مدل‌سازی مشترک موضوع و احساس در داده‌های متنی با استفاده از شبکه‌های عصبی

چکیده

امروزه در حوزه‌ی هوش مصنوعی ما به دنبال الگوریتم‌ها و ساختارهایی هستیم که با دقت بالا یک رفتار انسانی و یا فرا انسانی را با بیشترین سرعت ممکن انجام دهند. با گسترش اینترنت و وب، انواع مختلف رسانه‌های اجتماعی نظیر وبلاگ‌ها و شبکه‌های اجتماعی در به یک منبع بسیار عظیم از انواع مختلف داده به ویژه داده‌های متنی تبدیل شده‌اند. با پردازش این داده‌ها می‌توان اطلاعات سودمند و مفیدی در مورد مباحث مختلف، نظر افراد و احساس کلی جامعه بدست آورد. از این جهت داشتن مدل‌هایی که کاملا خودکار به تشخیص اطلاعات مفهومی و احساس در اسناد متنی بپردازند بسیار مفید است. روش‌های مدل‌سازی موضوع و استخراج اطلاعات مفهمومی از داده‌های متنی و همچنین تشخیص احساس، همواره از مهمترین مباحث مطرح شده در زمینه‌ی پردازش زبان طبیعی، و کاوش داده‌های متنی است. بیشتر مدل‌هایی که در این زمینه وجود دارند بر پایه‌ی روش‌های آماری و شبکه‌های بیزی هستند به طوری که در زمینه‌ی مدل‌سازی موضوع-احساس با استفاده از شبکه‌های عصبی تا به امروز هیچ رویکردی وجود ندارد. همچنین بیشتر رویکردهای موجود دارای محدویت‌هایی مانند پیچیدگی محاسباتی بالا هستند. در این مقاله یک ساختار جدید برای مدل‌سازی مشترک احساس-موضوع در داده‌های متنی بر پایه‌ی شبکه‌ی عصبی ماشین بلتزمن محدود پیشنهاد می‌گردد. مدل پیشنهاد شده پس از پیاده‌سازی با مدل‌های موجود مقایسه گردید. مشاهده می‌شود رویکرد پیشنهادی در بحث ارزیابی به عنوان بک مدل مولد، طبقه‌بندی احساس و بازیابی اطلاعات عملکرد بهتری نسبت به مدل‌های موجود دارد.

کلمات کلیدی: مدل‌سازی موضوع، آنالیز احساس، شبکه‌های عصبی، ماشین بلتزمن محدود، مدل احتمالاتی، الگوریتم واگرایی مقابله

۱ مقدمه

امروزه در تمام مباحث مربوط به هوش مصنوعی ما به دنبال روش‌ها، الگوریتم‌ها و ساختارهایی هستیم که بتوانند هرچه بهتر، به صورت خودکار و با دقت بالا یک رفتار انسانی و یا فرا انسانی را با بیشترین سرعت ممکن انجام دهند. اعمالی مانند دسته‌بندی، استخراج اطلاعات مفهومی، آنالیز و برچسب گذاری داده‌ها و از جمله فعالیت‌هایی می‌باشند که امروزه ما انجام بسیاری از آن‌ها را به ماشین‌ها واگذار می‌کنیم.

در بین انواع مختلف داده، داده‌های متنی دارای سهم عظیمی از نظر حجم و مقدار هستند. به خصوص با گسترش اینترنت و وب در دهه‌ی اخیر با سرعتی بسیار زیاد، انواع مختلف رسانه‌های اجتماعی نظیر وبلاگ‌ها، شبکه‌های اجتماعی و گروه‌های بحث در اینترنت به یک منبع بسیار عظیم و قوی از انواع مختلف داده و اطلاعات به ویژه داده‌های متنی تبدیل شده اند. با پردازش این داده‌ها می‌توان اطلاعات سودمند و مفیدی در مورد مباحث مختلف، نقطه نظر افراد و احساس کلی جامعه بدست آورد. فعالیت‌های انجام گرفته در زمینه کاوش داده‌ها به خصوص کاوش داده‌های متنی و همچنین پردازش زبان طبیعی بیشتر از هر زمینه‌ی دیگری به تلاش برای درک و فهم این حجم عظیم از داده‌های متنی مربوط می‌شوند.

حجم عظیمی از داده‌های متنی که بدون هیچ ساختار و قاعده و قانونی هستند و روز به روز مقدار آن‌ها با سرعت بسیاری چشمگیری در حال افزایش است. در این میان وجود الگوریتم‌ها و روش‌هایی که بتوانند به صورت خودکار با این حجم زیاد از داده‌های بدون ساختار ارتباط برقرار کرده و اطلاعات مفید و سودمند را از آن برای ما استخراج کنند بیش از پیش احساس می‌گردد.

تمرکز ما در ابن مقاله و روش پیشنهادی پردازش بر روی داده‌های متنی است. در تقابل با داده‌های متنی، هدف ما پیدا کردن توزیع موضوع‌های مختلف موجود در مجموعه‌ی اسناد پایگاه داده و همچنین توزیع کلمات و احساس همراه با هر موضوع با استفاده شبکه‌ی عصبی است. فرآیند مورد نظر در داده‌های متنی تحت عنوان مدل‌سازی موضوع شناخته می‌شود که در مباحث مربوط به هوش مصنوعی در دسته‌ی کارهای مربوط به یادگیری ماشین، پردازش زبان طبیعی، شبکه‌های عصبی مصنوعی و کاوش احساسات قرارمی‌گیرد. در بحث مدل‌سازی موضوع با استفاده از شبکه‌های عصبی در سال‌های اخیر تعداد اندکی روش ارائه شده است. اما در زمینه‌ی مدل‌سازی مشترک احساس و موضوع با استفاده از شبکه‌های عصبی تا کنون مدلی مطرح نشده و مورد آزمایش قرار نگرفته است. نتایج بهتر مدل‌های شبکه عصبی در بحث مدل‌سازی موضوع در مقایسه با روش‌های پیشین که از ساختارهای گرافی و مدل‌های بیزی استفاده می‌کردند، همچنین عدم وجود روشی برای تشخیص همزمان احساس و موضوع در داده‌های متنی با استفاده از شبکه‌های عصبی منجر به رویکرد پیشنهادی در این مقاله برای مدل‌سازی مشترک احساس و موضوع در داده‌های متنی بر پایه‌ی شبکه‌های عصبی گردید.

مدل‌سازی موضوع و استخراج اطلاعات مفهمومی از داده‌های متنی و همچنین تشخیص احساس از مهمترین مباحث مطرح شده در زمینه‌ی پردازش زبان طبیعی، و کاوش داده‌های متنی هستند. رویکردهای موجود در این زمینه با اجرا بر روی یک پایگاه داده از اسناد متنی به تشخیص و مدل‌سازی موضوع‌ها، احساسات و مفاهیم همراه با هر سند متنی می‌پردازند. تشخیص احساس برای هر سند و هر موضوع در بحث بازیابی اطلاعات نیز می‌تواند به اندازه تشخیص اطلاعات موجود در هر متن حائز اهمیت باشد. از این جهت داشتن مدل‌هایی که به صورت اتوماتیک و کاملا خودکار به مدل‌سازی موضوع و تشخیص اطلاعات مفهومی و احساس در اسناد بپردازند می‌تواند بسیار مفید باشد. بیشتر کارهایی که در این زمینه وجود دارند بر پایه‌ی رویکردهای آماری و شبکه‌های بیزی هستند که از محدودیت‌هایی مانند پیچیدگی محاسباتی بالا رنج می‌برند. در بحث شبکه‌های عصبی بر خلاف مدل‌های آماری، روشی برای مدل کردن موضوع و احساس به صورت همزمان و مشترک وجود ندارد. در این مقاله نیز در همین راستا یک رویکرد نوین بر پایه‌ی شبکه‌های عصبی مصنوعی برای مدل‌سازی همزمان موضوع و احساس در یک مجموعه از داده‌های متنی پیشنهاد می‌گردد. رویکرد پیشنهاد شده در این مقاله یک مدل نظارت شده‌ی مولد احتمالی بر پایه‌ی شبکه‌ی عصبی ماشین بلتزمن محدود است. برای آموزش در این مدل مانند سایر روش‌هایی که بر پایه‌ی ماشین بلتزمن محدود هستند از الگوریتم یادگیری واگرایی مقابله استفاده می‌شود.

ساختار بخش‌های بعدی در مقاله به این صورت است: ابتدا در بخش دوم به مرور کارهای پیشین در زمینه‌ی تخمین توزیع‌های احتمالی در داده‌های ورودی، مدل‌سازی موضوع، تشخیص احساس و مدل‌سازی احساس‌ـ موضوع در داده‌های متنی می‌پردازیم. در بخش سوم کلیات نظری و تئوری مدل پیشنهادی بیان می‌شوند. در این فصل با معرفی یک مدل معروف به عنوان پایه مدل جدید تعریف و قسمت‌های مختلف آن شرح داده می‌شوند و روابط مورد نیاز برای هر قسمت تعریف می‌شوند. در بخش چهارم این مقاله مراحل شبیه‌سازی مدل پیشنهادی و نتایج حاصل از آزمایش‌های بدست آمده ومفایسه با دیگر مدل‌هاارائه می‌گردد. در بخش پایانی، نتیجه‌گیری حاصل از این مقاله شرح داده خواهد شد. همچنین

راهکارهایی برای بهبود و توسعه مدل پیشنهادی ارائه خواهد شد.

۲ بررسی مدل‌های پیشین

در این بخش روش‌های موجود را از چندین زاویه مورد نقد و بررسی قرار می‌دهیم و بسته به ساختار، نحوه‌ی عملکرد، نوع داده‌ی ورودی و سیر تکاملی، آن‌ها را در چندین کلاس طبقه‌بندی می‌کنیم.

به طور کلی روش‌های مدل‌سازی موضوع به مدل‌هایی گفته می‌شود که یک چکیده از موضوعات موجود در یک سند یا مجموعه‌ای از اسناد را تشخیص داده و استخرج می‌کنند. در بررسی رویکردهای موجود از دید ساختاری، می‌توان آن‌ها را در دو گروه کلی طبقه‌بندی کرد. دسته‌ی اول مدل‌های گرافی و بیزی و دسته‌ی دوم مدل‌های بر پایه‌ی شبکه‌های عصبی. از نظر نحوه‌ی عملکرد مدل‌های پیشین را در سه کلاس مختلف قرار دارند. دسته‌ي اول روش‌هایی که تنها به مدل‌سازی موضوع می‌پردازند. دسته‌ی دوم روش‌هایی که به تشخیص احساس در داده‌های ورودی می‌پردازند. و در دسته‌ي سوم از نظر نحوه‌ی عملکرد روش‌هایی قرار دارند که به صورت همزمان به مدل‌سازی موضوع و احساس بر روی داده‌ی ورودی می‌پردازند. اگرچه باید توجه داشت که مدل‌های موجود در زمینه تشخیص احساس در دسته‌ی مدل‌های موضوعی قرار نمی‌گیرند و بیشتر شامل مدل‌هایی هستند که یک طبقه‌بندی دو حالته (مثبت و منفی) یا سه حالته (منفی، مثبت و بی طرف) را انجام می‌دهند.

