ارائه یک روش مبتنی بر گرایش معنایی برای طبقه‌بندی چندبرچسبی محتوای فیلم‌ها به‌کمک متون زیرنویس آن‌ها

فرید قنبری[1]، دانشجوی کارشناسی ارشد؛ محسن رحمانی[2]، استادیار

۱- دانشکده فنی و مهندسی - گروه مهندسی کامپیوتر - دانشگاه اراک - اراک - ایران - ghanbari.f.70@gmail.com
۲- دانشکده فنی و مهندسی - گروه مهندسی کامپیوتر - دانشگاه اراک - اراک - ایران - m-rahmani@araku.ac.ir

چکیده: پی‌بردن به محتوای تصاویر متحرک و ژانر آن‌ها، همواره امری پیچیده و مسئله‌ای باز برای پژوهشگران بوده است. فعالیت‌های متعددی توسط پژوهشگران برای پی‌بردن به محتوای فیلم‌ها انجام پذیرفته است. اکثر فعالیت‌های صورت‌گرفته در این زمینه با استفاده از پردازش صوت یا تصویر فیلم‌ها بوده است. اخیراً گروهی از پژوهشگران ایده استفاده از متون زیرنویس فیلم‌ها را برای پی‌بردن به محتوای فیلم‌ها مطرح نموده و پردازش متن را سریع‌تر و ساده‌تر از پردازش صوت و تصویر قلمداد نموده‌اند. در این مقاله یک روش مبتنی بر گرایش معنایی برای طبقه‌بندی ژانر در داده چندبرچسبی زیرنویس فیلم‌ها ارائه شده است. برای این‌کار ابتدا یک روش استخراج ویژگی برای استخراج ویژگی‌های یکتا از هر ژانر ارائه شده است. سپس روشی ارائه شده که در آن با محاسبه گرایش معنایی یک زیرنویس به هر ژانر، به پیش‌بینی چندبرچسبی ژانرهای زیرنویس پرداخته می‌شود. درنهایت نیز به‌کمک روش‌های استخراج قوانین باهم‌آیی، ارتباط بین ژانرها در داده خام کشف شده و به‌کمک این قوانین به اصلاح ژانرهای پیش‌بینی‌شده پرداخته می‌شود. نتایج به‌دست‌آمده، بهبود قابل توجه دقت روش پیشنهادی را نسبت به روش‌های پیشین به نمایش می‌گذارد.

واژه‌های کلیدی: طبقه‌بندی ژانر فیلم، متن‌کاوی، پردازش زبان طبیعی، گرایش معنایی، طبقه‌بندی چندبرچسبی، استخراج قوانین باهم‌آیی.

Presenting a Semantic Orientation Based Method for Multi-Label Classification of Movies Content Using Their Subtitle Texts

F. Ghanbari[1], MSc Student; M. Rahmani[2], Assistant Professor

1- Faculty of Engineering, Department of Computer Engineering, Arak University, Arak, Iran, Email: ghanbari.f.70@gmail.com
2- Faculty of Engineering, Department of Computer Engineering, Arak University, Arak, Iran, Email: m-rahmani@araku.ac.ir

Abstract: Understanding movies content and their genre, is always a complex and open issue to researchers. Several activities have been carried out by researchers to find out movies content. Most of the activities conducted in this area have been using audio processing or video processing. Recently a group of researchers have proposed the idea of using movies subtitle texts to understand movies content and considered text processing faster and easier than audio and image processing. In this paper a semantic orientation based method is presented for genre classification in multi-label data of movies subtitles. To do this, a feature extraction method is presented to extract unique features of each genre. Then a method is presented, in which with calculation of a subtitles semantic orientation to each genre, subtitles genres are predicted. Finally, using association rule mining methods, the relationship between genres in raw data is discovered and using these rules, predicted genres have been modified. Obtained results indicate significant improvement of proposed method compared to previous methods.

Keywords: Movie genre classification, text mining, natural language processing, semantic orientation, multi-label classification, association rule mining.

تاریخ ارسال مقاله: ۱۳۹۵/۰۴/۲۲
تاریخ اصلاح مقاله: ۱۳۹۵/۰۸/۰۸
تاریخ پذیرش مقاله: ۱۳۹۵/۰۹/۲۵
نام نویسنده مسئول: محسن رحمانی
نشانی نویسنده مسئول: ایران - اراک - سردشت - پردیس دانشگاه اراک - دانشکده فنی و مهندسی - گروه مهندسی کامپیوتر.

۱- مقدمه

یکی از استانداردهای دسته‌بندی فیلم‌ها، ژانر آن‌ها می‌باشد. ژانر یک فیلم نمایانگر اثری است که آن فیلم در مخاطب خود می‌گذارد. در مراجع مربوط به اطلاعات فیلم‌ها، استاندارد قطعی برای تعیین ژانر فیلم وجود ندارد و عمل تعیین ژانر فیلم توسط گروهی از داوران و با در نظر گرفتن نظرات بینندگان آن فیلم تعیین می‌گردد.

در دنیای امروز که با حجم قابل توجهی از تولید محتواهای تصویری روبرو هستیم، وجود راهکاری سریع و کارآمد برای تعیین محتوای فیلم‌ها بدون مشاهده کل فیلم، می‌تواند کاربردهای فراوانی را در پی داشته باشد. وجود روشی خودکار برای تشخیص ژانر فیلم، کمک مؤثری به کسب و کارهای وابسته به این حوزه و نیز به افراد در انتخاب محتوای مورد علاقه خود خواهد کرد.

از طرفی پی‌بردن به محتوای تصاویر متحرک، امری پیچیده و دشوار و مسئله‌ای باز برای پژوهشگران بوده است. فعالیت‌های متعددی برای پی‌بردن به ژانر فیلم‌ها انجام پذیرفته است. در میان این پژوهش‌ها پژوهشگرانی صدا را برای پردازش انتخاب نموده‌اند. پژوهشگرانی تصویر آنونس فیلم‌ها را مورد بررسی قرار داده‌اند و افرادی نیز به بررسی قاب‌های با مشابهت زیاد در فیلم‌ها پرداخته‌اند. اخیراً یک گروه دیگر از پژوهشگران در پژوهش خود [۱] پردازش سریع‌تر و آسان‌تر متن را نسبت به پردازش صوت و تصویر در فایل‌های چندرسانه‌ای، مورد توجه قرار داده‌اند و روشی برای تشخیص ژانر فیلم‌ها بر اساس زیرنویس آن ارائه کرده‌اند. همچنین در پژوهش ذکرشده، مجموعه داده زیرنویس فیلم‌ها با نام MetaSub معرفی شده است که شامل زیرنویس‌های انگلیسی فیلم‌ها با ژانرهای مختلف است.

از طرفی طبقه‌بندی ژانر فیلم‌ها به‌کمک زیرنویس آن‌ها با چالش‌های جدی روبه‌رو است. اولین مسئله این است که متون زیرنویس فیلم‌ها دارای یک گسستگی ذاتی می‌باشند. به‌گونه‌ای که ممکن است در بخشی از فیلم صحبتی انجام نشده باشد و بنابراین متنی نیز برای آن وجود نداشته باشد و در این حین، موضوع و سناریوی فیلم تغییر نموده باشد. این در حالی است که متن به‌صورت یکپارچه ارائه گردیده است و به‌صورت یکپارچه تجزیه‌وتحلیل می‌گردد. مسئله دیگر، تداخل بسیار زیاد ژانرهای فیلم‌ها است. چراکه در مراجع اطلاعات فیلم‌ها، برای یک فیلم چندین ژانر ذکر می‌شود. درواقع در تشخیص ژانر فیلم‌ها با یک داده چندبرچسبی مواجه هستیم. این امر، تداخل و خطا را در طبقه‌بندی ژانر فیلم‌ها افزایش می‌دهد. آخرین مسئله، سلیقه‌ای بودن تعیین ژانرها می‌باشد که در اکثر متون پژوهشی که در مورد طبقه‌بندی فیلم‌ها بر اساس ژانر آن‌ها منتشر گردیده است به این مطلب اشاره شده است. به این معنی که استاندارد خاصی برای طبقه‌بندی فیلم‌ها به ژانر آن‌ها وجود ندارد و این امر توسط مجموعه وسیعی از بینندگان تعیین می‌گردد و انسان‌ها نیز ممکن است درحدس ژانر فیلم‌ها با خطا عمل نمایند.

برای فائق آمدن بر این چالش‌ها، باید یک روش مناسب برای طبقه‌بندی چندبرچسبی ژانر فیلم‌ها مورد استفاده قرار گیرد. لذا تمرکز این مقاله، بر طبقه‌بندی چندبرچسبی ژانر زیرنویس فیلم‌ها خواهد بود. در یک مسئله چندبرچسبی هر نمونه می‌تواند بیش از یک برچسب کلاس را داشته باشد که برچسب‌ها نسبت به یکدیگر هیچ اولویتی ندارند [۲]. بنابراین در این داده‌ها، پیش‌بینی‌های صورت‌گرفته نیز باید به‌صورت چندبرچسبی باشد و روش‌های پیش‌بینی تک‌برچسبی قابل اعمال بر این داده‌ها نیستند.

تاکنون، تنها پژوهشی که طبقه‌بندی ژانر فیلم را به‌کمک متون زیرنویس آن انجام داده و در عین حال داده مورد استفاده آن چندبرچسبی است، پژوهش [۱] است. پژوهش یادشده، روشی برای استخراج ویژگی‌های یکتا برای هر ژانر ارائه می‌کند سپس با استفاده از تشخیص میزان نزدیکی لغات یک زیرنویس نامشخص با مجموعه ویژگی‌های یکتای هر ژانر، اقدام به پیشنهاد دو ژانر برای زیرنویس نامشخص می‌نماید. همچنین این پژوهش، برای افزایش دقت خود، اقدام به کاهش تعداد برچسب‌ها یا ژانرهای مسئله کرده و به حذف ژانرهایی پرداخته که فیلم‌های آن‌ها میزان اشتراک بیشتری با سایر ژانرها دارند.

