دانشکده مهندسي فناوري اطلاعات و کامپيوتر

رشته مهندسي کامپيوتر، گرايش هوش مصنوعي و رباتيکز

پايان‌نامه کارشناسی ارشد

تخمین شباهت معنایی متن با استفاده از روش‌های آماری

نگارنده: فرشته رياحی

استاد راهنما

دکتر مرتضی زاهدی

استاد مشاور

دکتر علیرضا تجری

بهمن ۱۳۹۹

مدیریت تحصیلات تکمیلی

باسمه‌تعالی

شماره:
تاریخ:

فرم شماره (۳) صورتجلسه نهایی دفاع از پایان نامه دوره کارشناسی ارشد

با نام و یاد خداوند متعال، ارزیابی جلسه دفاع از پایان نامه کارشناسی ارشد خانم: **فرشته ریاحی** با شماره دانشجویی **۹۶۰۷۷۲۴** رشته **مهندسی کامپیوتر** گرایش **هوش مصنوعی و رباتیکز** تحت عنوان **تخمین شباهت معنایی متن با استفاده از روش‌های آماری** که در تاریخ **۱۳۹۹/۱۱/۲۹** با حضور هیأت محترم داوران در دانشگاه صنعتی شاهرود برگزار گردید به شرح ذیل اعلام می گردد:

الف) درجه عالی: نمره ۲۰-۱۹ ☐ ب) درجه خیلی خوب: نمره ۱۸/۹۹ – ۱۸ ☑

ج) درجه خوب: نمره ۱۷/۹۹– ۱۶ ☐ د) درجه متوسط: نمره ۱۴-۱۵/۹۹ ☐

ه) کمتر از ۱۴ غیر قابل قبول و نیاز به دفاع مجدد دارد ☐

نوع تحقیق: نظری ☐ عملی ☐

عضو هیأت داوران	نام ونام خانوادگی	مرتبهٔ علمی	امضاء
۱- استادراهنمای اول	دکتر مرتضی زاهدی	استادیار	
۲- استادراهنمای دوم	---------	-------	-------
۳- استاد مشاور	دکتر علیرضا تجری	استادیار	
۴- نماینده تحصیلات تکمیلی	مهندس محسن فرهادی	مربی	
۵- استاد ممتحن اول	دکترمنصور فاتح	استادیار	
۶- استاد ممتحن دوم	دکتر محسن رضوانی	استادیار	

نام و نام خانوادگی رئیس دانشکده: **دکتر علیرضا الفی** تاریخ و امضاء و مهر دانشکده:

تقدیم

به پدرم که زندگی را از دستانش،
مادرم که محبت را از نگاهش آموختم
و خانواده‌ام به پاس ایثار و فداکاری‌هایشان.

تشکر و قدردانی

سپاس خدایی را که به من این توفیق را عطا فرمود تا بتوانم قدمی در راه علم و دانش بردارم.

بدین وسیله از زحمات و تلاش بی‌دریغ استاد محترم جناب آقای دکتر مرتضی زاهدی و جناب آقای دکتر علیرضا تجری و خانواده عزیزم صمیمانه سپاسگزاری می‌نمایم و همچنین از دوست عزیزم خانم الهام احمدی و دیگر دوستانی که هر کدام به نحوی در تهیه این مجموعه با اینجانب همکاری داشته‌اند تشکر نموده و موفقیت همه آن‌ها را از خداوند متعال خواهانم.

فرشته ریاحی
بهمن ۱۳۹۹

تعهد نامه

اینجانب **فرشته ریاحی** دانشجوی کارشناسی ارشد رشته **مهندسي كامپيوتر مهندسي فناوري اطلاعات و كامپيوتر** دانشگاه شاهرود، نویسنده پایان‌نامه با عنوان **تخمین شباهت معنایی متن با استفاده از روش‌های آماری** ، تحت راهنمایی **مرتضی زاهدی** متعهد می‌شوم:

- تحقیقات در این پایان‌نامه توسط اینجانب انجام شده است و از صحت و اصالت برخوردار است.
- در استفاده از نتایج پژوهش‌های دیگر پژوهش‌گران، به مرجع مورد استفاده استناد شده است.
- مطالب این پایان‌نامه، تا کنون توسط خود، یا فرد دیگری برای دریافت هیچ نوع مدرک یا امتیازی در هیچ‌جا ارایه نشده است.
- حقوق معنوی این اثر، به دانشگاه صنعتی شاهرود تعلق دارد، و مقالات مستخرج با نام "دانشگاه صنعتی شاهرود" یا "Shahrood University of Technology" به چاپ خواهد رسید.
- حقوق معنوی تمام افرادی که در به‌دست آوردن نتایج اصلی پایان‌نامه تاثیرگذار بوده‌اند، در مقالات مستخرج از پایان‌نامه رعایت می‌گردد.
- در تمام مراحل انجام این پایان‌نامه، در مواردی که از موجود زنده (یا بافت‌های آنها) استفاده شده است، ضوابط و اصول اخلاقی رعایت شده است.
- در تمام مراحل انجام این پایان‌نامه، در مواردی که به حوزه اطلاعات شخصی افراد دسترسی یافته (یا استفاده شده است)، اصل رازداری و اصول اخلاق انسانی رعایت شده است.

فرشته ریاحی
بهمن ۱۳۹۹

مالکیت نتایج و حق نشر

- تمام حقوق معنوی این اثر و محصولات آن (مقالات مستخرج، کتاب، برنامه‌های رایانه‌ای، نرم‌افزارها و تجهیزات ساخته شده) متعلق به دانشگاه صنعتی شاهرود می‌باشد. این مطلب باید به نحو مقتضی، در تولیدات علمی مربوطه ذکر شود.
- استفاده از اطلاعات و نتایج موجود در این پایان‌نامه بدون ذکر منبع مجاز نمی‌باشد.

چکیده

در دنیای امروز به دلیل افزایش حجم اطلاعات اینترنتی و فراوانی فرهنگ‌های مختلف دیجیتالی، وجود تشابه در داده‌های متنی به دلایل مختلفی افزایش پیدا کرده است، به همین دلیل تخمین میزان شباهت میان آن‌ها امری ضروری به حساب می‌آید. همچنین یافتن شباهت میان داده‌های متنی در زمینه‌هایی همچون سیستم‌های بازیابی اطلاعات، سرقت ادبی، داده‌کاوی، طبقه‌بندی اسناد و موارد بسیار دیگری مورد استفاده قرار می‌گیرد.

با ورود جمله یا سند یا متنی به سیستم، میزان شباهت آن با اسناد موجود مورد بررسی قرار می‌گیرد و در زمینه موردنظر استفاده می‌شود. تاکنون روش‌های مختلف پردازش زبان طبیعی و یادگیری ماشین برای محاسبه میزان شباهت داده‌های متنی معرفی شده‌اند. این روش‌ها به میزان دقت متفاوتی دست یافته‌اند، برای بهبود این معیار نیاز است تا پژوهش بهتری صورت گیرد. از جمله روش‌های آماری که جز روش‌های مبتنی بر پیکره می‌باشد و از ترکیب این روش با روش‌های دیگر می توان به نتایج جالب دست یافت.

در این پژوهش، قبل از تبدیل متن به فرمت قابل خواندن برای ماشین، پیش‌پردازش روی داده‌ها اعمال می‌شود تا ویژگی‌ها کاهش یابد و دقت اندازه‌گیری را بالا ببرد. پس از آن با استفاده از تخصیص پنهان دیریکلت هر کلمه به یک نمره همبستگی اختصاص می‌یابد و این روند براساس احتمالات تکرار می‌شود تا اختصاص نمره همبستگی به کلمات بهبود یابد و جمله در دسته مورد نظر قرار گیرد. برای بازنمایی جملات نیز از doc2bow استفاده شده است. سپس برای محاسبه شبیه‌ترین جملات از فاصله جنسن شانون که با مقایسه واگرایی توزیع برچسب مورد نظر بدست می‌آید استفاده شده است. در انتها ماشین بردار پشتیبان خطی و تابع پایه شعاعی را برای دسته‌بندی جملات مشابه به کار برده است. روش پیشنهادی به دقت دسته‌بندی ۸۹ درصد و نمره همبستگی ۰,۹۲ درصد روی پایگاه داده SICK دست پیدا کرده است و توانسته است ۳,۸ درصد دقت بهتری نسبت به پژوهش MaLSTM داشته باشد.

