علوم رایانش و فناوری اطلاعات
نشریه علمی انجمن کامپیوتر ایران
مجلد ۲۰، شماره ۱، ۱۴۰۱
صفحات ۱۱-۱۹
مقاله پژوهشی

پیش‌بینی شباهت متن با استفاده از یک شبکه عصبی سیامی مبتنی بر شبکه عمیق و ویژگی‌های شباهت لغوی

فریبا خلج[۱]، حسین عباسی مهر[۲*]

*نویسنده مسئول، دریافت: ۱۴۰۰/۰۸/۱۸، بازنگری: ۱۴۰۰/۱۲/۱۲، پذیرش: ۱۴۰۰/۱۲/۲۱

[۱و۲]دانشکده فناوری اطلاعات و مهندسی کامپیوتر، دانشگاه شهید مدنی آذربایجان، تبریز

چکیده

اندازه‌گیری شباهت متن یکی از اصلی‌ترین عملیات در کاربردهای مرتبط با متن نظیر بازیابی اطلاعات، خوشه‌بندی متن، سیستم‌های پرسش و پاسخ است. هدف این مطالعه ارائه رویکردی برای بهبود دقت مدل‌های یادگیری عمیق در اندازه‌گیری تشابه متون است. بدین منظور یک رویکرد ترکیبی مبتنی بر شبکه عصبی سیامی و ویژگی‌های شباهت لغوی ارائه می‌شود. شبکه سیامی پیشنهادی شامل دو زیر شبکه یکسان است که اجزای اصلی هر کدام از آن‌ها به صورت کلی شامل یک لایه تعبیه کلمات و شبکه عصبی عمیق است. با در نظر گرفتن سه نوع شبکه عصبی عمیق شامل شبکه عصبی پیچشی، شبکه حافظه کوتاه‌مدت طولانی و شبکه حافظه کوتاه‌مدت طولانی دوطرفه و همچنین دو نوع مدل تعبیه کلمات به همراه ویژگی‌های شباهت لغوی، گونه‌های مختلفی از مدل‌ها پیاده‌سازی می‌شود. نتایج آزمایش‌ها روی سه مجموعه داده مورد استفاده نشان می‌دهد مدل شبکه عصبی سیامی ترکیبی مبتنی بر شبکه پیچشی و ویژگی‌های لغوی بالاترین مقدار همبستگی پیرسون و کمترین مقدار میانگین مربع خطاها (MSE) را در بین مدل‌ها به دست می‌آورد. همچنین نتایج بدست آمده حاکی از عملکرد موفق مدل پیشنهادی نسبت به مدل‌های تحقیقات قبلی در معیارهای ضریب همبستگی و MSE است.

کلمات کلیدی: شباهت متن، یادگیری عمیق، شبکه سیامی، معیارهای تشابه متن

۱- مقدمه

پردازش داده‌های متنی با توجه به تولید روزافزون این داده‌ها در پلتفرم‌های مختلف از اهمیت زیادی برخوردار است. اندازه‌گیری شباهت متن یکی از مهم‌ترین عملیات در کاربردهای مختلف متن‌کاوی نظیر بازیابی اطلاعات، دسته‌بندی متن، خوشه‌بندی متن، خلاصه‌سازی متن و سیستم‌های پرسش و پاسخ است[۱-۳].

با توجه به اهمیت اندازه‌گیری شباهت متن، روش‌های مختلفی در تحقیقات قبلی در این راستا ارائه شده است [۱-۶]. این روش‌ها را می‌توان به صورت کلی به روش‌های سنتی و روش‌های مبتنی بر یادگیری عمیق دسته‌بندی کرد. روش‌های سنتی تشابه متن در سطح کلمات عمل می‌کنند. یافتن شباهت بین کلمات اولین گام برای محاسبه شباهت بین جملات، پاراگراف‌ها و اسناد است. شباهت کلمات می‌تواند به دو روش شباهت لغوی و معنایی باشد. کلمات به صورت لغوی مشابه هستند اگر ترتیب کاراکترهای موجود در آن‌ها شبیه باشد. همچنین کلمات از لحاظ معنایی مشابه هستند اگر یک کارکرد مشابه‌ی داشته باشند و در یک زمینه مشابه‌ی استفاده گردند. به صورت کلی روش‌های مبتنی بر شباهت لغوی[۱] با استفاده از الگوریتم‌های مقایسه رشته‌ها کار می‌کنند. همچنین روش‌های مبتنی بر شباهت معنایی بر اساس الگوریتم‌های مبتنی بر پیکره متنی[۲] و مبتنی بر دانش[۳] عمل می‌کنند [۸, ۷].

روش‌های نوین در حوزه اندازه‌گیری تشابه متن مبتنی بر تکنیک‌های شبکه عصبی و یادگیری عمیق هستند. روش‌های مبتنی بر یادگیری عمیق در ایجاد بازنمایی درستی از کلمات نتایج موفقیت آمیزی داشته‌اند [۲]. روش تعبیه کلمات[۴] [۹] کار بازنمایی کلمه را با در نظر گرفتن کلمات اطراف آن در یک پیکره متن انجام می‌دهند. خروجی یک روش تعبیه کلمه، یک بردار (معمولا ۳۰۰ عنصری از اعداد حقیقی) که معنی کلمه را منعکس می‌کند. کلماتی که دارای معانی مشابه‌ی هستند، بردارهای نزدیک‌تری را دارند. بسیاری از رویکردهای جدید از بردارهای تعبیه کلمه

برای اندازه‌گیری شباهت جملات استفاده می‌کنند که این رویکردها در دسته یادگیری بدون نظارت طبقه‌بندی می‌شوند [۱۱، ۱۰، ۲].

تکنیک‌های یادگیری عمیق شبکه مانند شبکه‌های عصبی بازگشتی[۵]، شبکه حافظه کوتاه مدت طولانی[۶] (LSTM)، در سال‌های اخیر توجه زیادی را به خود جلب کرده‌اند. با استفاده از این روش‌ها یک بازنمایی از جملات و متن‌ها با در نظر گرفتن ترتیب کلمات بدست می‌آید [۱۱]. شبکه‌های عصبی بازگشتی توانایی پردازش وابستگی بین کلمات را دارند و بازنمایی ایجاد شده توسط آن‌ها به اندازه‌گیری دقیق شباهت متون کمک می‌کند. در دسته دیگری از تحقیقات قبلی بازنمایی ایجاد شده توسط لایه‌های شبکه‌های عصبی بازگشتی اغلب در یک معماری شبکه عصبی سیامی[۷] مورد استفاده قرار گرفته است و نتایج خوبی در زمینه اندازه‌گیری شباهت متن حاصل شده است [۱۲-۱۴]. روش‌های ارائه شده در این تحقیقات جز دسته روش‌های یادگیری با نظارت هستند. شبکه عصبی سیامی یک نوع شبکه عصبی خاص است که در وظایف مربوط به اندازه‌گیری تشابه یا رابطه بین دو چیز قابل مقایسه (برای مثال دو تصویر) بسیار مورد استفاده قرار می‌گیرد. شبکه‌های سیامی شامل دو یا چند زیرشبکه یکسانی هستند که هر کدام از آن‌ها دارای پیکره بندی مشابه و پارامترهای مشترکی هستند [۱۱].