در این مقاله یک روش چندبرچسبی برای طبقه‌بندی ژانر فیلم‌ها به‌وسیله متون زیرنویس آن‌ها ارائه شده است. روش ارائه‌شده در این مقاله دارای سه گام اساسی است. در گام اول، روش جدیدی برای استخراج ویژگی‌های یکتا از هر ژانر (برچسب) ارائه شده است. این ویژگی‌ها به این منظور استخراج می‌گردند که ژانرها را از لحاظ مفهومی از یکدیگر متمایز سازند. بر خلاف روش [۱]، در روش ارائه‌شده در این مقاله برای استخراج ویژگی‌های یکتا، روابط به‌گونه‌ای طراحی شده‌اند که واژه‌های غیرمرتبط کم‌تری به مجموعه ویژگی‌های یکتای هر ژانر اضافه شود و واژه‌های مهم هر ژانر در مجموعه ویژگی‌های یکتای آن ژانر قرار گیرند. در گام دوم به ارائه یک روش جدید برای طبقه‌بندی چندبرچسبی پرداخته می‌شود. در این روش، با محاسبه گرایش معنایی یک زیرنویس به مجموعه ویژگی‌های یکتای هر ژانر، به پیش‌بینی چندبرچسبی ژانرهای زیرنویس پرداخته می‌شود. رابطه گرایش معنایی برای مسائل چندکلاسی و چندبرچسبی برای اولین بار در این مقاله طراحی شده است. همچنین این رابطه برخلاف روش مورد استفاده در پژوهش پیشین، به‌گونه‌ای طراحی شده که وابستگی بین برچسب‌های مختلف را نیز در نظر می‌گیرد. در گام سوم، برای اصلاح پیش‌بینی‌های صورت‌گرفته، یک روش مبتنی بر قوانین باهم‌آیی ارائه می‌گردد. همچنین نکته مهم دیگر در خصوص روش ارائه‌شده در این مقاله، این است که برخلاف پژوهش پیشین برای افزایش دقت، از سیاست کاهش ژانرهای مسئله استفاده نشده است.

لازم به ذکر است که در گام اول و دوم برای استخراج ویژگی‌های بهتر و انجام پیش‌بینی‌های صحیح‌تر، از روش‌های موجود در حوزه تحلیل احساسات[1] نیز بهره گرفته شده است. تحلیل احساسات یکی از

حوزه‌های مهم متن‌کاوی است که به بررسی نگرش احساسی متون می‌پردازد. مسائل موجود در این حوزه بررسی می‌کنند که یک جمله یا متن ورودی دارای کدام یک از احساس‌های مثبت، منفی و یا خنثی در یک موضوع خاص است. این عمل معمولاً بر نظرات مربوط به یک محصول، انجمن‌ها، بلاگ‌ها، اخبار و یا میکروبلاگ‌ها اعمال می‌شود [۳]. علت استفاده از روش‌های تحلیل احساسات در این مقاله، تأکید بر استفاده از مفاهیم در روش‌های این حوزه است. به‌عبارت‌دیگر در روش‌های تحلیل احساسات به معانی واژه‌ها و ویژگی‌ها، توجه بیش‌تری صورت می‌گیرد. به‌عنوان‌مثال برخی روش‌ها به استخراج جهت‌گیری معنایی[۲] لغات یا ویژگی‌ها می‌پردازند و برخی دیگر از روش‌ها به کمک الگوهای نحوی زبان، ویژگی‌هایی را از جملات استخراج می‌کنند که دربردارنده مفاهیم و احساسات بیش‌تری هستند.

در ادامه، مقاله به‌صورت زیر سازمان‌دهی شده است. در بخش ۲ به مرور کارهای مرتبط پرداخته شده است. معرفی داده مورد استفاده در مقاله در بخش ۳ صورت گرفته است. بخش ۴ به ارائه روش پیشنهادی برای طبقه‌بندی ژانر فیلم‌ها می‌پردازد. درنهایت نیز ارزیابی و تحلیل نتایج در بخش ۵ و نتیجه‌گیری کارهای صورت گرفته در بخش ۶ ارائه شده است.

۲- کارهای مرتبط

۲-۱- طبقه‌بندی ژانر فیلم به‌کمک پردازش صوت و تصویر آن

اکثر پژوهش‌های صورت گرفته برای تشخیص محتوا و طبقه‌بندی ژانر فیلم‌ها، به‌کمک پردازش صوت و تصویر آن‌ها بوده است. در این بخش برای نشان دادن اهمیت حوزه طبقه‌بندی ژانر فیلم‌ها، به مرور چند مورد از پژوهش‌های حوزه صوت و تصویر پرداخته می‌شود.

در [۴] طبقه‌بندی ژانرها با استفاده از بخش کوتاهی از فیلم صورت می‌پذیرد. روش ارائه‌شده در این پژوهش به این صورت است که هر فیلم، ابتدا به بخش‌هایی به نام قاب‌های کلیدی تقسیم می‌گردد. سپس ویژگی‌های هر قاب کلیدی با استفاده از مدل‌های تشخیص ویژگی GIST[۵] و CENTRIST [۶]، استخراج می‌گردد. سپس با تبدیل این ویژگی‌ها به بردار عددی، با استفاده از روش خوشه‌بندی K-means، اقدام به دسته‌بندی و کشف مشابهت فیلم‌ها می‌نمایند.

در [۷] طبقه‌بندی فیلم‌ها بر اساس ویژگی‌های صدا و تصویر فیلم صورت می‌پذیرد. در این پژوهش، فیلم‌ها به دو دسته پرتحرک[۳] و غیرپرتحرک طبقه‌بندی می‌گردند. سپس کلاس غیر پرتحرک را به سه زیرکلاس وحشت[۴]، کمدی[۵] و درام[۶] تقسیم‌بندی می‌نماید. کلاس پرتحرک نیز به دو زیرکلاس آتش‌سوزی، انفجار و سایر تقسیم‌بندی می‌گردد. برای دسته‌بندی فیلم‌ها به پرتحرک و غیرپرتحرک نیز از فاصله بصری و طول مدت سکانس‌ها استفاده می‌شود.

در [۸] به طبقه‌بندی فیلم‌ها به چهار ژانر کمدی، پرتحرک، درام و وحشت پرداخته شده است. برای انجام این کار ابتدا به استخراج ویژگی‌های محاسباتی از ویدئو همچون متوسط طول تصویر، واریانس رنگ و محتوای حرکت و روشنایی پرداخته شده است. سپس به‌کمک روش mean shift classification به طبقه‌بندی ژانرها پرداخته شده است.

در [۹] روشی برای تشخیص فیلم‌هایی با ژانر پویانمایی ارائه شده است. در این روش ویژگی‌هایی همچون پیچیدگی سیگنال تصاویر استخراج شده سپس به‌کمک طبقه‌بند SVM پیش‌بینی نهایی صورت می‌گیرد.

۲-۲- طبقه‌بندی چندبرچسبی ژانر فیلم به کمک متون زیرنویس آن

تاکنون، تنها پژوهش [۱] به طبقه‌بندی چندبرچسبی متون زیرنویس فیلم‌ها پرداخته است. در پژوهش یادشده، ابتدا این مسئله مطرح می‌شود که پردازش متن به‌مراتب سریع‌تر و آسان‌تر از پردازش صوت و تصویر در فایل‌های چندرسانه‌ای است. لذا برای تشخیص ژانر فیلم‌ها می‌توان از زیرنویس آن‌ها استفاده نمود. در این پژوهش ابتدا مجموعه داده چندبرچسبی زیرنویس فیلم‌ها با نام MetaSub معرفی می‌گردد. سپس برای افزایش دقت، در ابتدا کاهش تعداد برچسب یا ژانر انجام می‌شود و ژانرهایی که فیلم‌های آن‌ها اشتراک بیشتری با سایر ژانرها دارند حذف شده و درنهایت الگوریتم ارائه‌شده تنها بر روی پنج ژانر از هشت ژانر پرکاربرد مجموعه داده، اجرا و ارزیابی می‌گردد. همچنین این پژوهش روشی برای استخراج ویژگی‌های یکتا برای هر ژانر ارائه می‌کند. سپس با استفاده از تشخیص میزان نزدیکی لغات یک زیرنویس نامشخص با مجموعه ویژگی‌های یکتای هر ژانر، اقدام به پیشنهاد دو ژانر برای زیرنویس نامشخص می‌نماید.

۲-۳- طبقه‌بندی چندبرچسبی متون در حوزه‌های دیگر

در این بخش پژوهش‌هایی مورد بررسی قرار می‌گیرند، که به طبقه‌بندی چندبرچسبی متون در حوزه‌های مشابه پرداخته‌اند.

در [۱۰] روش BoosTexter برای طبقه‌بندی چندبرچسبی متون ارائه شده است. این روش که از خانواده الگوریتم‌های Boost است، بر روی داده‌های متنی و صوتی قابل اعمال است. روش ارائه‌شده در این مقاله دارای دو حالت است. در حالت اول، تنها به پیش‌بینی برچسب‌های صحیح پرداخته می‌شود. در حالت دوم، طبقه‌بند برچسب‌ها را رتبه‌بندی می‌کند و برچسب‌های صحیح رتبه بهتری دریافت می‌کنند. از این روش برای طبقه‌بندی چندبرچسبی متون اخبار خبرگزاری رویترز (مجموعه داده Reuters-21578) استفاده شده است.

در [۱۱] برای طبقه‌بندی چندبرچسبی متون، دو مدل مولد احتمالاتی با نام PMM[۷] ارائه شده است. در ارائه این مدل‌های مولد احتمالاتی، از مجموعه کلمات هر متن ورودی استفاده شده است؛ به‌طوری‌که ترتیب رخ دادن کلمات در متن حائز اهمیت نیست بلکه تنها وجود یا عدم وجود کلمه، موردنظر است. از این روش برای طبقه‌بندی متون صفحات وب‌سایت یاهو استفاده شده است.

در [۱۲] برای طبقه‌بندی چندبرچسبی متون، یک تکنیک جدید برای ترکیب ویژگی‌های حاصل از متن و ویژگی‌های حاصل از روابط بین کلاس‌ها، ارائه شده است. این تکنیک می‌تواند در کنار الگوریتم‌های طبقه‌بندی مورد استفاده قرار گیرد. همچنین در این مقاله، از این تکنیک در کنار الگوریتم SVM برای طبقه‌بندی چندبرچسبی استفاده شده است. این روش برای طبقه‌بندی چندبرچسبی مجموعه داده Patents و متون اخبار خبرگزاری رویترز استفاده شده است. مجموعه داده Patents شامل مجموعه متون زبان انگلیسی درباره اختراعات و کاربردهای آن‌ها است.

