دسترسی در سایتِ http://jnrm.srbiau.ac.ir

سال دهم، شماره پنجاهم، مهر و آبان ۱۴۰۳

شماره شاپا: ۲۵۸۸-۵۸۸X

دانشگاه آزاد اسلامی، واحد علوم و تحقیقات

پژوهش‌های نوین در ریاضی

بهبود نرخ صحت طبقه‌بندی داده‌های نامتوازن حجیم با الگوریتم‌های یادگیری عمیق

شکوفا مستوفی[۱]، سهراب کردرستمی[۲*]، امیر حسین رفاهی شیخانی[۳]، مرضیه فریدی ماسوله[۴]، سهیل شکری[۵]

[۱و۲و۳و۵] گروه ریاضی، دانشکده ریاضی و علوم کامپیوتر، واحد لاهیجان، دانشگاه آزاد اسلامی، لاهیجان، ایران

[۴] گروه کامپیوتر، دانشکده کامپیوتر و فناوری اطلاعات، دانشگاه احرار، رشت، ایران

تاریخ ارسال مقاله: ۱۴۰۱/۰۷/۲۰ تاریخ پذیرش مقاله: ۱۴۰۲/۰۷/۲۵

چکیده:

در دنیای نوین حجم گسترده‌ای از اطلاعات بصورت متنی و نامتوازن به محیط دیجیتال منتقل شده اند. از طرفی تحلیل و آنالیز داده های نامتوازن حجیم یک ضرورت در این محیط است. آنالیز داده های متنی توسط تکنیکهای یادگیری ماشین، بازیابی اطلاعات هوشمند، پردازش زبان طبیعی یا روشهای مرتبط دیگر ارائه شده اند اما صحت طبقه‌بندی این داده‌ها مشکلی هست که همچنان باقی می‌باشد. هدف از این مقاله ارائه یک سیستم بهبود نرخ صحت طبقه‌بندی داده‌های نامتوازن حجیم است. برای این منظور از الگوریتم‌های یادگیری عمیق جهت پردازش داده‌ها و تولید ویژگی و در نهایت انجام طبقه‌بندی استفاده شده است. داده‌هایی که در این تحقیق مورد تحلیل قرار گرفته‌اند شامل داده‌های حجیم متنی هستند. این روش شامل مجموعه پیش‌پردازش‌ها جهت آماده‌سازی داده و سپس استفاده از یک مدل جهت تولید بردارهای بازنمایی است. در این روش از دو نوع شبکه عمیق استفاده شده است: شبکه‌های کانولوشنی دو بعدی و شبکه‌های LSTM. نتایج بر اساس معیار های دقت و صحت نشان می‌دهد که شبکه‌های دو بعدی پیشنهادی روی مجموعه داده‌های متنی نتایج بهتری را از لحاظ هر دو معیار بیان شده نسبت به شبکه‌های بازگشتی به دست می‌آورند. همچنین تاثیر لایه‌های نرمال‌سازی و تولید بردارهای بازنمایی مورد بررسی قرار گرفته و مشاهده شده است که اهمیت این لایه‌ها به گونه‌ای است که در بعضی موارد می‌تواند تا ۱۵ درصد صحت طبقه‌بندی را افزایش بدهد. نهایتا مدل نهایی که یک مدل دو جریانی از ادغام ویژگی‌های شبکه‌های دو بعدی و بازگشتی است مورد بررسی قرار گرفته و مشاهده شده است که این نوع ادغام می‌تواند تا۲/۵درصد صحت مدل را بهبود ببخشد.

واژه‌های کلیدی: داده‌های نامتوازن، داده‌های حجیم، یادگیری عمیق، شبکه کانولوشن، شبکه LSTM

* عهده‌دار مکاتبات: Email: Sohrabkordrostami@gmail.com

۱- مقدمه

همانطور که می‌دانیم، اینترنت در سال‌های اخیر با سرعت فراوانی در حال رشد و توسعه است که همین امر، موجب بوجود آمدن و در دسترس بودن مقدار بسیار عظیمی از اطلاعات می‌شود. هرچند، به این علت که وب، بدون ساختار است و همچنین ماهیتی دینامیکی دارد، سازماندهی اطلاعات موجود در آن، امری دشوار است. بازیابی اطلاعات، روشی اساسی و بنیادی برای حل آشفتگی و بی نظمی موجود در شبکه است، که همین مساله، جستجوی اینترنتی را امری بسیار حیاتی می‌کند. داده‌های نامتعادل به مشکلات طبقه‌بندی اشاره دارد که در آن موارد نابرابر برای کلاس‌های مختلف داریم. داده های نامتوازن دارای توزیع نابرابر بین رده‌هایش می باشند و می توان با ایجاد مجموعه داده تبدیل‌شده آنها را آموزش داد و بر اساس مساله مورد نظر، طبقه بندی لازم روی داده ها را انجام داد. مطالعات نشان می‌دهد که اغلب دسته‌بندهای پایه، روی داده متوازن بهترعمل می‌کنند همچنین نشان دادند که ترکیب طبقه بندها در طبقه بندی و یادگیری داده های نامتوازن عملکرد بهتری دارند . از آنجایی که حجم اطلاعات الکترونیکی و آنلاین روز به روز بیشتر می شود، دسترسی سریع و صحیح به منابع مهم و مورد علاقه ، یکی از دغدغه های استفاده از این منبع اطلاعاتی بسیار بزرگ است. ارائه ابزارهایی که با بررسی متون بتواند تحلیلی روی آنها انجام دهد منجر به شکل گیری این زمینه در هوش مصنوعی شده که به متن کاوی معروف است. این حوزه ، تمام فعالیتهایی که به نوعی به دنبال کسب دانش از متن هستند را شامل می‌گردد. آنالیز داده های متنی توسط تکنیکهای یادگیری ماشین، بازیابی اطلاعات هوشمند ، پردازش زبان طبیعی یا روشهای مرتبط دیگر، همگی در زمره مقوله متن کاوی قرار می‌گیرند[1]. این تکنیکها در ابتدا در مورد داده های ساخت‌یافته به کار گرفته شدند و علمی به نام داده کاوی را بوجود آوردند. داده های ساخت‌یافته به داده هایی گفته می شود که بطور کاملاً مستقل از همدیگر ولی یکسان از لحاظ ساختاری در یک محل گردآوری شده اند. انواع بانکهای اطلاعاتی را می‌توان نمونه هایی از این دسته اطلاعات نام برد. در اینصورت مسئله داده کاوی عبارت از کسب اطلاعات و دانش از این مجموعه ساخت یافته است. اما در مورد متون، که عمدتاً غیر ساخت یافته یا نیمه ساخت یافته هستند ابتدا باید توسط روشهایی، آنها را ساختارمند نمود و سپس از این روشها برای استخراج اطلاعات و دانش استفاده کرد [2]. به هر حال استفاده از داده کاوی در مورد متن خود شاخه ای دیگر را در علوم هوش مصنوعی بوجود آورد به نام متن کاوی. از جمله فعالیتهای بسیار مهم در این زمینه، طبقه بندی (دسته بندی) متن می‌باشد. طبقه بندی متن، یعنی انتساب اسناد متنی بر اساس محتوی یک یا چند طبقه از قبل تعیین شده است [3]. از شبکه عصبی پیچشی برای حل مسئله طبقه بندی داده های مختلف استفاده می‌شود. مزیت استفاده از این شبکه عصبی این است که مرحله استخراج ویژگی در این شبکه تعبیه شده است و نیازی نیست صریحاً تلاشی در جهت استخراج و انتخاب ویژگی های مناسب صورت گیرد [4]. همچنین برخی محققین متن کاوی را به عنوان ابزاری برای جستجو کردن اطلاعات باکیفیت در حجم زیادی از داده‌ها تعریف می‌کنند. در متن کاوی به مجموعه ای از ارتباط داشتن، نوپدید بودن و موردتوجه بودن اشاره می‌شود. متن کاوی به عنوان آنالیز متفکرانه متن یا داده کاوی متن نیز بیان می‌شود. برای انجام فرایند متن کاوی زمینه های گوناگون تحقیقی وجود دارد. در حال حاضر، تعیین مناسب ترین روش ، جهت به دست آوردن حالت بهینه یک چالش بزرگ محسوب می شود. اگر فقط به داده در سطح وب، مانند شبکه های اجتماعی نگاه کنیم، می توان متوجه شد که چشم انداز داده ها به صورت غیر ساخت یافته و نامتوزان است. این داده ها برای شرکتها، دولتها، خدمات مالی، حوزه ی توسعه تجارت، سازمان های دفاع و پژوهشگران علمی بسیار ارزشمند است[5]. لذا هدف از این تحقیق، ارائه روشی جهت بهبود صحت طبقه بندی داده های نامتوازن حجیم متنی با الگوریتم های یادگیری عمیق می‌باشد. در این تحقیق خاص، از دو نوع متمایز از شبکه‌های عمیق ، شبکه‌های کانولوشنال دو بعدی و

شبکه‌های، استفاده می‌شود. یافته‌های مبتنی بر معیارهای دقت نشان می‌دهد که شبکه‌های دو بعدی پیشنهادی بر روی مجموعه داده‌های متنی نتایج برتر از نظر هر دو معیار نسبت به شبکه‌های بازگشتی به دست می‌آورند. این نتایج مبتنی بر این واقعیت است که شبکه‌های پیشنهادی از مجموعه داده‌های متنی استفاده می‌کنند. علاوه بر این، تأثیر لایه های نرمال‌سازی و همچنین توسعه بردارهای تعبیه شده بررسی شده است و نشان داده شده است که اهمیت این لایه‌ها به حدی است که در برخی شرایط ممکن است دقت طبقه‌بندی را تا ۱۵ درصد افزایش دهد. بررسی مدل نهایی که یک مدل دو جریانی است که ویژگی‌های شبکه‌های بازگشتی و شبکه‌های دو بعدی را ادغام می‌کند.