عملکرد موفق روش‌های یادگیری عمیق در پژوهش‌های پیشین مورد تایید قرار گرفته است [۱۵، ۱۴، ۱۱، ۱]. در این تحقیق به بررسی این موضوع می‌پردازیم که ترکیب ویژگی‌های بدست آمده از شبکه عصبی عمیق با ویژگی‌های شباهت لغوی منجر به بهبود عملکرد مدل یادگیری عمیق می‌شود یا خیر. در این راستا یک رویکرد ترکیبی مبتنی بر شبکه عصبی سیامی در این تحقیق ارائه می‌شود. در رویکرد ترکیبی ارائه شده با در نظر گرفتن لایه‌های عمیق شامل شبکه حافظه کوتاه-مدت طولانی، شبکه پیچشی[۸] و شبکه حافظه کوتاه-مدت طولانی دو طرفه[۹] و همچنین با استفاده از دو رویکرد تعبیه کلمات، مدل‌های مختلفی پیاده‌سازی شده و بر روی مجموعه داده N2C2[۱۰] اعمال شدند. نتایج ارزیابی مدل‌ها با استفاده از معیار همبستگی پیرسون[۱۱] و میانگین مربع خطاها[۱۲] نشان داد که مدل ترکیبی با استفاده از شبکه پیچشی بهترین نتیجه در پیش‌بینی شباهت را در بین مدل‌ها دارد. به صورت کلی مدل‌های ترکیبی نتیجه بهتری نسبت به مدل‌های پایه مبتنی بر شبکه عمیق دارند. همچنین مدل پیشنهادی نسبت به مدل‌های ارائه شده در [۱۵, ۱۴] عملکرد بهتری دارد.

ساختار ادامه این مقاله بدین‌گونه است. بخش ۲ ابتدا مرور ادبیات پیش‌بینی شباهت با استفاده از روش‌های یادگیری عمیق را ارائه می‌دهد و سپس به توصیف مدل‌های مورداستفاده می‌پردازد. در بخش ۳، ابتدا به بیان مسئله، و سپس به تشریح رویکرد پیشنهادی می‌پردازیم. بخش ۴ به توصیف داده‌ها و آزمایش‌ها پرداخته و نتایج مدل‌ها را مورد مقایسه و تحلیل قرار می‌دهد. در بخش ۵ نتیجه‌گیری و ارائه پیشنهاد کارهای آتی می‌پردازیم.

۲- کارهای مرتبط

در این بخش ابتدا به مرور کلی تحقیقات قبلی پرداخته و سپس به توصیف روش‌های یادگیری عمیق بکارگرفته شده در رویکرد پیشنهادی می پردازیم.

۲-۱- مروری بر روش‌های تخمین شباهت متن مبتنی بر تکنیک‌های یادگیری عمیق

روش‌های نوین در حوزه اندازه‌گیری تشابه متن با استفاده از تکنیک‌های شبکه عصبی و یادگیری عمیق ارائه شده‌اند. این روش‌ها به طورکلی به دو دسته روش‌های بی نظارت و با نظارت تقسیم می‌شوند. روش‌های بدون نظارت مبتنی بر یادگیری عمیق بر اساس بازنمایی ایجاد شده توسط یک روش تعبیه کلمه عمل می‌کنند و با داشتن بردار هر کلمه، بردار جملات را بدست آورده و با استفاده از معیارهای شباهت نظیر معیار کسینوسی، شباهت بین دو جمله یا متن را محاسبه می‌کنند [۱۶، ۱۱، ۱۰، ۲].

در دسته روش‌های بانظارت، معماری‌های مختلف شبکه عصبی عمیق جهت یادگیری مدل اندازه‌گیری تشابه مورد استفاده قرار گرفته است که در این میان شبکه‌های یادگیری عمیق مانند شبکه‌های عصبی بازگشتی، شبکه حافظه کوتاه مدت طولانی از روش‌های پرکاربرد در این زمینه محسوب می‌شوند. این شبکه‌ها بردار تعبیه کلمات مربوط به دو متن را به عنوان ورودی گرفته و بازنمایی منعکس‌کننده معنای آن دو متن را با استفاده از شبکه عمیق بدست آورده و بر اساس آن کار پیش‌بینی شباهت را انجام می‌دهند. این تحقیقات غالبا از یک معماری شبکه عصبی سیامی استفاده کرده و عملکرد موفقی در زمینه اندازه‌گیری شباهت متن به دست آورده‌اند [۱۳، ۱۲]. جدول ۱ چندین نمونه از تحقیقات مبتنی بر یادگیری عمیق را با ارائه اطلاعاتی درباره نوآوری روش ارائه شده، رویکرد و معیارها و ابزارهای مورد استفاده و مجموعه داده انتخاب شده جهت انجام آزمایشات نشان می‌دهد.

۲-۲- روش‌های مورداستفاده در این پژوهش

در این تحقیق از ۳ شبکه عصبی بازگشتی حافظه کوتاه-مدت طولانی، شبکه عصبی بازگشتی حافظه کوتاه-مدت طولانی دوطرفه و شبکه عصبی پیچشی استفاده می‌کنیم. در بخش‌های زیر هرکدام از این شبکه‌ها را به‌صورت مختصر توضیح می‌دهیم.

۲-۲-۱- شبکه کانولوشنی

شبکه عصبی پیچشی یکی از مهم‌ترین روش‌های یادگیری عمیق هستند که در آن‌ها چندین لایه با روشی قدرتمند آموزش می‌بینند این روش بسیار کارآمد بوده و یکی از رایج‌ترین روش‌ها در کاربردهای مختلف بینایی کامپیوتر است [۲۰, ۱۱]. یک لایه پیچشی داده‌های ورودی متن را دریافت کرده و با انجام عملیات کانولوشن با استفاده از کرنل‌های کانولوشن ویژگی‌های جدیدی از متن را استخراج می‌کند. هر لایه کانولوشن شامل یک کرنل (یک پنجره کوچک) است که روی داده‌ها حرکت کرده و از طریق انجام عملیات کانولوشنی ویژگی‌های جدید را محاسبه می‌کند [۱۱]. ویژگی‌های جدید قابلیت متمایزسازی بالایی نسبت به داده‌های خام ورودی داشته و باعث بهبود دقت پیش‌بینی می‌شود.

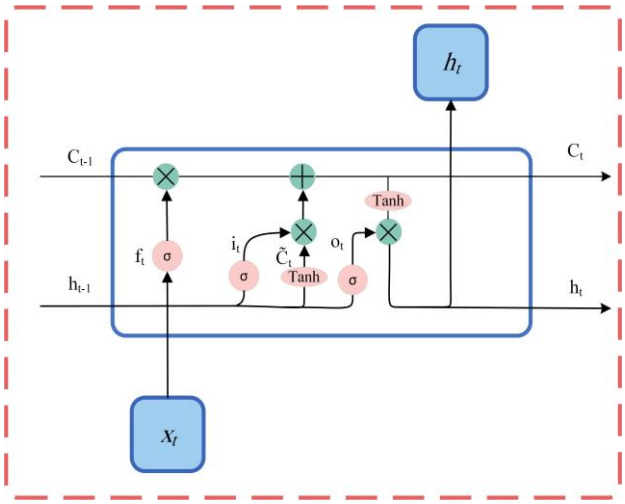
۲-۲-۲- شبکه LSTM

شبکه LSTM یک نوع شبکه عصبی بازگشتی بهبودیافته است که توسط هوچریتر و اشمیت [۲۰] توسعه داده شده است. این شبکه برای رفع مشکل ناپدید شدن گرادیان و عدم یادگیری توالی‌های طولانی در شبکه‌های عصبی بازگشتی معرفی شده است و توانایی به یادسپاری اطلاعات برای بازه‌های زمانی بلندمدت را دارد همچنین LSTM می‌تواند اطلاعات وابستگی‌های طولانی مدت در داده‌های متنی را استخراج کرده و به طور مناسب بین داده ورودی و خروجی نگاشت ایجاد کند [۱۱].

