در سند و نیز ارتباط میان آنها و هر نوع اطلاعات مفهومی دیگری که قابل استخراج است، از متن استخراج می شود. در این نوع نمایش دیگر با مستندات به عنوان یک موجودیت مواجه نیستیم بلکه با مفاهیمی که از این مستندات استخراج شده اند، رو به رو هستیم. قدم بعدی استخراج دانش از این فرمهای میانی نمایش اسناد است. بر اساس نحوه ی نمایش یک سند، روش استخراج دانش از یک سند متفاوت است. نمایش مبتنی بر سند، برای گروه بندی، طبقه بندی، تجسم‌سازی و نظایر این‌ها استفاده می شود، درحالیکه نمایش مبتنی بر مفهوم برای یافتن روابط میان مفاهیم، ساختن اتوماتیک تزاروس و آنتولوژی و نظایر آن بکار می رود.

۱.٤ رده‌بندی

هدف از رده‌بندی متون نسبت دادن کلاس‌های از پیش تعریف شده به اسناد متنی است. مثلا یک خبر جدید که وارد می‌شود بگوییم متعلق به کلاس ورزشی یا سیاسی یا هنری. برای رده‌بندی اسناد روش‌های گوناگونی به کار می‌روند. در رده‌بندی یک مجموعه آموزش از اسناد وجود دارد که برای این مجموعه کلاس‌ها مشخص است. با استفاده از این مجموعه مدل رده‌بندی مشخص می‌شود، سپس با استفاده از آن کلاس سند جدید که وارد می‌شود، مشخص می‌گردد. برای اندازه‌گیری کارایی مدل رده‌بندی، یک مجموعه تست که مستقل از مجموعه آموزش است در نظر گرفته می‌شود. و برچسب‌هایی که برای این اسناد توسط مدل تخمین زده می‌شود با برچسب واقعی اسناد مقایسه می‌شود. نسبت اسنادی که به درستی رده‌بندی شده اند به تعداد کل اسناد accuracy نامیده می‌شود سه معیار برای مقایسه رده‌بندی کننده‌ها استفاده می‌شود: ۱)precision: کسری از اسناد بازیابی شده‌ای که مربوط هستند ۲) recall: نشان دهنده کسری از اسناد مربوط بازیابی شده است.

۵. دسته‌بندی توییت‌ها

دسته‌بندی توییت‌ها به معنای برچسب زدن یک مفهوم به هر توییت و جایگذاری توییت‌های مشابه در یک دسته مشابه است. به عبارتی هر توییتی با توجه به دارا بودن کلمات و واحدهای معنایی مشابه مستتر در آن به یک دسته خاص اختصاص داده می‌شود. تمام توییت‌های مربوط به فوت مرتضی پاشایی پس از خوانش اولیه در پنج دسته و مقوله کلی شامل؛ ۱) تبلیغ در مورد آلبوم پاشایی. ۲) انعکاس عیادت هنرمندان و بازیگران. ۳)بازتاب مراسم تشییع پاشایی. ۴) پیگیری خبر سلامتی وضعیت پاشایی و دعا برای او ۵) فکاهی کردن و لوث کردن مرگ پاشایی، تقسیم و سپس مدل با دو الگوریتم **Naive Bayes** و **k –nn** بر روی تعداد **۴۶۴۲ توییت** مدل اجرا و سپس دقت مدل توسط دو اپراتور loop parameter و log ، برابر شکل ۳ محاسبه گردید. در گام دوم مدل بدست آمده برای رده بندی تمام توییت‌ها به تعداد ۱۷۷۱۴ مورد پردازش قرار گرفت. در گام سوم با روش نمونه‌گیری آماری احتمالی

تعدادی از کاربران را با توجه به فیلدهای اطلاعات بدست آمده از شبکه اجتماعی توییتر نسبت به فیلد هدف برابر جدول شماره ۴ توصیف گردیدند. اطلاعات این جدول نشان می‌دهد.

جدول شماره (۳) خروجی اپراتور log

K-nn			***naive Bayes***	
k-Nearest Neighbor	*Top k*	*accuracy*	*Top k*	*accuracy*
۳	۱۵۰۰	۰,۶۸۴۷۲۲	۱۵۰۰	۰,۵۶۵۹۷۲
۶	۲۰۰۰	۰,۶۸۰۵۵۶	۲۰۰۰	۰,۵۶۵۹۷۲
۱۰	۲۵۰۰	۰,۶۵۰۶۹۴	۲۵۰۰	۰,۵۶۵۹۷۲
۱۳	۳۰۰۰	۰,۶۳۸۸۸۹	۳۰۰۰	۰,۵۶۵۹۷۲
۱۷	۳۵۰۰	۰,۶۳۶۸۰۶	۳۵۰۰	۰,۵۶۵۹۷۲

اطلاعات جدول بالا دقت دو الگوریتم *naive Bayes* و *K-nn* *را با پارامترهای مختلف نزدیکترین همسایگی در شمارش گام-های ۳،۶،۱۰،۱۳،۱۷؛ بر حسب افزایش تعداد وزن کلمات مهم به نمایش می‌گذارد.*