کلمات کلیدی: شباهت معنایی متن، تخصیص پنهان دیریکلت، سند به کیسه کلمات، فاصله جنسن شانون، ماشین بردار پشتیبان

فهرست مطالب

فهرست تصاویر ق

فهرست جداول ش

۱ مقدمه ۱

۱.۱ مقدمه ۲

۲.۱ تعریف مسئله ۲

۳.۱ کاربردهای تخمین شباهت معنایی متن ۵

۴.۱ لزوم استفاده از تخمین شباهت معنایی متون ۷

۵.۱ اهداف تحقیق ۸

۶.۱ ساختار تحقیق ۸

۲ پیشینه تحقیق ۹

۱.۲ مقدمه ۱۰

۲.۲ تاریخچه شباهت معنایی متن ۱۰

۳.۲ پردازش زبان طبیعی ۱۱

۴.۲ ویژگی‌های مهم و مبانی نظری در شباهت معنایی ۱۲

۱.۴.۲ شبکه واژگان ۱۲

۲.۴.۲ مفهوم اطلاعات ۱۳

۳.۴.۲ ویکی‌پدیا ۱۳

۱.۳.۴.۲ تحلیل معنایی صریح مبتنی بر ویکی پدیا ۱۴

۴.۴.۲ مدلN_ گرم ۱۵

۵.۴.۲ طولانی ترین پیامد مشترک ۱۵

۶.۴.۲ فرکانس مدت ـ فرکانس اسناد معکوس ۱۵

۷.۴.۲ چارچوب کلمه به بردار ۱۶

۸.۴.۲ مدل کیسه ای از کلمات ۱۷

۹.۴.۲ تخصیص پنهان دیریکلت ۱۸
۱۰.۴.۲ فاصله جنسن شانون ۱۸
۱۱.۴.۲ ماشین بردار پشتیبان ۱۹
۱۲.۴.۲ تجزیه و تحلیل آماری ۲۰
۱۳.۴.۲ حافظه کوتاه مدت طولانی ۲۰
۵.۲ پیشینه تحقیق ۲۱
۱.۵.۲ روش‌های مبتنی بر پیکره ۲۱
۱.۱.۵.۲ روش‌هاي آماري ۲۱
۲.۱.۵.۲ روش‌های مبتنی بر یادگیری عمیق ۲۳
۲.۵.۲ روش‌های مبتنی بر معنا ۲۶
۳.۵.۲ روش‌های مبتنی بر رشته ۲۹
۴.۵.۲ روش‌های ترکیبی ۳۰
۵.۵.۲ سایر روش‌ها ۳۴
۶.۲ جمع‌بندی ۳۵

۳ روش تحقیق **۳۹**
۱.۳ مقدمه ۴۰
۲.۳ طرح مسئله ۴۰
۳.۳ روش پیشنهادی ۴۰
۴.۳ ارائه‌ی مدل کلی ۴۱
۱.۴.۳ پیش پردازش متن ۴۱
۲.۴.۳ استخراج ویژگی از متن ۴۵
۳.۴.۳ جزئیات روش جنسن شانون و دسته‌بندی ۴۷
۵.۳ جمع‌بندی ۴۹

۴ نتایج و تحلیل آن‌ها **۵۱**
۱.۴ مقدمه ۵۲
۲.۴ مجموعه دادگان مورد استفاده ۵۲
۳.۴ پیاده‌سازی ۵۳
۱.۳.۴ طبقه‌بندی نمرات شباهت به‌دست‌آمده ۵۵
۴.۴ آزمون و پارامترهای مورد نظر ۵۵
۵.۴ معیار ارزیابی ۵۷
۶.۴ نتایج ۵۹
۷.۴ جمع‌بندی ۶۲

۵ نتیجه‌گیری **۶۵**
۱.۵ نتیجه‌گیری ۶۶
۲.۵ کارهای آینده ۶۷

مراجع **۶۹**

فهرست تصاویر

۱.۱ روش‌های مختلف برای محاسبه تشابه کلمه به کلمه [۱] ۴

۱.۲ برآورد کارآمد از بازنمایی کلمات در فضای بردار ۱۷
۲.۲ نمای کلی مدل ۲۴

۱.۳ ساختار کلی روش پیشنهادی ۴۲
۲.۳ مراحل پیش‌پردازش متن ۴۵
۳.۳ مدل تخصیص پنهان دیریکلت [۷۴] ۴۷
۴.۳ مرز تصمیم برای بردار پشتیبان خطی[۷۵] ۵۰
۵.۳ مرز تصمیم برای داده‌های غیرخطی[۷۵] ۵۰

۱.۴ نمودار داده‌های پیش پردازش شده ۵۴
۲.۴ نمودارد‌سته‌بندی داده‌های آموزش ۶۰
۳.۴ نمودار دقت روش‌های بیان شده ۶۳

فهرست جداول

۱.۲ مقایسه برخی کارهای صورت گرفته در زمینه شباهت معنایی متن ۳۷

۱.۴ طبقه‌بندی مجموعه دادگان مورد استفاده براساس امتیازات ۵۳
۲.۴ طبقه‌بندی مجموعه دادگان مورد استفاده براساس برچسب گذاری موجود . . . ۵۳
۳.۴ نمونه‌ای از مجموعه داده‌ها همراه با ضریب همبستگی ۵۳
۴.۴ جملات مورد مقایسه ۴۱۴ و ۷۳۰ مجموعه دادگان ۵۶
۵.۴ روش‌های مقایسه شده با روش پژوهش ۵۷
۶.۴ نتایج طبقه‌بندی مجموعه دادگان آموزش براساس امتیازات ۵۹
۷.۴ ماشین بردار پشتیبان خطی ۶۰
۸.۴ ماشین بردار پشتیبان کرنل چند جمله‌ای ۶۱
۹.۴ ماشین بردار پشتیبان کرنل تابع پایه شعاعی ۶۱
۱۰.۴ مقایسه دقت بدست آمده روش‌های دسته‌بندی ۶۲
۱۱.۴ مقایسه روش ارائه شده با سایر روش‌ها ۶۲

فصل ۱

مقدمه

۱.۱ مقدمه

با رشد روزافزون اینترنت و افزایش قابلیت استفاده از آن، توانایی تعیین تشابه معنایی بین متون برای بسیاری از برنامه‌های کاربردی موردنیاز است. روزانه حجم عظیمی از داده‌های زبان طبیعی در اینترنت اضافه می‌شود. علاوه بر این، ادبیات انسانی در فرهنگ‌های مختلف دیجیتالی شده و در کتابخانه‌های دیجیتالی در دسترس قرار می‌گیرد. در میان این حجم انبوه از داده‌ها، داده‌های مشابه دیگری نیز به دلایل مختلفی مانند متفاوت بودن منابع خبری، زمان و مکان ایجاد گردیده است. مقدار بسیار زیادی از این داده‌ها به زبان طبیعی قالب‌بندی شده‌اند. این امر باعث می‌شود فن‌های پردازش زبان طبیعی[1] برای استفاده از این مقدار داده بسیار مهم باشند [۱]. بنابراین تخمین شباهت بین داده‌ها یک مسئله کاربردی و مهم می‌باشد. علاوه بر این، صحت سنجش تشابه داده‌ها نیز مهم است. ارزیابی دو بخش از متن توسط معیار ارزیابی مرتبی انجام می‌شود که اختلاف معنایی کامل تا همبستگی کامل معنایی را تشخیص می‌دهد [۲]. به‌عنوان حامل اصلی اطلاعات، متون نقش مهمی در عصر داده‌های بزرگ دارند. تعداد زیادی از داده‌های مبهم مانند واژه‌هایی که چندمعنا دارند، در متون وب چالش‌های بیشتری را برای درک خودکار متن به وجود می‌آورند. برای پرداختن به این مشکل، اندازه‌گیری تشابه معنایی پیشنهادشده است تا ارتباط معنایی بین کلمات را ازنظر شباهت معنایی و واژگانی به‌جای تشابه سطحی کلمات پیدا کند[۳].