روش‌های پیشین از نظر نوع داده‌ی ورودی در دو کلاس متفاوت قرار می‌گیرند. یک گروه روش‌هایی که تنها یک نوع داده را به عنوان ورودی قبول می‌کنند. منظور از یک مدل داده این است که روش‌های موجود توانایی عمل کردن به صورت همزمان بر روی چند مد مختلف از داده‌ها را ندارند، و داده‌های ورودی تنها باید یک حالت داشته باشند، مثلا تنها متن و یا تنها تصویر باشند و نمی‌توانند ترکیبی از این‌ها باشند. دسته‌ی دوم که آن‌ها را مدل‌های چندحالته می‌شناسیم مدل‌هایی هستند که با داده‌های چندوجهی کار می‌کنند. منظور از داده‌های چندوجهی آن‌هایی هستند که شامل ترکیب چند حالت مختلف از داده‌ها می‌شوند، برای مثال ترکیب متن و تصویر و یا ترکیب تصویر و صدا.

از نقطه‌نظر سیر تکاملی می‌توان روش‌های موجود را در سه سطح: یک مدل‌های تخمین‌زننده‌ی توزیع، دو روش‌های مدل‌سازی موضوع و سه رویکردهای تشخیص همزمان موضوع و احساس قرار داد. البته لازم به ذکر است که روش‌های تخمین توزیع که در اینجا مطرح می‌گردند و در بحث پردازش زبان طبیعی مورد استفاده قرار می‌گیرند به تنهایی در دسته‌ی مدل‌های موضوعی قرار نمی‌گیرند اما پایه و اساس بسیاری از مدل‌های موضوعی هستند.

در بحث مدل‌سازی موضوع، مدل‌های گرافی نسبت به روش‌های شبکه‌های عصبی از قدمت بیشتری برخوردار هستند. روش پایه‌ای که امروزه همچنان در مرورگرهای اینترنتی مورد استفاده قرار می‌گیرد، تکرار ترم‌ـ معکوس تکرار سند (tf-idf) نام دارد. این رویکرد در سال ۱۹۸۶ توسط Salton معرفی شد و در آن هر سند متنی به یک بردار اعدد حقیقی تبدیل می‌شود که شامل نسبت‌های تعداد تکرار کلمات مختلف است و از آن برای بازیابی اطلاعات استفاده می‌شود. برای غلبه بر محدودیت‌های tf-idf محققین حوزه‌ی IR چندین مدل کاهش بعد دیگر معرفی کردند که مهمترین آن‌ها مدل فهرست کردن معنایی نهفته (LSI) است که توسط Deerwester و همکاران در سال ۱۹۹۰ ارائه گردید. مدل LSI با استفاده از تجزیه مقدار منفرد بر روی ماتریس خروجی از مدل tf-idf یک زیرفضای خطی در فضای ویژگی‌های مدل

tf-idf شناسایی می‌کند. این روش منجر به کاهش قابل توجهی در مجموعه‌های بزرگ می‌گرد. همچنین Deerwester و همکاران ادعا کردند که ویژگی‌های بدست آماده توسط مدل LSI که در حقیقت یک ترکیب خطی از ویژگی‌های مدل tf-idf هستند، توانایی تشخیص بعضی از ویژگی‌های زبانی مانند مترادف و متضاد را دارند.

برای اثبات ادعاهای مطرح شده در مدل LSI و همچنین بررسی نقاط ضعف و قدرت این مدل، روش فهرست‌سازی معنایی نهفته‌ی احتمالاتی (pLSI) توسط Hofmann در سال ۱۹۹۹ معرفی شد. مدل pLSI یک مدل مولد احتمالاتی می‌باشد که از آن به عنوان یک مدل موضوعی نیز یاد می‌شود. در روش pLSI هر کلمه از یک موضوع خاص تولید می‌شود و کلمه‌های مختلف در داخل یک سند ممکن است از موضوع‌های مختلفی تولید شوند. مهم‌ترین مدلی که در دسته‌ی رویکردهایی گرافی وجود دارد مدل معروف تخصیص دیریکله‌ی پنهان (LDA) است که در سال ۲۰۰۳ توسط Blei و همکاران ارائه گردید و پس آن به عنوان پایه‌ی مدل‌سازی موضوعی در بخش مدل‌های گرافی قرار گرفت. در روش LDA مانند دیگر روش‌های مدل‌سازی موضوع، هر سند متنی به صورت یک توزیع مخلوط بر روی موضوع‌های مختلف که در آن هر موضوع به وسیله‌ی یک توزیع بر روی کلمه‌ها مشخص می‌شود در نظر گرفته می‌شود.

مدل ماشین بلتزمن محدود که به اختصار آن را RBM می‌نامیم یک شبکه عصبی دولایه‌ی (یک لایه‌ی قابل مشاهده ویک لایه‌ی پنهان) بدون‌نظارت برای تخمین توزیع داده‌های ورودی باینری است. RBM یک مدل احتمالاتی مولد است که اولین بار در سال ۱۹۸۶ توسط Smolensky و پس از آن در سال ۲۰۰۲ به شکل دیگری توسط Hinton معرفی گردید. مدل شبکه‌ی عصبی خودکاهشی تخمین‌زننده‌ی توزیع (NADE) که از مدل RBM الهام گرفته شده است، یک روش احتمالاتی مولد بدون‌نظارت برای مدل‌سازی احتمال داده‌های گسسته است که در سال ۲۰۱۱ توسط Larochelle و همکاران ارائه شد. یکی از محدودیت‌های روش RBM مناسب نبودن این مدل برای تخمین احتمال مشترک در ابعاد بالا است. این محدودیت در مدل NADE بدلیل استفاده از ایده‌ی شبکه‌های بیزین کاملا مشاهده‌پذیر برای محاسبه‌ی احتمال مرتفع گردیده است.

دسته‌ی دیگر مدل‌های موضوعی موجود از نظر ساختار آن‌هایی هستند که بر پایه‌ی شبکه‌های عصبی می‌باشند و اولین بار در سال ۲۰۰۹ توسط Hinton و Salakhutdinov تحت عنوان مدل سافتمکس تکرار شونده (RS) معرفی شدند. RS اولین روش مدل‌سازی موضوع بر پایه‌ی شبکه‌های عصبی و گسترش یافته‌ی مدل RBM است که از آن برای تشخیص توزیع موضوع‌های مختلف در داده‌های متنی استفاده می‌شود. مدل RBM به دلیل محدودیت‌هایی مانند محدود بودن به بردار ورودی باینری و در نظر گرفتن طول ثابت برای ورودی‌ها نمی‌تواند در تشخیص توزیع موضوع‌ها مورد استفاده قرار بگیرد، چرا که اولا کلمات باینری نیستند و دوما در یک مجموعه از داده‌های متنی طول اسناد با یکدیگر متفاوت هستند. پس از مدل ،RS شبکه‌ی عصبی خودکاهشی تخمین‌زننده‌ی توزیع سندی (DocNADE) یک روش بدون‌نظارت برای مدل‌سازی موضوع بر پایه‌ی شبکه‌های عصبی است در سال ۲۰۱۲ توسط Larochelle و Lauly با ترکیب مدل‌های NADE و RS معرفی شد.

تمام مدل‌های بررسی شده تا کنون تنها توانایی تشخیص موضوع از داده‌های متنی را داشتند. گروه دیگری از مدل‌های موضوعی وجود دارند که به صورت همزمان به تشخیص موضوع‌ها و احساس همراه با هرکدام می‌پردازند. در ادامه دو روبکرد بر پایه‌ی شبکه‌های بیزی که در این دسته قرار دارند را معرفی می‌کنیم. مدل یکی‌سازی احساس‌ـ موضوع (ASUM) در سال ۲۰۱۱ برای تشخیص موضوع‌ها و احساس همراه با آن‌ها در بازبینی‌های آنلاین توسط Jo و Oh معرفی شد. این

	Structure		Modeling Type			Data Input	
	Graphical	NN	DE	T	S/T	Unimodal	MultiModal
LSI	⋆			⋆		⋆	
pLSI	⋆			⋆		⋆	
LDA	⋆			⋆		⋆	
JST	⋆				⋆	⋆	
ASUM	⋆				⋆	⋆	
NADE		⋆	⋆			⋆	
RBM		⋆	⋆			⋆	
RS		⋆		⋆		⋆	
DocNADE		⋆		⋆		⋆	
SupDocNADE		⋆		⋆			⋆
Note: NN = Neural Network, DE = Distribution Estimator, T = Topic, S/T = Sentiment/Topic							

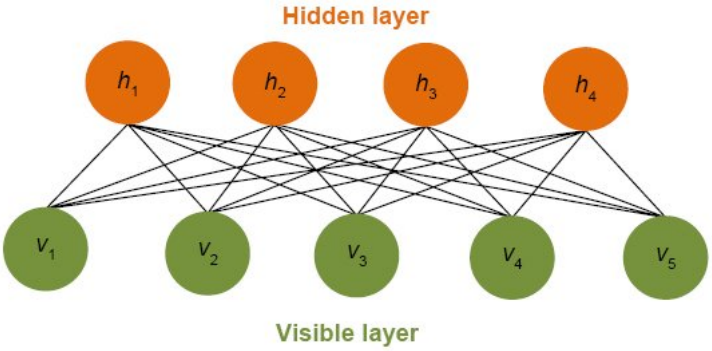
جدول ۱ : دسته‌بندی مدل‌های پیشین از نظر ساختار، نحوه‌ی عملکرد و نوع داده‌ی ورودی.