به جای استفاده از بردارهایی که از قبل آماده شده‌اند، با استفاده از بردارهای بازنمایی شده به عنوان جزئی از فرآیند یادگیری مدل حمایت شده است. از آنجایی که بردارهای موجود برای کاربردهای خاص دیگر آموزش داده شده‌اند، اما تحقیقات بر روی داده‌های نامتعادل متمرکز شده است، به این ترتیب اطمینان حاصل می‌کنیم که بردارهای تعبیه شده آموخته‌شده برای داده‌های نامتعادل استفاده شده در این تحقیق مناسب هستند. به این دلیل است که تحقیقات بر روی داده‌های نامتعادل متمرکز شده است. زمانی که نوبت به تولید بردارهای اصلی برای ایجاد شبکه‌های عمیق رسید، بر مدل GLoVe تکیه کردیم. در ابتدایی‌ترین شکل خود، GloVe یک مدل log-bilinear است که از کوچک‌ترین تابع هدف مربعی استفاده می‌کند. این تحقیق بیشتر بر روی شبکه‌های کانولوشنال عمیق و شبکه‌های بازگشت عمیق به عنوان انواع کلیدی شبکه‌های عمیق برای بررسی متمرکز میباشد. در پایان، دسته‌بندی داده‌های نامتعادل با کمک مدلی که در شبکه‌های عصبی عمیق با استفاده از مجموعه داده‌های آموزش داده شده بودند به‌دست‌آمد. مدل پیشنهادی بر اساس صحت، دقت و بی زمانی بودن آن که سه معیار مجزا هستند مورد ارزیابی قرار گرفت.

GloVe یکی دیگر از روش‌های رایج برای به‌دست آوردن جاسازی های از پیش آموزش دیده است. هدف GloVe دستیابی به دو هدف است: بردارهای کلمه ای ایجاد کند که معنی را در فضای برداری به تصویر می‌کشد، به جای اطلاعات محلی، از آمار شمارش جهانی بهره ببرد. طالب آنلاین زیادی برای توضیح مفهوم GloVe وجود دارد.

اطلاعات جهانی: word2vec به طور پیش فرض هیچ اطلاعات جهانی صریحی را در آن جاسازی نکرده است. GloVe با تخمین احتمال هم‌زمانی یک کلمه با کلمات دیگر، یک ماتریس هم‌روی جهانی ایجاد می‌کند. وجود این اطلاعات جهانی باعث می‌شود GloVe به طور ایده‌آل بهتر کار کند. اگرچه از نظر عملی، آنها تقریباً مشابه کار می کنند.در word2vec، مدل‌های Skipgram سعی می‌کند همزمانی یک پنجره را در یک زمان ثبت کنند. در Glove سعی می‌کند تعداد دفعات نمایش آمار کلی را ثبت کند.

حضور شبکه‌های عصبی: GloVe از شبکه‌های عصبی استفاده نمی‌کند در حالی که word2vec از آن استفاده می‌کند. در GloVe، تابع ضرر، تفاوت بین حاصلضرب جاسازی کلمات و گزارش احتمال وقوع همزمان است. در این تحقیق سعی می‌کنیم آن را کاهش دهیم و از SGD استفاده کنیم اما آن را همانند یک رگرسیون خطی حل می کنیم. در حالی که در مورد word2vec، کلمه را در زمینه آن آموزش می‌دهیم (skip-gram) یا با استفاده از یک شبکه عصبی ۱ لایه پنهان، متن را روی کلمه (کیسه پیوسته کلمات) آموزش می‌دهیم.

در این تحقیق در بخش ۲ به بررسی کارهای انجام شده پرداخته می شود، در ادامه در بخش ۳ معماری مورد استفاده در روش پیشنهادی و در بخش ۴ به ارزیابی و تحلیل روش پیشنهادی و مقایسه آن با روش های مشابه پرداخته می‌شود و در نهایت در بخش ۵ نتیجه گیری از روش پیشنهادی مطرح خواهد شد.

۲- پیشینه تحقیق

با توجه به اینکه موضوع تحقیق از طبقه بندی داده‌های نامتوازن حجیم با استفاده از الگوریتم‌های یادگیری عمیق است، برخی از کارهای حوزه طبقه بندی داده های نامتوازن حجیم با تمرکز روی روش های مبتنی بر الگوریتم‌های یادگیری عمیق بررسی شده‌اند.

با استفاده از دسته‌بندی Naive Bayes کار دسته‌بندی داده‌های نامتوازن متنی انجام شده است. در این مقاله کاهش بعد نمایش فضای برداری اسناد[۱] با استفاده از چند روش انتخاب ویژگی[۲] انجام شد. در این روش ویژگی‌هایی که در هر دسته اهمیت بیش‌تری داشته‌اند شناسایی و انتخاب شده‌اند. هم‌چنین با استفاده از قضیه‌ی تصویر کردن تابع توزیع احتمال[۳]، pdf به دست آمده از فضای با بعد پایین ویژگی‌های انتخابی به فضای با ابعاد بالا[۴] تمام ویژگی‌ها تبدیل شده است. در این مقاله از دو مجموعه‌داده‌ی معروف 20-NEWSGROUPS و REUTERS برای ارزیابی و مقایسه نتایج استفاده شده است. مزیت این روش در این است که در ضمن کاهش بعد و پیچیدگی محاسباتی، ویژگی‌های مهم برای تعیین هر دسته انتخاب شده و باقی می‌مانند. هم‌چنین روش پیشنهادی در این مقاله می‌تواند چندین روش انتخاب ویژگی را ترکیب و به‌طور همزمان مورد استفاده قرار دهد [6].

در مقاله [7]روش خوشه‌بندی[۵] ویژگی‌ها برای کاهش بعد و تسریع محاسبات مورد بررسی قرار گرفته است. با روش پیشنهادی در این مقاله نمایش مستندات در فضای برداری با استفاده از الگوریتم کیسه کلمات[۶] به دست می‌آید و سپس کاهش بعد این نمایش به صورت خودکار با استفاده از خوشه بندی ویژگی‌ها (کلمات) انجام می‌شود. در این روش نیازی نیست تعداد خوشه‌ها از قبل توسط کاربر انتخاب شود. از مزایای این روش این است که به زبان های مختلف بدون تغییر قابل اعمال است و در عین سریع بودن دقت بالایی دارد. در نهایت نویسندگان مقاله کارایی روش پیشنهادی خود را بر روی سه مجموعه‌داده‌ی 20 Newsgroups، RCV1 [۷] و Cade12 مورد بررسی قرار داده‌اند و ادعا کرده‌اند که مدل پیشنهادی آن‌ها در عین کاهش پیچیدگی محاسباتی، دقتی نزدیک به مدل‌های موجود دارد.

در مقاله [8] با استفاده از الگوریتم کلنی مورچه‌ها و هوش تجمعی[۸] تلاش کرده‌اند مسئله‌ی دسته‌بندی متون صفحات وب را حل کنند. در این مقاله نویسندگان تمرکز خود را بر صفحات فارسی گذاشته‌اند و با استفاده از الگوریتم Ant Miner II متون صفحات وب را دسته‌بندی کرده‌اند. هم‌چنین در این مقاله روشی برای پیش پردازش متون صفحات وب بدون در نظر گرفتن ویژگی‌های زبانی به‌کار گرفته شده است. در این پیش پردازش تلاش شده است، بخش‌هایی از صفحه وب که حاوی اطلاعات مفید برای دسته‌بندی آن نیست، حذف شود.

در مقاله [9] روشی برای بهبود الگوریتم دسته‌بندی در حالت چند دسته‌ای ارائه داده‌اند. در این روش که مجموعه‌ی توانی برچسب‌ها نام دارد، هر زیرمجموعه از دسته‌ها خود در قالب یک دسته جدید برچسب گذاری شده و برای آن یک مدل جداگانه آموزش داده می‌شود. در مراحل بعدی با استفاده از نتایج این مدل‌ها برای یک نمونه جدید کار برچسب‌گذاری انجام می‌شود. استفاده از این روش به علت داشتن پیچیدگی محاسباتی بالا در پروژه

---

[1] Documents
[2] Feature Selection
[3] PDF Projection Theorem
[4] high dimensional
[5] Clustering
[6] Bag of words
[7] Reuters Corpus Volume 1
[8] Swarm Intelligence

دسته‌بندی دامنه‌ها ضروری به نظر نمی‌رسد، اما ممکن است در آینده مشخص شود، به کارگیری این روش به افزایش دقت سیستم کمک می‌کند.