در [۱۳] برای طبقه‌بندی چندبرچسبی متون الگوریتم MLME، به‌کمک مدل‌سازی ارتباط متقابل بین کلاس‌ها و با استفاده از اصل حداکثر آنتروپی ارائه شده است. این روش درنهایت، طبقه‌بندی چندبرچسبی را به‌کمک محاسبه احتمالات شرطی انجام می‌دهد. روش ارائه‌شده در این مقاله برای طبقه‌بندی چندبرچسبی متون ایمیل‌ها و متون اخبار خبرگزاری رویترز استفاده شده است.

در [۱۴] برای طبقه‌بندی چندبرچسبی متون، از روش انتساب برچسب مبتنی بر آنتروپی استفاده شده است. در این روش بر اساس آنتروپی یک برچسب، وزن آن برچسب به یک سند چندبرچسبی انتساب داده می‌شود. این روش که ELA نام دارد برای طبقه‌بندی چندبرچسبی متون اخبار خبرگزاری رویترز مورد استفاده قرار گرفته است.

در [۱۵] برای طبقه‌بندی چندبرچسبی متون، یک الگوریتم پایه‌ای ارائه شده است. این روش نوعی از Maximum Margin Markov Network است که در آن سلسله‌مراتب طبقه‌بندی به‌صورت یک درخت مارکوف به نمایش گذاشته می‌شود. در این روش مسئله به زیرمسئله‌های کوچک‌تری تجزیه می‌شود که در این زیرمسئله‌ها برای یافتن پاسخ بهینه از روش بهینه‌سازی گرادیان صعودی استفاده می‌شود. برای آزمایش دقت این روش از مجموعه داده متون اخبار خبرگزاری رویترز استفاده شده است.

در [۱۶] برای طبقه‌بندی چندبرچسبی متون، بر استفاده از ارتباط موجود بین برچسب‌ها تاکید شده است. برای این منظور، دو مدل مبتنی بر گراف برای مدل کردن ارتباط بین برچسب‌ها ارائه شده است. مدل گرافی اول با نام CML عملیات یادگیری ماشین را بر روی تمامی جفت برچسب‌ها اعمال می‌کند و مدل گرافی دوم با نام CMLF عملیات یادگیری ماشین را بر روی سه‌تایی‌های *ویژگی-برچسب-برچسب* اعمال می‌کند. برای آزمایش دقت این روش از مجموعه داده متون اخبار خبرگزاری رویترز استفاده شده است.

۲-۴- طبقه‌بندی تک‌برچسبی متون

در مرحله پیش‌پردازش مسائل طبقه‌بندی چندبرچسبی متون، می‌توان از روش‌های موجود در حوزه طبقه‌بندی تک‌برچسبی متون نیز بهره برد. به‌همین منظور در این بخش به مرور چند نمونه از این روش‌ها پرداخته می‌شود. در [۱۷] به‌کمک روش‌های خوشه‌بندی، روشی برای بهبود اسناد بازیابی شده در پرس‌وجوها ارائه شده است. در این روش از شباهت مبتنی بر کلمه برای قرار دادن شبیه‌ترین اسناد کنار هم استفاده شده و برای رتبه‌بندی اسناد از تابع رتبه‌بندی TF-IDF بهره گرفته شده است. در [۱۸] روشی برای دسته‌بندی پرسش‌ها ارائه شده است. برای این کار از بردار ویژگی مجموعه کلمات[۸] پرسش‌ها استفاده شده است. همچنین در این روش برای غنی‌سازی بردار ویژگی مجموعه کلمات، ابتدا انواع پاسخ‌های مورد انتظار پرسش، تعیین می‌شوند. سپس بردار دودویی دیگری با طول برابر با تعداد انواع پاسخ‌ها، به بردار ویژگی اولیه افزوده می‌شود.

۳- جمع‌آوری و انتخاب داده

روش‌های ارائه‌شده در این مقاله بر روی مجموعه داده زیرنویس فیلم‌ها اعمال شده‌اند. این مجموعه داده با نام MetaSub در [۱] معرفی و ارائه شده است. این مجموعه داده شامل ۵۸۱۲ زیرنویس فیلم به‌زبان انگلیسی و فراداده‌های مرتبط به هر فیلم است. مهم‌ترین فراداده مرتبط با هر فیلم، ژانرهای آن هستند. این ژانرها از پایگاه داده اینترنتی فیلم‌ها[۹] استخراج شده‌اند. پایگاه داده اینترنتی فیلم‌ها به‌کمک طیف وسیعی از بینندگان و استفاده از نظرات آنان، حداقل یک ژانر را به هر یک از فیلم‌هایش اختصاص می‌دهد. درواقع در پایگاه داده اینترنتی فیلم‌ها برای هر فیلم یک تا چند ژانر معرفی شده است که هیچیک بر دیگری برتری نداشته و همه به‌عنوان ژانرهای اصلی فیلم شناخته می‌شوند. همین عامل سبب شده است که یک فیلم چند برچسب به‌عنوان ژانر داشته باشد. بنابراین مجموعه داده MetaSub یک مجموعه داده چندبرچسبی است که برای هر فیلم چند ژانر را معرفی می‌کند. جدول ۱ مشخص می‌کند که به‌ازای هر ژانر چه تعداد فیلم در مجموعه داده MetaSub وجود دارد. همچنین این جدول اطلاعاتی را درباره داده خام در اختیار قرار می‌دهد.

در این مقاله از بین ژانرهای موجود در مجموعه داده خام، ژانرهایی برای یک فیلم انتخاب شده است که تعداد فیلم‌های متناظر آن‌ها بیش از ۹۰۰ فیلم باشد. یعنی ژانرهای انتخابی برای فیلم‌ها، هشت ژانر پرتکرارتر جدول ۱ خواهند بود. این هشت ژانر عبارت‌اند از: درام، پرتحرک، کمدی، عاشقانه[۱۰]، جنایی[۱۱]، ماجراجویانه[۱۲]، هیجان‌انگیز[۱۳] و وحشت[۱۴]. این کار به این سبب صورت می‌گیرد که در فرآیند یادگیری و آزمون توسط الگوریتم‌های یادگیری، داده کافی در دسترس باشد. درنهایت سهم هر فیلم از بین هشت ژانر معرفی‌شده، حداقل یک و حداکثر سه ژانر است. سهم برخی از فیلم‌ها نیز دو ژانر بوده است.

جدول ۲ نمونه‌ای از داده‌های نهایی مورد بررسی در این مقاله و اطلاعات آماری در خصوص این داده را به نمایش می‌گذارد. اختلاف بسیار کم متوسط تعداد ژانر برای یک فیلم در داده نهایی مورد استفاده در مقاله (جدول ۲) و داده خام (جدول ۱) نشان می‌دهد که ژانرهای هشتگانه انتخاب شده، ژانرهای اساسی و پرکاربرد این مجموعه داده هستند.

جدول ۱: تعداد فیلم‌های متناظر با هر ژانر در مجموعه داده خام زیرنویس فیلم‌ها

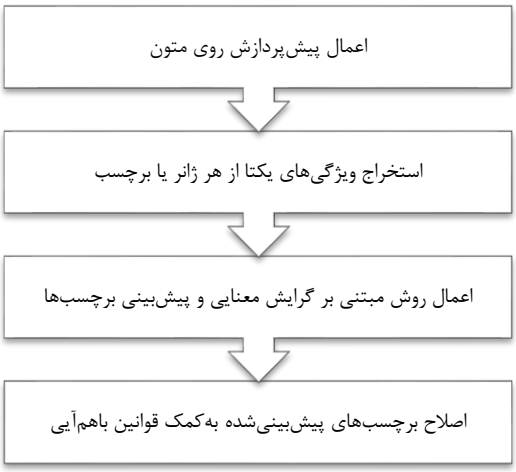
تعداد	ژانر	تعداد	ژانر
۱۸۹	Family	**۲۹۰۰**	**Drama**
۱۱۶	War	**۱۶۴۲**	**Action**
۱۰۹	History	**۱۶۲۳**	**Comedy**
۹۷	Musical	**۱۵۹۰**	**Romance**
۸۲	Biography	**۱۵۲۰**	**Crime**
۶۵	Film-Noir	**۱۰۸۴**	**Adventure**
۵۹	Western	**۱۰۶۲**	**Thriller**
۵۵	Music	**۹۲۶**	**Horror**
۵۲	Short	۶۲۰	Mystery
۴۲	Sport	۴۱۷	Sci-Fi
۱۷	Documentary	۳۵۶	Fantasy
۴	Game-Show	۲۰۹	Animation
تعداد فیلم‌های یکتا: **۵۸۱۲**			
تعداد کل ژانرها (برچسب‌ها): **۲۴**			
متوسط تعداد ژانر (برچسب) برای هر فیلم: **۲،۵۵**			
حداقل و حداکثر تعداد ژانر (برچسب) برای یک فیلم: **حداقل ۱ و حداکثر ۳**			

جدول ۲: نمونه‌ای از داده نهایی مورد استفاده در مقاله و اطلاعات آماری درباره آن

شناسه فیلم	ژانر ۱	ژانر ۲	ژانر ۳
۱	Action	Thriller	–
۲	Romance	Crime	Drama
۳	Horror	–	–
۴	Adventure	Comedy	–
۵	Action	Adventure	Drama
تعداد فیلم‌های یکتا: **۵۸۱۲**			
تعداد کل ژانرها (برچسب‌ها): **۸**			
متوسط تعداد ژانر (برچسب) برای هر فیلم: **۲،۱۲**			
حداقل و حداکثر تعداد ژانر (برچسب) برای یک فیلم: **حداقل ۱ و حداکثر ۳**			

۴- روش پیشنهادی

در این بخش به توضیح روش پیشنهادی برای طبقه‌بندی چندبرچسبی متون زیرنویس فیلم‌ها پرداخته خواهد شد. شکل ۱، فرآیند روش پیشنهادی را به نمایش می‌گذارد.