مدل گسترش یافته‌ی مدل LDA است و در گروه مدل‌های احتمالاتی گرافی مولد قرار می‌گیرد. در مدل ASUM ما برای هر سند یک توزیع چندجمله‌ای احساسی و برای هر یک از احساس‌ها یک توزیع چندجمله‌ای موضوعی داریم و فرض بر این است که هر جمله در داخل هر سند دارای یک برچسب احساس و یک موضوع است. پس از روش ASUM، مدل نظارت‌شده‌ی ضعیف تشخیص مشترک احساس‌ـموضوع (JST) در سال ۲۰۱۲ توسط Lin و همکاران معرفی شد. مدل JST یک مدل احتمالاتی مولد گرافی و گسترش یافته‌ی مدل LDA است که علاوه بر تشخیص موضوع به تشخیص احساس از داده‌های متنی نیز می‌پردازد. خاصیت نظارت‌شده‌ی ضعیف باعث می‌شود که در مقایسه با سایر مدل‌ها، JST به راحتی قابل انتقال به یک دامنه‌ی دیگر بدون کاهش محسوس در کارایی که در سایر مدل ها این اتفاق رخ می‌دهد باشد.

مدل‌های چندحالته (Multimodal) دسته‌ای از مدل‌های موضوعی هستند که داده‌ی ورودی در آن‌ها ترکیبی از چند حالت مختلف داده است. مدل نظارت‌شده‌ی شبکه‌ی عصبی خودکاهشی تخمین‌زننده‌ی توزیع سندی (SupDocNADE) یک روبکرد چندحالته است که در سال ۲۰۱۴ توسط Zheng و همکاران معرفی شد. این مدل گسترش یافته‌ی مدل DocNADE است. در مدل SupDocNADE داده‌های وردی تنها اسناد متنی نیستند. ورودی در این مدل تصاویر به همراه توضیح کوتاهی در مورد هر تصویر است که مدل ترکیب این دو نوع داده در کنار یکدیگر را یاد گرفته و در کاربرد مورد نظر از آن استفاده می‌کند. در جدول ۱ به صورت خلاصه ویژگی‌های مدل‌های معرفی شده نشان داده شده و این مدل‌ها در سه کلاس مختلف دسته‌بندی گردیده‌اند.

۳ مدل‌سازی مشترک موضوع و احساس با شبکه‌های عصبی

مدل پایه برای روش پیشنهادی در این مقاله شبکه عصبی RBM است که در شکل ۱ نشان داده شده است. RBM یک مدل بدون نظارت برای داده‌های باینری است که در دسته‌ی مدل‌های مولد احتمالاتی قرار می‌گیرد. در این مدل با بیشینه



شکل ۱: ماشین بلتزمن محدود RBM

کردن یک تابع انرژی، یا کمینه کردن مقدار منفی آن که به صورت رابطه ۱ تعریف می‌شود، توزیع‌های احتمالی موجود در داده‌های ورودی یاد گرفته می‌شود و از داده‌های ورودی ویژگی استخراج می‌گردد. در رابطه‌ی ۱، $\theta = \{W, \mathbf{a}, \mathbf{b}\}$ مجموعه‌ی پارامترهای مدل است. W_{D*H} ماتریس وزن بین لایه‌ی ورودی و لایه‌ی پنهان است، که در آن D سایز بردار ورودی و H سایز لایه‌ی پنهان هستند. $\mathbf{a}$ بردار بایاس لایه‌ی ورودی با سایز D و $\mathbf{b}$ بردار بایاس لایه‌ی پنهان با سایز H است. در ساختار RBM توزیع‌های شرطی برای لایه‌های قابل مشاهده و پنهان به شکل روابط ۲ و ۳ هستند.

$$E(\mathbf{v}, \mathbf{h}) = -\sum_i \sum_j v_i W_{ij} h_j - \sum_i v_i a_i - \sum_j h_j b_j \quad (۱)$$

$$p(\mathbf{h}|\mathbf{v}) = sigmoid(\mathbf{v}W + \mathbf{b}) \quad (۲)$$

$$p(\mathbf{v}|\mathbf{h}) = sigmoid(W\mathbf{h} + \mathbf{a}) \quad (۳)$$

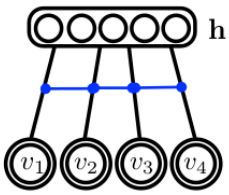
در مدل RBM احتمال هر ترکیب $(\mathbf{v}, \mathbf{h})$ از رابطه‌ی ۴ بدست می‌آید که در آن $Z(\theta)$ تابع قسمت‌بندی است که مقدار آن با استفاده از رابطه‌ی ۵ محاسبه می‌شود و تضمین می‌کند که مقدار بدست آمده برای هر ترکیب $(\mathbf{v}, \mathbf{h})$ در رابطه‌ی ۴ یک مقدار صحیح احتمالی (بین ۰ و ۱) است. در این مدل احتمال هر بردار ورودی از رابطه‌ی ۶ بدست می‌آید.

$$p(\mathbf{v}, \mathbf{h}) = \frac{1}{Z(\theta)} e^{-E(\mathbf{v},\mathbf{h})} \quad (۴)$$

$$Z = \sum_{\mathbf{v},\mathbf{h}} e^{-E(\mathbf{v},\mathbf{h})} \quad (۵)$$

$$p(\mathbf{v}) = \sum_h \frac{1}{Z(\theta)} e^{-E(\mathbf{v},\mathbf{h})} \quad (۶)$$

محدود بودن به حالت باینری برای داده‌های ورودی و همچنین طول ثابت برای آن‌ها دو محدودیت اساسی در مدل RBM استاندارد هستند. فرض کنید در مساله‌ای که با آن سرو کار داریم هر داده‌ی ورودی دارای D ویژگی است که هر یک از این



شکل ۲: مدل Replicated Softmax

ویژگی‌ها می‌توانند K مقدار داشته باشند. مدل RBM استاندارد توانایی کار کردن با یک چنین داده‌های ورودی را ندارد چرا که در آن داده‌های ورودی تنها می‌توانند یک بردار با طول ثابت و شامل ۰ و ۱ باشند. در این مدل جدید هر داده‌ی ورودی به صورت ماتریسی با سایز K*D در نظر گرفته می‌شود که همان‌طور که بیان شد D طول بردار ورودی یا همان تعداد ویژگی‌های مساله و K ماکسیمم مقداری است که هر ویژگی می‌تواند داشته باشد. در این صورت تابع انرژی برای حالت $\{\mathbf{V}, \mathbf{h}\}$ به شکل رابطه‌ی ۵ تعریف می‌شود. همچنین رابطه‌ی ۳ به شکل رابطه‌ی ۸ تبدیل می‌شود. دلیل استفاده از تابع Softmax در رابطه‌ی ۸ به جای تابع سیگموید که در مدل RBM استاندارد از آن استفاده می‌شود تغییر در ساختار ورودی و لایه‌ی قابل مشاهده است. با توجه به اینکه هر ویژگی تنها یک مقدار از K مقدار ممکن را می‌تواند داشته باشد، لذا برای هر ستون در این ماتریس از تابع Softmax استفاده می‌گردد و یک توزیع اختمال چند جمله‌ای بدست می‌آید. سپس با تولید نمونه از این توزیع چند جمله‌ای مقدار آن ویژگی تعیین می‌گردد.

$$E(\mathbf{V}, \mathbf{h}) = -\sum_{i=1}^{D}\sum_{j=1}^{H}\sum_{k=1}^{K} W_{ijk} h_j v_{ik} - \sum_{i=1}^{D}\sum_{k=1}^{K} v_{ik} a_{ik} - \sum_{j=1}^{H} h_j b_j \qquad (۷)$$

$$p(v_{ik} = 1|\mathbf{h}) = \frac{exp(a_{ik} + \sum_{j=1}^{H} h_j W_{ijk})}{\sum_{k=1}^{K} exp(a_{ik} + \sum_{j=1}^{H} h_j W_{ijk})} \qquad (۸)$$

برای مدل‌سازی موضوع پیش از آموزش مدل ابتدا یک لغت‌نامه از تمام کلمات متمایز در مجموعه اسناد ساخته می‌شود. حال در بردار ودرودی مقدار هر ویژگی برابر با اندیس یکی از کلمات دیکشنری است. به بیان دیگر هر سند ورودی پس از انجام پیش پردازش‌های لازم به یک دنباله از کلمات تبدیل می‌شود که هر کدام از این کلمات برابر با یکی از کلمات دیکشنری هستند. به این ترتیب در ماتریس ورودی به مدل که یک ماتریس به اندازه‌ی $K * D$ است، K برابر با سایز دیکشنری و D نشان دهنده‌ی طول سند متنی است. در این حالت برای هر ستون مقدار سطر متناظر با اندیس آن کلمه در دیکشنری برابر با ۱ می‌شود و دیگر درایه‌های آن ستون همچنان صفر باقی می‌مانند. در ابن مدل که Hinton و Salakhutdinov آن را RS نامیدند و در شکل ۲ نشان داده شده است، برای مدل کردن داده‌های متنی برای هر سند یک شبکه‌ی جدا می‌سازیم که به تعداد کلمات همان سند دارای واحد Softmax است. در این حالت ورودی دیگر یک ماتریس باینری نخواهد بود و به صورت برداری از تعداد کلمات موجود در آن سند است که می‌توانیم ترتیب را در آن‌ها نادیده بگیریم. رابطه‌ی محاسبه‌ی انرژی در این حالت به شکل رابطه‌ی ۹ است که در آن $\hat{v}^k = \sum_{i=1}^{D} v_{ik}$ است. به عبارت دیگر $\hat{\mathbf{v}}$ برداری است با طول K، که K همان سایز دیکشنری است و از محاسبه‌ی حاصل جمع سطرهای ماتریس باینری ورودی بدست می‌آید. در این حالت روابط شرطی محاسبه‌ی لایه‌ی قابل مشاهده و لایه‌ی پنهان به شکل روابط ۱۰

و ۱۱ هستند. همانطور که مشاهده می‌شود در روابط ۹ و ۱۱ ترم بایاس برای لایه‌ی پنهان با سایز سند جاری نیز متناسب است. وجود این تناسب در پیاده‌سازی‌های تجربی و هنگامی که اسناد با طول‌های متفاوت در مجموعه اسناد وجود دارد بسیار حیاتی است.