۲.۱ تعریف مسئله

در پردازش زبان طبیعی، شباهت معنایی[2] نقش مهمی دارد و یکی از وظایف اساسی برای بسیاری از برنامه‌های پردازش زبان طبیعی و مناطق مرتبط با آن است [۴]. شباهت معنایی متن[3] درصدد تخمین میزان هم‌ارزی معنایی بین دو قطعه متن است. هدف از یافتن شباهت بین متون، محاسبه میزان نزدیک بودن دو جمله است؛ هم ازنظر معنایی و هم واژگانی. شباهت معنایی متن می‌خواهد درجه همگرایی بین دو جمله را ارزیابی کند [۵]. با محاسبه شباهت یک جفت جمله، نمره سطح همبستگی بین آن دو جمله به دست می‌آید [۶]. هدف این است که شباهت معنایی بین دو جمله داده‌شده در یک مقیاس از ۰ تا ۵ را اندازه‌گیری کنیم؛ ۰ نشان می‌دهد که معنایی جملات کاملاً مستقل و ۵ به این معنی است که یک معادل معنایی کامل وجود دارد [۷]. اندازه‌گیری شباهت جملات به سه دسته تقسیم‌شده‌اند: روش‌های مبتنی بر آمار، روش‌های مبتنی بر معنا و روش‌های ترکیبی. روش‌های آماری با استفاده از محاسبه کلمات مشترک در دنباله رشته‌ها شباهت جملات را

[1] Natural language processing(NLP)
[2] semantic similarity(SS)
[3] semantic textual similarity(STS)

محاسبه می‌کنند [۸]. روش‌های مبتنی بر معنا، معنای واقعی پشت کلمات یا کل عبارت را در نظر می‌گیرد و روش‌های ترکیبی از ترکیب این دو روش استفاده می‌کند.

شباهت معنایی متن و استنباط زبان طبیعی دو وظیفه محبوب زبان طبیعی است که برای محک زدن مدل‌های بیان جمله درجایی که دو جمله باهم جفت شده‌اند، استفاده می‌شود[۹]. جملات می‌تواند ازنظر لغوی یا معنایی مشابه باشد. تشابه واژگانی به معنای شباهت مبتنی بر رشته است. بااین‌حال، شباهت معنایی نشانگر معنای مشابه در جمله است. شباهت یابی مبتنی بر معنا که به معنی جملات بستگی دارد، رویکردهای مختلفی دارد. این روش‌ها فن‌های مختلفی را برای مقایسه دو جمله ازنظر معنایی به کار می‌گیرند. رویکرد اول، مبتنی بر مجموعه نوشته‌ها (پیکره)[۱] شباهت کلمات را بر اساس تجزیه‌وتحلیل آماری متون بزرگ می‌یابد. علاوه بر این، برای تجزیه‌وتحلیل یک گروه بزرگ برای نشان دادن معانی کلمات از یادگیری عمیق می‌تواند استفاده کند. رویکرد دوم، مبتنی بر دانش، به شبکه معنایی برای کلمات بستگی دارد. معنی کلمات و روابط بین کلمات در این شبکه معنایی گنجانده‌شده است. بیشترین کاربرد معنای خالص شبکه واژگان[۲] است. رویکرد سوم، مبتنی بر ساختار، از اطلاعات ساختار یک جمله برای به دست آوردن معنی این جمله استفاده می‌کند. جمله مشابه باید دارای ساختار اساسی مشابه باشد [۱].

در پردازش زبان طبیعی، شباهت معنایی نقش مهمی دارد و یکی از وظایف اساسی برای بسیاری از برنامه‌های پردازش زبان طبیعی و مناطق مرتبط با آن است. شباهت متنی معنایی را می‌توان با استفاده از متریک بیش از مجموعه‌ای از اسناد تعریف کرد و ایده این است که شباهت معنایی بین آن‌ها را پیدا کنید. شباهت بین اسناد مبتنی بر روابط مستقیم و غیرمستقیم بین آن‌هاست . این روابط را می‌توان باوجود روابط معنایی در بین آن‌ها سنجید و شناخت[۴].

شباهت معنایی یک مورد خاص از ارتباط است که به شباهت بین مفاهیم مربوط می‌شود. علاوه بر این، روش‌های اندازه‌گیری شباهت معنایی دو مفهوم را به‌عنوان ورودی در نظر می‌گیرند و یک نمره عددی را برمی‌گردانند که میزان آن‌ها مشابه است. در حقیقت، چندین اثر در مورد اقدامات شباهت معنایی در سال‌های گذشته ارائه‌شده است. چهار خانواده مختلف به شرح زیر ذکرشده است:

۱. اقدامات مبتنی بر مسیر[۱۰]؛[۱۱]؛[۱۲] شباهت‌ها را بافاصله معنایی مفاهیم در یک هستی‌شناسی معین ارزیابی می‌کنند. این نوع اقدامات در محاسبات ساده در بسیاری از مناطق مورداستفاده گسترده قرارگرفته است.

۲. اقدامات مبتنی بر محتوای اطلاعات [۱۳]؛[۱۴]؛[۱۵]؛[۱۶]؛[۱۷] شباهت‌های مفهوم اطلاعات[۳] را در یک هستی‌شناسی مشخص می‌کنند. مفهوم اطلاعات بر پایه این فرض استوار است که موجودات خاص و ویژه، مفهوم اطلاعات بیشتری نسبت به موارد کلی و انتزاعی ارائه می‌دهند.

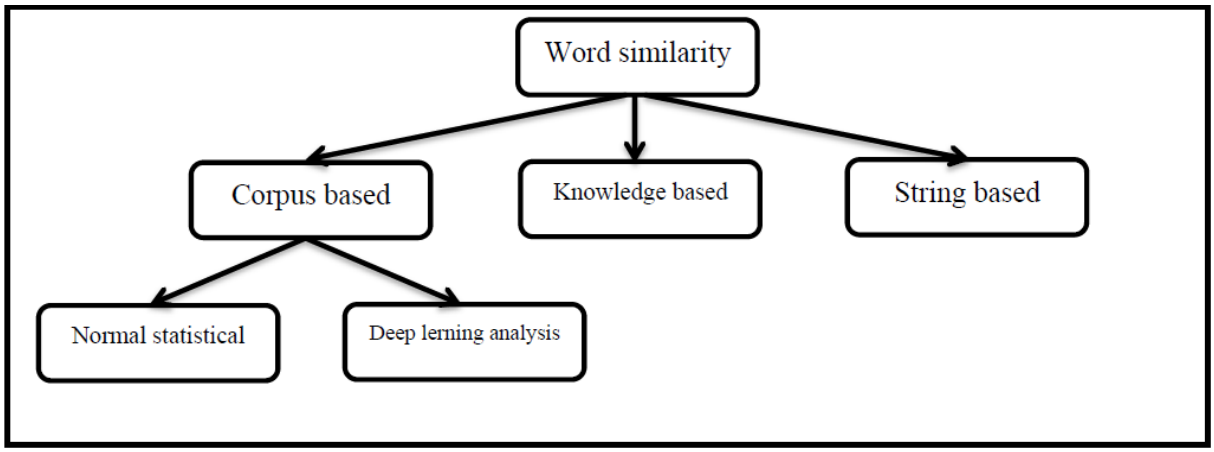
۳. اقدامات مبتنی بر ویژگی [۱۸]؛[۱۹]؛[۲۰]؛[۲۱]؛[۲۲] شباهت‌ها را با استفاده از مجموع وزنی مشترک و ویژگی‌های غیرمعمول مفاهیم ارزیابی می‌کنند.

---

[1] corpuse
[2] wordnet
[3] Information Content(IC)

۴. اقدامات ترکیبی حداقل دو تا از اقدامات فوق را برای ادغام مزایا ترکیب می‌کنند که منجر به وابستگی به پارامترها می‌شود [۲۳].

سه روش اصلی برای اندازه‌گیری شباهت بین کلمات وجود دارد ۱. تشابه کلمات مبتنی بر مجموعه نوشته‌ها ۲. دانش مبتنی بر کلمات ۳. شباهت کلمات مبتنی بر رشته که تشابه مبتنی بر پیکره خود دو روش مختلف دارد اول استفاده از آنالیز آماری عادی و دوم استفاده از یادگیری عمیق می‌باشد. پیدا کردن شباهت بین کلمات، هسته تشابه جمله است. در ادبیات معیارهای زیادی برای محاسبه شباهت کلمه به کلمه وجود دارد. سلسله‌مراتب برای روش‌های مورداستفاده برای اندازه‌گیری شباهت کلمه در شکل ۱.۱ نشان داده‌شده است[۱].