مطابق شکل ۱ LSTM از چهار بخش سلول حافظه C، دروازه ورودی i، دروازه فراموشی f، و دروازه خروجی o تشکیل شده است. دروازه ورودی i_t تشخیص می‌دهد که از کدام مقدار ورودی باید برای بهبود حافظه استفاده شود. اینکه تا چه حدی اطلاعات حافظه فعلی فراموش شود توسط دروازه فراموشی، f_t تعیین می‌شود. دروازه خروجی o_t اطلاعات خروجی واحد LSTM را تنظیم می‌کند.

جدول ۱- بررسی روش‌های شباهت متن

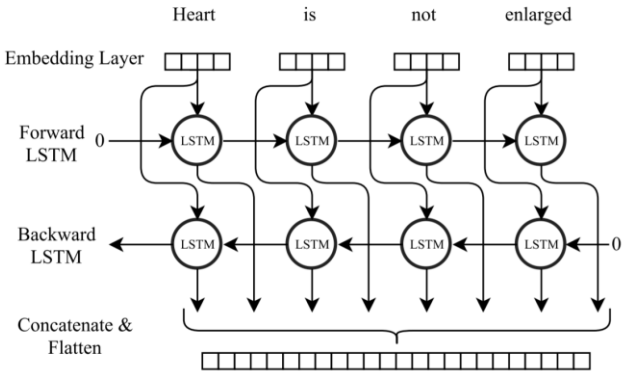
مرجع	نوآوری	رویکرد	مجموعه داده
[۱۵]	ارائه یک شبکه عصبی سیامی و استفاده از مکانیزم توجه برای بهبود اندازه‌گیری شباهت متن	شبکه عصبی بازگشتی دوطرفه مکانیزم خود توجه	N2C2 CDD-ful CDD-ref
[۲]	استفاده از بردار تعبیه کلمات در کنار معیارهای سنتی	بردار تعبیه کلمات بردار ترتیب کلمات شباهت ساختاری	مجموعه داده معیار شباهت معنایی با فرهنگ لغت Collins Cobuild Microsoft Research Li2006
[۱۴]	ارائه یک رویکرد ترکیبی مبتنی بر شبکه عصبی سیامی	شبکه LSTM معیارهای سنتی نظیر Dice، جاکارد	مجموعه جملات پزشکی و جملاتی از زبان پرتغالی
[۱۷]	ارائه یک روش محاسبه شباهت معنایی با استفاده از ویکی‌پدیا برای از بین بردن محدودیت‌های اطلاعات ناکافی و از دست دادن اطلاعات معنایی است	محتوای اطلاعات ویژگی‌های مفاهیم	R&G M&C WS353-Sim
[۵]	استفاده از پایگاه‌داده واژگانی و آماره corpus برای تعیین شباهت معنایی کلمات	پایگاه‌داده واژگانی WordNet بردار معنایی محتوای اطلاعاتی کلمه بردار ترتیب	Pilot Short Text Semantic Similarity Benchmark Data Set
[۱]	خلاصه‌سازی متن با استفاده از یادگیری عمیق	تجزیه‌وتحلیل واژگانی تحلیل‌گر کلمه به کلمه و جمله به جمله WordNet پایگاه‌داده الگوریتم فاصله Levenshtein	چندین مجموعه متون انگلیسی و بنگالی
[۴]	روشی برای محاسبه تشابه متن کوتاه بر اساس اطلاعات معنایی و نحوی	ترکیب اطلاعات معنایی و نحوی بردار معنایی پویا تجزیه و تحلیل ساختار با درخت تجزیه	۲۴ مجموعه داده تشابه متون کوتاه
[۱۸]	محاسبه تشابه متن بر اساس توالی‌های مداوم	توالی کامل کلمات معماری شبکه عصبی برای آموزش پارامترهای TextFlow برای کارهای خاص	۸ مجموعه داده شامل سؤالات StackOverflow و انواع داده‌های دیگر
[۱۹]	محاسبه شباهت متن معنایی با استفاده از ویژگی‌های غنی	مدل Support Vector Regression ویژگی‌های WordNet-based Word2Vec-based Corpus-based Alignment-based Literal-based	SemEval 2015



شکل ۱- ساختاریک واحد شبکه حافظه کوتاه-مدت طولانی [۲۰]

۲-۲-۳- LSTM دوطرفه

با توصیف جزئیات مربوط به هر واحد از LSTM در بخش قبلی در این بخش به توصیف LSTM دوطرفه می‌پردازیم. این مدل از دو LSTM یک‌طرفه جداگانه تشکیل شده است که در مسائل مدل‌سازی ترتیبی عملکرد مناسبی از خود نشان داده است و همچنین یکی از مدل‌های پرکاربرد در طبقه‌بندی متون محسوب می‌شود. شکل ۲ معماری یک شبکه LSTM دو طرفه را نشان می‌دهد. همانطور که در شکل مشاهده می‌شود با داشتن یک جمله به عنوان ورودی، Forward LSTM محاسبات رو به جلو را انجام می‌دهد، یعنی از ابتدای جمله شروع می‌کند و محاسبات را انجام داده و خروجی را در هر نقطه زمانی (به ازای هر کلمه)، بر اساس اطلاعات کلمات مجاور بدست می‌آورد. همچنین Backward LSTM از انتهای یک جمله شروع کرده و در جهت عکس محاسبات را انجام می‌دهد و در نهایت بازنمایی حاصل از هر دو LSTM ترکیب می‌شود.



شکل ۲- معماری یک شبکه LSTM دوطرفه [۲۱]

۳- روش پیشنهادی

در بخش ۲ خلاصه‌ای از رویکردهای مختلف برای یافتن شباهت بین متون بیان شد. استفاده از شبکه‌های عصبی عمیق جهت آموزش مدل‌های اندازه‌گیری تشابه یکی از رویکردهای نوین در سال‌های اخیر است. در این مطالعه به دنبال ارائه رویکردی برای بهبود دقت مدل‌های یادگیری عمیق در اندازه‌گیری تشابه متون هستیم. بدین ترتیب با به‌کارگیری لایه‌های مختلف یادگیری عمیق نظیر شبکه پیچشی، حافظه کوتاه‌مدت طولانی و همچنین حافظه کوتاه‌مدت طولانی دوطرفه و با گنجاندن ویژگی‌های شباهت لغوی نظیر شباهت کسینوسی، یک روش شبکه عصبی سیامی ترکیبی برای محاسبه شباهت بین متن ارائه می‌کنیم. در این بخش ابتدا به توصیف مسئله شباهت متن می‌پردازیم سپس رویکرد پیشنهادی تحقیق را تشریح می‌کنیم.