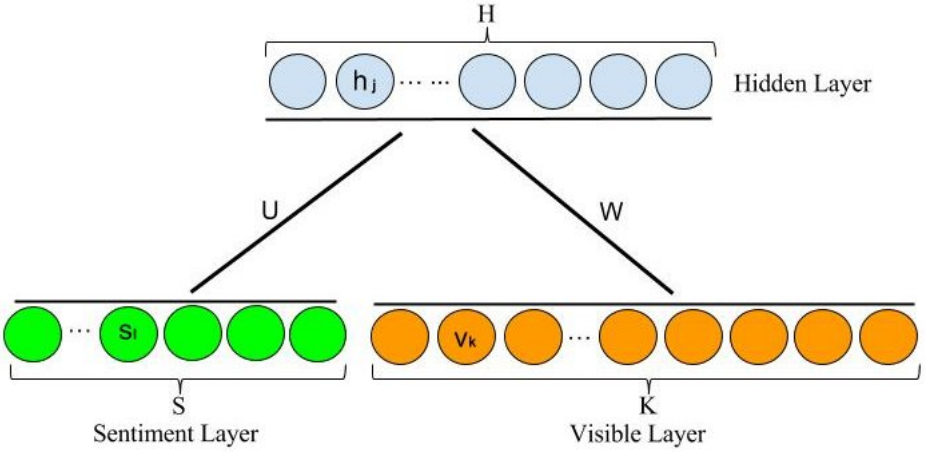
$$E(\mathbf{v},\mathbf{h}) = -\sum_{j=1}^{H}\sum_{k=1}^{K} W_{jk} h_j \hat{v}_k - \sum_{k=1}^{K} v_k a_k - D\sum_{j=1}^{H} h_j b_j \qquad (۹)$$

$$p(v_i = w|\mathbf{h}) = \frac{exp(a_w + \sum_{j=1}^{H} h_j W_{wj})}{\sum_{k=1}^{K} exp(a_w + \sum_{j=1}^{H} W_{wj})} \qquad (۱۰)$$

$$p(h_j = 1|\mathbf{v}) = \sigma\left(Db_j + \sum_{k=1}^{K} W_{kj}\hat{v}_k\right) \qquad (۱۱)$$

مدل‌ها و حالت‌های معرفی شده تاکنون رویکردهایی هستند که پایه‌ی و اساس ساختار پیشنهادی در این مقاله هستند و با گسترش آن‌ها به شکلی که در ادامه بیان می‌شود مدل پیشنهادی در این مقاله بدست می‌آید. روبکرد معرفی شده در این مفاله یک مدل مولد احتمالاتی نظارت شده بر پابه‌ی شبکه‌ی عصبی برای مدل‌سازی موضوع و احساس در داده‌های متنی است که در شکل ۳ نشان داده شده است. مشاهده می‌گردد که این روش نیز یک ساختار دو لایه دارد که در سمت لایه‌ی قابل مشاهده‌ی آن یک بردار متناظر با برچسب هر سند یا تعداد کلاس‌های موجود که در این پژوهش ما آن را به عنوان بردار متناظر با احساس هر سند تعبیر می‌کنیم به ساختار مدل اضافه شده است. بردار ورودی در این ساختار در قسمت قابل مشاهده یک بردار با طولی ثابت و به اندازه‌ی سایز دیکشنری یا همان تعداد کلمات متمایز در متن است که در آن تعداد تکرار کلمات مشخص شده است. در نظر گرفتن یک ماتریس ۲ بعدی برای هر سند ورودی به این معنی است که جایگاه هر کلمه در متن دارای اهمیت است و ترتیب کلمات در هر سند درنظر گرفته می‌شود که این امر موجب بزرگ شدن فضای پارامترهای مساله (سه بعدی شدن ماتریس وزن و دو بعدی شدن ماتریس بایاس برای لایه‌ی قابل مشاهده) و کند شدن فرآیند آموزش می‌گردد. علاوه بر این در بحث مدل‌سازی موضوع حضور و عدم حضور کلمات به همراه فرکانس تکرار آن‌ها دارای اهمیت است نه محل قرار گرفتن هر کلمه در متن، چرا که تمام مدل‌های بررسی شده در این پژوهش و همچنین مدل پیشنهادی بر اساس کیسه کلمات رفتار می‌کنند که در آن ترتیب کلمات در نظر گرفته نمی‌شود.

در مدل پیشنهادی در این پژوهش و همچنین مدل RS هر سند به صورت یک بردار شامل تعداد کلمات به مدل وارد می‌شود. در این ساختار مانند آنچه که در شکل ۳ نشان داده شده است طول بردار ورودی برابر با سایز دیکشنری یا همان تعداد کلمات متمایز در مجموعه سند در نظر گرفته می‌شود که درایه‌های آن تعداد تکرار هر کلمه از دیکشنری در سند جاری را نشان می‌دهند. در این حالت در واقع وزن‌ها برای هر کلمه به اشتراک گذاشته می‌شوند، فارغ از اینکه این کلمه در کجای سند ورودی قرار دارد. به طور مثال برای کلمه‌ی nام دیکشنری یک وزن و یک بایاس تعریف می‌گردد و این کلمه در هر جای سند ورودی قرار داشته باشد وزنش تغییری نخواهد کرد و در فرآیند آموزش تنها یک وزن و یک بایاس برای هر کلمه یاد گرفته می‌شود. با توجه به ساختار مدل که در شکل ۳ نشان داده شده است، برای هر سند متنی یک بردار باینری که نشان دهنده احساس سند جاری است به عنوان ورودی به شبکه وارد می‌شود، و توزیع‌های موجود برروی کلمات مختلف در هر موضوع و همچنین احساس مرتبط با آن‌ها توسط مدل در لایه‌ی پنهان استخراج می‌شوند.



شکل ۳: مدل پیشنهادی مولد احتمالی احساس/موضوع

برای محاسبه‌ی انرژی در مدل پیشنهادی ترم‌های مربوط به وزن و بایاس لایه‌ی احساس نیز در بدست آوردن مقدار نهایی مشارکت دارند. پس از محاسبه‌ی مقدار انرژی به کمک رابطه‌ی ۱۲، با استفاده از فرمول ۱۳ احتمالی که مدل به هر سند و لایه‌ی احساس همراه با آن اختصاص می‌دهد، محاسبه می‌گردد. برای آموزش این مدل و بروزرسانی پارامترهای شبکه که شامل ماتریس‌های وزن بین لایه‌ی قابل‌مشاهده و پنهان و همچنین لایه‌ی احساس و پنهان هستند و همچنین بایاس‌های هر سه لایه از الگوریتم CD به شکل رابطه‌ی ۱۴ استفاده می‌شود. در رابطه ۱۲، $\theta = \{W, U, \mathbf{a}, \mathbf{b}, \mathbf{c}\}$ مجموعه پارامترهای مدل است که در آن W_{K*H} ماتریس وزن بین بردار قابل مشاهده و لایه‌ی پنهان، U_{S*H} ماتریس وزن بین لایه‌ی احساس و لایه‌ی پنهان و $\mathbf{a}$، $\mathbf{b}$ و $\mathbf{c}$ به ترتیب بردارهای بایاس لایه‌ی قابل مشاهده، پنهان و احساس هستند. لازم به ذکر است که K و H مانند آنچه که بیش از این ذکر کردیم به ترتیب سایز دیکشنری و طول لایه‌ی پنهان هستند و S به عنوان تعداد احساس موجود یا سایز بردار احساس تعریف می‌شود.

$$E(\mathbf{V}, \mathbf{s}, \mathbf{h}) = -\sum_{j=1}^{H}\sum_{k=1}^{K} W_{kj} h_j \hat{v}_k - \sum_{j=1}^{H}\sum_{l=1}^{S} U_{lj} h_j s_l - \sum_{k=1}^{K} v_k a_k - \sum_{l=1}^{S} s_l c_l - D\sum_{j=1}^{H} h_j b_j \quad (۱۲)$$

$$p(\mathbf{v}, \mathbf{s}, \mathbf{h}) = \frac{1}{Z} e^{-E(\mathbf{v},\mathbf{s},\mathbf{h})} \Rightarrow p(\mathbf{v}, \mathbf{s}) = \frac{1}{Z}\sum_{h} e^{-E(\mathbf{v},\mathbf{s},\mathbf{h})} \ , Z = \sum_{\mathbf{v}}\sum_{\mathbf{s}}\sum_{\mathbf{h}} e^{-E(\mathbf{v},\mathbf{s},\mathbf{h})} \quad (۱۳)$$

$$\triangle\theta = \alpha\left(E_{P_{data}}[\theta] - E_{P_{model}}[\theta]\right) \Rightarrow \theta_{t+1} = \theta_t + \triangle\theta \quad (۱۴)$$

در مدل پیشنهادی مقادیر هر یک از لایه‌های قابل مشاهده، احساس و پنهان به کمک روابط ۱۵ تا ۱۷ محاسبه می‌شوند. در اینجا چون مقدار لایه‌ی پنهان به هر دو مقدار لایه‌ی قابل مشاهده و احساس وابسته است، لذا مشاهده می‌شود که در رابطه‌ی ۱۷ برای مقدار لایه‌ی پنهان از یک توزیع شرطی که وابسته به هر دو مقدار لایه‌های قابل مشاهده و احساس

است نمونه گرفته می‌شود. اما با توجه به اینکه با داشتن مقدار لایه‌ی پنهان، بردارهای قابل مشاهده و احساس از یکدیگر مستقل شرطی هستند لذا در روابط ۱۵ و ۱۶ مقدار این دو بردار از یک توزیع شرطی که تنها به مقدار بردار پنهان وابسته است نمونه گرفته می‌شوند.

$$p(v_i = w|\mathbf{h}) = \frac{exp(a_w + \sum_{j=1}^{H} W_{wj} h_j)}{\sum_{k=1}^{K} exp(a_w + \sum_{j=1}^{H} W_{wj} h_j)} \quad (۱۵)$$

$$p(s_l = 1|\mathbf{h}) = \frac{exp(c_l + \sum_{j=1}^{H} U_{lj} h_j)}{\sum_{l=1}^{S} exp(c_l + \sum_{j=1}^{H} U_{lj} h_j)} \quad (۱۶)$$

$$p(h_j = 1|\mathbf{v}, \mathbf{s}) = \sigma\left(Db_j + \sum_{k=1}^{K} W_{kj} \hat{v}_k + \sum_{l=1}^{S} U_{lj} s_l\right) \quad (۱۷)$$

با توجه به خصوصیات بیان شده برای تابع Softmax، لذا همان‌طور که مشاهده می‌شود، در روابط ۱۵ و ۱۶ برای محاسبه‌ی مقادیر لایه‌های قابل مشاهده و احساس از یک تابع Softmax استفاده می‌گردد. در فرآیند آموزش با استفاده از الگوریتم CD برای بدست آوردن مقدار بازسازی شده از لایه‌ی قابل مشاهده مشروط به بردار پنهان از رابطه‌ی ۱۵ که به صورت Softmax است، استفاده می‌شود. در واقع دلیل اینکه این رابطه و رابطه‌ی ۱۶ برای لایه‌ی احساس به فرم تابع Softmax هستند همین امر می‌باشد، که پس از محاسبه‌ی مقادیر این لایه‌ها مشروط به بردار پنهان نیاز به تولید نمونه و نمونه‌برداری از این مقادیر بدست آماده داریم. در نتیجه استفاده از تابع Softmax برای ما تضمین می‌کند که مقادیر محاسبه شده برای این دو بردار یک توزیع احتمالی چندجمله‌ای خواهد بود که می‌توان به راحتی از آن نمونه تولید کرد.