یک روش ترکیبی برای دسته‌بندی اسناد با استفاده از روش Naïve Bayes و SVM ارائه کرده‌اند. مزیت این روش در این است که علاوه بر داشتن دقت بالاتر نسبت به روش‌های Naïve Bayes و TFIDF/SVM مدت زمان آموزش مدل نیز در آن بسیار کم‌تر است. نویسندگان این مقاله عملکرد روش پیشنهادی خود را بر روی مجموعه‌داده‌ی 20 NewsGroup آزمایش کرده‌اند. در این روش ابتدا با استفاده از روش Naïve Bayes احتمال تعلق سند به هریک از دسته‌های موجود، محاسبه می‌شود. سپس این بردار احتمالات به عنوان ورودی دسته‌بندی کننده‌ی SVM در نظر گرفته شده و دسته‌بندی نهایی توسط مدل SVM انجام می‌شود [10].

یک بستر برای طبقه بندی داده های نامتوازن به نام Sigspace ارائه شده است. در این روش دسته بندی اسناد، برای هر دسته با توجه به داده‌های آموزشی، یک الگوی کلی استخراج می‌شود. از این الگوها که با نام امضا شناخته می‌شوند برای دسته بندی متون استفاده می‌شود. برای تبدیل متون به بردار از روش Word2Vec و برای کاهش بعد آن از معیار TFIDF استفاده شده است. یادگیری الگوهای مربوط به هر کلاس نیز با استفاده از روش‌های خوشه بندی مثل K-Means ، SOM [1] و مدل‌های مخلوط گوسی [2] استفاده شده است. از مزایای این مدل این است که قابلیت یادگیری افزایشی، توزیع شده و موازی را دارد و می‌تواند برای روی بسترهای تحلیل کلان داده مثل Apache Spark و با بهره گیری از کتابخانه Spark MLlib پیاده سازی شوند. نویسنده این تحقیق مدعی شده است که این روش پیشنهادی توانسته است بر روی مجموعه‌داده‌ی 20-NewsGroup که از مجموعه‌داده‌های معروف برای ارزیابی دسته‌بندی اسناد است، در مقایسه با روش‌های موجود عملکرد بهتری داشته باشد [8].

دسته‌بندی مطالب و متون پزشکی که روزانه در خصوص بیماری‌های متنوع در مراجع مختلف به چاپ می‌رسد، از ملزومات مهم در حوزه داده های نامتوازن خواهد بود. مقاله [11] اشاره می‌کند که روش‌های مبتنی بر وب کاوی کمک شایانی به استخراج مطالب مفید و مرتبط در این حوزه می‌نماید. در این مقاله از روش‌های مبتنی بر ماشین‌های پشتیبان برای دسته‌بندی متون به زبان غیر انگلیسی استفاده شده است. این روش توانمندی خود را در دسته‌بندی داده‌ها با ابعاد بالا به‌خوبی نشان داده است و از یک و یا چندین ابر صفحه برای دسته‌بندی غیرخطی ویژگی‌ها استفاده می‌نماید. ابتدا انواع روش‌های آماده‌سازی متون جهت دسته‌بندی، بویژه در زبان‌های غیر انگلیسی مدنظر قرارگرفته است (Tokenization و Lemmatization) که باعث می‌شوند مراحل یادگیری دقیق‌تر و سریع‌تر انجام شود. از انواع روش‌ها در دسته‌بندی نیز نام برده شده است اما روش SVM به‌عنوان یک روش مبتنی بر یادگیری مورداستفاده قرارگرفته و بر اساس دیتاست های موردنظر نویسنده، نتایج مناسبی ارائه نموده است.

به دسته‌بندی محتوایی وب‌سایت‌های دارای ساختار خاص، همانند وب دایرکتوری‌ها می‌پردازد و دسته‌بندی را بر روی سایت‌های موجود در وب دایرکتوری yahoo که شامل تعداد بسیار زیادی متن از صفحات وب است، انجام می‌دهد. وب دایرکتوری دارای تعداد بسیار زیاد و متنوعی از وب‌سایت‌ها با دسته‌بندی موضوعی محتوایی متفاوت است، درنتیجه دیتاستی که منطبق بر این وب دایرکتوری‌ها ایجاد شود نیز دارای تعداد زیاد و متنوع از داده‌های ناسازگار سلسله مراتبی خواهد بود. تمامی روش‌های سنتی در دسته‌بندی وب‌سایت‌ها برای داده‌های با حجم کمتر و قوانین مشخص ارائه‌شده‌اند و برای داده‌های با حجم و ساختارهای سلسله مراتبی پیچیده‌تر همانند وب دایرکتوری‌ها مناسب نخواهند بود. در این مقاله دیتاست از وب‌سایت‌های دایرکتوری یاهو و در پنج حوزه مختلف بر

[1] Self-Organizing Maps
[2] Gaussian Mixture Model

اساس برچسب‌های تعریف‌شده در آن، مورداستفاده قرارگرفته است. در ادامه، دیتاست بر اساس تعداد برچسب‌های موجود برای وب‌سایت‌ها به سه قسمت تقسیم شده است و در هر بخش الگوریتم‌های یادگیری ماشین همانند SVM برای دسته‌بندی وب‌سایت استفاده شده است. درنهایت با استفاده از روش‌های یادگیری ensemble، به بالاترین دقت ممکن در این دیتاست دست‌یافته است. این روش دسته‌بندی برای موتورهای جستجو نیز کاربردهای مهمی خواهند داشت. برای اطمینان از نتایج نهایی، الگوریتم مقاله با دیتاست دیگری که حاوی تعداد کمتری از دسته‌های داده‌ای است مورد ارزیابی قرارگرفته است و این نتیجه حاصل‌شده است که هم برای داده‌هایی پیچیده مانند وب دایرکتوری و هم برای داده‌هایی در اندازه کوچک‌تر (دیتاست DMOZ) مدل موردنظر به‌خوبی جوابگو خواهد بود و دقت مناسبی را در دسته‌بندی ارائه می‌نماید [10].

در مقاله [13] اشاره می‌نماید که به دلیل حجم بالای داده ها و محتوای تولیدشده توسط کاربران در فضای وب روش‌های سنتی کارایی خود را ازدست‌داده‌اند، همانند روش‌هایی که عمدتاً برای اهداف یادگیری با نظارت به‌طور گسترده مورداستفاده قرار می‌گیرند. درنتیجه لازم خواهد بود الگوریتم‌هایی گسترش یابند که برچسب‌های متنوع از دامنه‌های مختلف را دسته‌بندی نموده و یا حتی عملکرد مناسبی در دسته‌بندی داده‌های بدون برچسب ارائه نمایند. در این مقاله یک شبکه دو مرحله‌ای برای یادگیری لایه‌های مختلف، ارائه‌شده است. برای آموزش لایه‌ها و عملکرد بهتر یادگیری، از روش‌های انتقال یادگیری و ادغام دامنه استفاده‌شده است. در مراحل آموزش برای تحلیل متقابل دامنه‌ها از شبکه عصبی استفاده‌شده است. بر اساس محتویات دیتاست‌های مبدأ می‌توان یک شبکه را آموزش داد و سپس از روش‌های انتقال یادگیری برای دسته‌بندی محتوای وب‌سایت‌های مقصد که هیچ برچسب مشخصی ندارند، استفاده نمود. از ترکیب انواع الگوریتم‌ها برای مقایسه با الگوریتم مقاله استفاده شده و ارزیابی بروی دیتاست انجام شده است (دیتاست جمع‌آوری‌شده برای بررسی اثر متقابل دامنه بر هم توسط آمازون)بوده است. درواقع نویسنده توانسته است به‌عنوان مثال با دسته‌بندی محتوای مربوط به کتاب‌های مورد بازدید، شبکه‌ای را آموزش دهد که بتواند در دامنه‌های مدنظر کاربر، محتواهای مربوط به مشاهده انواع DVD ها را تشخیص دهد و از روش‌های انتقال یادگیری منطبق بر شبکه‌های عصبی استفاده نماید. در ارزیابی‌ها مشخص‌شده این الگوریتم از عملکرد مناسبی برخورداراست. این روش می‌تواند در تحلیل احساسی وب‌سایت‌ها نیز مورداستفاده قرار گیرد.

در مقاله [14]یک روش خودکار مبتنی بر وب کاوی برای دسته‌بندی وب‌سایت بر اساس شبکه‌های عصبی ارائه شده است. این دسته‌بندی بر اساس محتوای وب‌سایت‌ها است و جهت ساخت بردار ویژگی‌ها از روش Boolean استفاده‌شده است بردار ویژگی ۱۲۸ مؤلفه‌ای است که برای انتخاب این ویژگی‌ها از روش IG[1] بهره برده شده است. در این مقاله طبقه‌بندی CMAC مدلی برای طبقه‌بندی متون مبتنی بر محتوا است که قادر به یادگیری از پروفایل کاربران در استفاده از وب‌سایت‌ها است. نتایج مبتنی بر دیتاست نشان می‌دهد مدل پیشنهادی این مقاله دارای یادگیری سریعی است و در مقایسه با سایر الگوریتم‌های طبقه‌بندی (SVM) دقت بیشتری را در دیتاست ها ارائه می‌دهد. همچنین حافظه لازم برای این مدل در مسائل با ابعاد بزرگ مانند طبقه‌بندی متون که در آن تعداد ویژگی‌ها بسیار زیاد است، عملکرد بهتری دارد.