شکل ۱.۱: روش‌های مختلف برای محاسبه تشابه کلمه به کلمه [۱]

با توجه به منابع لغوی مورداستفاده، رویکردهای موجود برای محاسبه شباهت معنایی را می‌توان به رویکردهای مبتنی بر پیکره و رویکردهای دانش‌محور طبقه‌بندی کرد. رویکردهای مبتنی بر پیکره معنای واژگانی را با استخراج بردارهای کلمه‌ای از خصوصیات توزیع کلمات متنی آن‌ها در یک کالبد، مانند پیکره ملی انگلیس، ویکی‌پدیا و موتور جستجوی وب یاد می‌گیرند. با توجه به داده‌های متن گسترده، چنین رویکردهایی برای تصفیه اطلاعات معنایی واژگان پنهان در یک فضای بردار، مانند تحلیل معنایی صریح [1] (گابیلوویچ و مارکوویچ ۲۰۰۷) به فن‌های کاهش بعد تکیه می‌کنند. به‌عنوان یک جایگزین، لو و همکاران(۲۰۱۵ ، ۲۰۱۶) مدل شبکه عصبی بسته‌های کیسه کلمات[2] را آموزش می‌دهد تا بردارهای کلمه‌ای کم بعدی برای اندازه‌گیری شباهت معنایی واژگانی یاد بگیرد. بااین‌حال، رویکردهای مبتنی بر پیکره معمولاً از پیچیدگی محاسباتی بالایی برای شرکت‌های پیش‌پردازش و کاهش ابعاد بردارهای کلمه برخوردار هستند. علاوه بر این، رویکردهای مبتنی بر بدنه برای محاسبه ارتباط معنایی نسبت به شباهت معنایی کاربردی‌تر است. درحالی‌که، رویکردهای دانش‌بنیان که

[1] Explicit Semantic Analysis (ESA)
[2] Continuous Bag-of-Words (CBOW)

عمدتاً به منابع دانش ساختاریافته بستگی دارند، می‌توانند با بهره‌گیری از ساختار ذاتی پایه‌های دانش سلسله مراتبی، مانند هستی‌شناسی، اصطلاحات، فرهنگ لغت، طبقه‌بندی، بر پیچیدگی محاسباتی غلبه کنند. بیشتر رویکردهای هنری مبتنی بر یک پایگاه داده شبکه واژگان برای اندازه‌گیری شباهت معنایی مفهوم ارائه‌شده است(لی و همکاران ۲۰۰۳ ؛ سانچز و همکاران ۲۰۱۱ ؛ لیو و همکاران ۲۰۱۲ ؛ منگ و همکاران ۲۰۱۳ ؛ گائو و همکاران ۲۰۱۵). در مطالعه لو و همکاران (۲۰۱۶)، شبکه واژگان به‌عنوان کمکی برای تقویت معنایی بردارهای کلمه مورداستفاده قرار می‌گیرد، زیرا ثابت‌شده است که در مطالعات مبتنی بر دانش قبلی مؤثر است [۳].

سنجش شباهت معنایی متن مربوط به تعیین میزان تشابه معنایی بین جفت متون است. چنین اقدامات شباهت معنایی متن، توجه زیادی به تحقیقات در پردازش زبان طبیعی، درک زبان طبیعی و جوامع بازیابی اطلاعات جلب کرده است. اولین روش‌ها برای اندازه‌گیری شباهت بین متون طولانی، به‌عنوان‌مثال، برای طبقه‌بندی اسناد، یا بین متن کوتاه و متن طولانی، به‌عنوان‌مثال، برای پرس‌وجو یک سند در یک سیستم جستجو ارائه‌شده است. راحت‌ترین راه برای نشان دادن یک سند توسط آن روش‌های کیف کلمات یک مدل فضای برداری است که در آن هر سند به‌عنوان کیف کلمات در نظر گرفته می‌شود و توسط یک بردار ارائه می‌شود. تشابه کسینوسی، به‌عنوان یک اقدام سنتی بردار که در بازیابی اطلاعات مورداستفاده قرار می‌گیرد، شناخته می‌شود، یک اقدام محبوب برای ارزیابی شباهت بین متون است[۲۴].

۳.۱ کاربردهای تخمین شباهت معنایی متن

شباهت معنایی یک فن گسترده است که در بسیاری از زمینه‌ها ازجمله بازیابی اطلاعات، کشف دانش، هوش مصنوعی، علوم شناختی، پاسخ دادن به پرسشنامه سامانه‌ای، سامانه‌های پیشنهاددهنده، ارزیابی عملکرد، داده‌کاوی، مدیریت دانش، پردازش زبان طبیعی، زبانشناسی محاسباتی و استخراج اطلاعات کاربرد دارد. علاوه بر این‌ها شباهت معنایی متن قادر به ارائه یک چارچوب یکپارچه برای ارزیابی مؤلفه‌های معنایی بیرونی چندگانه ازجمله ابهام در معانی جملات برچسب‌گذاری معنایی متن و کیفیت نام‌گذاری برای مراجع است. بازیابی اطلاعات، اطلاعات و داده‌های ازدست‌رفته را بازیابی می‌نماید. داده‌کاوی فرآیند کشف الگوهای موجود در مجموعه‌ی بزرگی از داده‌ها است.

فرآیند کشف دانش از پایگاه داده شامل مراحل زیر می‌باشد:

۱. انتخاب ۲. پیش‌پردازش ۳. انتقال و تبدیل ۴. داده‌کاوی ۵. تفسیر و ارزیابی

می‌توان این فرآیند را به‌صورت زیر نیز بیان نمود:

۱. شناخت کسب‌وکار ۲. شناخت داده‌ها ۳. فراهم آوردن داده‌ها ۴. مدل‌سازی ۵. ارزیابی ۶. گسترش.

در حالت کلی این گام‌ها به‌صورت زیر قابل خلاصه می‌باشند:

۱. پیش‌پردازش ۲. داده‌کاوی ۳. اعتبار سنجی نتایج

اقدامات شباهت احکام نقش مهمی در تحقیقات و برنامه‌های مربوط به متن در زمینه‌هایی مانند متن‌کاوی، بازیابی صفحه وب و سیستم گفتگو ایفا می‌کند[۴].

در بازیابی صفحه وب، شباهت جملات به‌عنوان یکی از بهترین فن‌های بهبود اثربخشی بازیابی، اثبات‌شده است که در آن از عناوین برای ارائه اسناد در صفحه پیدا کردن صفحه استفاده می‌شود. در بازیابی تصویر از وب، استفاده از متن کوتاه پیرامون تصاویر می‌تواند به‌دقت بازیابی بالاتری نسبت به استفاده از کل سندی که در آن تصویر تعبیه‌شده است، دست یابد. در متن‌کاوی، از شباهت جمله به‌عنوان معیاری برای کشف دانش غیب از پایگاه داده‌های متنی استفاده می‌شود. علاوه بر این، ترکیب شباهت متن کوتاه برای برنامه‌هایی مانند جمع‌بندی متن، طبقه‌بندی متن و ترجمه ماشین تفکیک مفهوم کلمه و برای خلاصه‌سازی استخراج، خوشه‌بندی متن، همچنین تشخیص ناهنجاری‌ها مفید است. این کاربردهای نمونه نشان می‌دهد که محاسبه شباهت جمله به یک مؤلفه عمومی برای جامعه تحقیقاتی درگیر در بازنمایی و کشف دانش مرتبط با متن تبدیل‌شده است. اقدامات شباهت متن همچنین برای ارزیابی انسجام متن مفید بوده است[۲۵].

یکی از اولین کاربردهای شباهت متن شاید مدل برداری در بازیابی اطلاعات باشد، جایی که سندی که بیشتر مربوط به یک پرس‌وجو ورودی است با رتبه‌بندی اسناد در یک مجموعه به ترتیب معکوس از شباهت آن‌ها با پرس‌وجو داده‌شده مشخص می‌شود. یک چالش بزرگ در زبانشناسی محاسباتی، ساختن مدل‌های نمایندگی برای فعال کردن درک طبیعی است. به‌منظور آموزش و ارزیابی این روش‌ها، جامعه چندین چالش و مجموعه داده ازجمله ماشین‌آلات درک، پاسخگویی به سؤالات خودکار پیشنهاد کرده است. پاسخگویی کوتاه درجه‌بندی، استنباط زبان طبیعی[1] و تشابه متن معنایی. در این کارها، سیستم درسی زبان طبیعی باید دو قطعه متن را جفت کند و سپس خروجی مانند ارتباط بین یک سؤال و یک متن را (درک دستگاه)، یک سؤال و یک پاسخ (پاسخ به سؤال) ارائه دهد[۹].