۳-۱- مسئله شباهت متن

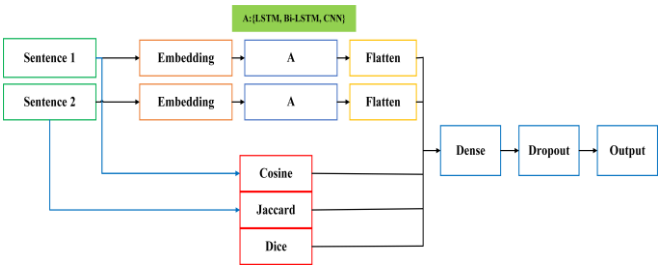
در این بخش مسئله شباهت متن را با درنظرگرفتن شکل زیر توضیح می‌دهیم. ورودی الگوریتم پیش‌بینی شباهت، دو متن S_1 و S_2 است. به‌ازای دو متن ورودی یک امتیاز شباهت بین ۰ تا ۵ وجود دارد که امتیاز ۰ نشان‌دهنده عدم تشابه و امتیاز ۵ بیانگر کاملاً مشابه است (جدول۲). هدف الگوریتم، پیش‌بینی شباهت دو متن بر اساس مدل آموزش دیده است.

جدول ۲- مثال‌هایی از جفت جملات و امتیاز شباهت آن‌ها: امتیاز صفر نشان دهنده عدم شباهت بین دو جمله و امتیاز ۵ نشان دهنده شباهت کامل بین دو جمله است.

متن ۱ (S1)	متن ۲ (S2)	امتیاز شباهت
When visiting a restaurant, the patient typically does add salt to the meal.	A letter has been sent for the patient to call and schedule.	۰
Gastrointestinal: Negative for abdominal (belly) pain or cramping and diarrhea.	Negative for abdominal pain, diarrhea, nausea and vomiting.	۵

۳-۲- رویکرد پیشنهادی

شکل ۳ ساختار کلی رویکرد پیشنهادی این تحقیق برای محاسبه شباهت بین متن را نشان می‌دهد. همانطور که اشاره شد، رویکرد پیشنهادی عبارت است از یک معماری شبکه عصبی سیامی که شامل ویژگی‌های شباهت رایج است. اجزای اصلی رویکرد پیشنهادی عبارتند از: ۱) لایه ورودی ۲) لایه تعبیه کلمات ۳) لایه اصلی ۴) لایه تماما متصل ۵) لایه خروجی. ابتدا عملیات پیش پردازش روی جملات انجام می‌گیرد. این عملیات شامل جداسازی کلمات و حذف کلمات توقف است. جملات پیش پردازش شده در ادامه عملیات وارد لایه تعبیه کلمات می‌شوند که به هر کلمه یک بردار حقیقی تعبیه کلمات نسبت داده می‌شود. برای بدست آوردن بردار تعبیه هر کلمه علاوه بر استفاده از لایه تعبیه کتابخانه کراس، از مدل‌های از پیش آموزش دیده در این حوزه نیز استفاده می‌شود. در ادامه به ازای هر جمله بردارهای کلمات آن ادغام شده و بردار مربوط به هر جمله تشکیل می‌شود که این بردارها وارد لایه‌های شبکه عصبی عمیق می‌شوند. همانطور که در شکل ۳ مشاهده می‌شود، در این تحقیق از سه لایه حافظه کوتاه-مدت طولانی، حافظه کوتاه-مدت طولانی دوطرفه و شبکه پیچشی استفاده می‌کنیم. وظیفه این لایه ایجاد یک بازنمایی جدید از هر یک از جملات ورودی است طوری که بتواند معنای موجود در جمله را بیان کند. به جهت اینکه از معماری شبکه عصبی سیامی استفاده شده است، متناظر با هر زیرشبکه در شبکه سیامی، یک نوع لایه استفاده شده است (لایه A). همچنین در مدل پیشنهادی سه ویژگی شباهت لغوی رایج شامل معیار شباهت کسینوسی[۱۳]، جاکارد[۱۴]، دایس[۱۵] جهت بهبود مدل محاسبه شده (مطابق فرمول های (۱) تا (۳)) و با ادغام با بردارهای خروجی مسطح شده لایه‌های اصلی وارد یک لایه تماما متصل[۱۶] می‌شود. در ادامه یک لایه حذف تصادفی[۱۷] در مدل قرار داده می‌شود که وظیفه آن جلوگیری از بیش برازش مدل است. خروجی لایه تماما متصل پس از عبور از لایه حذف تصادفی وارد لایه خروجی می‌شود که کار پیش‌بینی شباهت بین دو جمله را انجام می‌دهد.



شکل ۳- رویکرد پیشنهادی برای پیش بینی تشابه جملات

ویژگی‌های لغوی مطابق با فرمول (۱) محاسبه می‌شوند. در این فرمول، S_1 و S_2 به ترتیب مجموعه کلمات جمله ۱ و جمله ۲ در شکل ۳ می‌باشند.

$$\text{Cosine}(S_1, S_2) = \frac{|S_1 \cap S_2|}{\sqrt{|S_1|}\sqrt{|S_2|}} \quad (۱)$$

$$\text{Jaccard}(S_1, S_2) = \frac{|S_1 \cap S_2|}{|S_1 \cup S_2|} \quad (۲)$$

$$\text{Dice}(S_1, S_2) = \frac{2 \times |S_1 \cap S_2|}{|S_1| + |S_2|} \quad (۳)$$

۴- آزمایش‌های عملی و نتایج

در این بخش ابتدا به توصیف مجموعه داده‌های مورداستفاده جهت انجام آزمایش‌ها پرداخته سپس معیارهای اندازه‌گیری عملکرد مدل‌ها و تنظیمات مدل‌ها و نتایج را بیان می‌کنیم. در این مطالعه از کتابخانه Keras [۱۱] جهت پیاده‌سازی مدل‌ها استفاده می‌کنیم.

۴-۱- مجموعه داده‌ها

در این مطالعه، از سه مجموعه داده استاندارد پزشکی که هر کدام شامل مجموعه‌ای از جفت جملات با نمرات شباهت/ مرتبط بودن است، استفاده می‌کنیم. توصیف مجموعه داده‌ها در جدول ۳ بیان شده است.