در مقاله [15] که مروری بر جزییات کار تحقیقاتی انجام شده در Textflow می‌باشد، جزییات مراحل انجام طبقه بندی داده‌های نامتوازن حجیم متنی به تشریح آمده است. حال ازآنجاکه محتوای این مقاله از گزارش بر روی چگونگی دسته‌بندی متون تمرکز یافته است، از بیان جزییات انجام شده در پردازش زبان طبیعی در Textflow صرف‌نظر کرده و تنها به تشریح چگونگی دسته‌بندی پلتفرم مذکور پرداخته خواهد شد. دسته‌بندی انجام شده در

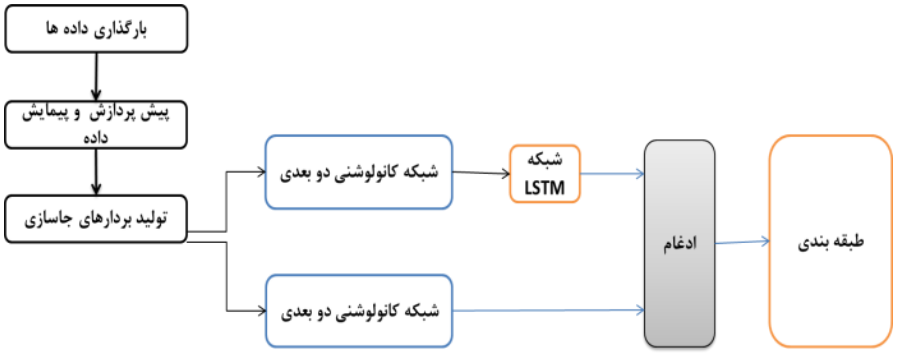
[1] Information Gain

این مقاله از نوع SUPERVISED می‌باشد که با تکیه بر کتابخانه‌های LATINO، NLTK و scikit-learn پیاده‌سازی شده است. نویسنده با استفاده از کتابخانه LATINO اقدام به یکپارچه‌سازی کلاسیفرهای Maximum Entropy و kNN نموده است. همچنین با استفاده از کتابخانه NLTK اقدام به پیاده‌سازی نمونه بهینه‌تری از Naive Bayes نموده و در پایان نیز با استفاده از کتابخانه scikit-learn از کلاسیفرهایی همچون درخت تصمیم، SVM، Gaussian Naive Bayes Classifier وSVM Linear استفاده کرده است.

بطور خلاصه تحقیقات انجام‌شده در زمینه بهبود نرخ صحت طبقه‌بندی داده‌های نامتوازن حجیم با الگوریتم‌های مختلف مورد بررسی قرار گرفت. روش‌هایی که تاکنون ارائه شده‌اند یا دقت کافی را ندارند و یا بسیار زمان بر و پیچیده می‌باشند که مقرون به صرفه نمی‌باشد.

۳- مدل‌سازی

مدل‌های یادگیری عمیق، دسته‌ای از مدل‌ها هستند که می‌توانند سلسله مراتبی از ویژگی‌ها را با ساخت ویژگی‌های سطح بالا از روی ویژگی‌های سطح پایین، یاد بگیرند و از این طریق استخراج ویژگی را خودکار کنند. این ماشین های یادگیری به هر دو صورت با ناظر و بی‌ناظر می‌توانند به کار برده شوند و در هر دو حالت نیز نتایج قابل رقابتی در حوزه‌های تشخیص و پردازش سیگنال نشان داده‌اند. شبکه‌های عصبی کانولوشنی، دسته‌ای از مدل‌های عمیق هستند که در آن فیلترهای قابل آموزش و عملگرهای max pooling به صورت یک در میان روی بردارهای ورودی اعمال می‌شوند و باعث ایجاد یک سلسله مراتب از ویژگی‌ها با افزایش پیچیدگی می‌شوند. نشان داده شده است که اگر این مدل‌ها با تنظیمات خاصی آموزش دیده شوند، می‌توانند بدون تکیه بر ویژگی‌های دستی، نتایج پیشرویی را در زمینه‌های پردازش سیگنال به دست آورند. معماری‌های چند فازی و ادغام ویژگی‌های مختلف نیز به نوبه خود باعث بهبود بیشتر این نتایج شده‌اند. هسته اصلی شبکه‌های کانولوشنی فیلترهای کانولوشن است که روی کل بردار ورودی عمل می‌کنند. ساختار نهایی روش پیشنهادی طبق نمودار جریان شکل ۱ خواهد بود. این ساختار شامل روش دو بعدی پیشنهادی و همچنین ماژولهای اضافه شده جهت بهبود کارایی است.



شکل۱: نمودارجریان حل مساله

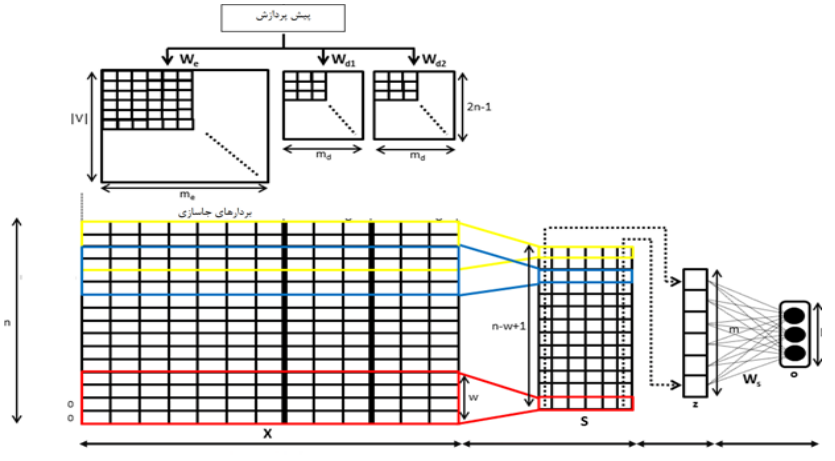
در ادامه به ترتیب مراحل موجود در نمودار جریان فوق تشریح خواهد شد.

در این روش پیشنهادی، ترجیح می‌دهیم که بردارهای بازنمایی نیز بخشی از فرآیند یادگیری مدل باشند و از بردارهای تولید شده آماده استفاده نشود. با این روش مطمن خواهیم شد که بردارهای بازنمایی یادگرفته شده مناسب داده های استفاده شده در این تحقیق خواهد بود زیرا بردارهای موجود آموزش داده شده برای کاربردهای خاص دیگری هستند ولی در این تحقیق روی داده های نامتوازن تمرکز شده است. برای تولید بردارهای اولیه جهت فرآهم سازی برای شبکه‌های عمیق از مدل GLoVe استفاده خواهیم کرد. GloVe اساسا یک مدل log-bilinear با

تابع هدف کوچکترین مربع است. هسته اصلی این مدل بر این شهود استوار است که نرخ احتمال رویداد همزمان چند رکورد باهم، پتانسیل استخراج برخی از ویژگی ها را فراهم می‌کند.

۱-۳: شبکه‌های کانولوشنی و بردارهای بازنمایی

همانطور که از ابتدا شرح داده شد، در این تحقیق از شبکه‌های عمیق جهت یادگیری داده های نامتوازن در متون مختلف استفاده خواهد شد. داده‌های متنی که دارای اطلاعات معنایی و نگارشی هستند از طریق یک روش کارآمد و با استفاده از شبکه‌های کانولوشنی دو بعدی مانند GoogleNet و AlexNet و همچنین با تقویت آن‌ها توسط شبکه‌های دارای حافظه LSTM یادگیری شده‌اند [16]. نوآوری روش پیشنهادی در این مساله است که علاوه بر این که از شبکه‌های کانولوشنی دو بعدی برای یادگیری بردارهای بازنمایی شده عبارات استفاده میشود، از ویژگی‌های این شبکه‌های دو بعدی جهت آموزش شبکه‌های LSTM بهره گرفته شده است. نهایتا پیش‌بینی انجام گرفته توسط هر دو نوع شبکه در تصمیم نهایی اعمال شده است. این کار در لایه ادغام صورت گرفته است. لذا فرآیند کلی این مرحله بصورت خلاصه به این شکل صورت می‌گیرد که در ابتدا بردارهای بازنمایی چندین کلمه در یک عبارت که در مراحل قبل تولید شده‌اند در یک آرایش دو بعدی کنار یکدیگر قرار خواهند گرفت ونتیجه این کار تعدادی ماتریس دو بعدی خواهد بود ، سپس از این ماتریس‌های دو بعدی جهت آموزش شبکه‌های دو بعدی عمیق استفاده شده است. خروجی این شبکه‌های عمیق، یک آرایه از نورون‌های کاملا متصل است که یک بردار یک بعدی را تشکیل می‌دهد. سپس این بردار به یک لایه softmax تزریق می‌شود و یک بردار از احتمالات برای تصمیم‌گیری در مورد برچسب عبارت تولید می‌شود. در راه حل پیشنهادی ، از هر دو خروجی استفاده می‌شود. با توجه به نمودار جریان رسم شده در ابتدای این بخش، در قسمت شبکه‌های عمیق ، از بردار یک بعدی خروجی و از شبکه‌های کانولوشنی دو بعدی، به عنوان بردار ویژگی جهت آموزش شبکه‌های LSTM استفاده شده است. خروجی این شبکه‌های LSTM نیز نهایتا یک بردار احتمالات است برای تصمیم‌گیری در مورد برچسب عبارت است [17]. سپس این دو بردار احتمالات که شامل، بردار خروجی شبکه کانولوشنی دو بعدی و بردار خروجی شبکه LSTM میباشد ، در آخرین مرحله ادغام می‌شوند و نهایتا برچسب عبارت پیش‌بینی می‌شود. این لایه از یک فرآیند تصمیم گیری جهت دخالت دادن نتیجه هر دو شبکه استفاده می‌کند. در ادامه جزئیات عملکرد شبکه‌ها بیان می‌شود. شکل ۲ نحوه قرارگیری بردارهای بازنمایی جهت آموزش یک شبکه کانولوشنی دو بعدی نشان داده شده است.