شباهت معنایی همچنین در بسیاری از برنامه‌های وب معنایی مانند استخراج جامعه، تولید هستی‌شناسی و تفکیک موجودیت نقش دارد. همچنین برای جستجوی توییتر مفید است، جایی که لازم است با دقت اندازه‌گیری ارتباط معنایی بین مفاهیم یا موجودات انجام شود. در بازیابی اطلاعات یکی از مشکلات اصلی بازیابی مجموعه‌ای از اسناد و بازیابی تصاویر توسط زیرنویس‌ها است که ازلحاظ معنایی با یک پرس‌وجو کاربر مشخص‌شده در موتور جستجوگر وب مرتبط است. در زمینه بانک اطلاعاتی نیز می‌توان از شباهت متن برای تطبیق طرح‌واره برای حل ناهمگونی معنایی برای سیستم به اشتراک‌گذاری داده‌ها، سیستم ادغام داده‌ها، سیستم ارسال پیام و سیستم مدیریت داده همتا به نظیر استفاده کرد. همچنین برای عملیات پیوند رابطه در دیتابیس که صفات پیوند متنی به یکدیگر شبیه هستند مفید است. همچنین دارای صحت دامنه برنامه است ازجمله ادغام و جستجوی داده‌ها از منابع ناهمگن، تمیز کردن داده‌ها و استخراج داده‌ها. سنجش تشابه معنایی متن، بین کلمات، اصطلاحات، جملات، پاراگراف و سند نقش مهمی در علم کامپیوتر و زبان محاسباتی ایفا می‌کند. این برنامه همچنین در چندین زمینه کاربردهای زیادی دارد، مانند

[1] Natural language inference (NLI)

داده‌ورزی زیست پزشکی و ژئوی اطلاعات. علاوه بر زبان‌شناسی، محاسبات معنایی نیز در زمینه‌های خاص حوزه، مانند زیست پزشکی ظاهر می‌شود. شباهت معنایی در برنامه‌های دیگر مانند ترجمه و تفسیر و ترجمه حافظه نیز نقش اساسی دارد. بااین‌حال، محاسبه تشابه معنایی بین جملات یک کار پیچیده و دشوار است. باگذشت سال‌ها، کارهای مشترک SemEval برای تنظیم دقیق و کامل کردن این اقدامات تشابه و کشف ماهیت معنا در زبان کارکرده است.

از دیگر کاربردهای شباهت معنایی:

- اصلاح اشتباهات نوشتاری که از آن با عنوان غلط‌یاب املایی یاد می‌شود.
- شناسایی متون تقلبی یا دزدی علمی و ادبی که در مقالات لاتین با عنوان سرقت ادبی[1] مطرح است.
- یافتن عبارات جایگزین در موتورهای جست‌وجو جهت افزایش دقت و صحت در بازیابی اسناد.
- تشخیص تقلب در کد نرم‌افزار.
- استخراج پاسخ در سامانه‌های پرسش و پاسخ.
- طبقه‌بندی اسناد بر اساس موضوع یا اندازه متن در فایل متن.

۴.۱ لزوم استفاده از تخمین شباهت معنایی متون

با توجه به کاربردهای که در بخش قبل برای تخمین شباهت معنایی ذکر شد و همچنین رشد نمایی تعداد متون، لزوم استفاده از یک روش مکانیزه برای به دست آوردن تشابه بین متون ضروری به نظر می‌رسد. روزانه هزاران خبر به یک خبرگزاری ارسال می‌شود، میلیون‌ها ایمیل ردوبدل می‌شود، متون تقلبی یا دزدی علمی و سرقت ادبی زیادی صورت می‌گیرد. همه موارد برشمرده لزوم استفاده از روش‌های مکانیزه را بیش‌ازپیش نمایان می‌نماید.

امروزه رشد اطلاعات در رسانه‌های دیجیتالی تجزیه‌وتحلیل مقادیر زیادی از داده‌های متنی را تشویق می‌کند. این موردتوجه محققان علوم داده و هوش مصنوعی و همچنین صنعت اینترنت را به خود جلب می‌کند. کاربران اینترنت مسئولیت بخش قابل‌توجهی از این رشد را بر عهده‌دارند. آن‌ها اطلاعاتی را در شبکه وارد می‌کنند که برای به اشتراک‌گذاری دانش نیز استفاده می‌شود. بخش مهمی از این دانش در مخازن ازجمله انجمن‌های پرسش و پاسخ، روزنامه‌های دیجیتالی و دائره‌المعارف‌های دیجیتالی یافت می‌شود. با توجه به تکثیر بی‌شمار اطلاعات در این مخازن، نگرانی‌های زیادی در مورد نحوه تغذیه و مصرف دانش کاربران وجود دارد. برخی از این نگرانی‌ها شامل حذف موارد اضافی در انجمن‌های پاسخگو به سؤال یا بهره‌برداری از افزونگی‌ها برای ارزیابی اطمینان از اخبار در رسانه‌ها یا به‌سادگی فشرده‌سازی اندازه متن است. دستیابی به این پردازش گسترده اطلاعات برای

[1] plagiarism

داوران انسانی به‌وضوح غیرممکن است. در این سناریو، روش‌های آماری پردازش زبان طبیعی کمک قابل‌توجهی است[۲۶].

۵.۱ اهداف تحقیق

با توجه به کاربردهای شباهت یابی و لزوم به‌کارگیری آن‌که در بخش‌های قبلی به آن اشاره شد، هدف این تحقیق تخمین شباهت معنایی مجموعه‌ای از متون با روش‌های آماری و دسته‌بندی با ماشین بردار پشتیبان است. هدف اصلی روش‌های شباهت متن بهبود توانایی رایانه‌ها در درک زبان بشری است که می‌تواند از طریق روش‌های مختلف برای بیان معنایی مشابه باشد. یکی از اهداف اصلی کار شباهت معنایی متن ایجاد یک چارچوب یکپارچه با ترکیب چندین مؤلفه معنایی مستقل به‌منظور یافتن تأثیر آن‌ها در چندین کار پردازش زبان طبیعی است تدوین چنین چارچوبی یک مسئله مهم تحقیقاتی است. یکی از چالش‌های پیش‌رو در روش‌های بیان‌شده در این زمینه رسیدن به دقت حداکثری نتایج به‌دست‌آمده از تخمین شباهت است. همچنین ضریب همبستگی نیز یکی دیگر از معیارهایی است که باید به بهبود آن توجه کرد. بنابراین ما درصدد هستیم تا این معیارها را بهبود بدهیم.

۶.۱ ساختار تحقیق

در این فصل به تعریف شباهت معنایی پرداختیم و کاربردهای آن و هدف از انجام این پژوهش را بیان نمودیم. در فصل دوم این تحقیق به معرفی انواع مختلف روش‌های شباهت یابی اسناد متنی و ارزیابی کلی الگوریتم‌های معرفی‌شده به همراه اشکال یا فرمول‌ها خواهیم پرداخت. این روش‌ها غالباً در سه دسته کلی روش‌های مبتنی بر پیکره و روش‌های مبتنی بر معنا و مبتنی بر رشته قرار می‌گیرند که هر یک مزایا و معایب خاص خود را دارا می‌باشند.

در فصل سوم به بررسی روش پیشنهادی خود در این زمینه می‌پردازیم و مدل خود را ارائه می‌دهیم. همچنین به شرح جزئیات تحقیق پرداخته و روش خود را به تفضیل بیان خواهیم کرد.

در فصل چهارم نتایج به‌دست‌آمده از روش ارائه‌شده و جزئیات پیاده‌سازی بیان خواهد شد و با روش‌های قبلی بیان‌شده در این زمینه مقایسه خواهیم کرد.

در فصل پنجم در خصوص نتایج به‌دست‌آمده در فصل چهار بحث و بررسی خواهیم کرد و اینکه آیا نتایج حاصل یک جواب عمومی می‌باشد یا اینکه در حوزه خاصی قابل‌استفاده است. همچنین پیشنهاداتی را که برای بهبود آن مفید است در آینده بررسی گردد، ارائه خواهیم کرد.