شکل ۱: فرآیند روش پیشنهادی

۴-۱- اعمال پیش‌پردازش روی متون به‌کمک

همان‌گونه که قبلاً نیز اشاره شد در این مقاله سعی شده است که از روش‌های موجود در زمینه تحلیل احساسات برای استخراج ویژگی‌ها استفاده شود. چراکه روش‌های موجود در این حوزه سعی می‌کنند که به بار معنایی و مفاهیم کلمات توجه بیش‌تری داشته باشند. ازجمله روش‌هایی که در فعالیت‌های حوزه تحلیل احساسات برای استخراج ویژگی صورت می‌گیرد استخراج بخش‌های سخن[۱۵] مؤثر، بر مبنای الگوهای پردازش زبان طبیعی است. منظور از استخراج یا برچسب‌گذاری بخش‌های سخن، مشخص نمودن این مسئله است که هر یک از کلمات یک جمله دارای کدام یک از نقش‌های اسم، فعل، قید، صفت و غیره هستند. استخراج بخش‌های سخن مؤثر به این ترتیب صورت می‌گیرد که پس از استخراج جملات در یک متن، برچسب‌گذاری بخش‌های سخن در هریک از جملات صورت می‌گیرد و درنهایت اسم‌ها، صفات و قیدها از میان بخش‌های سخن مختلف انتخاب می‌شوند. علت انتخاب صفات، قیود و اسامی این است که برطبق پژوهش‌های متعدد انجام‌پذیرفته در زمینه‌ی تحلیل احساسات ازجمله [۲۲-۱۹]، صفات، قیود و اسامی، نقش پررنگ‌تری در انتقال مفاهیم و احساسات دارند.

در ادامه نیز در راستای کاهش ویژگی‌های مورد استفاده، از روش‌های رایج در متن‌کاوی استفاده شده است. این روش‌ها شامل حذف واژه‌های بی‌اثر[۱۶] و حذف کلمات با طول کوچک‌تر یا مساوی سه حرف هستند. در گام بعدی پیش‌پردازش، به‌جهت همسان‌سازی کلمات، ریشه کلمات استخراج شده و جایگزین کلمات اصلی شده است. همچنین تمامی حروف کلمات به حروف کوچک تبدیل شده‌اند. لازم به ذکر است که برای برچسب‌گذاری بخش‌های سخن و استخراج ریشه کلمات[۱۷] از ابزار پردازش زبان طبیعی Stanford CoreNLP [۲۳] استفاده شده است. همچنین برای حذف واژه‌های بی‌اثر، از فهرست واژه‌های بی‌اثر موجود در همین ابزار بهره گرفته شده است. در الگوریتم

شکل ۲، روش پیش‌پردازش متون خام برای انتخاب ویژگی‌های اولیه بیان شده است.

مدل: پیش‌پردازش متون خام برای انتخاب ویژگی‌های اولیه
ورودی: تمامی متون خام موجود در مجموعه داده

شبه کد روش مورد نظر:

```
FOR each Subtitle in Data Set do
    Extract Subtitle Sentences
    FOR each sentence in Subtitle do
        Apply POS tagging
        Extract Nouns, Adjectives and Adverbs
        Remove Stop Words
        Remove words with length ≤ 3
        Lemmatize extracted words
        Convert case of extracted words to lowercase
    END FOR
END FOR
```

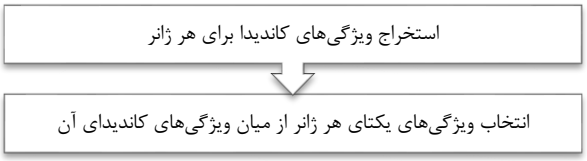
خروجی: مجموعه ویژگی‌های تک‌کلمه‌ای برای هر متن

شکل ۲: الگوریتم پیش‌پردازش متون خام برای انتخاب ویژگی‌های اولیه

۴-۲- استخراج ویژگی‌های یکتا از هر ژانر یا برچسب

پیش‌پردازش‌های صورت‌گرفته و ویژگی‌های استخراج‌شده در مرحله قبل، برای طبقه‌بندی متون چندبرچسبی کافی نیستند. چراکه اشتراک و همپوشانی موجود در ویژگی‌های استخراج‌شده از داده‌های چندبرچسبی باعث می‌شود که کلاس‌ها به یکدیگر نزدیک‌تر گردند. به‌عبارت‌دیگر در داده‌های چندبرچسبی این مقاله، یک متن همزمان در چند کلاس قرار می‌گیرد و همین امر باعث می‌شود که ویژگی‌های مشترک فراوانی در بین کلاس‌های مختلف وجود داشته باشد. این عامل سبب می‌شود که دقت طبقه‌بندی کاهش پیدا کند. از این رو، ویژگی‌های انتخابی باید به‌گونه‌ای باشند که تا حد امکان کلاس‌ها را از لحاظ مفهومی از یکدیگر جدا سازند.

در این بخش به ارائه روشی برای استخراج ویژگی‌های یکتا برای هر ژانر پرداخته می‌شود. باید به این مسئله توجه داشت که منظور از ویژگی‌های یکتا، ویژگی‌هایی هستند که ژانرها را از لحاظ مفهومی از یکدیگر جدا می‌کنند و ممکن است ژانرهایی که از لحاظ مفهومی به هم نزدیک هستند چند ویژگی مشترک نیز داشته باشند. ایده استفاده از ویژگی‌های یکتا برای هر ژانر در [۱] مطرح شده است. اما در پژوهش جاری به طراحی روابط جدیدی برای استخراج ویژگی‌های یکتا برای هر ژانر پرداخته می‌شود. در روش ارائه‌شده در این مقاله، برخلاف روش [۱]، برای استخراج ویژگی‌های یکتا، روابط به‌گونه‌ای طراحی شده‌اند که واژه‌های غیرمرتبط کم‌تری به مجموعه ویژگی‌های یکتای هر ژانر اضافه شود و واژه‌های مهم هر ژانر در مجموعه ویژگی‌های یکتای آن ژانر قرار گیرند. به‌همین منظور ابتدا باید مجموعه آموزش و آزمون ایجاد شود. به این ترتیب که به‌کمک نمونه‌گیری تصادفی، ۲۰ درصد از کل داده به‌عنوان مجموعه آزمون و ۸۰ درصد باقی‌مانده به‌عنوان مجموعه آموزش در نظر گرفته می‌شود. سپس از میان تمامی فیلم‌هایی که در مجموعه آموزش، در یک ژانر (کلاس) قرار گرفته‌اند، ویژگی‌های یکتایی برای آن ژانر (کلاس) استخراج می‌گردد. شکل ۳ مراحل استخراج ویژگی‌های یکتا برای هر ژانر را به نمایش می‌گذارد.



شکل ۳: فرآیند استخراج ویژگی‌های یکتا برای هر ژانر

۴-۲-۱- استخراج ویژگی‌های کاندیدا برای هر ژانر

برای انجام این کار، در رابطه (۱) معیار حد آستانه ویژگی کاندیدا[۱۸] (CFT) را برای یک کلمه تعریف می‌کنیم. به‌کمک این رابطه ویژگی‌های کاندیدا برای یک ژانر قابل استخراج هستند. درواقع به‌کمک این رابطه می‌توان گفت که کلمه w عضو ویژگی‌های کاندیدای ژانر G محسوب می‌شود، اگر معیار CFT(w,G) عددی بزرگ‌تر از حد آستانه T باشد.

$$CFT(w, MainGenre) = (GenreCount - 1) \times \left(\frac{Freq(w, MainGenre) / MainGenreMovieCount}{\sum_{i=1}^{RemGenreCount} \frac{Freq(w, Genre_i)}{Genre_i MovieCount}} \right) \quad (۱)$$

در رابطه (۱)، ضریب GenreCount مشخص‌کننده تعداد کل کلاس‌ها می‌باشد که در این پژوهش کلاس‌های مسئله، ژانرها می‌باشند. متغیر MainGenre، ژانر اصلی است، به‌عبارت دیگر نمایانگر ژانری است که قصد دارد کلمه‌ی w را به‌عنوان یکی از ویژگی‌های کاندیدای خود معرفی نماید. تابع $Freq(w, G)$ مشخص‌کننده‌ی نرخ تکرار کلمه‌ی w در فیلم‌های دارای ژانر G در مجموعه آموزش می‌باشد و MainGenreMovieCount مشخص کننده تعداد فیلم‌هایی است که در مجموعه آموزش داری ژانر MainGenre هستند. متغیر RemGenreCount معرف تعداد ژانرهای باقی‌مانده می‌باشد به‌عبارت‌دیگر، این متغیر نشان‌دهنده تمامی ژانرها به‌جز ژانر اصلی است. متغیر $Genre_i MovieCount$ نیز بیانگر تعداد فیلم‌هایی است که در مجموعه آموزش دارای ژانر $Genre_i$ هستند. همان‌طور که قبلاً ذکر شد کلمه w عضو ویژگی‌های کاندیدای یک کلاس محسوب می‌شود اگر معیار CFT(w) در آن کلاس عددی بزرگ‌تر از حد آستانه T باشد. اگر حد آستانه T مقداری کوچک باشد، کلمات آزادانه‌تر به لیست ویژگی‌های کاندیدا اضافه می‌گردند، بنابراین شاهد لغات نامربوط به ژانر خواهیم بود. از سوی دیگر با افزایش حد آستانه T کلمات به‌صورت سخت‌گیرانه‌تری به لیست ویژگی‌های کاندیدا اضافه می‌گردند، در این حالت نیز ممکن است برخی کلمات مهم ژانر حذف گردند. بنابراین بایستی تا حد امکان، حد وسطی از هر دو بعد را به‌عنوان مقدار حد

آستانه T در نظر بگیریم. حد آستانه T باید متناسب با نوع مسئلهبهصورت تجربی ارزیابی و انتخاب شود.

۴-۲-۲- انتخاب ویژگی‌های یکتای هر ژانر از میان ویژگی‌های کاندیدای آن

پس از استخراج ویژگی‌های کاندیدا برای هر ژانر، حال باید به استخراج ویژگی‌های نهایی ژانرها بپردازیم. درواقع این ویژگی‌های نهایی به‌عنوان ویژگی‌های یکتای هر ژانر شناخته خواهند شد. برای این کار، به‌ازای هر یک از ویژگی‌های کاندیدای یک ژانر، به محاسبه معیار حد آستانه ویژگی یکتا[19] (UFT) پرداخته می‌شود. این معیار در رابطه (۲) معرفی شده است. پس از محاسبه معیار حد آستانه ویژگی یکتا برای ویژگی‌های کاندیدای هر ژانر، ویژگی‌های کاندیدایی که مقدار متناظرشان در رابطه (۲) بزرگ‌تر یا مساوی حد آستانه H باشد، به‌عنوان ویژگی‌های یکتای ژانر معرفی می‌گردند.

$$UFT\left(w,Genre_i\right)=\frac{Freq\left(w,Genre_i\right)}{Genre_i MovieCount} \qquad (۲)$$

در این رابطه w یک ویژگی کاندیدای ژانر $Genre_i$ است. همچنین $Freq(w, Genre_i)$ نشان‌دهنده تعداد دفعاتی است که در مجموعه آموزش، کلمه یا ویژگی w، در فیلم‌های مربوط به $Genre_i$ تکرار شده است. همچنین $Genre_i MovieCount$ نشان‌دهنده تعداد فیلم‌هایی است که در مجموعه آموزش متعلق به ژانر $Genre_i$ هستند.