جدول ۳: توصیف مجموعه داده‌های مورد استفاده

مجموعه داده	تعداد نمونه‌های آموزشی	تعداد نمونه‌های تست
N2C2[18]	۱۶۵۵	۴۱۲
CDD-ful [۲۲]	۲۵۷۱	۵۰۹
CDD-ref [۲۲]	۲۵۸۸	۵۰۹

۴-۲- معیارهای اندازه‌گیری

در این تحقیق، دقت مدل‌ها با استفاده از معیارهای همبستگی پیرسون و معیار MSE محاسبه می‌شود. با درنظرگرفتن n جفت جمله $P = (P_1, P_2, \cdots, P_n)$ و مجموعه‌های $Y = (y_1, y_2, \cdots, y_n)$ و $\hat{Y} = (\hat{y}_1, \hat{y}_2, \cdots, \hat{y}_n)$ به ترتیب به عنوان امتیازات شباهت واقعی و پیش‌بینی شده مجموعه P، معیار همبستگی پیرسون با استفاده از رابطه (۲) محاسبه می‌شود:

$$r = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}\sqrt{\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2}} \quad (۴)$$

همچنین معیار میانگین مربع خطاها با رابطه (۳) محاسبه می‌شود.

$$MSE = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n} \quad (۵)$$

که در رابطه‌های (۴) و (۵) $\hat{y}_i, y_i$ به ترتیب نشان‌دهنده امتیاز شباهت واقعی و پیش‌بینی شده جفت جمله p_i است. در رابطه (۲) $\bar{y}$ و $\bar{\hat{y}}$ به ترتیب مقادیر میانگین مجموعه Y و $\hat{Y}$ را نشان می‌دهند [۱۱].

۴-۳- تنظیمات آزمایش‌ها

تنظیم هایپرپارامترها نقش مهمی در عملکرد مدل‌های یادگیری عمیق دارد. پارامترها و هایپرپارامترهای مورداستفاده در هر لایه در جدول ۴ توصیف شده است.

جدول ۴- تنظیمات لایه‌ها در رویکرد پیشنهادی

مشخصه شبکه		مقدار پارامتر /هایپرپارامتر
ورودی		۲ متن به طول ۱۰۰ کلمه
خروجی		پیش‌بینی شباهت دو جمله در بازه [۵-۰]
ابعاد لایه تعبیه		۳۰۰
لایه اصلی	LSTM	تعداد نورون‌ها: ۳۰۰
	Bi-LSTM	تعداد نورون‌ها: ۳۰۰
	CNN	تعداد فیلترها در لایه کانولوشن: ۳۲ اندازه هر فیلتر: ۳، ۶ تعداد کلمات ادغامی در لایه ماکزیمم ادغام (pool_size): ۲
ابعاد لایه تماماً متصل		۵۰
نرخ لایه حذف تصادفی		۰/۲۵
تابع فعال‌ساز		RELU[19]
تابع بهینه‌ساز		آدام[۲۰]
تابع زیان[۲۱]		MSE
تعداد تکرارهای شبکه[۲۲]		۱۰-۱۰۰

۴-۴- نتایج

در این بخش با در نظر گرفتن رویکرد پیشنهادی (شکل ۳) و بکارگیری دو لایه تعبیه شامل لایه تعبیه کتابخانه Keras و لایه تعبیه از پیش آموزش دیده Wiki-PubMed-PMC[23] و همچنین استفاده از انواع مختلف لایه‌های شبکه عصبی (بلوک A)، مدل‌های مختلفی را روی سه مجموعه داده مورد آزمایش قرار دادیم. در جدول‌های نتایج عملکرد مدل‌ها، لایه تعبیه کراس با عدد ۱ و لایه تعبیه تعبیه از پیش آموزش دیده Wiki-PubMed-PMC با عدد ۲ کدگذاری شده است.

۴-۴-۱- نتایج مدل‌ها روی مجموعه داده N2C2

نتایج به‌کارگیری مدل‌ها روی مجموعه داده N2C2 در جدول‌های ۵ و ۶ آورده شده است. جدول ۵ نتایج روش‌های یادگیری عمیق بدون استفاده از ویژگی‌های لغوی شامل Dice, Jaccard, Cosine را نشان می‌دهد. همچنین جدول ۶ نتایج روش‌های ترکیبی شامل ویژگی‌های لغوی اشاره شده را نمایش می‌دهد. مطابق جدول ۵ از بین روش‌های استفاده شده، مدل SNN_Bi-LSTM_no_feat با لایه تعبیه Keras، بالاترین مقدار همبستگی (۰/۶۷۵۹) و پایین‌ترین خطا MSE (۱/۲۹۱۳) را در پیش‌بینی شباهت جملات به دست آورده است. همچنین مدل SNN_LSTM_no_feat با لایه تعبیه Keras، نتایج نزدیک‌تری نسبت به مدل برتر بدست آورده است. بررسی نتایج مدل‌ها در جدول ۵ بیان‌گر آن است که استفاده از مدل از پیش آموزش دیده در لایه تعبیه باعث بهبود در نتایج هیچ یک از مدل‌ها نشده است.

در جدول ۶ نتایج مربوط به مدل‌های ترکیبی نمایش داده شده است. همان‌طور که مشخص است مدل شبکه عصبی سیامی ترکیبی مبتنی بر شبکه پیچشی و ویژگی‌های لغوی به همراه لایه تعبیه کراس (SNN_CNN_feat) بالاترین مقدار همبستگی (۰/۷۳۲۴) و کمترین مقدار خطای MSE (۱/۲۳۱۰) را به‌دست آورده است. پس از این مدل، به ترتیب مدل‌های مبتنی بر شبکه حافظه کوتاه-مدت طولانی دوطرفه (SNN_Bi-LSTM_feat) و شبکه حافظه کوتاه-مدت طولانی (SNN_LSTM_feat) بهترین نتایج را به‌دست آورده‌اند. مقایسه نتایج جدول‌های ۵ و ۶ نشان می‌دهد که افزودن ویژگی‌های لغوی باعث بهبود نتایج هر سه مدل شده است و میزان بهبود برای مدل مبتنی بر شبکه عصبی پیچشی بیشتر از سایر مدل‌ها است. میزان بهبود برای این مدل بر مبنای معیار همبستگی برابر با ۰.۱۰۹۱ و همچنین میزان کاهش خطای این مدل در مقایسه با مدل غیر ترکیبی متناظر آن، برابر با ۰.۴۷۰۶است.

جدول ۵- نتایج روش‌ها بدون استفاده از ویژگی‌های لغوی بر اساس معیارهای همبستگی و MSE

روش	لایه تعبیه	همبستگی پیرسون	MSE
SNN_LSTM_no_feat	۱	۰/۶۷۰۲	۱/۲۹۹۶
	۲	۰/۶۲۹۴	۱/۵۱۲۹
SNN_ Bi-LSTM_no_feat	**۱**	**۰/۶۷۵۹**	**۱/۲۹۱۳**
	۲	۰/۶۰۲۰	۱/۵۱۶۶
SNN_ CNN_no_feat	۱	۰/۶۲۳۳	۱/۷۰۱۶
	۲	۰/۶۲۶۸	۱/۴۹۵۲

جدول ۶- نتایج روش‌ها با ترکیب ویژگی‌های لغوی بر اساس معیارهای همبستگی و MSE

روش	لایه تعبیه	همبستگی پیرسون	MSE
SNN_LSTM_feat	۱	۰/۶۸۴۱	۱/۳۳۵۸
	۲	۰/۶۳۳۵	۱/۳۸۲۲

SNN_ Bi-LSTM_feat	۱	۰/۶۹۸۷	۱/۳۳۳۳
	۲	۰/۶۵۳۵	۱/۴۲۰۸
SNN_ CNN_feat	**۱**	**۰/۷۳۲۴**	**۱/۲۳۱۰**
	۲	۰/۶۳۶۱	۱/۴۶۸۵

۴-۴-۲- نتایج مجموعه داده CDD-ful

در این بخش نتایج مربوط به مدل‌های استفاده شده روی مجموعه داده CDD-ful را توصیف می‌کنیم. جدول‌های ۷ و ۸ به ترتیب نتایج مدل‌ها را برای حالت پایه و ترکیبی نمایش می‌دهد. در معیار همبستگی پیرسون نتایج روش‌های پایه نشان از برتری روش شبکه پیچشی با مدل تعبیه کلمات PubMed-PMC دارد. همچنین در معیار MSE روش LSTM دوطرفه با لایه تعبیه کراس بهترین عملکرد را دارد.