فصل ۲

پیشینه تحقیق

۱.۲ مقدمه

این فصل به‌مرور کارهای انجام‌شده در ارتباط با شباهت معنایی جملات و متون کوتاه و طولانی می‌پردازد. روش‌های مختلفی در این زمینه وجود دارد که هرکدام مزایا و معایب خاص خود را دارند. به‌طورکلی روش‌های شباهت یابی می‌تواند بر اساس ساختار متن یا معنای آن انجام بگیرد. ابتدا با مبانی نظری در زمینه شباهت یابی آشنا می‌شویم، سپس مروری بر کارهایی که در این زمینه انجام‌شده است خواهیم داشت. در انتها به جمع‌بندی کارهای بیان‌شده در این فصل می‌پردازیم و مقایسه‌ای بین آن‌ها انجام خواهیم داد.

۲.۲ تاریخچه شباهت معنایی متن

شناسایی شباهت معنایی متن در متن‌های کوتاه در سال ۲۰۰۶ گزارش شد که در [۲۷]؛[۲۸] ارائه گردید. پس‌ازآن، تمرکز روی اسناد بزرگ یا کلمات فردی تغییر یافت. از سال ۲۰۱۲، وظیفه تشابه معنایی نه‌تنها برای یافتن شباهت بین دو متن محدود نیست، بلکه برای ایجاد نمره تشابه توسط کارهای مختلف [۲۹]؛[۳۰] نیز ایجاد شد.

در سال ۲۰۱۲ اولین کار آزمایشی در SemEval 2012، به‌عنوان بخشی از اجلاس SEM 2012، با موفقیت بزرگ برگزار شد: ۳۵ گروه با ۸۸ اجرای سیستم شرکت کردند. علاوه بر این، یک کارگاه پشتیبانی DARPA در دانشگاه کلمبیا برگزار گردید. در سال ۲۰۱۳، شباهت معنایی متن به‌عنوان کار مشترک رسمی اجلاس SEM 2013 انتخاب شد. بر این اساس در STS 2013، دو وظیفه تنظیم گردید: وظیفه اصلی هسته، که مشابه کار ۲۰۱۲ است. و یک کار آزمایشی در نوع شباهت تایپ‌شده بین سوابق نیمه ساختاریافته [۳۱]؛ سیستم UKP در SemEval 2012 بهترین کار را در شباهت معنایی متن انجام داده است.

در سال ۲۰۱۲ استفاده از شبکه‌های عصبی برای پردازش زبان طبیعی موردتوجه بسیاری قرار گرفت، یکی از مقالات اصلی که از شبکه‌های معلولیت در کارهای سنتی پردازش زبان طبیعی استفاده کرده است (کولبرت و همکاران، ۲۰۱۱) می‌باشد. در سال‌های گذشته مدل‌های شبکه عصبی برای کارهای مختلفی از ترجمه ماشین (زو و همکاران، ۲۰۱۳ ؛ چو و همکاران، ۲۰۱۴) برای پاسخگویی به سؤال استفاده‌شده‌اند (وستون و همکاران، ۲۰۱۵)[۳۲].

وظایف مشترک STS از سال ۲۰۱۲ به‌طور سالیانه برگزار می‌شود و محل برگزاری ارزیابی الگوریتم‌ها و مدل‌های پیشرفته‌ای است. در طول این دوره، مجموعه‌ای متنوع از ژانرها و منابع داده‌شده موردبررسی قرارگرفته است (سرفصل‌های خبری، توضیحات تصویری و فیلم، واژه‌نامه‌هایی از منابع لغوی ازجمله شبکه واژگان، FrameNet، OntoNote، انجمن‌های گفتگوی وب و مجموعه داده‌های پرسش‌وپاسخ) [۲].

هرسال STS به یکی از محبوب‌ترین کارها در رقابت SemEval تعلق دارد. بهترین سیستم STS در [۳۳] از شباهت واژگانی و تحلیل معنایی صریح استفاده کرد. در SemEval 2013، بهترین مدل از مدل‌های معنایی مانند تحلیل معنایی، منابع اطلاعات خارجی (شبکه واژگان) و فن‌های تطابق N ـ گرم استفاده کردند. برای SemEval 2014 و 2015 بهترین سیستم از طریق [۳۴]؛[۳۵]؛[۳۶] معرفی شدند. آن‌ها الگوریتم جدیدی را معرفی کردند که کلمات رابین دو جمله قرار می‌دهد. آن‌ها نشان دادند که از این رویکرد مي‌توان براي شباهت معنایی متن نیز استفاده کرد.

روش‌های شباهت معنایی متن را می‌توان به‌طورکلی به دودسته تقسیم کرد: روش‌های کنترل نشده و روش‌های نظارت‌شده. روش‌های کنترل نشده لی و همکاران [۳۷] و میهالچه و همکاران [۳۸] شاید نمایندگان آغازگر تحقیقات شباهت معنایی متن بانفوذ قابل‌توجهی باشند. روش بی‌نظیر سلطان و همکاران [۳۹] تأثیرات عمده‌ای را در کارهای SemEval STS از سال ۲۰۱۴ داشته است. در سال‌های اخیر، سامانه‌های تحت نظارت STS مدل‌های یادگیری عمیق را با فن‌های پردازش زبان طبیعی سنتی ترکیب کرده‌اند و بهترین عملکرد را به دست آورده‌اند.

در حال حاضر الگوریتمی دقیق‌تر است که به شباهت انسانی نزدیک‌تر باشد. با استفاده از مجموعه‌ای از معیارها به‌عنوان شباهت رشته‌ای کلمه و طولانی‌ترین الگوریتم متعارف مشترک برای محاسبه شباهت معنایی بین جملات استفاده می‌شود[۴۰].

۳.۲ پردازش زبان طبیعی

پردازش زبان‌های طبیعی یکی از زیرشاخه‌های بااهمیت در حوزه گسترده علوم رایانه و هوش مصنوعی است که به تعامل بین کامپیوتر و زبان‌های طبیعی (انسانی) می‌پردازد؛ بنابراین پردازش زبان‌های طبیعی بر ارتباط انسان و رایانه متمرکز است. پس چالش اصلی و عمده در این زمینه درک زبان طبیعی و ماشینی کردن فرایند درک و برداشت مفاهیم بیان‌شده با یک‌زبان طبیعی انسانی است. به تعریف دقیق‌تر، پردازش زبان‌های طبیعی عبارت است از استفاده از رایانه برای پردازش زبان گفتاری و زبان نوشتاری. بدین معنی که رایانه‌ها را قادر سازیم که گفتار یا نوشتار تولیدشده در قالب و ساختار یک‌زبان طبیعی را تحلیل و درک نموده یا آن را تولید نمایند. در این صورت، با استفاده از آن می‌توان به ترجمه‌ی زبان‌ها پرداخت، از صفحات وب و بانک‌های اطلاعاتیِ نوشتاری جهت پاسخ دادن به پرسش‌ها استفاده کرد، یا با دستگاه‌ها، مثلاً برای مشورت گرفتن به گفت‌وگو پرداخت. هدف اصلی در پردازش زبان طبیعی، ایجاد تئوری‌های محاسباتی از زبان، با استفاده از الگوریتم‌ها و ساختارهای داده‌ای موجود در علوم رایانه است. بدیهی است که در راستای تحقق این هدف، نیاز به دانشی وسیع از زبان است و علاوه بر محققان علوم رایانه، نیاز به دانش زبان‌شناسان نیز در این حوزه می‌باشد. با پردازش اطلاعات زبانی می‌توان آمار موردنیاز برای کار با زبان طبیعی را استخراج کرد. کاربردهای پردازش زبان طبیعی به دودسته کلی قابل‌تقسیم است: کاربردهای نوشتاری و کاربردهای گفتاری. از کاربردهای نوشتاری آن می‌توان به استخراج اطلاعاتی خاص از یک متن، ترجمه یک متن به زبانی دیگر یا یافتن مستنداتی خاص در یک پایگاه داده نوشتاری (مثلاً یافتن

کتاب‌های مرتبط به هم در یک کتابخانه) اشاره کرد. نمونه‌هایی از کاربردهای گفتاری پردازش زبان عبارت‌اند از: سامانه‌های پرسش و پاسخ انسان با رایانه، سرویس‌های اتوماتیک ارتباط با مشتری از طریق تلفن، سامانه‌های آموزش به فراگیران یا سامانه‌های کنترلی توسط صدا [۴۱]؛[۴۲].

۴.۲ ویژگی‌های مهم و مبانی نظری در شباهت معنایی

در این بخش؛ برای سنجش شباهت بین دو جمله ابتدا با مبانی نظری در این زمینه آشنا می‌شویم.