همان‌گونه که اشاره شد، دو حد آستانه T و H باید متناسب با نوع مسئله به‌صورت تجربی، ارزیابی و انتخاب شوند. لذا برای تعیین مقدار این دو حد آستانه، تمامی مقادیر بازه [۰/۵ ,۲] با فاصله ۰/۱، مورد آزمایش قرار گرفت. در نهایت مقدار ۱/۶ برای حد آستانه T، و مقدار ۰/۷ برای حد آستانه H، باعث ایجاد بهترین نتیجه از لحاظ دقت طبقه‌بندی نهایی شد. درنهایت الگوریتم استخراج ویژگی‌های یکتا برای هر ژانر را می‌توان در قالب الگوریتم شکل ۴ بیان نمود.

مدل: استخراج ویژگی‌های یکتا برای هر ژانر
ورودی: متون مجموعه آموزش (پیش‌پردازش شده توسط الگوریتم شکل ۲)

شبه کد روش مورد نظر:

```
Load Train Set
FOR each Genre G in Train Set do
    FOR each subtitle S in G do
        FOR each word W in S do
            Calculate CFT(W,G)
            IF(CFT(W,G)>1.6) do
                Add W to G.CandidateFeaturesList
                Calculate UFT(W,G)
                IF( UFT(W,G)≥ 0.7) do
                    Add W to G.UniqueFeaturesList
                END IF
            END IF
        END FOR
    END FOR
END FOR
```

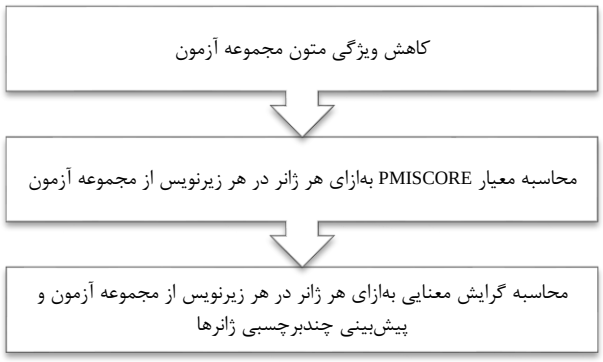
خروجی: ویژگی‌های یکتا برای هر ژانر

شکل ۴: الگوریتم استخراج ویژگی‌های یکتا برای هر ژانر

لازم به ذکر است، از ویژگی‌های یکتای استخراج‌شده برای هر ژانر در مرحله یادگیری استفاده خواهد شد.

۴-۳- اعمال روش مبتنی بر گرایش معنایی و پیش‌بینی برچسب‌ها

در این بخش یک روش مبتنی بر گرایش معنایی برای پیش‌بینی برچسب‌ها ارائه می‌شود. فرآیند پیش‌بینی برچسب‌ها و نحوه اعمال آن بر مجموعه آزمون در شکل ۵ به نمایش گذاشته شده است.



شکل ۵: فرآیند روش مبتنی بر گرایش معنایی

۴-۳-۱- کاهش ویژگی متون مجموعه آزمون

در مرحله پیش‌پردازش داده‌ها، برای تمامی زیرنویس‌های مجموعه آزمون، ویژگی‌های تک‌کلمه‌ای استخراج شد. حال در این بخش، برای هر زیرنویس از مجموعه آزمون، از بین ویژگی‌های تک‌کلمه‌ای استخراج شده برای هر زیرنویس، ویژگی‌هایی که تعداد تکرار آن‌ها بیشتر از دو مرتبه در هر زیرنویس باشند به‌عنوان کلمات مناسب آن زیرنویس انتخاب می‌گردند.

۴-۳-۲- محاسبه معیار PMISCORE به‌ازای هر ژانر در هر زیرنویس از مجموعه آزمون

در گام بعدی معیار PMISCORE به‌گونه‌ای طراحی شده است تا بتوان به‌کمک آن در یک زیرنویس به هر ژانر یک امتیاز مشخص اختصاص داد. برای انجام این کار کلمات انتخاب شده برای هر زیرنویس در مرحله قبل، با کلمات انتخاب‌شده به‌عنوان ویژگی‌های یکتای هر ژانر، به‌صورت دوبه‌دو در نظر گرفته شده و طبق رابطه (۳) معیار PMISCORE بین آن‌ها محاسبه می‌شود.

$$PMISCORE\left(S,G\right)=\left(\sum_{i=0}^{S.count}\sum_{j=0}^{G.count}PMI\left(S_i,G_j\right)\right)/G.count \qquad (۳)$$

در این رابطه، S یک زیرنویس از مجموعه آزمون و G یکی از هشت ژانر موجود در مجموعه داده است. همچنین S.Count تعداد لغات زیرنویس S و G.Count تعداد ویژگی‌های یکتای ژانر G می‌باشد. در رابطه (۳) معیاری بنام PMI بین دو متغیر S_i و G_j محاسبه شده است. متغیر S_i کلمه i ام از مجموعه لغات زیرنویس S و G_j ویژگی j ام از مجموعه

ویژگی‌های یکتای ژانر G است. در ادامه به توضیح نحوه محاسبه معیار PMI پرداخته خواهد شد.

معیار اطلاعات متقابل نقطه‌به‌نقطه[۲۰] یا PMI معیاری است که به‌کمک آن در برخی از پژوهش‌های حوزه تحلیل احساسات همچون پژوهش [۲۴]، شباهت معنایی بین دو کلمه محاسبه می‌شود. اگر x یک کلمه از زیرنویس مورد بررسی و y_i یکی از ویژگی‌های یکتای ژانر i ام باشد، معیار اطلاعات متقابل نقطه‌به‌نقطه موسوم به PMI با توجه به رابطه (۴) محاسبه می‌شود [۲۴].

$$PMI(x,y_i)=\log_2\left(\frac{P(x,y_i)}{P(x)P(y_i)}\right) \quad (۴)$$

در این رابطه برای هر دو کلمه‌ی x و y_i دو احتمال تعریف شده است، که در ادامه روش محاسبه‌ی هر یک توضیح داده خواهد شد. احتمال $P(x,y_i)$ نمایانگر حضور مشترک دو کلمه‌ی x و y_i می‌باشد. در مسئله‌ی موردبحث، x کلمه‌ای از متن زیرنویس است و y_i یکی از ویژگی‌های یکتای ژانر i ام می‌باشد. به‌ازای هر دو کلمه x و y_i ، در تمامی متونی که در مجموعه آموزش با ژانر *i* شناسایی شده‌اند، $P(x,y_i)$ به‌صورت رابطه (۵) محاسبه می‌گردد.

$$P(x,y_i)=\frac{Occurance(x,y,Genre_i,TrainSet)}{NumberOfDocs(Genre_i,TrainSet)} \quad (۵)$$

که دراین رابطه تابع $Occurance(x,y,Genre_i,TrainSet)$ نشان‌دهنده تعداد زیرنویس‌های ژانر i ام در مجموعه آموزش است که هم شامل کلمه x و هم شامل کلمه y باشند. همچنین تابع $NumberOfDocs(Genre_i,TrainSet)$ نشان‌دهنده تعداد زیرنویس‌های ژانر i ام در مجموعه آموزش است.

دو احتمال $P(x)$ و $P(y_i)$ بیانگر حضور کلمه‌ی x یا y از متن زیرنویس، در متون زیرنویس ژانر i می‌باشند که به‌صورت رابطه (۶) و رابطه (۷) محاسبه می‌گردند.

$$P(x)=\frac{Occurance(x,Genre_i,TrainSet)}{NumberOfDocs(Genre_i,TrainSet)} \quad (۶)$$

که در این رابطه تابع $Occurance(x,Genre_i,TrainSet)$ نشان‌دهنده تعداد زیرنویس‌های ژانر i ام در مجموعه آموزش است که شامل کلمه x باشند. همچنین تابع $NumberOfDocs(Genre_i,TrainSet)$ نشان‌دهنده تعداد زیرنویس‌های ژانر iام در مجموعه آموزش است.

$$P(y_i)=\frac{Occurance(y,Genre_i,TrainSet)}{NumberOfDocs(Genre_i,TrainSet)} \quad (۷)$$

که در این رابطه تابع $Occurance(y,Genre_i,TrainSet)$ نشان‌دهنده تعداد زیرنویس‌های ژانر i ام در مجموعه آموزش است که شامل کلمه y باشند. همچنین تابع $NumberOfDocs(Genre_i,TrainSet)$ نشان‌دهنده تعداد زیرنویس‌های ژانر i ام در مجموعه آموزش است.

در ادامه به توضیح چند نکته در ارتباط با رابطه (۳) می‌پردازیم. همان طور که در رابطه (۳) مشخص است، رابطه PMISCORE درنهایت به تعداد ویژگی‌های یکتای ژانر (G.count) تقسیم شده است. این کار بدان جهت است که تعداد ویژگی‌های یکتای ژانرها با یکدیگر متفاوت است و عمل تقسیم بدان‌جهت صورت گرفته که یک نرمال‌سازی در رابطه صورت پذیرد. لازم به ذکر است این رابطه به‌ازای تمامی ژانرهای هشتگانه موجود در مجموعه داده محاسبه می‌شود. بنابراین در هر زیرنویس از مجموعه آزمون، به‌ازای هر ژانر یک معیار PMISCORE محاسبه خواهد شد.