جدول ۷- نتایج روش‌ها بدون استفاده از ویژگی‌های لغوی بر اساس معیارهای همبستگی و MSE - مجموعه داده CDD-ful

روش	لایه تعبیه	همبستگی پیرسون	MSE
SNN_LSTM_no_feat	۱	۰/۶۹۵۰	۱/۱۷۱۰
	۲	۰/۶۷۳۶	۱/۱۸۸۶
SNN_ Bi-LSTM_no_feat	۱	۰/۷۰۸۲	**۱/۰۶۶۴**
	۲	۰/۶۸۸۲	۱/۱۶۱۰
SNN_ CNN_no_feat	۱	۰/۷۱۵۹	۱/۲۷۳۳
	۲	**۰/۷۱۹۹**	۱/۲۴۵۲

نتایج مدل‌های ترکیبی توصیف شده در جدول ۸ نشان می‌دهد که در مدل‌های ترکیبی که در هر دو معیار همبستگی پیرسون و خطای MSE، روش شبکه پیچشی با لایه تعبیه Keras عملکرد بهتری نسبت به سایر مدل‌ها داشته است. همچنین روش LSTM دوطرفه عملکرد نزدیک‌تری نسبت به مدل برتر از خود نشان می‌دهد.

جدول ۸ - نتایج روش‌ها با ترکیب ویژگی‌های لغوی با لایه تعبیه کراس، بر اساس معیارهای همبستگی و MSE – مجموعه داده CDD-ful

روش	لایه تعبیه	همبستگی پیرسون	MSE
SNN_ LSTM_Feat	۱	۰/۶۹۸۷	۱/۰۹۴۸
	۲	۰/۶۹۹۱	۱/۱۴۰۷
SNN_Bi-LSTM_Feat	۱	۰/۷۳۰۲	۱/۰۰۰۴
	۲	۰/۷۱۴۲	۱/۰۹۴۴
SNN_CNN_Feat	**۱**	**۰/۷۳۴۲**	**۰/۹۸۴۵**
	۲	۰/۷۲۷۰	۱/۱۶۴۱

مطابق جدول ۸ از بین روش‌های استفاده شده، مدل SNN_CNN_Feat با لایه تعبیه کراس، بالاترین مقدار همبستگی (۰/۷۳۴۲) و پایین‌ترین خطا MSE (۰/۹۸۴۵) را در پیش‌بینی شباهت جملات به‌دست آورده است. پس از این مدل، مدل ترکیبی مبتنی بر شبکه حافظه کوتاه‌مدت طولانی دوطرفه SNN_Bi-LSTM_Feat با لایه تعبیه کراس بهترین نتیجه را به‌دست می‌آورد.

۴-۴-۳- نتایج مجموعه داده CDD-ref

برای این مجموعه داده نتایج مدل‌های پایه و ترکیبی به ترتیب در جدول‌های ۹ و ۱۰ نمایش داده شده است. در بین روش‌های پایه از نظر معیار همبستگی پیرسون، روش LSTM با مدل تعبیه کلمات Wiki- PubMed-PMC بهترین عملکرد را دارد. همچنین این روش در معیار MSE بهتر از سایر مدل‌ها عمل می‌کند.

جدول ۹- نتایج روش‌ها بدون استفاده از ویژگی‌های لغوی بر اساس معیارهای همبستگی و MSE - مجموعه داده CDD-ref

روش	لایه تعبیه	همبستگی پیرسون	MSE
SNN_LSTM_no_feat	۱	۰/۵۷۱۶	۱/۰۲۶۵
	۲	**۰۶۱۴۳**	**۰/۹۱۵۷**
SNN_ Bi-LSTM_no_feat	۱	۰/۵۷۰۴	۱/۰۰۹۹
	۲	۰/۵۶۲۹	۱/۰۶۹۶
SNN_ CNN_no_feat	۱	۰/۵۶۱۰	۰/۹۸۹۲
	۲	۰/۵۷۶۳	۱/۳۱۳۴

همچنین نتایج حاصل از مدل‌های ترکیبی در جدول ۱۰ حاکی از آن است که در هر دو معیار همبستگی پیرسون و خطای MSE، روش شبکه پیچشی با لایه تعبیه Keras بهترین عملکرد را کسب می‌کند.

جدول ۱۰ - نتایج روش‌ها با ترکیب ویژگی‌های لغوی با لایه تعبیه کراس، بر اساس معیارهای همبستگی و MSE – مجموعه داده CDD-ref

روش	لایه تعبیه	همبستگی پیرسون	MSE
SNN_ LSTM_Feat	۱	۰/۵۹۱۰	۰/۹۶۵۴
	۲	۰/۶۰۲۴	۰/۹۵۸۳
SNN_BiLSTM_Feat	۱	۰/۵۹۷۴	۰/۹۶۸۴
	۲	۰/۵۸۴۵	۱/۰۶۴
SNN_CNN_Feat	**۱**	**۰/۶۸۱۷**	**۰/۷۷۰۶**
	۲	۰/۵۸۹۴	۱/۲۲

۴-۵- مقایسه نتایج مدل‌ها با روش‌های ارائه شده در تحقیقات کنونی

برای بررسی بیشتر عملکرد مدل‌ها روی سه مجموعه داده مورد استفاده، در جدول‌های ۱۱ ، ۱۲ و۱۳ نتایج مدل برتر بدست آمده در این مطالعه را با نتایج ارائه شده در مقالات لی و همکاران [۱۵] و دی سوزو و همکاران [۱۴] بر اساس معیارهای همبستگی پیرسون و خطای MSE مقایسه می‌کنیم. با بررسی نتایج موجود در جدول ۱۱ به این نتیجه می‌رسیم که مدل بدست آمده در این مطالعه دارای عملکرد بهتری از نظر معیارهای همبستگی پیرسون و MSE می‌باشد. مدل بدست آمده از روش پیشنهادی در این تحقیق بر مبنای معیار همبستگی پیرسون با اختلاف ۰.۰۹۲۴ و ۰.۰۷۵۴ به ترتیب قوی‌تر از مدل پیشنهادی دی سوزو و همکاران [۱۴] و همچنین مدل ارائه شده توسط لی و همکاران [۱۵] است. همچنین با درنظر گرفتن معیار MSE، مدل بدست آمده در این تحقیق بهتر از روش‌های ارائه شده در لی و همکاران [۱۵] و دی سوزو و همکاران [۱۴] بوده و به ترتیب با اختلاف ۰.۳۵ و ۰/۲۸۲۰ نسبت به این مدل‌ها پیشی گرفته است.