۴ آزمایش‌ها و ارزیابی مدل

در این بخش رویکرد معرفی شده در این مقاله مورد ارزیابی و آزمایش قرار می‌گیرد و نتایج بدست آمده در آزمایش‌های گوناگون گزارش کرده و مورد تحلیل و بررسی قرار می‌دهیم.

۱.۴ معرفی پایگاه داده‌ها

برای انجام آزمایش‌ها و ارزیابی از چند پایگاه داده‌ی معبار در بحث مدل‌سازی موضوع و تشخیص احساس استفاده می‌کنیم که در ادامه آن‌ها را معرفی می‌کنیم.

پایگاه داده‌ی بازبینی فیلم (MR) پس از استفاده در کار Pang و همکاران تبدیل به یک معیار در بحث تشخیص احساس گردیده است. ورژن ۲ از این پایگاه داده که ما در آزمایش‌های خود از آن استفاده می‌کنیم شامل ۱۰۰۰ بازبینی مثبت از فیلم‌های مختلف و ۱۰۰۰ بازبینی منفی می‌باشد. این بازبینی‌ها از سایت پایگاه داده‌ی اینترنتی فیلم (IMDB) جمع‌آوری شده‌اند. میانگین طول هر بازبینی در این پایگاه داده ۳۰ جمله است.

Data Set	Dictionary Size	Num of Train	Num of Test	Avg Docs Length	Std Deviation
Movie Review	2000	1000	1000	90.18	40.23
Movie Review	10000	1000	1000	186.35	81.33
Movie Review	24916	1000	1000	299.75	126.51

جدول ۲: اطلاعات آماری پایگاه داده‌ی Movie Review

پایگاه داده‌ی ۲۰ گروه خبری (۲۰NG) یکی از دیتاست‌های معروف در بحث مدل‌سازی موضوع است. این پایگاه داده شامل ۱۸۷۸۶ سند متنی است که از مخازن گروه‌های خبری Usenet جمع‌آوری شده‌اند. این مجموعه سند به ۲۰ گروه خبری مختلف تقسیم می‌شود که هر کدام از این ۲۰ گروه مربوط به یک موضوع خاص هستند. از مجموع ۱۸۷۸۶ سند موجود در این پایگاه داده، ۱۱۲۸۴ سند برای مجموعه‌ی آموزش و ۷۵۰۲ سند برای مجموعه‌ی تست در نظر گرفته می‌شوند. در این پایگاه داده ۲۰۰۰ کلمه‌ای که بیشترین تکرار را دارند جدا شده و به عنوان دیکشنری در نظر گرفته می‌شوند.

پایگاه داده‌ی احساس چند دامنه (MDS) اولین بار توسط Blitzer و همکاران در سال ۲۰۰۷ مورد استفاده قرار گرفت. این پایگاه داده شامل بازبینی‌های نوشته شده در مورد چهار نوع مختلف از محصولات سایت آمازون است که جمع‌آوری شده‌اند. بازبینی‌های موجود در این دیتاست مربوط به چهار گروه کتاب، دی‌وی‌دی، وسایل الکترونیکی و وسایل آشپزخانه هستند. برای هر یک از این چهار دسته ۱۰۰۰ بازبینی مثبت و ۱۰۰۰ بازبینی منفی در MDS وجود دارد.

۲.۴ آماده‌سازی پایگاه داده‌ها

در این قسمت مراحل آماده‌سازی پایگاه داده‌ها برای استفاده در بخش‌های آینده به صورت کامل توضیح داده می‌شود.

پس از انجام پیش‌پردازش‌های متنی (حذف کلمات توقف، ریشه‌یابی لغوی و نحوی) در پایگاه داده‌ی MR، و تبدیل هر سند متنی به دنباله‌ای از کلمات برای ساخت دیکشنری علاوه بر استفاده از دو دیکشنری معروف در بحث مدل‌سازی موضوع، یک دیکشنری نیز از داده‌های پیش‌پردازش شده ساخته می‌شود. برای ساخت این لغت‌نامه‌ی واژگان تمام اسناد را به صورت کامل پیمایش کردیم تا کلمات متمایز در آن‌ها مشخص گردند. تعداد کلمات متمایز در این حالت ۲۴۹۱۶ عدد است. در نتیجه سایز دیکشنری در این حالت برابر با ۲۴۹۱۶ در نظر گرفته می‌شود. همان‌طور که بیان گردید از دو دیکشنری دیگر با سایزهای ۲۰۰۰ و ۱۰۰۰۰ کلمه‌ی متمایز برای ساخت فایل lib-svm برای پایگاه داده MR نیز استفاده کردیم. این دو دیکشنری به ترتیب مربوط به دو پایگاه داده‌ی ۲۰NG و توده‌ی اسناد رویتر ورژن ۱ (RCV۱) هستند که از پایگاه داده‌های معیار در بحث مدل‌سازی موضوعی هستند. اطلاعات آماری بدست آماده پس از انجام مراحل گفته شده در جدول ۲ نشان داده شده است.

همان‌طور که بیان شد پایگاه داده‌ی MR شامل ۲۰۰۰ سند است که ۱۰۰۰ عدد از این اسناد مثبت و ۱۰۰۰ سند باقی مانده دارای برچسب احساس منفی می‌باشند. مانند آنچه که در مدل JST استفاده شده است در اینجا ما نیز این پایگاه داده را به دو دسته‌ی آموزش و آزمون تقسیم می‌کنیم. تعداد اسناد در هر یک از این دو گروه ۱۰۰۰ می‌باشد که ۵۰۰ عدد از آن‌ها مثبت و ۵۰۰تای دیگر دارای برچسب منفی هستند.

۳.۴ لغت‌نامه‌ی احساس

منظور از لغت‌نامه‌ی احساس یک دیکشنری عمومی از پیش ساخته شده است که در آن به ازای هر کلمه برای هر یک از برچسب‌های احساس مثبت، منفی و بی‌طرف وزنی بین ۰ تا ۱ وجود دارد به گونه‌ای که مجموع این مقادیر برای هر کلمه برابر با ۱ است. در این نقاله از یک لغت‌نامه‌ی احساس به نام MPQA استفاده می‌کنیم. این دیکشنری احساسی شامل ۴۰۵۳ کلمه است که در مقابل هر لغت یک بردار ۳تایی وجود دارد که عدد اول نشان دهنده‌ی وزن بی‌طرفی، عدد دوم نشان دهنده‌ی وزن مثبت و عدد سوم نشان دهنده‌ی وزن منفی برای آن لغت است. در مجموع در این لغت‌نامه ۱۵۱۱ کلمه‌ی مثبت و ۲۵۴۲ کلمه‌ی منفی وجود دارد.

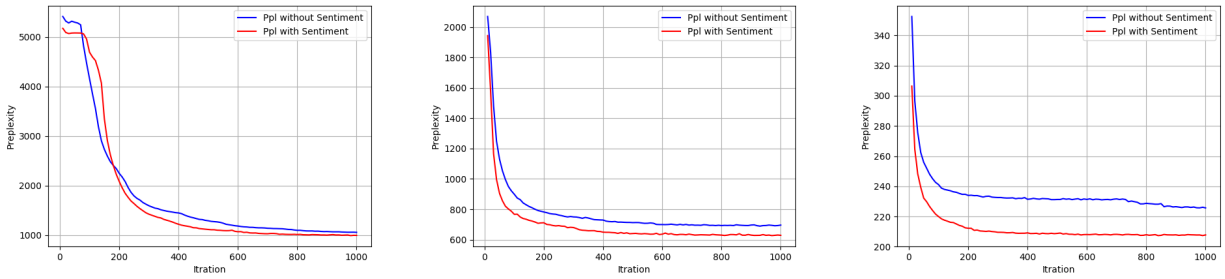
۴.۴ جزئیات آموزش مدل پیشنهادی

برای پیاده‌سازی رویکرد پیشنهادی از زبان برنامه‌نویسی پایتون ورژن ۲,۷ در محیط سیستم عامل لینوکس استفاده شده است. برای پیاده‌سازی ابتدا مدل RS شبیه‌سازی گردید، و پس از ارزیابی و حصول اطمینان از صحت مدل پیاده‌سازی شده شبیه‌سازی انجام شده را به مدل پیشنهادی در این مقاله گسترش دادیم. سپس از پایگاه داده‌ی MR در سه حالت مختلف به عنوان داده‌ی ورودی به مدل برای فرآیند آموزش و آزمون استفاده کردیم و نتایج بدست آمده از آزمایشات انجام شده را در بخش‌های بعدی شرح می دهیم. برای آموزش مدل در سه حالات موجود، یعنی با استفاده از دیکشنری های با سایز ۲۰۰۰، ۱۰۰۰۰ و ۲۴۹۱۶ ما از الگوریتم CD با مرتبه‌ی ۱ استفاده کردیم. همچنین در هر سه حالت، مدل را به ازای ۱۰۰۰ تکرار بر روي كل داده‌هاي آموزش با Batch سایز ۱ آموزش دادیم و نتایج بدست آمده برای هر حالت را در ۲ مرحله، یکی در تکرار ۲۰۰ام و یکی در زمان پایان فرآیند آموزش (تکرار ۱۰۰۰ام) ثبت کردیم. پارامتر دیگری که در آموزش مدل دخیل است تعداد واحدهاي لایه‌ی پنهان یا همان تعداد موضوع‌ها است که می توانند متغیر باشند. در استفاده از هر ۳ دیکشنری، رویکرد پیشنهادی و همچنین مدل RS را به ازای $h = \{5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 60, 70, 80, 90\}$ آموزش دادیم. برای تمام حالت‌ها از مقدار $alfa = 0.001$ برای ضریب یادگیری استفاده کردیم. پارامترهاي $W, U, \mathbf{a}, \mathbf{c}$ که به ترتیب وزن بین لایه‌ی قابل مشاهده و پنهان، وزن بین لایه‌ی احساس و پنهان، بایاس لایه‌ی قابل مشاهده و بایاس لایه‌ی احساس هستند را با استفاده از مقادیری که به صورت تصادفی از یک توزیع گوسی با میانگین ۰ و واریانس ۱ بدست آوردیم، مقدار دهی کردیم. همچنین مقدار اولیه‌ی برای بایاس لایه‌ی پنهان که آن را با $\mathbf{b}$ نشان دادیم را برابر صفر قرار دادیم.