۴-۳-۳- محاسبه گرایش معنایی به‌ازای هر ژانر در هر زیرنویس از مجموعه آزمون و پیش‌بینی چندبرچسبی ژانرها

در برخی از روش‌های تحلیل احساسات ازجمله پژوهش [۲۴] به محاسبه گرایش معنایی ویژگی‌های استخراج‌شده پرداخته می‌شود. همان‌گونه که قبلاً اشاره شد مسائل حوزه تحلیل احساسات دوکلاسه (مثبت و منفی) و تک‌برچسبی هستند و معیارهای محاسبه گرایش معنایی در آن‌ها برای حالت دوکلاسه و تک‌برچسبی طراحی شده‌اند. در این مقاله، با الگو گرفتن از روابط موجود در این حوزه، رابطه جدیدی با نام PMI_SO برای محاسبه گرایش معنایی در مسائل چندکلاسه طراحی شده است. برای انجام این‌کار، پس از محاسبه PMISOCRE ژانرهای مختلف برای یک زیرنویس، به محاسبه PMI_SO برای هر ژانر پرداخته می‌شود. رابطه (۸) نحوه محاسبه PMI_SO را برای زیرنویس S و ژانر G به نمایش می‌گذارد.

$$PMI\_SO(S,G)=PMISCORE(S,G)-\left(\frac{1}{RemGenreCount}\sum_{i=1}^{RemGenreCount}PMISCORE(S,G_i)\right) \quad (۸)$$

که در آن RemGenreCount تعداد ژانرهای باقی‌مانده است (منظور تمامی ژانرها به‌غیر از ژانر G است).

بنابراین در هر فیلم به‌ازای هر یک از ژانرهای هشتگانه یک معیار PMI_SO محاسبه می‌شود. از ویژگی‌های بسیار مهم این رابطه این است که وابستگی بین برچسب‌ها را هم در نظر می‌گیرد. همچنین این رابطه، نشان‌دهنده گرایش معنایی یک زیرنویس به ژانرهای مختلف است. مثبت بودن این رابطه برای یک ژانر در یک زیرنویس، نشان‌دهنده گرایش زیرنویس به آن ژانر است و هر چه این عدد مثبت بزرگ‌تر باشد میزان گرایش به آن ژانر بیشتر است. همچنین منفی بودن این رابطه برای یک ژانر در یک زیرنویس، نشان دهنده عدم گرایش زیرنویس به آن ژانر است و هرچه این عدد منفی بزرگ‌تر باشد میزان عدم گرایش به آن ژانر بیش‌تر است. بنابراین برای پیش‌بینی چندبرچسبی ژانرها به‌کمک رابطه گرایش معنایی به‌صورت زیر عمل می‌کنیم:

اگر در یک زیرنویس از مجموعه آزمون، معیار PMI_SO برای ژانر G عددی بزرگ‌تر از صفر باشد آن زیرنویس دارای ژانر G است و در غیر این صورت آن زیرنویس دارای ژانر G نیست. درواقع در هر فیلم، به‌ازای هر یک از ژانرهای هشتگانه یک پیش‌بینی صورت می‌گیرد و

گفته می‌شود که یک فیلم دارای کدام یک از این هشت ژانر هست و دارای کدام ژانر نیست.

لازم به ذکر است که اگر فیلمی وجود داشته باشد که دارای هیچ ژانری نیست یا حتی دارای ژانری است که در مجموعه داده آموزش ما وجود ندارد؛ باتوجه به اینکه روش ارائه‌شده در این مقاله میزان گرایش معنایی یک فیلم را به هشت ژانر پرتکرار مجموعه داده محاسبه می‌کند و درنهایت پیش‌بینی می‌کند که فیلم ورودی دارای کدام یک از این هشت ژانر هست و دارای کدام ژانر نیست، بنابراین این روش این قدرت را دارد که پیش‌بینی کند که فیلم ورودی دارای هیچ‌یک از این هشت ژانر پرتکرار نیست. لذا در مواجهه با فیلم‌های بدون ژانر یا فیلم‌هایی که دارای ژانری خارج از مجموعه داده هستند، دچار مشکل نمی‌شود.

بنابراین الگوریتم پیش‌بینی چندبرچسبی ژانرها با روش مبتنی بر گرایش معنایی را می‌توان در قالب الگوریتم شکل ۶ بیان نمود.

مدل: پیش‌بینی چندبرچسبی ژانرها با روش مبتنی بر گرایش معنایی

ورودی:

۱. متون مجموعه آزمون (پیش‌پردازش‌شده توسط الگوریتم شکل ۲)

۲. ویژگی‌های یکتای ژانرها

شبه کد روش مورد نظر:

```
FOR each Subtitle S in Test Set do
        FOR each word W1 in S do
          IF (W1.frequency>2) do
                FOR each genre G do
                    FOR each word W2 in G.UniqueFeaturesList do
                            PMISCORE(S,G) += PMI(W1,W2_G)
                    END FOR
                END FOR
          END IF
        END FOR
        FOR each genre G do
                Calculate PMI_SO(S,G)
                IF (PMI_SO(S,G)>0) do
                        S is a film with genre G
                ELSE
                        S is not a film with genre G
                END IF
        END FOR
END FOR
```

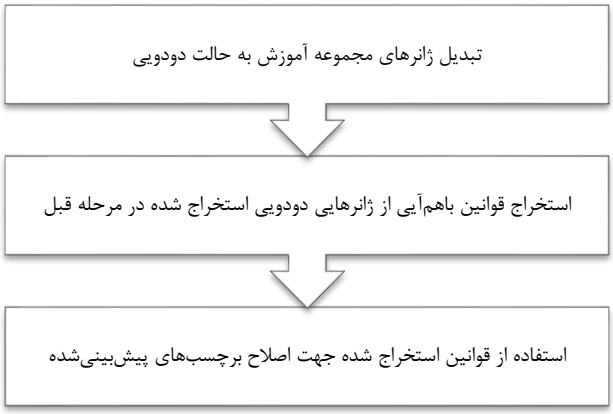
خروجی: برچسب‌های پیش‌بینی‌شده برای هر نمونه از مجموعه آزمون

شکل ۶: الگوریتم پیش‌بینی چندبرچسبی ژانرها با روش مبتنی بر گرایش معنایی

۴-۴- اصلاح برچسب‌های پیش‌بینی‌شده به‌کمک قوانین باهم‌آیی

در مسائل چندبرچسبی باید به این نکته توجه داشت که بین برچسب‌ها ارتباط و وابستگی وجود دارد. بنابراین کشف ارتباط بین برچسب‌ها می‌تواند نقش مؤثری در بهبود برچسب‌های پیش‌بینی‌شده در روش‌های طبقه‌بندی چندبرچسبی ایفا کند. یکی از روش‌های مناسب برای کشف ارتباط میان برچسب‌ها استفاده از قوانین باهم‌آیی است. تاکنون پژوهش‌های متعددی از قوانین باهم‌آیی برای بهبود دقت روش‌های چندبرچسبی خود بهره برده‌اند. در [۲۵]، در فاز پیش‌پردازش داده‌ها، با استفاده از قوانین باهم‌آیی به کاهش برچسب‌های داده اصلی پرداخته شده است و از این طریق به بهبود طبقه‌بندی نهایی کمک شده است. در [۲۶] به استخراج قوانین باهم‌آیی پرداخته شده و از این قوانین در مرحله طبقه‌بندی و در ورودی طبقه‌بند پایه PART که مبتنی بر قانون است، استفاده شده است. در [۲] به‌کمک قوانین باهم‌آیی به استخراج قوانین مثبت و منفی پرداخته شده و به کمک این قوانین به اصلاح برچسب‌هایی که توسط طبقه‌بند پیش‌بینی نشده یا به اشتباه پیش‌بینی شده‌اند، پرداخته شده است.

در این مقاله با الهام گرفتن از روش‌های موجود، روشی برای اصلاح برچسب‌های پیش‌بینی‌شده ارائه شده است. در واقع روش ارائه‌شده یک فرآیند پس‌پردازش تلقی می‌شود. فرآیند روش ارائه شده، در شکل ۷ به نمایش گذاشته شده است.



شکل ۷: فرآیند اصلاح برچسب‌های پیش‌بینی شده به کمک قوانین باهم‌آیی

۴-۴-۱- تبدیل ژانرهای مجموعه آموزش به حالت دودویی

در مجموعه آموزش تنها به این مسئله اشاره شده است که هر زیرنویس داری چه ژانرهایی است و لیست ژانرهایی که هر زیرنویس شامل آن‌هاست در مقابل زیرنویس ذکر شده است. اما در استخراج قوانین، دانش به این مسئله که یک زیرنویس دارای چه ژانرهایی نیست هم یک دانش مهم تلقی می‌شود. درواقع این دانش که هر زیرنویس دارای کدام یک از ژانرها است و دارای کدام یک از ژانرها نیست، دانش مفیدتری نسبت به حالتی است که تنها بدانیم یک زیرنویس دارای چه ژانرهایی است. لذا برای دستیابی به دانش بیشتر اقدام به تبدیل ژانرهای مجموعه آموزش به حالت دودویی می‌نماییم.

در داده اصلی در مجموع، هشت ژانر در کل داده‌ها وجود دارد. بنابراین جدول دودویی تشکیل‌شده باید دارای هشت ستون مربوط به هر ژانر باشد که هر ستون مشخص می‌کند که یک زیرنویس، دارای ژانر مربوط به آن ستون هست یا خیر. شکل ۸ نمونه‌ای از تبدیل ژانرهای مجموعه آموزش را به حالت دودویی به نمایش می‌گذارد.

شناسه فیلم	ژانر ۱	ژانر ۲	ژانر ۳
۱	Action	Thriller	–
۲	Romance	Crime	Drama

الف) نمونه‌ای از جدول آموزش قبل از تبدیل به حالت دودویی

شناسه فیلم	Action	Adventure	Comedy	Crime	Drama	Horror	Romance	Thriller
۱	Yes	No	No	No	No	No	No	Yes
۲	No	No	No	Yes	Yes	No	Yes	No

ب) نمونه‌ای از جدول آموزش پس از تبدیل به حالت دودویی

شکل ۸: روند تبدیل جدول آموزش به حالت دودویی

۴-۴-۲- استخراج قوانین باهم‌آیی از ژانرهایی دودویی استخراج‌شده در مرحله قبل

در این بخش به استخراج قوانین باهم‌آیی به‌کمک ژانرهایی دودویی استخراج‌شده در مرحله قبل پرداخته می‌شود. جدولی که از آن به‌عنوان ورودی الگوریتم استخراج قوانین باهم‌آیی استفاده می‌شود مشابه شکل ۸–ب است. برای استخراج قوانین، از الگوریتم APRIORI با ورودی Confidence بزرگ‌تر از ۰/۹۵ و Support بزرگ‌تر از ۰/۲ استفاده شده است. پس از استخراج قوانین، قوانینی که قابل نتیجه‌گیری از سایر قوانین بودند، حذف شدند. در نهایت قوانین ذکرشده در جدول ۳ به‌عنوان قوانین نهایی جهت اصلاح برچسب‌های پیش‌بینی‌شده معرفی می‌گردند.

