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 [۱]. 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 مفاله 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 احساس_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.

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. در بحث مدل‌سازی موضوع، مدل‌های گرافی نسبت به مدل‌های شبکه‌های عصبی از قدمت بیشتری برخوردار هستند. مهم‌ترین مدلی که در این دسته 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. 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 معرفی شدند.

از نظر نحوه‌ی عملکرد مدل‌های پیشین را در سه کلاس مختلف قرار می‌دهیم. دسته‌ي اول روش‌هایی که تنها به مدل‌سازی موضوع می‌پردازند و آن‌ها را به عنوان روش‌های مدل‌سازی موضوعی معرفی می‌کنیم. دسته‌ی دوم روش‌هایی که تنها به تشخیص احساس و دانش مفهومی از داده‌های ورودی می‌پردازند. اگرچه باید توجه داشت که مدل‌های موجود در زمینه تشخیص احساس در دسته‌ی مدل‌های موضوعی قرار نمی‌گیرند و بیشتر شامل مدل‌های یادگیری ماشین می‌باشند که یک طبقه‌بندی دو حالته (مثبت و منفی) یا سه حالته (منفی، مثبت و بی طرف) را انجام می‌دهند. و در دسته‌ي سوم از نظر نحوه‌ی عملکرد روش‌هایی را بررسی می‌کنیم که به صورت همزمان به مدل‌سازی موضوع و احساس بر روی داده‌ی ورودی می‌پردازند.

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.

	Structure		Modeling Type		Data Input	
	Graphical	Neural Net	Topic	Sentiment/Topic	Unimodal	MultiModal
LSI	⋆		⋆		⋆	
pLSI	⋆		⋆		⋆	
LDA	⋆		⋆		⋆	
JST	⋆			⋆	⋆	
ASUM	⋆			⋆	⋆	
NADE		⋆			⋆	
RBM		⋆			⋆	
RS		⋆	⋆		⋆	
DocNADE		⋆	⋆		⋆	
SupDocNADE		⋆	⋆			⋆

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

۳ مدل‌سازی مشترک موضوع و احساس با شبکه‌های عصبی

۴ آزمایش‌ها و ارزیابی مدل

۵ نتیجه‌گیری

مراجع

[1] Lin, Chenghua, He, Yulan, Everson, Richard, and Ruger, Stefan. Weakly supervised joint sentiment-topic detection from text. *IEEE Transactions on Knowledge and Data engineering*, 24(6):1134–1145, 2012.