۵.۴ مدل‌سازی اسناد و ارزیابی به عنوان یک مدل مولد

در این بخش رویکرد پیشنهادی را به عنوان یک مدل مولد احتمالاتی با مدل RS در تخمین احتمال برای مشاهده‌ی سندهای پایگاه داده‌های آموزش و تست با استفاده از هر سه دیکشنری مورد ارزیابی قرار داده و با تحلیل نتایج بدست آمده نشان می‌دهیم که روش پیشنهادی نسبت به روش RS یک روش بهتر در تخمین احتمال برای سندهای دیده نشده و آزمون است.

برای ارزیابی احتمال محاسبه شده برای مشاهده‌ی اسناد در فرآیند مدل‌سازی مجموعه سند، از یک معیار به نام سرگشتگی



(آ) با استفاده از دیکشنری با سایز ۲۰۰۰ (ب) با استفاده از دیکشنری با سایز ۱۰۰۰۰ (ج) با استفاده از دیکشنری با سایز ۲۴۹۱۶

شکل ۴: ارزیابی تغییرات سرگشتگی در فرآیند آموزش برروی پایگاه داده‌ی MR برای مدل پیشنهادی و مدل RS

استفاده می‌شود. در مباحث مربوط به NLP معیار سرگشتگی پارامتری است که از آن برای مقایسه‌ی مدل‌های احتمالاتی مختلف استفاده می‌شود. با توجه به فرمول محاسبه‌ی مقدار سرگشتگی که در رابطه‌ی ۱۸ نشان داده شده است می‌توان گفت که مقدار این معیار برابر است با معکوس میانگین درست‌نمایی بدست آماده برای هر سند در مقیاس لگاریتمی به ازای تمام کلمات مجموعه اسناد. در یک فرآیند مدل‌سازی و با استفاده از یک مدل احتمالی مناسب مقدار سرگشتی باید به صورت پیوسته و یکنوا کاهش یابد و مدلی که مقدار سرگشتگی کمتری بر روی پایگاه داده آزمون داشته باشد در بحث مدل‌سازی اسناد به عنوان مدل بهتری شناخته می‌شود.

$$Perplexity = exp\left(-\frac{\sum_{n=1}^{N}\log p(\mathbf{v}_n)}{\sum_{n=1}^{N} D_n}\right) \qquad (۱۸)$$

در شکل ۴ قسمت‌های ۴(آ) تا ۴(ج) نمودار تغییرات سرگشتگی در فرآیند آموزش برای مدل پیشنهادی و مدل RS در حالت‌های مختلف از پایگاه داده‌ی MR نشان داده شده است. همان‌طور که مشاهده می‌گردد در هر ۳ نمودار رویکرد پیشنهادی در این پژوهش که یک مدل مشترک احساس موضوع است نسبت به مدل RS که یک رویکرد موضوعی است با کاهش بهتری در مقدار سرگشتگی همراه است. برای هر سه حالت می‌توان مشاهده کرد که مقدار افت سرگشتگی در ابتدای فرایند آموزش نسبت به مراحل پایانی با سرعت بیشتری همراه بوده است، به صورتی که از تکرار ۲۰۰ام تا به انتها مقدار سرگشتی با تغیرات آنچنانی همراه نبوده است. مشاهده‌ی این ویژگی در فرایند آموزش موجب گردید که ما هر دو مدل را برای هر سه حالت مختلف پایگاه داده به ازای دو مقدار ۲۰۰ و ۱۰۰۰ چرخه، آموزش داده و از نتایج بدست آماده برای تست مدل بر روی پایگاه داده آزمون استفاده کنیم. با دقت در نمودارهای شکل ۴ می‌توان نتیجه گرفت که با اضافه کردن و در نظر گرفتن احساس و ساخت یک مدل مشترک احتمالاتی مولد، مانند آنچه که در این مقاله انجام دادیم، در مرحله‌ی آموزش برای مدل‌سازی اسناد مقدار سرگشتگی با افت بیشتری همراه می‌شود و در نتیجه روش احتمالاتی مناسب‌تری برای مدل‌سازی اسناد ساخته می‌شود.

مقادیر محاسبه شده برای سرگشتگی که در جدول ۳ نشان داده شده است نیز دلیلی بر اثبات ادعای ما نسبت به بهتر بودن رویکرد پیشنهادی در فرآیند مدل‌سازی به عنوان یک مدل مولد است. در جدول ۳ مقدار سرگشتگی برای داده‌های تست در پایگاه داده‌ی MR به‌ازای هر ۳ دیکشنری مورد استفاده و ۲ تکرار ۲۰۰ و ۱۰۰۰ برای هر کدام محاسبه شده

است. با توجه به مقادیر بدست آمده برای سرگشتگی برروی پایگاه داده‌ی تست برای مدل پیشنهادی در این پژوهش در مقایسه با مدل RS در جدول ۳، مشاهده می‌کنیم که در تمامی حالت‌ها رویکرد پیشنهادی مقدار کمتری را برای سرگشتگی محاسبه کرده است. لذا در تایید آنچه که گفتیم نتیجه گرفته می‌شود که راهکار پیشنهادی که با اضافه کردن یک لایه برای احساس نیز همراه است منجر به ساخت یک رویکرد احتمالی مناسب برای مدل‌سازی اسناد است که در مقایسه با مدل RS نتایج بهتری در بحث مدل‌سازی موضوع بدست می‌دهد.

۶.۴ بازیابی اطلاعات

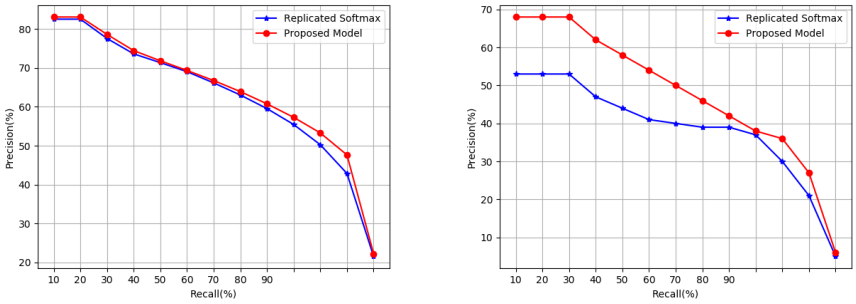
با توجه به اینکه رویکرد پیشنهادی در این پژوهش یک روش مولد برای مدل‌سازی همزمان احساس و موضوع است، لذا گام نخست برای ارزیابی این مدل در بحث بازیابی اطلاعات استفاده از پایگاه داده‌ای است که علاوه بر برچسب احساس برای اسناد، دارای برچسب موضوع برای هر سند نیز باشد. با توجه به عدم وجود یک چنین پایگاه داده‌ای، در این پژوهش ۲ پایگاه داده که همزمان شامل برچسب احساس و برچسب موضوع هستند ساخته شده است.

اولین پایگاه داده‌ی احساس‌ـموضوع ساخته شده در این قسمت با تخصیص برچسب احساس به دیتاست ۲۰NG ساخته می‌شود. برای اضافه کردن احساس به این مجموعه، برای هر سند به شمارش تعداد کلمات با قطبیت مشخص احساسی با استفاده از لغت‌نامه‌ی احساس MPQA کردیم. سپس برای هر سند اگر تعداد کلمات مثبت بیشتر بود به آن سند برچسب مثبت اختصاص دادیم و برعکس.

پایگاه داده‌ی دومی که برای ارزیابی رویکرد پیشنهادی در این پژوهش در بحث بازیابی اطلاعات ساخته می‌شود، از ترکیب چند دیتاست بدست می‌آید. پایگاه داده‌های MR و MDS در بخش ۱.۴ معرفی شدند. هر کدام از این ۵ پایگاه داده (MDS شامل ۴ بخش مختلف با ۲۰۰۰ سند در هر بخش است) تنها شامل برچسب احساس هستند. می‌توان هر کدام از این مجموعه اسناد را به صورت یک موضوع خاص در نظر گرفت. به عبارت دیگر با کنار هم قرار دادن این پایگاه داده‌ها می‌توان یک پایگاه داده بزرگتر ایجاد کرد. این دیتاست جدید ساخته شده شامل ۱۰۰۰۰ سند است که ۵۰۰۰تا از آن‌ها دارای برچسب مثبت و ۵۰۰۰تای دیگر دارای برچسب منفی هستند. همچنین این پایگاه داده‌ی جدید شامل ۵ موضوع مختلف که بازبینی فیلم، کتاب، دی‌وی‌دی، وسایل آشپزخانه و وسایل الکترونیکی هستند،می‌شود. پس اتمام مرحله‌ی پیش‌پردازش هر کدام از این اسناد با استفاده از دیکشنری پایگاه داده‌ی ۲۰NG (۲۰۰۰ کلمه) به فایل

TestSet Type	Num of Docs	Num of Epoch	Ppl without Sentiment	Ppl with Sentiment
MR by 2000	1000	200	400.77	**393.69**
MR by 2000	1000	1000	423.89	**406.74**
MR by 10000	1000	200	1553.52	**1529.42**
MR by 10000	1000	1000	2028.69	**1871.57**
MR by 24916	1000	200	4237.65	**3898.67**
MR by 24916	1000	1000	5842.39	**5824.97**

جدول ۳: تخمین سرگشتگی برای پایگاه داده‌ی Movie Review با استفاده از مدل پیشنهادی



(آ) پایگاه داده‌ی ۲۰ News Groups (ب) پایگاه داده‌ی MRMDS

شکل ۵: بازیابی اطلاعات با استفاده ۲ پایگاه داده‌ی ۲۰ News Groups و MRMDS برای رویکرد پیشنهادی و مدل RS

libsvm تبدیل شدند. از ۱۰۰۰۰ سند نتیجه که با ترکیب این ۵ پایگاه داده بدست می‌آید، ۷۵۰۰ سند برای مجموعه آموزش با توزیع مساوی از نظر برچسب احساس (۳۷۵۰ سند مثبت و ۳۷۵۰ سند منفی) و موضوع (۱۵۰۰ سند از هر موضوع که ۷۵۰تای آن مثبت و ۷۵۰تای دیگر منفی هستند) انتخاب شدند و مابقی مجموعه‌ی تست را که شامل ۲۵۰۰ سند (۵۰۰ سند از هر موضوع که ۲۵۰تای آن مثبت و ۲۵۰تای آن منفی هستند) است، تشکیل می‌دهند. این پایگاه داده‌ی ساخته شده را به اختصار MRMDS نام‌گذاری می‌کنیم.