جدول ۳: قوانین استخراج‌شده برای اصلاح برچسب‌های پیش‌بینی‌شده

شناسه قانون	قانون	confidence	support
۱	Action=No, Adventure=No, Crime=No, Horror=No, Thriller=No ==> Romance=Yes	۱	۰/۲۳
۲	Romance=Yes ==> Horror=No	۰/۹۸	۰/۲۷
۳	Romance=Yes ==> Thriller=No	۰/۹۷	۰/۲۷
۴	Romance=Yes ==> Adventure=No	۰/۹۶	۰/۲۷
۵	Romance=Yes ==> Action=No	۰/۹۵	۰/۲۷
۶	Comedy=Yes ==> Thriller=No	۰/۹۶	۰/۲۸
۷	Crime=Yes ==> Horror=No, Adventure=No	۰/۹۶	۰/۲۶
۸	Action=Yes ==> Horror=No, Romance=No	۰/۹۵	۰/۲۶

۴-۴-۳- استفاده از قوانین استخراج‌شده جهت اصلاح برچسب‌های پیش‌بینی شده

در این بخش به توضیح نحوه اصلاح برچسب‌های پیش‌بینی‌شده در روش پیشنهادی، به‌کمک قوانین استخراج‌شده در مرحله قبل خواهیم پرداخت. برای این کار یک قاعده کلی مطرح می‌شود. اگر سمت چپ یک قانون، در پیش‌بینی P صدق کند و سمت راست قانون، Yes بودن ژانر G را به نمایش بگذارد، درصورتی‌که ژانر G عضو مجموعه ژانرهای پیش‌بینی‌شده توسط پیش‌بینی P نباشد، ژانر G باید به مجموعه ژانرهای پیش‌بینی‌شده اضافه شود. همچنین اگر سمت چپ یک قانون، در پیش‌بینی P صدق کند و سمت راست قانون، No بودن ژانر G را به نمایش بگذارد، درصورتی‌که ژانر G عضو مجموعه ژانرهای پیش‌بینی‌شده توسط پیش‌بینی P باشد، ژانر G باید از مجموعه ژانرهای پیش‌بینی‌شده حذف شود.

۵- ارزیابی و تحلیل نتایج

در مسائل طبقه‌بندی چندبرچسبی نمی‌توان از معیارهای طبقه‌بندی تک‌برچسبی برای ارزیابی بهره برد. لذا معیارهای ارزیابی در طبقه‌بندی چندبرچسبی باید مجدداً تعریف شوند. حال اگر فرض کنیم که D مجموعه داده آزمون، H طبقه‌بند، Y_i مجموعه برچسب‌های رکورد i ام از مجموعه آزمون و Z_i مجموعه برچسب‌های پیش‌بینی‌شده توسط طبقه‌بند H برای رکورد i ام از مجموعه آزمون باشد، آنگاه معیار Precision یا دقت به‌صورت رابطه (۹) تعریف می‌شود [۱۴].

$$\Pr ecision(H,D) = \frac{1}{|D|}\sum_{i=1}^{|D|}\frac{|Y_i \cap Z_i|}{|Z_i|} \qquad (۹)$$

پیش از ارزیابی روش‌ها برای سادگی کار ابتدا دو تعریف ارائه داده می‌شود:

أ. **پیش‌بینی مثبت:** اگر در یک پیش‌بینی گفته شود که یک فیلم دارای چه ژانرهایی است، یک پیش‌بینی مثبت رخ داده است. برچسب‌های پیش‌بینی‌شده در این روش، برچسب مثبت نامیده می‌شوند.

ب. **پیش‌بینی منفی:** اگر در یک پیش‌بینی گفته شود که یک فیلم دارای چه ژانرهایی نیست، یک پیش‌بینی منفی رخ داده است. برچسب‌های پیش‌بینی‌شده در این روش، برچسب منفی نامیده می‌شوند.

همان‌طور که در بخش‌های قبلی ذکر شد روش پیشنهادی، برای یک فیلم هشت پیش‌بینی انجام می‌دهد که همان تعداد کل برچسب‌های مسئله است. این پیش‌بینی‌ها هم دارای پیش‌بینی مثبت و هم دارای پیش‌بینی منفی هستند. بنابراین ارزیابی روش پیشنهادی می‌تواند در دو حالت رخ بدهد: حالتی که تنها پیش‌بینی‌های مثبت در نظر گرفته می‌شود و حالت دوم که علاوه بر پیش‌بینی‌های مثبت، پیش‌بینی‌های منفی نیز در نظر گرفته می‌شوند. حال با توجه به دو حالت ذکرشده، به ارزیابی روش پیشنهادی و روش‌های ارائه‌شده در گذشته پرداخته می‌شود.

۵-۱- مقایسه دقت روش پیشنهادی با روش پیشین طبقه‌بندی چندبرچسبی زیرنویس فیلم

همان‌گونه که قبلاً اشاره شد، تنها پژوهش موجود در حوزه طبقه‌بندی چندبرچسبی متون زیرنویس فیلم‌ها، پژوهش [۱] است. بنابراین در این بخش به مقایسه روش پیشنهادی با این پژوهش پرداخته می‌شود. نتایج ارزیابی در جدول ۴ قابل مشاهده است. تمامی روش‌ها در حالتی مورد ارزیابی قرار گرفته‌اند که تعداد برچسب‌های مسئله، هشت برچسب بوده است.

در نتایج جدول ۴ مشاهده می‌شود که روش پیشنهادی، بدون اعمال اصلاح برچسب‌ها نیز افزایش دقت چشمگیری در مقایسه با روش پیشین دارد. به‌عبارت‌دیگر بخش اعظمی از بهبود دقت روش پیشنهادی مربوط به روش طبقه‌بندی چندبرچسبی ارائه شده است اما پس از اصلاح برچسب‌های روش پیشنهادی به‌کمک قوانین باهم‌آیی نیز، دقت بهبود یافته است. علاوه‌براین، اصلاح برچسب‌ها سبب شده که معیارهای متوسط برچسب‌های پیش‌بینی‌شده و حداقل و حداکثر برچسب‌های پیش‌بینی شده به معیارهای متناظر در داده اصلی (جدول ۲) نزدیک‌تر شوند. همان‌طور که قبلاً نیز اشاره شد در روش ارائه‌شده در [۱] برای افزایش دقت، در ابتدا کاهش تعداد برچسب یا ژانر انجام شده است و ژانرهایی که فیلم‌های آن‌ها اشتراک بیشتری با سایر ژانرها داشتند حذف شده و در نهایت روش ارائه‌شده تنها بر روی پنج ژانر از هشت ژانر پرکاربرد مجموعه داده، اجرا و ارزیابی شده است. در جدول ارزیابی مشاهده می‌شود که دقت این روش در مسئله هشت‌برچسبی و بدون اعمال سیاست کاهش ژانر، با افت چشمگیری روبه‌رو شده است.

۵-۲- مقایسه دقت روش پیشنهادی با روش‌های پیشین طبقه‌بندی چندبرچسبی سایر حوزه‌ها

در این بخش، برای افزایش اعتبار ارزیابی‌های صورت‌گرفته، روش پیشنهادی با پژوهش‌هایی که به طبقه‌بندی چندبرچسبی متون در حوزه‌های دیگر پرداخته‌اند، مقایسه می‌شود. اکثر روش‌های طبقه‌بندی چندبرچسبی متون، که در گذشته ارائه شده‌اند، برای ارزیابی روش خود از مجموعه داده چندبرچسبی متون اخبار خبرگزاری رویترز (Reuters-21578) استفاده کرده‌اند. بنابراین برای مقایسه روش پیشنهادی با روش‌های طبقه‌بندی چندبرچسبی سایرحوزه‌ها، روش پیشنهادی بر مجموعه داده چندبرچسبی Reuters-21578 اعمال شده و سپس به مقایسه نتایج پرداخته می‌شود. نتایج ارزیابی در جدول ۵ قابل مشاهده است.

جدول۴: مقایسه دقت روش پیشنهادی با روش پیشین طبقه‌بندی چندبرچسبی زیرنویس فیلم

روش	اصلاح برچسب با قوانین باهم‌آیی	نوع پیش‌بینی	دقت (درصد)	متوسط تعداد برچسب‌های پیش‌بینی‌شده		تعداد برچسب‌های مثبت پیش‌بینی‌شده		تعداد برچسب‌های منفی پیش‌بینی‌شده	
				برچسب مثبت	برچسب منفی	حداقل	حداکثر	حداقل	حداکثر
روش [۱]	خیر	مثبت	۱۹/۸۴	۲	–	۲	۲	–	–
روش پیشنهادی	خیر	مثبت	۳۳/۹۰	۳/۵۷	–	۱	۸	–	–
	بله	مثبت	۳۸/۱۲	۲/۶۴	–	۱	۴	–	–
	خیر	مثبت و منفی	۵۸/۹۸	۳/۵۷	۴/۴۳	۱	۸	صفر	۷
	بله	مثبت و منفی	۶۶/۲۵	۲/۶۴	۵/۳۶	۱	۴	۴	۷

جدول۵: مقایسه دقت روش پیشنهادی با روش‌های پیشین طبقه‌بندی چندبرچسبی سایر حوزه‌ها (مجموعه داده Reuters-21578)

روش پیشین	معیار استفاده شده برای ارزیابی دقت	مقدار دقت روش پیشین	مقدار دقت روش پیشنهادی
[۱۲]BoosTexter	Precision	۹۳/۴۰	۹۲/۷۸
[۱۴]	Precision	۹۲/۴۱	۹۲/۷۸
[۱۵]MLME	Accuracy AC (Precision)	۸۸/۵۷	۹۲/۷۸
[۱۶] ELA	micro F_1	۸۶/۵۰	۸۶/۷۱
[۱۷]	Precision | F_1	۹۴/۶ | ۷۶/۳	۹۲/۷۸ | ۸۶/۷۱
[۱۸] CMLF	micro F_1	۸۶/۵۹	۸۶/۷۱

نتایج مقایسه‌ها در جدول ۵، نشان می‌دهد که روش پیشنهادی در اکثر موارد دقت بهتری نسبت به روش‌های پیشین داشته است.