جدول شماره (۴) تحلیل آماری ویژگی‌های کاربران با دسته‌بندی توییت‌ها

t u favourites	***t u statuses***	***t u followers***	***rt u favourites***	***rt u followers***	***rt u friends***	***class***
14722	21921	1324	40459	10283	1083	***1***
12553	52868	1824	33156	13873	427	***2***
19150	36640	1987	37324	18453	509	***3***
24088	26618	2579	39610	5581	528	***4***
30248	26599	1691	62430	5192	439	***5***

اطلاعات جدول بالا همبستگی بین ویژگی‌های کاربران شامل؛(t u favourites، کاربرانی که بیشترین توییت‌ها را بعنوان علاقمندی انتخاب نمودند)، (t u statuses، کاربرانی بیشترین توییت را در پدیده فوت پاشایی منتشر نمودند)، (t u followers، کسانی بیشترین دنبال کننده دارند)، (rt u favourites، کاربرانی که دارای بیشترین علاقمندی بوده و توییت‌های دیگر کاربران را بازتوییت نموده‌اند)، (rt u followers، کاربرانی که بیشترین دنبال کننده را در شبکه اجتماعی توییتر داشته و اقدام به بازتوییت توییت‌های دیگر کاربران نموده‌اند)، (rt u friends، کاربرانی که دارای بیشترین دوست یا اعضا می‌باشند) و (class، رده‌بندی مقوله‌های توییت‌ها را نشان می‌دهد) با رده‌بندی توییت‌های انتشاریافته با پدیده فوت مرتضی پاشایی را نشان می‌دهد.

٦. نتیجه‌گیری

پدیده فوت مرتضی پاشایی باعث بروز احساسات و عواطف پیش‌بینی نشده در جامعه ایرانی گردید. بسیاری از کارشناسان، جامعه-شناسان و روانشناسان سعی نمودند با کمک گرفتن از نظریات کلان و فردگرایانه این پدیده را تبیین و توصیف نمایند. نوشتار حاضر به کمک دو تکنیک تحلیل محتوا یا برچسب زنی و داده‌کاوی محتوای انتشاریافته در شبکه اجتماعی توییتر، در بازه زمانی بستری مرتضی پاشایی در بیمارستان بهمن تا ده روز پس از فوت وی را تحلیل و بررسی نموده است. محققین این نوشتار محتوای بارگذاری

شده را در پنج مقوله شامل؛ ۱) تبلیغ در مورد آلبوم پاشایی. ۲) انعکاس عیادت هنرمندان و بازیگران. ۳)بازتاب مراسم تشییع پاشایی. ۴) پیگیری خبر سلامتی وضعیت پاشایی و دعا برای او ۵) فکاهی کردن و لوث کردن مرگ پاشایی رده بندی نموده و سپس به کمک علم داده‌کاوی ارتباط و همبستگی بین ویژگی‌های کاربران با مقولات مستتر فعالیت کاربران در توییت‌های انتشاریافته را بازنمایی نموده است و به این نتیجه رسیده است. تعداد کاربرانی که با قصد لوث و فکاهی نمودن این پدیده فعالیت داشته‌اند، بیشترین تعداد کاربران را شامل گردیده و از طرفی بیشترین تعداد توییت را بعنوان علاقمندی انتخاب نموده‌اند. همچنین کاربرانی که پیگیر وضعیت سلامتی وی در دوران بیماری بوده‌اند و بیشترین فعالیت را در بازتاب مراسم تشییع وی داشته‌اند دارای بیشترین دنباله کننده توسط کاربران دیگر بوده‌اند. افرادی که بیشترین اعضا و دوست را دارا بوده، دارای بیشترین لینک تبلیغ آلبوم و تصاویر مرتضی پاشایی بوده‌اند.

۷. مراجع

[Kara_۰۲] Haralampos Karanikas, et.al. *An Approach to Text Mining using Information Extraction*, ۲۰۰۰

[kanya_۰۷] N. Kanya*, S. Geetha *"INFORMATION EXTRACTION -A TEXT MININGAPPROACH"* ۲۰۰۷ produced IEEE

[Nahm_۰۵] Raymond J. Mooney and Un Yong Nahm *"Text mining with Informatin Exteraction"* ,۲۰۰۵

[Rajman_۹۷] M. Rajman. "Text Mining, Knowledge extraction from unstructured texual data". *Proc. of EUROSTAT Conference*, Francfort (Deutchland), may, ۱۹۹۷

[Book] Data mining Concepts and Techniques: jiawei Han and Micheline kamber

[Dumais_۹۸] S. Dumais, J. Platt, D. Heckerman, and M. Sahami. Inductive learning algorithms and representations for text categorization. In *۷th Int. Conf. on Information and Knowledge Managment*, ۱۹۹۸.

[Fayyad_۹۶] U. M. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. Knowledge discovery and data mining: Towards a unifying framework. In *Knowledge Discovery and Data Mining*, pages ۸۲–۸۸, ۱۹۹۶.

[Feldman_۹۵] R. Feldman and I. Dagan. Kdt - knowledge discovery in texts. In *Proc. of the First Int. Conf. on Knowledge Discovery (KDD)*, pages ۱۱۲–۱۱۷, ۱۹۹۵.