هدف از ابن ارزیابی مشاهده تاثیر در نظر گرفتن احساس برای بازیابی اطلاعات با استفاده از ساختار پیشنهادی در این پژوهش است. برای ارزیابی مورد نظر از نمودار صحت در برابر بازیابی استفاده می‌کنیم. این نمودار به عنوان معروف‌ترین معیار در بحث ارزیابی بازیابی اطلاعات و مقایسه‌ی روش‌های مختلف در این زمینه شناخته می‌شود. برای رسم این نمودار از مقادیر مختلف صحت و بازیابی که توسط هر مدل بدست می‌آیند در برابر یکدیگر استفاده می‌شود.

نمودارهای شکل ۵ نتایج حاصل از ارزیابی بازیابی اطلاعات برای رویکرد پیشنهادی در این پژوهش و همچنین مدل RS را نشان می‌دهند. همان‌طور که مشاهده می‌شود برای هر ۲ نمودار شکل ۵ بخصوص نمودار ۵(آ) روش پیشنهادی در این پژوهش عملکرد بهتری را در مقایسه با مدل RS در بحث بازیابی اطلاعات داشته است. برای محاسبه مقادیر صحت و بازیابی و رسم نمودارهای شکل ۵ به این صورت عمل شده است که، ابتدا بر روی هر کدام از پایگاه داده‌ها مدل پیشنهادی در این پژوهش بدون در نظر گرفتن برچسب موضوع و تنها با برچسب احساس و لایه‌ی مخفی با سایز ۵۰، و همچنین روش RS بدون در نظر گرفتن برچسب‌های احساس و موضوع و لابه‌ی مخفی با سایز ۵۰ به ازای ۵۰۰ تکرار آموزش داده شده‌اند. در مرحله‌ی بعدی برای تک‌تک سندهای مجموعه‌ی تست در هر پایگاه داده مقدار شباهت کسینوسی هر سند با تمام اسناد پایگاه داده‌ی آموزش محاسبه شده و مقادیر دقت و بازیابی به دست آمده‌اند. در ادامه مقادیر بدست آمده برای صحت برای کل پایگاه داده‌ی تست میانگین گرفته می‌شوند و نمودارهای ۵(آ) و ۵(ب) رسم می‌شوند.

	Total Number	Numb of Positive Words	Num of Negative Words
NG(2000)	155	100	55
RCV(10000)	950	447	503
MR(24916)	3114	1242	1872

جدول ۴: فراوانی‌های بدست آمده از مقایسه‌ی کلمات مشترک بین لغت‌نامه‌ی احساسی MPQA با سه دیکشنری واژگان

۷.۴ مجسم‌سازی موضوع‌ها و ارزیابی دقت در محاسبه‌ی آن‌ها

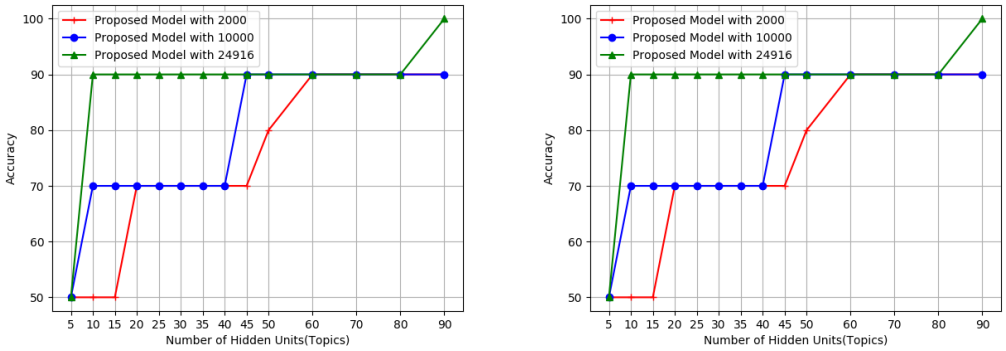
در این بخش با استفاده از لغت‌نامه‌ی احساس MPQA دقت موضوع‌های یاد گرفته شده توسط مدل را از نظر برچسب احساسی مورد ارزیابی قرار می‌دهیم. ایده‌ی ارزیابی مطرح شده در این بخش از آزمایش‌های انجام شده بر روی مدل‌های معروفی در زمینه‌ی مدل‌سازی موضوعی همچون DocNADE و LDA گرفته شده است.

با توجه به ساختار رویکرد پیشنهادی که در بخش ۳ توضیح داده شد، می‌دانیم که هر یک از واحدهای لایه‌ی پنهان به تمام واحدها چه در لایه‌ی قابل مشاهده و چه در لایه‌ی احساس متصل هستند. هر واحد در لایه‌ی احساس برابر با یک برچسب احساسی و هر واحد در لایه‌ی قابل مشاهده متناظر با یک کلمه است. از آنجا که در بحث مدل‌سازی موضوعی اسناد متنی، هر موضوع را به صورت یک توزیع احتمالی چند جمله‌ای بر روی تمام کلمات دیکشنری معرفی کردیم لذا می‌دانیم که هر واحد در لایه‌ی پنهان با یک وزن مشخص به تمام کلمات دیکشنری در لایه قابل مشاهده متصل است. این وزن برای هر کلمه نشان دهنده‌ی مقدار اهمیت آن کلمه در آن موضوع است.

ابتدا برای هر سه حالت مختلف از پایگاه داده تعداد کلمات مشترک با لغت‌نامه‌ی احساس MPQA را محاسبه کردیم. جدول ۴ نتایج مربوط به این عمل را نشان می‌دهد. سپس به ازای هر ۳ حالت از پایگاه داده و ۲ تکرار مختلف برای مرحله‌ی آموزش و همچنین تعداد موضوع‌های مختلف مراحل زیر را به ترتیب انجام دادیم:

۱. محاسبه‌ی مجموع وزن‌های کلمات مثبت و منفی برای هر موضوع با استفاده از لغت‌نامه‌ی احساس و ماتریس وزن بین لایه‌ی قابل مشاهده و پنهان.
۲. محاسبه‌ی تفاضل مقادیر حساب شده در مرحله ۱ برای هر موضوع و مرتب کردن مقادیر حاصل به صورت نزولی.
۳. انتخاب ۵ موضوع از ابتدای لیست مرتب (مثبت‌ترین موضوع‌ها) و تخصیص برچسب مثبت به آن‌ها، و ۵ موضوع از انتهای لیست مرتب (منفی‌ترین موضوع‌ها) و تخصیص برچسب منفی به آن‌ها.
۴. مقایسه‌ی برچسب تخصیص داده شده به هر موضوع با وزن‌های متناظر با آن موضوع در اتصال به لایه‌ی احساس و محاسبه‌ی دقت.

در مرحله‌ی ۴ام منظور از مقایسه‌ی برچسب تخصیص داده شده به هر موضوع با وزن لایه‌ی احساس به این صورت است که اگر به یک موضوع در مرحله‌ی ۳ برچسب مثبت اختصاص داده شد، باید وزن متناظر با برچسب احساس مثبت برای آن موضوع در لایه‌ی احساس بیشتر از وزن منفی برای همان موضوع باشد و بر عکس.نمودارهای ۶(آ) و ۶(ب) نتایج حاصل از این ارزیابی را به ازای ۲۰۰ و ۱۰۰۰ مرحله آموزش نشان می‌دهند. همان‌طور که مشاهده می‌گردد برای هر ۳ حالت مختلف از پایگاه و تعداد موضوع‌های مختلف تفاوتی بین دقت محاسبه شده برای ۲۰۰ و ۱۰۰۰ مرحله‌ی آموزش



(آ) برای ۲۰۰ تکرار (ب) برای ۱۰۰۰ تکرار

شکل ۶: ارزیابی دقت در تخصیص احساس به موضوع‌ها برای ۲۰۰ و ۱۰۰۰ مرحله آموزش

وجود ندارد. اما با دقت در این نمودارها مشاهده می‌گردد که با بزرگ شدن سایز دیکشنری دقت مدل در تخصیص برچسب احساسی به موضوع‌ها نیز افزایش می‌یابد.

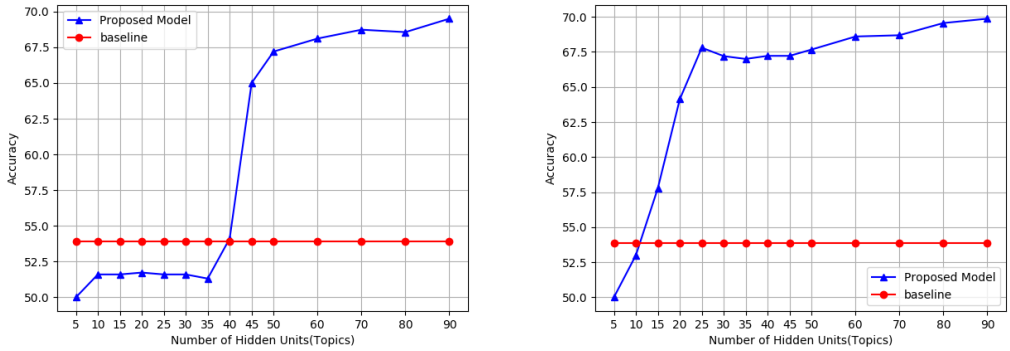
مقایسه‌ی اطلاعات موجود در جدول ۴ برای دیکشنری‌های مختلف با نمودارهای شکل ۶ علت افزایش دقت به ازای افزایش سایز دیکشنری را برای ما توجیه می‌کند. مشاهده می‌شود که با بزرگ شدن اندازه‌ی دیکشنری تعداد کلمه‌های مشترک بین آن و لغت‌نامه‌ی احساس نیز افزیش پیدا می‌کند و این امر سبب می‌گردد که در فرایند آموزش موضوع‌های مثبت و منفی بیشتر از یکدیگر تفکیک شده و در نتیجه دقت مدل در یادگیری و تخصیص برچسب احساس به موضوع‌ها افزایش پیدا می‌کند.

۸.۴ طبقه‌بندی احساسی اسناد

در این بخش نتایج حاصل از طبقه‌بندی احساس با استفاده از رویکرد پیشنهادی در این مقاله را بر روی پایگاه داده‌ی MR ارزیابی و گزارش می‌کنیم. برای مقایسه‌ی نتایج بدست آمده در بحث طبقه‌بندی احساس با استفاده از مدل پیشنهادی، از یک روش پایه که بر اساس شمارش تعداد کلمات است برای ارزیابی دقت در حالت‌های مختلف بهره می‌بریم. همچنین از نتایج بدست آماده برای طبقه‌بندی احساس با استفاده از چند روش معروف نظارت شده مانند ماشین بردار پشتیبان (SVM)، و دو شبکه عصبی (یک شبکه عصبی با مقادیر اولیه‌ی تصادفی و یک شبکه عصبی با مقادیر اولیه‌ی یادگرفته شده توسط رویکرد پیشنهادی) به منظور ارزیابی پارامترهای یاد گرفته شده توسط مدل استفاده می‌کنیم.