۵-۳- نتایج جانبی

بخش دیگری از نتایج این مقاله مربوط به ویژگی‌های یکتای استخراج‌شده برای هر ژانر است. جدول ۶ تعداد ویژگی‌های یکتای استخراج‌شده برای هر ژانر را به نمایش می‌گذارد.

جدول۶: تعداد ویژگی‌های یکتای استخراج شده برای هر ژانر

ژانر یا برچسب	تعداد ویژگی‌های یکتا
Action	۳۳
Adventure	۶۱
Comedy	۳۶
Crime	۴۲
Drama	۲۰
Horror	۱۴
Romance	۵۱
Thriller	۴

به جهت بزرگ بودن تعداد ویژگی‌ها، تنها ۱۰ ویژگی اول برای هر ژانر در جدول ۷ به نمایش گذاشته شده است. همان‌گونه که قبلاً نیز اشاره شد، ژانرهایی که از لحاظ مفهمی به هم نزدیک‌تر هستند ممکن است دارای چند ویژگی مشترک نیز باشند.

۶- نتیجه‌گیری

در این مقاله یک روش مبتنی بر گرایش معنایی برای طبقه‌بندی ژانر فیلم‌ها به‌کمک زیرنویس آن‌ها ارائه شد. برای این‌کار ابتدا به استخراج ویژگی‌های یکتا برای هر ژانر پرداخته شد. سپس میزان گرایش معنایی یک زیرنویس به ویژگی‌های یکتای هر ژانر محاسبه شده و ژانرهایی که دارای گرایش معنایی بزرگ‌تر از صفر بودند به‌عنوان ژانرهای مثبت فیلم و ژانرهایی که دارای گرایش معنایی کوچک‌تر از صفر بودند به‌عنوان ژانرهای منفی فیلم معرفی شدند. درنهایت نیز به‌کمک قوانین باهم‌آیی، ژانرهای پیش‌بینی‌شده اصلاح گردیدند. نتایج ارزیابی، افزایش دقت روش پیشنهادی را نسبت به روش‌های پیشین به نمایش می‌گذارد. همچنین پس از اصلاح برچسب‌ها به‌کمک قوانین باهم‌آیی، معیارهایی همچون متوسط تعداد ژانر پیش‌بینی‌شده و نیز حداقل و حداکثر تعداد ژانرهای پیش‌بینی‌شده برای یک فیلم به معیارهای متناظر در داده اصلی نزدیک‌تر شدند.

به‌عنوان کارهای آتی می‌توان در مرحله انتخاب ویژگی‌های کاندیدا برای هر ژانر، از معیارهای شباهت و بررسی معنایی جملات نیز بهره برد. در این حالت تمرکز ویژگی‌ها بر معنای انسانی افزایش یافته و این امر می‌تواند سبب بهبود دقت شود.

مراجع

[۱] علی صمیمی، *طبقه‌بندی و تحلیل احساسات در متون با استفاده از الگوریتم های یادگیری ماشین*، پایان‌نامه کارشناسی ارشد مهندسی کامپیوتر گرایش نرم افزار، دانشکده فنی و مهندسی، دانشگاه اراک، ۱۳۹۳.

[۲] خلیل غفوری‌پور و زهرا میرمؤمن، «بهبود رده‌بندی چندبرچسبی بر اساس استخراج قوانین انجمنی مثبت و منفی از روی برچسب‌ها»، *بیست و یکمین کنفرانس ملی سالانه انجمن کامپیوتر ایران، پژوهشگاه دانش‌های بنیادی*، تهران، ۱۳۹۴.

[3] K. Ravi and V. Ravi, "A survey on opinion mining and sentiment analysis: Tasks, approaches and applications," *Knowledge-Based Systems*, vol. 89, pp. 14-46, 2015.

[4] H. Zhou, T. Hermans, A. V. Karandikar, and J. M. Rehg, "Movie genre classification via scene categorization," *Proceedings of the 18th ACM international conference on Multimedia*, Firenze, Italy, pp. 747-750, 2010.

[5] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *International journal of computer vision*, vol. 42, pp. 145-175, 2001.

[6] J. Wu and J. M. Rehg, "Where am I: Place instance and category recognition using spatial PACT," *IEEE Conference on Computer Vision and Pattern Recognition*, Anchorage, Alaska, USA, pp. 1-8, 2008.

جدول ۷: نمونه‌ای از ویژگی‌های یکتای استخراج شده برای هر ژانر

ژانر Action	ژانر Adventure	ژانر Comedy	ژانر Crime	ژانر Drama	ژانر Horror	ژانر Romance	ژانر Thriller
brother	captain	great	money	german	body	girl	information
master	king	love	shit	denny	blood	love	detective
boss	ship	honey	police	crane	ghost	happy	evidence
captain	master	show	case	catherine	evil	miss	agent
power	water	yeah	boss	marie	professor	beautiful	–
officer	lord	actually	murder	shirley	spirit	heart	–
ship	fire	dinner	office	anne	animal	dinner	–
chief	power	wonderful	street	nigger	monster	darling	–
general	human	ball	bank	pierre	soul	wonderful	–
fight	horse	george	frank	nigga	patient	music	–
weapon	earth	cool	crime	allah	victim	evening	–

[7] Z. Rasheed and M. Shah, "Movie genre classification by exploiting audio-visual features of previews," *Proceedings of 16th International Conference on Pattern Recognition*, Quebec, Canada, pp. 1086-1089, 2002.
[8] Z. Rasheed, Y. Sheikh, and M. Shah, "On the use of computable features for film classification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 15, pp. 52-64, 2005.
[9] T. I. Ianeva, A. P. de Vries, and H. Rohrig, "Detecting cartoons: a case study in automatic video-genre classification," *IEEE International Conference on Multimedia and Expo*, Baltimore, USA, 2003.
[10] R. E. Schapire and Y. Singer, "BoosTexter: A boosting-based system for text categorization," *Machine Learning*, vol. 39, pp. 135-168, 2000.
[11] N. Ueda and K. Saito, "Parametric mixture models for multi-labeled text," *16th Annual Neural Information Processing Systems (NIPS) Conference*, Vancouver, Canada, pp. 721-728, 2002.
[12] S. Godbole and S. Sarawagi, "Discriminative methods for multi-labeled classification," *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, Sydney, Australia, pp. 22-30, 2004.
[13] S. Zhu, X. Ji, W. Xu, and Y. Gong, "Multi-labeled classification using maximum entropy method," *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Salvador, Brazil, pp. 274-281, 2005.
[14] W. Chen, J. Yan, B. Zhang, Z. Chen, and Q. Yang, "Document transformation for multi-label feature selection in text categorization," *7th IEEE International Conference on Data Mining (ICDM 2007)*, Omaha, Nebraska, USA, pp. 451-456, 2007.
[15] J. Rousu, C. Saunders, S. Szedmak, and J. Shawe-Taylor, "Kernel-based learning of hierarchical multilabel classification models," *Journal of Machine Learning Research*, vol. 7, pp. 1601-1626, 2006.
[16] N. Ghamrawi and A. McCallum, "Collective multi-label classification," *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, Bremen, Germany, pp. 195-200, 2005.
[۱۷] رضا خدایی، محمدعلی باالفر، و سیدناصر رضوی، «اثربخشی بسط پرس‌وجو مبتنی بر خوشه‌بندی اسناد شبه‌بازخورد با الگوریتم K-NN»، *مجله مهندسی برق دانشگاه تبریز*، دوره ۴۶، شماره ۱، صفحه ۱۵۱-۱۴۳، ۱۳۹۵.
[۱۸] سیده زهرا آفتابی و محمدعلی زارع‌چاهوکی، «کاهش شکاف معنایی در دسته‌بندی پرسش‌ها با بهره‌گیری از قوانین طبقه‌بندی»، *مجله مهندسی برق دانشگاه تبریز*، دوره ۴۶، شماره ۳، صفحه ۲۴-۱۳، ۱۳۹۵.
[19] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Foundations and trends in information retrieval*, vol. 2, pp. 1-135, 2008.
[20] S. Modak and A. C. Mondal, "A Study on Sentiment Analysis," *International Journal of Advanced Research in Computer Science & Technology (IJARCST)*, vol. 2, pp. 284-288, 2014.
[21] F. Benamara, C. Cesarano, A. Picariello, D. R. Recupero, and V. S. Subrahmanian, "Sentiment Analysis: Adjectives and Adverbs are better than Adjectives Alone," *International Conference on Weblogs and Social Media (ICWSM)*, Boulder, USA, 2007.
[22] J. Wiebe, "Learning subjective adjectives from corpora," *Innovative Applications of Artificial Intelligence Conferences (IAAI)*, Austin, Texas, pp. 735-740, 2000.
[23] C. D. Manning, M. Surdeanu, J. Bauer, J. R. Finkel, S. Bethard, and D .McClosky, "The stanford corenlp natural language processing toolkit," *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 55-60, 2014.
[24] L. S. Chen, C.-H. Liu, and H.-J. Chiu, "A neural network based approach for sentiment classification in the blogosphere," *Journal of Informetrics*, vol. 5, pp. 313-322, 2011.
[25] F. Charte, A. Rivera, M. J. del Jesus, and F. Herrera, "Improving multi-label classifiers via label reduction with association rules," *International Conference on Hybrid Artificial Intelligence Systems*, pp. 188-199, 2012.
[26] R. Alazaidah, F. Thabtah, and Q. Al-Radaideh, "A multi-label classification approach based on correlations among labels," *International Journal of Advanced Computer Science and Applications*, vol. 6, pp. 52-59, 2015.

زیرنویس‌ها

[1] Sentiment Analysis
[2] Semantic Polarity
[3] Action
[4] Horror
[5] Comedy
[6] Drama
[7] Parametric Mixture Model
[8] Bag of Word
[9] Internet Movie Database
[10] Romance
[11] Crime
[12] Adventure
[13] Thriller
[14] Horror
[15] Part of Speech (POS)
[16] Stop Words
[17] Lemmatize
[18] Candidate Feature Threshold (CFT)
[19] Unique Feature Threshold (UFT)
[20] Pointwise Mutual Information (PMI)