شکل ۲: نمونه ای از آرایش بردارهای بازنمایی جهت آموزش شبکه کانولوشنی

ورودی شبکه‌های دو بعدی مانند GoogleNet به دلیل داشتن فیلترهای دو بعدی، باید دو بعدی باشد. لذا ابتدا باید داده‌های نامتوازن به یک شکل دو بعدی تبدیل شوند. برای این منظور در این تحقیق عبارات به یک الگوی دو بعدی برای نمایش یک جمله نگاشت می‌شوند. این الگو به این صورت است که تمامی کلمات یک جمله را بصورت طرحی ستونی، مرتب و غیر هم‌پوشان در یک ماتریس دوبعدی نشان می دهد. به این ترتیب کل داده‌های نامتوازن تبدیل به یک ماتریسِ تبدیل می‌شود و در نتیجه برای ورود به شبکه‌های دو بعدی مذکور مناسب می‌باشد. همچنین شبکه‌های دوبعدیِ مذکور، به طور خاص شبکه GoogleNet به دلیل تعداد لایه‌های شبکه و مجموعه فیلترهایی که یاد می‌گیرند محدودیتی روی اندازه ماتریس ورودی ایجاد می‌کنند به این صورت است که اندازه طول و عرض ماتریس ورودی نباید کمتر از ۲۰۰ واحد باشد. این حداقل اندازه برای شبکه‌ای مانند AlexNet که تعداد لایه‌ها و پارامترهای کمتری دارد مناسب است. از طرفی اندازه‌های خیلی بزرگ، زمان یادگیری و تعداد پارامترها را بسیار زیاد خواهد کرد. این عامل در شکل‌دهی الگوی ماتریس دو بعدی تاثیر گذار خواهد بود.

شبکه‌های عصبی کانولوشنی باید توسط ورودی‌هایی با اندازه یکسان تغذیه شوند. لذا اندازه الگوی ماتریسی ایجادشده برای تمامی عبارات متنی باید هم‌اندازه باشد. در این حالت ممکن است فیلتر لبه‌هایی را که ناشی از تغییر داده نامتوازن است یاد بگیرد در حالی که این نوع داده بی‌ارزش است. دلیل بی‌ارزش بودن این داده این است که ارتباط یک رکورد تنها با رکوردهای قبل و بعد خود بصورت نقطه به نقطه است و در نقاط همسایگی اشتراک معنی داری از لحاظ زمانی ندارند. ما جهت جلوگیری از یادگیری این نقاط، دور هر بردار بازنمایی یک padding به اندازه نصف طول فیلتر کانولوشن ایجاد کرده‌ایم. این کار باعث می‌شود که فیلتر کانولوشن در حاشیه‌های بردارها در یک ناحیه کاملا یکنواخت قرار بگیرد. این paddingها نهایتا به اندازه نهایی الگوی ماتریسی اضافه می‌شود. ویژگی مهمی که شبکه‌های کانولوشنی از آن برخوردارند مقاوم بودن در مقابل جابه‌جایی در ماتریس است. در روش پیشنهادی نیز، در صورت تغییر داده های نامتوازن در ماتریس، این انتقال در همان راستا در کلیه بردارها اتفاق خواهد افتاد و در نتیجه کل الگوی ماتریسی دچار این جابه‌جایی می‌شود. لذا این روش نیز در مقابل جابه‌جایی مقاوم خواهد بود.

۲-۳- ادغام

با توجه به توضیح روش پیشنهادی همچنین توضیحاتی که در مورد اهمیت استفاده از ساختارهای چند جریانی و ادغام ویژگی‌های مختلف در بخش قبل بیان شد، در این تحقیق نیز یک روش جهت بهره‌گیری از نتایج مختلف و ادغام آن‌ها ارائه شده است. این نوع ادغام به ادغام‌های دیرهنگام معروف هستند زیرا در قسمت نهایی مدل، نتایج را با هم ترکیب می‌کنند. نوع دیگری از ادغام، ادغام‌های زودهنگام هستند که در مراحل ابتدایی فرآیند ویژگی‌ها با همدیگر ادغام می‌شوند. جهت انجام ادغام از احتمالاتی که لایه softmax هر شبکه تولید می‌کند استفاده شده است. به طور دقیق‌تر هر شبکه به صورت مجزا آموزش داده می‌شود و سپس در هنگام پیش‌بینی برچسبِ یک عبارت، ابتدا احتمالاتی که توسط هر شبکه برای آن عبارت تولید می‌شود، ضرب درایه به درایه[1] می‌شوند و نهایتا ماکزیمم این احتمالات جدید به عنوان پیش‌بینی انجام گرفته برای عبارت ورودی در نظر گرفته می‌شود. اگر فرض کنیم که $\vec{P}_{2d}(C \mid x)$ احتمالات تولید شده برای یک عبارت ورودی توسط شبکه دو بعدی پیشنهادی و $\vec{P}_{LSTM}(C \mid x)$ احتمالات تولید شده برای همان عبارت توسط شبکه LSTM باشد، آن‌گاه

[1] Element wise multiplication

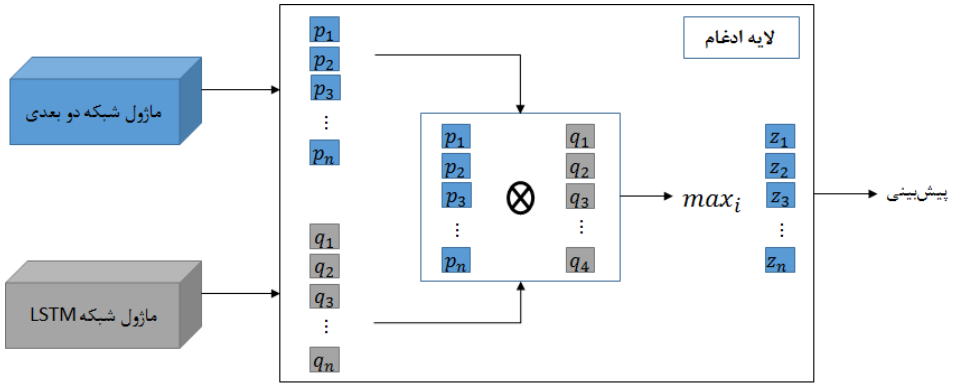
$\vec{P}_{new}(C \mid x)$ امتیازهای پیش‌بینی شده برای آن عبارت بر اساس رابطه زیر خواهد بود. C مجموعه برچسب‌های مجموعه داده است [17]. رابطه ۱ در بصورت زیر می باشد.

$$
\begin{aligned}
&\vec{P}_{2d}(C=i \mid x,W,b) = \\
&Soft\max(Wx+b) = \frac{e^{w_i x+b_i}}{\sum_j e^{w_j x+b_j}} \\
&\vec{P}_{LSTM}(C=i \mid x,W,b) = \\
&soft\max(Wx+b) = \frac{e^{w_i x+b_i}}{\sum_j e^{w_j x+b_j}} \\
&\vec{P}_{new}(C=i \mid x,W,b) = \\
&\vec{P}_{2d}(C=i \mid x,W,b) \otimes \vec{P}_{3d}(C=i \mid x,W,b) \\
&\vec{P}_{2d} = \left[P^1_{2d}, P^2_{2d}, P^3_{2d}, \ldots, P^1_{2d}\right], \\
&\vec{P}_{3d} = \left[P^1_{3d}, P^2_{3d}, P^3_{3d}, \ldots, P^1_{3d}\right]
\end{aligned}
\qquad (۱)
$$

علامت $\otimes$، علامت ضرب درایه به درایه است و 1برابر تعداد دسته‌ها است. در نهایت برچسب پیش‌بینی شده برای عبارات مورد نظر بصورت رابطه زیر خواهد بود [17]. نتیجه در رابطه ۲ نشان داده شده است.

$$C_{\Pr edict} = \arg\max_i P(C=i \mid x,W,b) \quad (۲)$$

خوبی این روش سهولت در پیاده‌سازی است. این سهولت از آن‌جا ناشی می‌شود که دو شبکه بصورت جداگانه آموزش داده می‌شوند و در نتیجه تداخلی بین الگوریتم‌های back propagation در دو شبکه پیش نمی‌آید. شکل ۳ لایه ادغام را نشان می‌دهد.



شکل ۳: نحوه عملکرد لایه ادغام جهت ادغام احتمالات تولید شده توسط شبکه دو بعدی و شبکه LSTM

در این بخش یک روش جدید جهت بهبود صحت طبقه بندی داده های نامتوازن معرفی گردید. این روش شامل مجموعه پیش‌پردازش‌ها جهت آماده سازی داده و سپس استفاده از یک مدل جهت تولید بردارهای بازنمایی است. مرحله اصلی این تحقیق شامل یادگیری عبارت ، با استفاده از شبکه‌های عمیق است. در این روش از دو نوع شبکه عمیق استفاده شده است: شبکه‌های کانولوشنی دو بعدی و شبکه‌های LSTM. نوآوری این روش در دو نکته است:

۱- ابتدا از شبکه‌های کانولوشنی دو بعدی جهت تولید یک سری بردارهای ویژگی یک بعدی برای آموزش شبکه‌های LSTM استفاده شده است .