همچنین مطابق جدول ۱۲، روی مجموعه داده CDD-ful، مدل برتر بدست آمده توسط روش پیشنهادی (SNN_CNN_feat) عملکرد بهتری نسبت به روش ارائه شده در لی و همکاران [۱۵] از خود نشان می‌دهد. به طور خاص در معیار همبستگی پیرسون با اختلاف ۰/۰۲۰۹ بهتر از روش لی و همکاران [۱۵] عمل می‌کند. همچنین در معیار MSE با اختلاف ۰/۰۷۲۵ نسبت به آن روش پیشی گرفته است.

علاوه بر این، بر اساس نتایج جدول ۱۳ در مجموعه داده CDD-ref نیز مدل حاصل از روش پیشنهادی این مطالعه، عملکرد بهتری نسبت به مدل پیشنهادی در لی و همکاران [۱۵] از خود نشان می‌دهد.

جدول ۱۱- مقایسه معیارهای همبستگی و MSE روش پیشنهادی با نتایج مقالات لی و همکاران و دی سوزو و همکاران – مجموعه داده N2C2

روش	همبستگی پیرسون	MSE
لی و همکاران [۱۵]	۰/۶۵۶	۱/۵۱۲
دی سوزو و همکاران [۱۴]	۰/۶۴	۱/۵۸
SNN_ CNN_feat	**۰/۷۳۲۴**	**۱/۲۳**

جدول ۱۲- مقایسه معیارهای همبستگی و MSE روش پیشنهادی با نتایج مقالات لی و همکاران– مجموعه داده CDD-ful

روش	همبستگی پیرسون	MSE
لی و همکاران [۱۵]	۰/۷۱۳	۱/۰۵۷
SNN_CNN_feat	**۰/۷۳۴۲**	**۰/۹۸۴۵**

جدول ۱۳- مقایسه معیارهای همبستگی و MSE روش پیشنهادی با نتایج مقالات لی و همکاران– مجموعه داده CDD-ref

روش	معیار همبستگی پیرسون	معیار MSE
لی و همکاران [۱۵]	۰/۶۶۴	۰/۸۰۳
SNN_CNN_feat	**۰/۶۸۱۷**	**۰/۷۷۰۶**

۴-۶- نمایش چند نمونه از نتایج پیش‌بینی شباهت با استفاده از مدل پیشنهادی

در جدول ۱۴ برای بررسی بیشتر دقت مدل‌ها، خروجی پیش‌بینی شباهت چند نمونه از جفت جملات نمایش داده شده است. همان‌طور که در این جدول قابل مشاهده است، مدل پیشنهادی در این مطالعه امتیاز شباهت نزدیکتری به امتیاز شباهت واقعی در مقایسه با روش پیشنهادی لی و همکاران بدست می‌آورد.

جدول ۱۴ - خروجی پیش‌بینی شباهت چند نمونه از جفت جملات

ردیف	جمله	امتیاز شباهت واقعی	امتیاز شباهت روش لی و همکاران [۱۵]	امتیاز شباهت به‌دست آمده با روش پیشنهادی
۱	S1: Moreover, UBA-like domain is required for binding ubiquitylated-protein substrates, UIM motif is responsible for the binding to cullin RING ligases (CRLs), and UBX domain is essential for p97 binding. S2: With up to 240 complexes in human cells, CRLs constitute the largest group of ubiquitin E3 ligases, accounting for >40% of all ubiquitin ligases and ~20% of protein degradation via the proteasome. For p97 this could mean an expansion in potential ubiquitylated substrates that require its function for their degradation.	۱	۳/۴	**۲/۶۶**
۲	S1: Periplasmic xylose-binding component of the ABC-type transport systems that belong to a family of pentose/hexose sugar-binding proteins of the type I periplasmic binding protein (PBP1) superfamily, which consists of two alpha/beta globular domains connected by a three-stranded hinge. S2: Parallel evolution of ligand specificity between LacI/GalR family repressors and periplasmic sugar-binding proteins.	۱/۵	۱/۵	۱/۵
۳	S1: Kalirin and Trio are encoded by separate genes in mammals and by a single one in invertebrates. S2: In mammals, Kalirin and Trio are encoded by separate genes, but invertebrates have a single homologous gene.	۵	۱/۶	**۲/۹۱**

۵- نتیجه‌گیری

اندازه‌گیری دقیق شباهت بین متون اهمیت زیادی در بسیاری از کاربردهای مرتبط با متن نظیر سیستم‌های پرسش‌وپاسخ، خلاصه‌سازی متون، بازیابی اطلاعات دارد.

در این تحقیق یک رویکرد ترکیبی مبتنی بر شبکه عصبی سیامی و ویژگی‌های شباهت لغوی ارائه شد. شبکه سیامی پیشنهادی شامل دو زیر شبکه یکسان است که اجزای اصلی هر کدام از آن‌ها به‌صورت کلی شامل یک لایه تعبیه کلمات، شبکه عصبی عمیق است. با در نظر گرفتن سه نوع شبکه عصبی عمیق شامل , LSTM

Bi-LSTM, CNN و همچنین دو نوع لایه تعبیه کلمات به همراه ویژگی‌های شباهت لغوی، چندین مدل پیاده‌سازی شد. نتایج آزمایشات بر سه مجموعه داده نشان داد که تمامی مدل‌های ترکیبی نسبت به مدل‌های پایه خود دارای مقدار همبستگی پیرسون بالاتر و خطای MSE کمتری هستند. همچنین مدل شبکه سیامی ترکیبی مبتنی بر شبکه عمیق پیچشی عملکرد بهتری نسبت به سایر مدل‌ها بر مبنای هر دو معیار مقدار همبستگی پیرسون و خطای MSE از خود نشان داد. مقایسه نتایج این تحقیق با نتایج مقالات دیگر نشان می‌دهد که مدل بدست آمده در این تحقیق دارای عملکرد بهتر از نظر معیارهای همبستگی پیرسون و MSE در تخمین شباهت جملات می‌باشد. استفاده از روش تعبیه کلمه BERT به همراه مکانیزم خودتوجه از کارهای آتی این پژوهش می‌باشد.

۶- تشکر و قدردانی

از حمایت مالی وزارت علوم، تحقیقات و فناوری از این تحقیق در قالب شماره پارسا ۰۰۰۳۱۹-۰۱-۹۹-۰۵ قدردانی می‌نماییم. همچنین از حمایت مالی پارک فناوری اطلاعات و ارتباطات از این تحقیق در قالب شماره پارسا ۰۰۰۳۱۹-۰۱-۹۹-۰۵ قدردانی می‌نماییم.

۷- مراجع

[1] S. Abujar, M. Hasan, and S. A. Hossain, "Sentence similarity estimation for text summarization using deep learning," Proc. The 2nd International Conference on Data Engineering and Communication Technology, 2019.

[2] M. Farouk, "Measuring text similarity based on structure and word embedding," *Cognit Syst Res,* vol. 63, pp. 1-10, 2020.

[3] Z. Li, H. Chen, and H. Chen, "Biomedical Text Similarity Evaluation Using Attention Mechanism and Siamese Neural Network," *IEEE Access*, vol. 9, pp. 105002-105011, 2021.

[4] J. Yang, Y. Li, C. Gao, and Y. Zhang, "Measuring the short text similarity based on semantic and syntactic information," *Future Generation Computer Systems*, vol. 114, pp. 169-180, 2021.