شبکه عصبی‌های استفاده شده در هر دو حالت (مقدار دهی تصادفی و مقدار دهی با پارامترهای یاد گرفته شده توسط مدل پیشنهادی) از دسته شبکه‌های MLP هستند. در لایه‌ی اول برای هر دو حالت تعداد نورون‌ها برابر با تعداد موضوع‌ها و در لایه‌ی دوم تعداد نورون‌ها برابر با تعداد احساس‌ها هستند. برای هردوی این شبکه‌ها از تابع خطای Cross Entropy استفاده شده است. همچنین در لایه‌ی اول این شبکه‌ها از تابع فعال‌ساز tanh و در لایه‌ی دوم از تابع Softmax استفاده شده است.

برای محاسبه‌ی دقت در مدل پایه برای هر سند در پایگاه داده‌ی تست شروع به شمارش کلمات با قطبیت مشخص



(آ) با استفاده از دیکشنری با سایز ۲۰۰۰ (ب) با استفاده از دیکشنری با سایز ۱۰۰۰۰

شکل ۷: طبقه‌بندی احساس در پایگاه داده‌ی MR با استفاده از مدل پیشنهادی و مدل پایه برای موضوع‌های مختلف

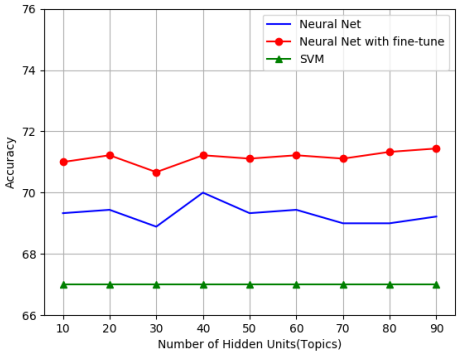
احساسی می‌کنیم. به بیان دیگر برای هر سند تعداد کلمات مثبت و تعداد کلمات منفی را با استفاده از لغت‌نامه‌ی احساس MPQA محاسبه می‌کنیم. بعد از محاسبه‌ی تعداد لغات مثبت و منفی برای هر سند اگر این مقدار برای کلمات مثبت در یک سند بیشتر از کلمات منفی بود به سند مورد نظر برچسب مثبت اختصاص دادیم و اگر تعداد کلمات منفی بیشتر بود آن سند را در دسته سندهای منفی دسته‌بندی می‌کنیم.

برای طبقه‌بندی احساس به کمک رویکرد پیشنهادی در این مقاله به این صورت عمل می‌کنیم که در ابتد برای هر سند متنی با استفاده از رابطه‌ی ۱۱ مقدار لایه‌ی مخفی متناظر با آن را بدست می آوریم. مرحله‌ی بعدی محاسبه‌ی لایه‌ی احساس متناظر با سند جاری با استفاده از رابطه‌ی ۱۶ است. چون مقدار این لایه از یک تابع Softmax بدست می‌آید لذا به فرم یک توزیع احتمالی است که مجموع درایه‌های آن برابر با ۱ است. با بدست آوردن مقدار این لایه‌ی احساس سپس برای تخصیص برچسب به مقادیر این لایه نگاه می‌کنیم و مقدار متناظر با هراحساس که بزرگتر بود برچسب آن سند را برابر با آن احساس انتخاب می کنیم.

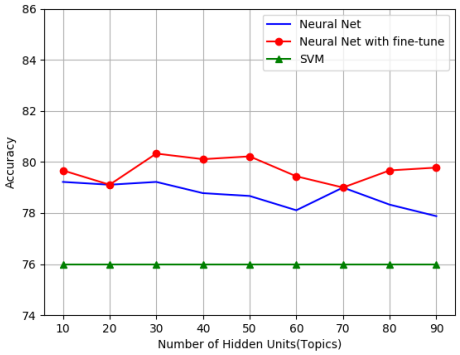
نتایج بدست آمده از طبقه‌بندی احساس با استفاده از رویکرد پیشنهادی و مدل پایه برای ۲ حالت مختلف از پایگاه داده در شکل ۷ نشان داده شده است. برای محاسبه‌ی دقت در طبقه‌بندی احساس با استفاده از مدل پیشنهادی در این پژوهش برای هر ۲ حالت مختلف از پایگاه داده و تعداد موضوع های مختلف از مدل‌هایی که به ازای ۱۰۰۰ تکرار آموزش دیده‌اند استفاده شده است. همچنین برای مقدار دهی اولیه‌ی شبکه عصبی با پارامترهای یاد گرفته شده توسط رویکرد پیشنهادی، از مقادیر بدست آمده برای پارامترها (ماتریس وزن و بایاس) در این حالت (۱۰۰۰ چرخه‌ی آموزش) استفاده کردیم.

همانطور که در شکل‌های ۷(آ) و ۷(ب) مشاهده می گردد در هر ۲ حالت دقت طبقه‌بندی برای مدل پیشنهادی با افزایش تعداد موضوع‌ها رو به افزایش بوده است. همچنین در هر ۲ حالت دقت طبقه‌بندی با استفاده از مدل پیشنهادی با افزایش تعداد موضوع‌ها با اختلاف بسیار زیادی از دقت بدست آمده توسط مدل پایه بهتر است.

شکل ۸ دقت نتایج حاصل از طبقه‌بندی احساس با استفاده از ۳ مدل مختلف را نشان می‌دهد. با دقت در نمودارهای شکل ۸ مشاهده می‌شود که در هر دو حالت مورد نظر برای پایگاه داده، دقت بدست آمده با استفاده از شبکه عصبی با مقدار دهی اولیه توسط پارامترهای یاد گرفته شده در روش پیشنهادی، از دقت بدست آمده توسط هر دو روش دیگر بهتر



(آ) با استفاده از دیکشنری با سایز ۲۰۰۰



(ب) با استفاده از دیکشنری با سایز ۱۰۰۰۰

شکل ۸: طبقه‌بندی احساس در پایگاه داده‌ی MR با استفاده از مدل‌های شبکه عصبی با مقدار دهی اولیه برای وزن‌ها و بایاس‌ها، شبکه عصبی، SVM و Logistic Regression

است.

در حالت استفاده از دیکشنری ۱۰۰۰۰تایی همان‌طور که در شکل ۸(ب) مشخص است، تنها زمانی که تعداد موضوع‌ها برابر با ۲۰ و ۷۰ هستند دقت هر دو شبکه عصبی با هم برابر است. در سایر موارد شبکه با مقدار دهی اولیه‌ی پارامترها نسبت به هر ۳ مدل دیگر نتایج بهتری داشته است. به طور کلی نتیجه گرفته می‌شود که در فرآیند طبقه‌بندی احساس، شبکه‌ی عصبی که مقادیر آن توسط پارامترهای یاد گرفته شده توسط مدل پیشنهادی مقدار دهی اولیه می‌شوند از عملکرد بهتری نسبت به سایر مدل‌ها برخوردار است.

۵ نتیجه‌گیری

در این مقاله مدلی نوین بر پایه‌ی شبکه‌های عصبی برای مدل‌سازی مشترک موضوع واحساس در داده‌های متنی پیشنهاد شده است. بررسی‌های انجام شده نشان داد که در زمینه‌ی مدل‌سازی مشترک موضوع واحساس تنها دو مدل ASUM و JST وجود دارند که با آن‌ها نیز به صورت کامل آشنا شدیم. گسترش شبکه‌های عصبی در سال‌های اخیر و استفاده‌ی فراوان از آن‌ها در بخش‌ها و زمینه‌های مختلف، همچنین عدم وجود ساختاری بر پایه‌ی شبکه‌های عصبی در زمینه‌ی مدل‌سازی مشترک احساس و موضوع به همراه کمبودها و کاستی‌های مدل‌های موجود در این زمینه دلایل اصلی نویسندگان این مقاله برای ساخت رویکردی جدید در این زمینه بوده است.

دراین مقاله یک رویکرد نظارت شده با استفاده از شبکه‌های عصبی برای مدل‌سازی مشترک موضوع واحساس در داده‌های متنی پیشنهاد شد. این ساختار که در دسته روش‌های احتمالاتی مولد دسته‌بندی می‌گردد بر پایه‌ی شبکه‌ی عصبی ماشین بلتزمن محدود است و با گسترش مدل RS ایجاد می‌شود. در رویکرد پیشنهادی یک لایه‌ی جدید با ماهیت توزیع احتمالاتی چندجمله‌ای به مدل اضافه شده و منجر به یادگیری ویژگی‌های بهتر و متمایز کننده‌تری برای هر سند در لایه‌ی مخفی می‌شود. برای یادگیری و آموزش در این مدل از الگوریتم واگرایی مقابله استفاده می‌کنیم که یک روش تقریبی برای

تخمین گرادیان است.

برای ارزیابی مدل پیشنهادی از پایگاه داده‌های بازبینی فیلم (MR)، ۲۰ گروه خبری (۲۰NG) و احساس چند دامنه (MDS) که همگی از پایگاه داده‌های شاخص در بحث مدل‌سازی موضوع و احساس در داده‌های متنی هستند، استفاده کردیم. با استفاده از یک معیار معروف برای ارزیابی مدل‌های مولد به نام سرگشتگی، رویکرد پیشنهادی در این پروژه را در فرآیند مدل‌سازی اسناد متنی ارزیابی کردیم. با توجه به نتایج بدست آمده در این بخش ادعا می‌کنیم که با در نظر گرفتن احساس موجود در اسناد و ایجاد ساختاری مانند آنچه که ما در این مقاله انجام دادیم، یک مدل مولد بهتر برای مدل‌سازی اسناد ساخته می‌شود. همچنین رویکرد پیشنهادی در این مقاله را در بحث بازیابی اطلاعات در مقایسه با مدل RS مورد ارزیابی قرار دادیم. با انجام آزمایش بر روی ۲ پایگاه داده‌ی مختلف و مقایسه‌ی نتایج مشاهده گردید که روش پیشنهادی نتایج بهتری را در بحث بازیابی اطلاعات بر روی داده‌های متنی دارد و از دقت بالاتری در این زمینه برخوردار است.