۲- نتایج احتمالات هر دو شبکه کانولوشنی و شبکه LSTM در پیش‌بینی نهایی دخالت داده شده است که این کار از طریق یک لایه تحت عنوان لایه ادغام انجام گرفته است.

در ادامه با استفاده از دروازه Selection سعی بر انتخاب اطلاعاتی داریم که در گذشته درستی آنها اثبات شده است و باید به درستی در حافظه ثبت شده و توسط شبکه به عنوان یک پروسه یادگیری عمیق جزء پیش بینی ها قلمداد گردد. در حقیقت در این بخش تلاش می کنیم در هر بار اجرای سیستم با اطلاعات جدید، پیش بینی های جدی با کمترین خطای ممکن شکل بگیرند. بنابراین در حالی که تمامی پیش بینی ها و قوانین قبلی را داخل بخش Memory حفظ می کنیم، دروازه ای به نام Selection را به سیستم اضافه می کنیم. این دروازه هم مانند دو دروازه قبلی شبکه عصبی منتسب به خود را دارد و می تواند از اطلاعات جدید و از پیش بینی های مرحله قبلی در راستای اینکه کدام اطلاعات فعلا در بخش Memory نگه داری شوند و کدام بخش از اطلاعات به بخش پیش بینی اضافه شوند، بهره ببرد.

در نهایت ، پس از مرحله قبلی برای پیاده سازی شبکه LSTM که به صورت کامل قادر به استفاده از دانش قبلی جهت یادگیری موارد پیوسته به هم و سلسله مراتبی باشد، دروازه Ignoring را به شبکه اضافه می کنیم. طبیعتا این دروازه هم شبکه عصبی و تابع فعالیت خود را دارا می باشد. به عبارت شفاف‌تر نیاز داریم تا برخی از احتمالات اشتباه و یا کم اهمیت توسط شبکه LSTM نادیده گرفته شوند. در حقیقت در این قسمت به دنبال طراحی یک فیلتر هستیم که شبکه توسط آن قادر به فیلتر کردن احتمالات کم اهمیت تر باشد و احتمالات مفیدتر برای بررسی بیشتر به دروازه ای دیگر و ذخیره شدن در Memory ارسال گردد.

۳-۳- معیارهای ارزیابی

در این تحقیق به منظور ارزیابی روش پیشنهادی، از معیارهای زیر استفاده شده است. این معیارها به گونه‌ای انتخاب شده‌اند که میزان تشخیص صحیح، میزان تشخیص نادرست هر عبارت و زمان را پوشش دهد [18].

دقت : معیار دقت مشخص می‌کند که روش با چه دقتی داده های نامتوازن را شناسایی کرده و عبارات را از هم تفکیک نموده است. به عبارتی این معیار تعیین می‌کند که روش به چه میزان ، نمونه‌ها را از دسته های مختلف به دسته‌ی خودشان تخصیص می‌دهد.

$$\Pr ecision_M = \frac{\sum_{i=1}^{1} \frac{tp_i}{tp_i + fp_i}}{1} \qquad (۳)$$

صحت: معیار صحت مشخص می‌کند که روش با چه دقتی نمونه‌های هم نوع را شناسایی می‌کند. یعنی به چه میزان نمونه‌های هم‌خانواده را به خانواده خودش تخصیص می‌دهد.

$$\mathrm{Re}call_M = \frac{\sum_{i=1}^{1} \frac{tp_i}{tp_i + fn_i}}{1} \qquad (۴)$$

زمان: معیار زمان را در دو فاز می‌توان گزارش کرد. فاز اول مرحله آموزش شبکه عصبی است که از روی مجموعه‌های آموزشی عبارات مختلف را یادگیری می‌کند و فاز دوم مرحله آزمایش است که برچسب یک عبارت ورودی را پیش‌بینی می‌کند. هر کدام از این زمان‌ها اهمیت خاص خود را دارند. لذا در آزمایشات هر دوی این زمان‌ها گزارش خواهد شد. در جدول‌ها و نمودارها از زمان آموزش تحت عنوان زمان۱ و از زمان آزمایش یک نمونه تحت عنوان زمان۲ یاد خواهد شد. لازم به ذکر است که زمان آموزش تا لحظه همگرا شدن شبکه لحاظ شده است.

در این رابطه‌ها tp_i مثبت درست برای دسته fp_i, i ، مثبت نادرست برای دسته fn_i, i منفی نادرست برای دسته $1, i$ تعداد گروه‌ها (دسته‌ها) ی مختلف است.

۳ -۴- نتایج

برای جلوگیری از فرابرازش مدل‌ها از روش drop out استفاده شده است. این روش که یک روش عمومی سازی۴ است با صفر کردن مقادیر نورون‌هایی که بصورت تصادفی در لایه‌های مشخصی انتخاب می‌شوند، از ایجاد الگوهای تکراری در اثر مشاهده داده آموزش جلوگیری می‌کند. در این تحقیق برای هر دو شبکه دو بعدی و بازگشتی، drop out را به لایه‌های کاملا متصل به جز آخرین لایه کاملا متصل و با نرخ حذف ۵۰ درصد اعمال کرده‌ایم. علت حذف آخرین لایه کاملا متصل از الگوریتم drop out این است که این لایه در حقیقت نقش یک لایه softmax را دارد که به تعداد دسته‌های مجموعه داده نورون دارد و هر نورون احتمال عضویت یک نمونه به هر یک از دسته‌ها را تعیین می‌کند. ادامه روند ارزیابی این‌گونه خواهد بود که ابتدا نتایج مدل دوبعدی پیشنهادی بر اساس الگوی پیش-پردازش شده روی مجموعه داده معرفی شده ارائه خواهد شد. این نتایج بر اساس سه معیار ارزیابی که ذکر شده اند بررسی شده است. همچنین تاثیر لایه‌های مختلف پیش‌پردازش روی نتایج مورد بررسی قرار خواهد گرفت، سپس کارکرد این مدل در مقابل شبکه‌های بازگشتی مقایسه خواهد شد. نهایتا در مورد پارامترهای انتخابی برای شبکه و بصری‌سازی ویژگی‌های لایه‌های مختلف شبکه بحث خواهد شد.

۴- نتایج مدلسازی

۴-۱ اثر لایه‌های پیش‌پردازش

لایه‌های نرمال سازی و تولید بردارهای بازنمایی جهت مقاوم سازی مدل در مقابل نویزِ موجود در کلمات مختلف و کاهش افت دقت استفاده شده‌اند. این لایه‌ها مخصوصا در مجموعه داده Quran که وضعیت محتوایی به ازای هر سوره تغییر می‌کند و در کل مجموعه‌داده لحاظ شده‌است اثر چشم‌گیری داشته‌اند.

جدول۱ میزان تاثیر این لایه‌ها و دقت نهایی شبکه را نشان می‌دهد. با توجه به معیارهای ارزیابی نتایج به دست آمده به شرح زیر می‌باشند. در دو مدل بررسی صورت گرفته است، مدل Google Net در بحث دقت با پیش فرض ۶۳/۵ و نرمال سازی ۷۰ ، تولید بردار بازنمایی ۶۶ و در نرمال سازی تولید بردار بازنمایی ۷۲/۵ می باشد و مدل AlexNet در بحث دقت با پیش فرض ۶۰/۶ و نرمال سازی ۶۹ ، تولید بردار بازنمایی۶۴/۱ و نرمال‌سازی تولید بردار بازنمایی ۷۰/۵ می‌باشد. مدل Google Net در بحث صحت با پیش فرض ۵۰ ، نرمال‌سازی ۶۴/۲ و تولید بردار بازنمایی ۵۳ و نرمال‌سازی تولید بردار بازنمایی ۶۶/۵ می باشد .در مدل AlexNet در بحث صحت با پیش فرض ۴۹ ، نرمال سازی ۵۹/۸ ، تولید بردار بازنمایی ۵۴ و نرمال‌سازی تولید بردار بازنمایی ۶۳/۵ می باشد.

جدول۱: تاثیر لایه‌های نرمال‌سازی و تولید بردارهای بازنمایی در معیارهای دقت و صحت برای مجموعه داده

مدل	دقت				صحت			
	پیش فرض	نرمال-سازی	تولید بردار بازنمایی	نرمال-سازی + تولید بردار بازنمایی	پیش فرض	نرمال-سازی	تولید بردار بازنمایی	نرمال-سازی + تولید بردار بازنمایی
GoogleNet	۶۳/۵	۷۰	۶۶	۷۲/۵	۵۰	۶۴/۲	۵۳	۶۶/۵
AlexNet	۶۰/۶	۶۹	۶۴/۱	۷۰/۵	۴۹	۵۹/۸	۵۴	۶۳/۵

در بخش بعد نتایج گزارش شده برای مدل‌های دو بعدی با نتایج به دست آمده برای شبکه‌های بازگشتی مقایسه خواهد شد.