[5] A. Pawar and V. Mago, "Calculating the similarity between words and sentences using a lexical database and corpus statistics," arXiv preprint arXiv:1802.05667, 2018.

[6] W. H. Gomaa and A. A. Fahmy, "A survey of text similarity approaches," *International Journal of Computer Applications*, vol. 68, no. 13, pp. 13-18, 2013.

[7] M. Han, X. Zhang, X. Yuan, J. Jiang, W. Yun, and C. Gao, "A survey on the techniques, applications, and performance of short text semantic similarity," *Concurrency and Computation: Practice and Experience*, vol. 33, no. 5, p. e5971, 2021.

[8] D. W. Prakoso, A. Abdi, and C. Amrit, "Short text similarity measurement methods: a review," Soft Computing, pp. 1-25, 2021.

[9] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," arXiv preprint arXiv:1301.3781, 2013.

[10] T. Kenter and M. De Rijke, "Short text similarity with word embeddings," *Proc. 24th ACM international on conference on information and knowledge management*, 2015.

[11] T. Ranasinghe, C. Orăsan, and R. Mitkov, "Semantic textual similarity with siamese neural networks," *Proc. International Conference on Recent Advances in Natural Language Processing (RANLP 2019),* 2019.

[12] J. Mueller and A. Thyagarajan, "Siamese recurrent architectures for learning sentence similarity," *Proc. AAAI Conference on Artificial Intelligence*, 2016.

[13] P. Neculoiu, M. Versteegh, and M. Rotaru, "Learning text similarity with siamese recurrent networks," *Proc. 1st Workshop on Representation Learning for NLP*, 2016.

[14] J. V. A. de Souza, L. E. S. E. Oliveira, Y. B. Gumiel, D. R. Carvalho, and C. M. C. Moro, "Exploiting Siamese Neural Networks on Short Text Similarity Tasks for Multiple Domains and Languages," *Proc. Computational Processing of the Portuguese Language*, Cham, 2020.

[15] Z. Li, H. Lin, W. Zheng, M. M. Tadesse, Z. Yang, and J. Wang, "Interactive self-attentive siamese network for biomedical sentence similarity," *IEEE Access*, vol. 8, pp. 84093-84104, 2020.

[16] M. Farouk, "Sentence Semantic Similarity based on Word Embedding and WordNet," *Proc. 13th International Conference on Computer Engineering and Systems (ICCES),* 2018.

[17] R. Qu, Y. Fang, W. Bai, and Y. Jiang, "Computing semantic similarity based on novel models of semantic representation using Wikipedia," *Information Processing & Management*, vol. 54, no. 6, pp. 1002-1021, 2018.

[18] Y. M'rabet, H. Kilicoglu, and D. Demner-Fushman, "TextFlow: A text similarity measure based on continuous sequences," *Proc. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers), 2017.

[19] Y. Liu, C.-J. Sun, L. Lin, X. Wang, and Y. Zhao, "Computing semantic text similarity using rich features," *Proc. Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation*, 2015.

[20] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput*, vol. 9, no. 8, pp. 1735-1780, 1997.

[21] S. Cornegruta, R. Bakewell, S. Withey, and G. Montana, "Modelling radiological language with bidirectional long short-term memory networks," arXiv preprint arXiv:1609.08409, 2016.

[22] R. Islamaj, W. J. Wilbur, N. Xie, N. R. Gonzales, N. Thanki, R. Yamashita, et al., "PubMed Text Similarity Model and its application to curation efforts in the Conserved Domain Database," *Database*, vol. 2019, 2019.

فریبا خلج تحصیلات خود را در مقطع کارشناسی، رشته مهندسی نرم افزار از دانشگاه زنجان به اتمام رساند. همچنین مدرک کارشناسی ارشد رشته‌ی مهندسی کامپیوتر گرایش نرم افزار در دانشگاه شهید مدنی آذربایجان را در سال ۱۴۰۰دریافت نمود. وی در دوره‌ی ارشد به عنوان شاگرد ممتاز انتخاب شده است. زمینه مورد علاقه تحقیقات وی، حوزه‌ی یادگیری عمیق و پردازش زبان طبیعی می‌باشد.

آدرس پست الکترونیکی ایشان عبارت است از:

fariba.71.khalaj@gmail.com

حسین عباسی مهر مدرک کارشناسی خود را در رشته مهندسی فناوری اطلاعات از دانشگاه شهید مدنی آذربایجان در سال ۱۳۸۸ اخذ نمود. همچنین وی مدارک کارشناسی ارشد و دکتری خود را در رشته مهندسی فناوری اطلاعات به ترتیب در سال‌های ۱۳۹۰ و ۱۳۹۵ از دانشگاه صنعتی خواجه نصیرالدین طوسی دریافت نمود. هم اکنون وی استادیار دانشکده فناوری اطلاعات و مهندسی کامپیوتر دانشگاه شهید مدنی

آذربایجان است و زمینه‌های تحقیقاتی مورد علاقه ایشان داده‌کاوی، متن‌کاوی، و تحلیل سری‌های زمانی است.
آدرس پست الکترونیکی ایشان عبارت است از:
abbasimehr@azaruniv.ac.ir

[1] Lexical
[2] Corpus-based
[3] Knowledge-based
[4] Word embedding
[5] Recurrent neural networks
[6] Long Short-term Memory (LSTM)
[7] Siamese Neural Network
[8] Convolutional Neural Network (CNN)
[9] Bi-LSTM
[10] https://n2c2.dbmi.hms.harvard.edu/track1
[11] Pearson correlation
[12] Mean Squared Error
[13] Cosine
[14] Jaccard
[15] Dice
[16] Dense
[17] Dropout
[18] https://n2c2.dbmi.hms.harvard.edu
[19] Rectified Linear Unit
[20] Adaptive Moment Estimation (Adam)
[21] Loss Function
[22] Epoch
[23] http://evexdb.org/pmresources/

The CSI Journal on Computing Science and Information Technology
Vol. 20, No.1, 2022

Text similarity prediction with Siamese of deep neural network and lexical similarity features

Fariba Khalaj[1], Hossein Abbasimehr[2]

[1,2] Faculty of Information Technology and Computer Engineering, Azarbaijan Shahid Madani University, Tabriz, Iran

Abstract

Measuring text similarity is crucial for many natural language processing applications such as Information retrieval, document clustering, question and answer systems. The aim of this study is to provide an approach to improve the accuracy of deep learning models in measuring text similarity. For this purpose, a hybrid approach based on Siamese neural network and lexical similarity features is developed. The proposed Siamese network consists of two identical sub-networks, the main components of each of which generally include a word embedding layer and a deep neural network for semantic representation. By considering three types of deep neural networks including convolutional neural network, short long-term memory network, and the bi-directional short long-term memory network, as well as two types of word embedding with lexical similarity features, different variants of models, are implemented. The experimental results on three public datasets show that the combined Siamese neural network model based on convolutional network and lexical features achieves the highest Pearson correlation value and the lowest MSE error value among the models. Also, in terms of both Pearson correlation and MSE measures, our model outperforms the existing methods.

Keywords: Text Similarity; deep learning; Siamese neural network; text similarity measures