۱.۴.۲ شبکه واژگان

شبکه واژگان پایگاه داده‌ای لغوی براي زبان انگليسي بشمار می‌آید که در سال ۱۹۸۵ میلادی پیاده‌سازی آن شروع شد و مراحل تکامل آن همچنان ادامه دارد. این منبع لغت حاصل تحقیقات انجام‌شده در آزمایشگاه علوم شناختي دانشگاه پرینستون و در ارتباط با مدل‌سازی دانش لغوی انسان است و به‌طور هم‌زمان مزایای یک فرهنگ جامع و فرهنگ لغت را دارا است. شبکه واژگان درواقع شبکه معنایی از تمام کلمات زبان انگلیسی است که در حال حاضر به‌صورت متن‌باز قابل‌استفاده است.

جملات در شبکه واژگان با توجه به تعاریف کلی که توسط اصطلاحات ارائه می‌شوند، به‌عنوان گروه‌های مترادف یعنی نحو[1] یا مفهوم[2] سازماندهی می‌شوند. با توجه به خصوصیات فنی و عناصر محاسباتی، رویکردهای مبتنی بر شبکه واژگان می‌توانند در سه دسته اصلی قرار بگیرند: مبتنی بر مسافت مسیری، مبتنی بر محتوای اطلاعات و مبتنی بر ویژگی‌ها. رویکردهای مبتنی بر فاصله مسیری لبه‌های جداکننده دو مفهوم، همراه با عمق مفهوم، چگالی محلی، نوع پیوند و استحکام پیوند است.

شبکه واژگان، یک پایگاه داده آنلاین واژگان انگلیسی است که به‌طور گسترده به‌عنوان دانش قبلی برای محاسبه شباهت معنایی بین دو مفهوم مورداستفاده قرار می‌گیرد. این واژگان را به اسم، افعال، صفت‌ها و قیدها به‌عنوان قالب مجموعه مترادف تقسیم می‌کند. هر مفهوم شامل گروهی از کلمات است که به معنای مشترک است. همگام‌سازی‌ها در شبکه واژگان به‌طورکلی با ۱۱ نوع روابط معنایی نظیر: هایپونامی / هایپرنیوم is-a، مرامی گری / هلمونی (بخشی از عضو ـ عضو، ماده ـ ماده) و antonymy (کلمه متضاد) و غیره به ارث رسیده‌اند. تقریباً ٪۸۰ روابط is-aهستند. بنابراین، شبکه واژگان را می‌توان به‌عنوان ساختار سلسله مراتبی مانند درخت در نظر گرفت که در آن کلیت مفهوم به‌صورت یکنواخت با عمق آن‌ها کاهش می‌یابد. روابط ارثی معمولاً برای محاسبه شباهت معنایی مورداستفاده قرار می‌گیرد، درحالی‌که روابط دیگر همچون کلمه شناسی / هلو نامی و آنتونی نیز برای ارزیابی ارتباط معنایی مناسب‌تر هستند. آثار مرتبط با استفاده از ساختار سلسله

[1] synset
[2] concept

مراتبی و خصوصیات معنایی شبکه واژگان برای اندازه‌گیری شباهت معنایی، که شامل طول لبه، عمق و تراکم است. رابطه is-a از شبکه واژگان است که عمدتاً برای محاسبه کوتاه‌ترین طول مسیر در اقدامات شباهت استفاده می‌شود از طرف دیگر، توپولوژی مفاهیم در شبکه واژگان در تعیین مفهوم اطلاعات ذاتی موجود در مفاهیم کمک می‌کند که می‌تواند به‌عنوان اقدامات ارزیابی کمکی برای بهبود دقت مورداستفاده قرار گیرد[۳].

۲.۴.۲ مفهوم اطلاعات

در تئوری اطلاعات، محتوای اطلاعات، سوابق متغیر یا سیگنال تصادفی، میزان اطلاعات به‌دست‌آمده هنگام نمونه‌برداری است. به‌طور رسمی، محتوای اطلاعات متغیر تصادفی است که بدون توجه به اندازه‌گیری متغیر تصادفی برای هر رویداد در تئوری احتمال تعریف‌شده است. درواقع محتوای اطلاعات در واحد اطلاعات بیان می‌شود. ارزش مورد انتظار سوابق، آنتروپی نظری اطلاعات است، میانگین اطلاعاتی که یک ناظر انتظار دارد در هنگام نمونه‌گیری از متغیر تصادفی، در مورد یک سیستم به دست آورد. با توجه به یک متغیر تصادفی x با عملکرد توده احتمال $PX(x)$، اطلاعات خود اندازه‌گیری x به‌عنوان نتیجه x به‌عنوان $IX(x) := -logPX(x)] = log(\frac{1}{PX(x)})$ تعریف‌شده است. به‌طور گسترده یک واقعه E داده می‌شود با احتمال P، محتوای اطلاعات به‌طور مشابه تعریف می‌شود:

$$I(E) := -log[Pr(E)] = -log(P) \qquad (۱.۲)$$

به‌طورکلی، پایه لگاریتمی انتخابی برای بسیاری از خصوصیات نظری اطلاعات اهمیت ندارد. بااین‌حال، واحدهای مختلف اطلاعات بر اساس گزینه‌های محبوب پایه تعیین می‌شوند. اگر پایه لگاریتمی ۲ باشد، این واحد به نام شانون نامیده می‌شود اما از بیت نیز استفاده می‌شود. اگر پایه لگاریتم طبیعی باشد، واحدی که به آن "nat" کوتاه گفته می‌شود. اگر قرار باشد که لگاریتم پایه ۱۰ باشد، به واحدها یا نمودارهای رقمی یا رقم اعشار گفته می‌شود. آنتروپی شانون از متغیر تصادفی x این‌گونه تعریف‌شده است.

$$H(X) = \Sigma_x - PX(x)\log PX(x) = \Sigma_x PX(x)IX(x) \stackrel{\text{def}}{=\!=} E[IX(x)] \qquad (۲.۲)$$

برابر است با تعریف محتوای اطلاعات موردانتظار اندازه‌گیری x [۴۳]؛[۴۴].

۳.۴.۲ ویکی‌پدیا[1]

ویکی‌پدیا به بزرگ‌ترین دائره‌المعارف ارائه امکانات مانند چندزبانه، پوشش گسترده و به‌روز تبدیل‌شده است و به‌صورت رایگان قابل‌دسترس همگان می‌باشد. واحد اصلی در ویکی‌پدیا مقاله است که

[1] Wikipedia

به‌صورت متن آزاد نوشته‌شده است که به دنبال یک مجموعه جامع از دستورالعمل‌های سرمقاله و ساختاری جهت ارتقاء سازگاری و انسجام است. هر مقاله یک مفهوم واحد را توصیف می‌کند و یک مقاله واحد برای هر مفهوم وجود دارد. عنوان هر مقاله عبارت مختصری است که در آن اصطلاح‌نامه متداول استفاده‌شده است و همچنین عنوان مفهوم هم نامیده می‌شود. مقالات با مقالات دیگر پیوند دارند. هایپر لینک، لینک همسایه نامیده می‌شود. لینک‌های ویکی‌پدیا به‌عنوان یک منبع اضافی مترادف که با تغییر مسیر هدایت می‌شود مفید نیستند. مقالات ذکرشده از دیگر ورودی‌های دانشنامه، به آن‌ها از طریق لینک‌های داخلی اشاره دارد، که این مقالات مرجع هستند[۲۳] .

در ویکی‌پدیا، اقدامات مبتنی بر مسیر و اقدامات مبتنی بر مفهوم اطلاعات از ساختار طبقه‌بندی برای ارزیابی شباهت معنایی استفاده می‌کنند. ساختار طبقه‌بندی به نام نمودار دسته‌بندی ویکی‌پدیا[۱] شامل حدود ۱۵۰۰۰۰۰ مفهوم است که بسیار کمتر از تعداد مفاهیم ویکی‌پدیا است. از خصوصیات سنتی مانند نحو، واژه‌نامه‌ها و همسایگان استفاده می‌شود که اقدامات مبتنی بر ویژگی‌ها به‌طور جهانی برای ارزیابی شباهت معنایی بین مفاهیم اتخاذ می‌شوند. ویکی‌پدیا، یک دانشنامه مشترک آنلاین است که می‌تواند به‌عنوان یک دانش خارجی برای تجزیه‌وتحلیل متن کوتاه استفاده شود. ویکی‌پدیا ساختار پیوندی متراکم و پوشش گسترده‌ای از اشخاص ازجمله اشخاص دارای نام، اشخاص خاص دامنه و اشخاص نوظهور دارد. علاوه بر این، داده‌های زباله ویکی‌پدیا را می‌توان آزادانه از وب به دست آورد. این مزایا محققان و توسعه‌دهندگان را ترغیب کرده است که از آن برای برنامه‌های کاربردی استفاده کنند[۴۵].