۴-۲ مقایسه شبکه‌های دو بعدی با شبکه‌های بازگشتی حافظه بلند

هدف از ارائه روش دو بعدی پیشنهادی، بهره‌گیری از شبکه‌های عصبی کانولوشنی دو بعدی موجود جهت کلاس بندی داده های نامتوازن حجیم بوده است. این شبکه‌های دو بعدی به دلیل داشتن پیچیدگی محاسباتی پایین‌تر نسبت به شبکه‌های بازگشتی در مقیاس مشابه، زمان آموزش کمتری خواهند داشت و همچنین به دلیل طراحی مناسب معماری این شبکه‌ها، دقت دسته‌بندی داده‌های نامتوازن حجیم توسط این شبکه‌ها از شبکه‌های بازگشتی حافظه دار بیشتر خواهد بود. در این بخش مقایسه‌ای بین عملکرد روش دو بعدی پیشنهادی و مدل‌های بازگشتی انجام گرفته است. برای مدل‌های دو بعدی از همان دو شبکه GoogleNet و AlexNet استفاده شده است و برای مدل‌های حافظه دار از دو مدل LSTM [17] و GRU (واحد بازگشتی دروازه‌ای) [18] بهره گرفته شده است. معماری این دو شبکه نیز در پیوست قابل مشاهده است. همانند شبکه‌های دو بعدی، تفاوت این دو شبکه نیز در عمق آن‌ها، اندازه فیلترهای کانولوشن و لایه‌های کاملا متصل است و این دو شبکه با پیکربندی یکسانی آموزش داده می‌شوند.

جدول۲ نتایج به دست آمده را برای شبکه‌های دو بعدی و بازگشتی مقایسه کرده است. همانطور که مشاهده می‌شود شبکه‌های دو بعدی هم دارای زمان آموزش پایین‌تر و هم دارای دقت تشخیص بالاتری نسبت به نمونه‌های بازگشتی هستند.

جدول۲: مقایسه مدل‌های دو بعدی با مدل‌های بازگشتی

مدل	معیار			
	دقت	صحت	زمان۱	زمان۲
GoogleNet	۷۷	۶۹	۳۶۵	۵
AlexNet	۷۵	۶۵	۱۰۲	۳/۵
LSTM	۷۲	۶۶	۵۶۹	۵/۵
GRU	۷۰	۶۵/۶	۴۰۴	۴

۴-۳ مقایسه با روش‌های دیگر

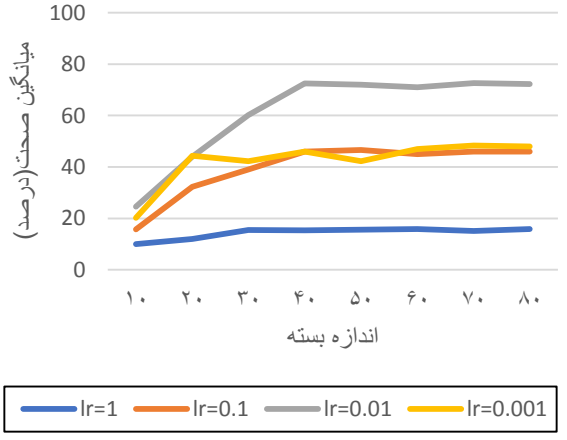
در این بخش نتایج به دست آمده از مدل نهایی پیشنهادی برای مجموعه‌داده‌های نامتوازن با آخرین کارهای انجام گرفته مقایسه شده‌است. برای این منظور، روش پیشنهادی با یک الگوریتم مرسوم در مباحث پردازش متن مقایسه شده است. این روش کلاسیک از الگوریتم Naïve Bayes جهت دسته‌بندی متون استفاده می‌کند. روش کار به این شکل است که ابتدا متن موجود، در این مقاله مجموعه داده های نامتوازن حجیم با دسته‌بندی توحیدی و غیر توحیدی، با یک رویکرد TF-IDF پیش پردازش می‌شوند و این رویکرد یک مجموعه ویژگی‌های غنی از متن که مبتنی بر فرکانس رویداد هر کلمه است استخراج می‌کند. سپس روی این ویژگی‌ها یک الگوریتم Naïve Bayes دوجمله‌ای، چون برچسب‌های مجموعه داده دو گزینه ای است، اعمال می‌شود تا الگوی مشاهده این ویژگی‌ها با توجه به محتوای معنایی یادگرفته شود. در این تحقیق، این الگوریتم را پیاده سازی کرده و نتایج را با روش پیشنهادی مقایسه کرده‌ایم. جدول زیر نتایج این مقایسه را نشان می‌دهد.

جدول ۳: مقایسه دقت روش پیشنهادی با روش Naïve Bayes به دست آمده روی مجموعه داده معرفی شده

مجموعه داده	دقت روش پیشنهادی (٪)	بیش‌ترین دقت گذشته (٪)
Quran Dataset	۷۵	۷۷/۵

۴-۴ جستجوی پارامترها

پیکربندی که تاکنون برای شبکه‌ها استفاده شد برای شبکه‌های دو بعدی و بازگشتی برای همه آزمایش‌ها یکسان بوده است. این پیکربندی شامل پارامترهای متعددی است که هر کدام تاثیر مستقیم یا غیر مستقیمی در کارایی شبکه دارند. اما دو پارامتری که در شبکه‌های عصبی نقش مستقیم و بسیار مهمی در خروجی شبکه‌ها دارند، نرخ یادگیری و اندازه بسته است. در این پژوهش این دو پارامتر از طریق انجام یک جست‌وجوی شبکه‌ای انتخاب شده‌اند. در شکل۴ نتیجه این جست‌وجوی شبکه‌ای برای شبکه‌ی GoogleNet در حالت ارزیابی بین‌شخصی نشان داده شده است. همچنین این بررسی روی مجموعه‌داده Quran انجام گرفته است. لازم به ذکر است که این دو پارامتر علاوه بر دقت ، بر زمان یادگیری نیز تاثیر گذار هستند و با آن رابطه عکس دارند. به طوری که کاهش نرخ یادگیری باعث افزایش زمان یادگیری می‌شود و بالعکس و از طرفی افزایش اندازه بسته باعث کاهش سرعت یادگیری می‌شود .



شکل ۴: جست‌وجوی شبکه‌ای برای انتخاب پارامترهای اندازه بسته و نرخ یادگیری

در این بخش نتایج به دست آمده از روش دوبعدی پیشنهاد شده ارائه شد. این نتایج بر اساس سه معیار دقت، صحت و زمان مورد بررسی قرار گرفتند. از این نتایج مشاهده شد که شبکه‌های دو بعدی پیشنهادی روی مجموعه داده‌های نامتوازن نتایج بهتری را از لحاظ هر سه معیار بیان شده نسبت به شبکه‌های بازگشتی به دست می‌دهند. همچنین تاثیر لایه‌های نرمال‌سازی و تولید بردارهای بازنمایی مورد بررسی قرار گرفت و مشاهده شد که اهمیت این لایه‌ها به گونه‌ای است که در بعضی موارد می‌تواند تا ۲۰ درصد دقت مدل را افزایش بدهد. نهایتا مدل نهایی که یک مدل دو جریانی از ادغام ویژگی‌های شبکه‌های دو بعدی و بازگشتی است مورد بررسی قرار گرفته و مشاهده شد که این نوع ادغام می‌تواند تا ۲/۵درصد دقت مدل را بهبود ببخشد.

۵ – نتیجه گیری

در این مقاله، ترجیح داده شد که بردارهای بازنمایی نیز بخشی از فرآیند یادگیری مدل باشند و از بردارهای تولید شده آماده استفاده نشد. با این روش مطمن خواهیم شد که بردارهای بازنمایی یادگرفته شده مناسب داده‌های نامتوازن استفاده شده هستند؛ زیرا بردارهای موجود آموزش داده شده برای کاربردهای خاص دیگری هستند ولی این تحقیق روی داده های نامتوازن تمرکز داشت. برای تولید بردارهای اولیه جهت فراهم سازی برای شبکه‌های عمیق از مدل GLoVe استفاده کردیم. GloVe اساسا یک مدل log-bilinear با تابع هدف کوچکترین مربع است. عمده شبکه‌های عمیقی که در این تحقیق بررسی شدند، شبکه‌های کانولوشنی عمیق و شبکه‌های بازگشتی عمیق بود. در نهایت هدف این بود که شبکه‌های عمیق با استفاده از مجموعه داده گردآوری شده آموزش داده شوند و سپس از این مدل آموزش دیده جهت طبقه بندی داده های نامتوازن استفاده شود. برای ارزیابی مدل پیشنهادی از سه معیار دقت، صحت و زمان استفاده شده است. همچنین این ارزیابی‌ها در دو حالت درون‌سوره‌ای و بین سوره‌ای انجام گرفته است. ارزیابی‌ها نشان داده‌است که مدل پیشنهادی روی مجموعه‌داده های نامتوازن تا ۴ درصد دقت پیش‌بینی را بهبود داده است. تاثیر لایه‌های اضافه شده به مدل نیز به طور جداگانه مورد بررسی قرار گرفت و مشاهده شد نرمال‌سازی و استفاده از مولد بردارهای بازنمایی تا بیش از ۱۰ درصد دقت مدل را بهبود بخشیده‌اند. در نهایت مشاهده شد که ادغام ویژگی‌های شبکه‌های دو بعدی و بازگشتی می‌تواند تا ۲/۵ درصد دقت مدل را ببهود ببخشد.

فهرست منابع

[1] Jang, J., Kim, Y., Choi, K. and Suh, S., 2021. Sequential targeting: A continual learning approach for data imbalance in text classification. Expert Systems with Applications 179: 115067.

[2]Tarekegn, A., Giacobini, M. and Michalak, K., 2021. A Review of Methods for Imbalanced Multi-Label Classification. Pattern Recognition 118:107965.

[3]Luo, X., 2021. Efficient english text classification using selected machine learning techniques. Alexandria Engineering Journal: 60(3): 3401-3409.

[4]BaniAsadi, A. and Babaali, B., 2020. Power Quality Disturbances Classification Using Identity Feature Vector and Support Vector Machine. Journal of Soft Computing and Information Technology 9(2): 151-164.

[5]Golestanifar, B. and Chalechale, A., 2021. Determination of Mental States from Texts Using Evolutionary Imperialist Competitive Algorithm and Convolution Neural Networks. Journal of Soft Computing and Information Technology 10(1): 13-23.

[6]Xiao, Y., Li, Y., Yuan, J., Guo, S., Xiao, Y. and Li, Z., 2021. History-based attention in Seq2Seq model for multi-label text classification. Knowledge-Based Systems 224: p.107094.

[7]Bhumika, P.S.S.S. and Nayyar, P.A., 2013. A review paper on algorithms used for text classification. International Journal of Application or Innovation in Engineering & Management 3(2): 90-99.

[8]Singh, J.N. and Dwivedi, S.K., 2012. Analysis of vector space model in information retrieval. International Journal of Computer Application (IJCA):14-18.

[9]Ting, S.L., Ip, W.H. and Tsang, A.H., 2011. Is Naive Bayes a good classifier for document classification. International Journal of Software Engineering and Its Applications 5(3): 37-46.

[10]Kim, S.B., Han, K.S., Rim, H.C. and Myaeng, S.H., 2006. Some effective techniques for naive bayes text classification. IEEE transactions on knowledge and data engineering: 18(11): 1457-1466.

[11]Li, Z., Zhang, Y., Wei, Y., Wu, Y. and Yang, Q., 2017, August. End-to-End Adversarial Memory Network for Cross-domain Sentiment Classification. In IJCAI (pp. 2237-2243).

[12]Fang, W., Luo, H., Xu, S., Love, P.E., Lu, Z. and Ye, C., 2020. Automated text classification of near-misses from safety reports: An improved deep learning approach. Advanced Engineering Informatics 44: 101060.

[13]Chen, J., Huang, H., Tian, S. and Qu, Y., 2009. Feature selection for text classification with Naïve Bayes. Expert Systems with Applications 36(3): 5432-5435.

[14]Sun, A., Lim, E.P. and Liu, Y., 2009. On strategies for imbalanced text classification using SVM: A comparative study. Decision Support Systems 48(1): 191-201.

[15]Thirumala, K., et al., 2019, A classification method for multiple power quality disturbances using EWT based adaptive filtering and multiclass SVM, Neurocomputing. 334: p. 265-274

[16]Goel, K., Vohra, R. and Bakshi, A., 2014, September. A novel feature selection and extraction technique for classification. In 2014 14th International Conference on Frontiers in Handwriting Recognition :104-109. IEEE.

[17]Chen, C. and Dai, J., 2021. Mitigating backdoor attacks in lstm-based text classification systems by backdoor keyword identification. Neurocomputing 452: 253-262.

[18]Li, Y., Guo, H., Zhang, Q., Gu, M. and Yang, J., 2018. Imbalanced text sentiment classification using universal and domain-specific knowledge. Knowledge-Based Systems 160: 1-15.

[19]Chen, Y.H., Zheng, Y.F., Pan, J.F. and Yang, N., 2013, November. A hybrid text classification method based on K-congener-nearest-neighbors and hypersphere support vector machine. In 2013 International Conference on Information Technology and Applications (pp. 493-497). IEEE.

[20]Cristian, P. and Elena, B.M., 2019. Dealing with Data Imbalance in Text Classification [J]. Procedia Computer Science 159: 736-745.

[21]Pop, I., 2006. An approach of the Naive Bayes classifier for the document classification. General Mathematics, 14(4): 135-138.

[22]Thabtah, F., Hammoud, S., Kamalov, F. and Gonsalves, A., 2020. Data imbalance in classification: Experimental evaluation. Information Sciences, 513: 429-441.

[23]Tsatsaronis, G. and Panagiotopoulou, V., 2009, April. A generalized vector space model for text retrieval based on semantic relatedness. In Proceedings of the Student Research Workshop at EACL 2009 (pp. 70-78).

[24]Atefeh BaniAsadi, bagher babaali.2020, Power Quality Disturbances Classification Using Identity Feature Vector and Support Vector Machine,Journal Of Soft Computing and Information Technology, pp. 151-164.

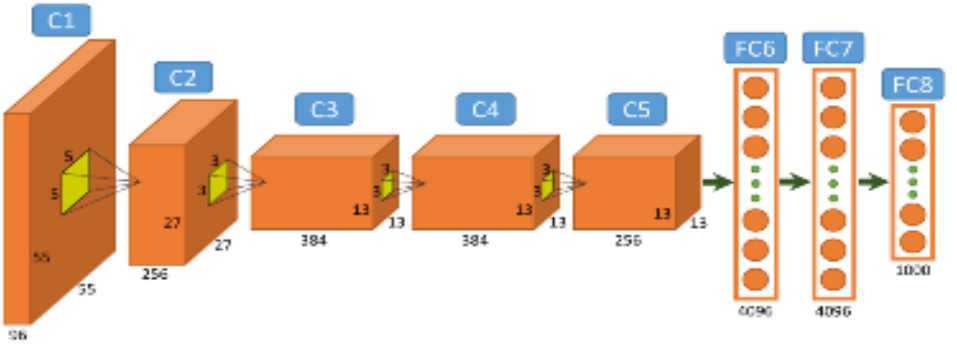
[25]Beniwal, R. K., Saini, M. K., Nayyar, A., Qureshi, B., & Aggarwal, A, 2021, A critical analysis of methodologies for detection and classification of power quality events in smart grid. IEEE Access, 9, 83507–83534.

[26]M. Buda et al. October 2018**,**A systematic study of the class imbalance problem in convolutional neural networks, Neural Networks,Volume 106, Pages 249-259.

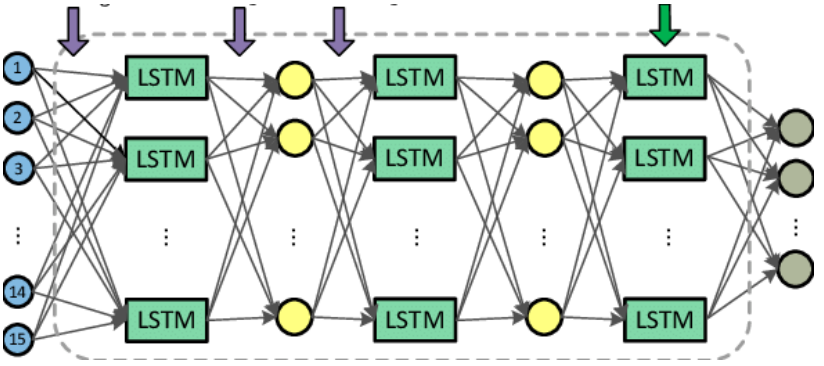
[27]S.G. Burdisso et al., 2019,A text classification framework for simple and effective early depression detection over social media streams, Neural Networks, Volume 133, Expert Systems With Applications, Elsevier.

پیوست

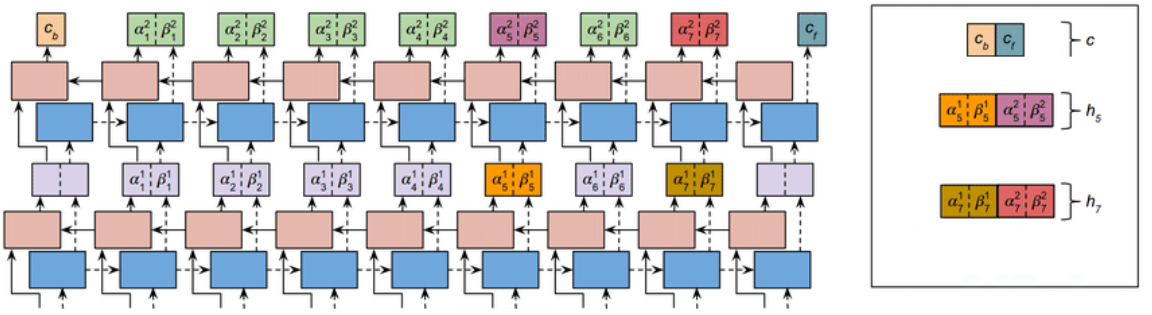
در زیر معماری شبکه‌های GoogleNet و AlexNet نشان داده شده است.



شکل الف- ۱: معماری شبکه AlexNet



شکل الف- ۲: معماری شبکه LSTM



شکل الف- ۳: معماری شبکه GRU. شبکه پایینی در شکل همان شبکه رزولوشن پایین یا GRU است

شکل الف- ۴: معماری شبکه GoogleNet