۱.۳.۴.۲ تحلیل معنایی صریح مبتنی بر ویکی پدیا [۲]

تحلیل معنایی صریح مبتنی بر ویکی‌پدیا روشی گسترده است که برای اندازه‌گیری شباهت معنایی بین متون با هر طول استفاده می‌شود. تحلیل معنایی صریح برداری از اشخاص مرتبط با ویکی‌پدیا را برای متن معین به‌عنوان بیان معنایی آن ایجاد می‌کند و از وکتور برای اندازه‌گیری شباهت استفاده می‌کند. یافتن واحدهای مرتبط با ویکی‌پدیا از یک متن به‌طورکلی شامل چندین زیر مشخص مانند استخراج اصطلاحات کلیدی، یافتن نهادهای مرتبط برای هر اصطلاح کلیدی و تجمع وزن اشخاص مرتبط است. به‌منظور حل مشکل، تحلیل معنایی صریح بردارهای وزنی اشخاص مرتبط با هر کلمه را فقط بر اساس قانون اکثریت خلاصه می‌کند. رویکرد جمع‌بندی بردارها برای متون کوتاه در دنیای واقعی مناسب نیست که در آن اصطلاحات کلیدی و نامربوط در تعداد بسیار کمی بار مشاهده می‌شوند. به دلیل عدم اطلاع کافی، قانون اکثریت به‌خوبی کار نمی‌کند. در چنین شرایطی، تمرکز روی اصطلاحات کلیدی در ضمن فیلتر کردن اصطلاحات پرسروصدا اهمیت دارد[۴۵].

---

[1] Wikipedia kategor Graph (WCG)

[2] **Explicit semantic analysis**

۴.۴.۲ مدلN ـ گرم[1]

مدل N ـ گرم نوعی مدل زبانی احتمالاتی است که برای پیش‌بینی قلم بعدی در دنباله‌ای به شکل مرتبه‌ی $(n-1)$ مارکوف به کار می‌رود. امروزه مدل‌های N ـ گرم در احتمالات، نظریه‌ی ارتباطات، زبان‌شناسی رایانشی (مثلاً پردازش آماری زبان طبیعی) و فشرده‌سازی داده‌ها کاربرد گسترده دارند. یک N ـ گرم دنباله‌ای از n کلمه است که به‌صورت متوالی در یک متن ظاهر می‌شوند. به‌طورمعمول جملات با موضوعات یکسان دارایN ـ گرم‌های مشابه هستند. به N ـ گرم با اندازه‌های یک، دو، و سه به ترتیب تک گرم، دوگرم، و سه گرم گفته می‌شود و نام‌گذاری به همین منوال ادامه می‌یابد[۴۶].

۵.۴.۲ طولانی ترین پیامد مشترک[2]

طولانی‌ترین پیامد مشترک طول طولانی‌ترین پیامد را در هر دو جمله می‌یابد. پیامد دنباله‌ای است که به همان ترتیب نسبی ظاهر می‌شود، اما لزوماً متناقض نیست. به‌عنوان‌مثال، abc، abg، bdf، aeg، acefg و غیره پیامدهای abcdefg هستند. توجه کنید که LCS لزوماً منحصربه‌فرد نیست؛ به‌عنوان‌مثال LCS ، ABC وACB هم AB و AC هستند. درواقع، مسئله LCS اغلب برای یافتن کلیه پیامدهای مشترک با حداکثر طول تعریف می‌شود[۴۷].

الگوریتم طولانی‌ترین پیامد مشترک برای یافتن طولانی‌ترین بستر یک‌رشته استفاده می‌شود. این دو رشته را باهم مقایسه می‌کند و بر اساس طولانی‌ترین زنجیره مشترک کاراکترها، شباهت را می‌یابد. این می‌تواند به شرح زیر اندازه‌گیری شود:

$$LCSubstr(S۱, S۲) = LCSuff(S۱_۱ \dots i, S۲_۲ \dots j), max, ۱ < i < m, ۱ < j < n \qquad (۳.۲)$$

که در آن m و n طول دو رشته است و $LCSuff$ تابعی است که طولانی‌ترین پسوندهای رایج از پیشوندهای ممکنS1 وS2را می‌یابد[۴].

۶.۴.۲ فرکانس مدت ـ فرکانس اسناد معکوس[3]

فرکانس مدت ـ فرکانس اسناد معکوس در بازیابی اطلاعات، یک آمار عددی است که میزان اهمیت یک کلمه نسبت به یک سند در یک مجموعه‌ای از اسناد را نشان می‌دهد. درواقع هدف این سیستم وزن دهی، نشان دادن اهمیت کلمه در متن است. درواقع Tf-idf روشی است که از طریق آن وزن لگاریتمی کلمه کلیدی به دست می‌آید و تمرکز اصلی آن‌ها صرفاً روی تعداد دفعات تکرار کلمه کلیدی است و بنا بر تنظیمات آن‌ها، هر زمان که تعداد تکرار کلمه کلیدی به‌صورت درصدی قابل‌قبول از کل متن ظاهر شود، آن متن را مناسب و هدفمند در نظر می‌گیرند درحالی‌که این امر اشتباه و یا

[1] N-gram
[2] Longest Common Substring(LCS)
[3] term frequency - inverse document frequency (Tf-idf)

حداقل روش ناقص است. Tf-idf از دو عبارت Tf به معنای محدوده فرکانس نرمال و idf که فرکانس محتوایی معکوس نامیده می‌شود تشکیل‌شده است. مقدار Tf-idf به‌تناسب تعداد تکرار کلمه در سند افزایش می‌یابد و توسط تعداد اسنادی که در مجموعه هستند و شامل کلمه نیز می‌باشند متعادل می‌شود. به این معنی که اگر کلمه‌ای در بسیاری از متون ظاهر شود احتمالاً کلمه‌ای متداول است و ارزش چندانی در ارزیابی متن ندارد.

برای به دست آوردن ضریب Tf-idf می‌بایست هرکدام از این دو عبارت را به‌صورت جداگانه محاسبه نموده و حاصل دو عبارت را در هم ضرب کنیم تا نتیجه حاصله، فراوانی وزنی کلمه کلیدی را به ما نشان دهد. فرمول دو عبارت فوق به شرح زیر می‌باشد:

Tf = عبارت است از تقسیم تعداد تکرار کلمه بر تعداد کل کلمات محتوا

idf = عبارت است از لگاریتم تقسیم تعداد کل محتوا بر محتواهایی که شامل کلمه موردنظر هستند

علت استفاده از لگاریتم در این فرمول این است که قصد داریم درون کل محتواهای سایت نرمال‌سازی انجام دهیم.

$$W_x, y = tf_x, y \times \log \frac{N}{df_x} \qquad (۴.۲)$$

۷.۴.۲ چارچوب کلمه به بردار [1]

یک روش آماری برای یادگیری کارآمد یک کلمه مستقل تعبیه‌شده از یک متن است که توسط توماس میکولو و همکاران توسعه یافت. در سال ۲۰۱۳ در گوگل به‌عنوان پاسخی برای آموزش کارآمد شبکه آموزش مبتنی بر شبکه عصبی و از آن زمان به‌عنوان استاندارد بالفعل برای توسعه تعبیه کلمه از قبل آموزش‌دیده مورداستفاده قرارگرفته است. علاوه بر این، کار شامل تجزیه‌وتحلیل بردارهای آموخته‌شده و اکتشاف ریاضی بردار در بازنمایی کلمات است.

ما می‌دانیم که این بازنمایی‌ها به‌طور شگفت‌آور در ضبط منظم نحوی و معنایی در زبان خوب هستند و این‌که هر رابطه با یک جبهه بردار خاص رابطه مشخص می‌شود که این اجازه را می‌دهد تا استدلال بردار محور مبتنی بر جبران خسارت بین کلمات نتیجه دهد.

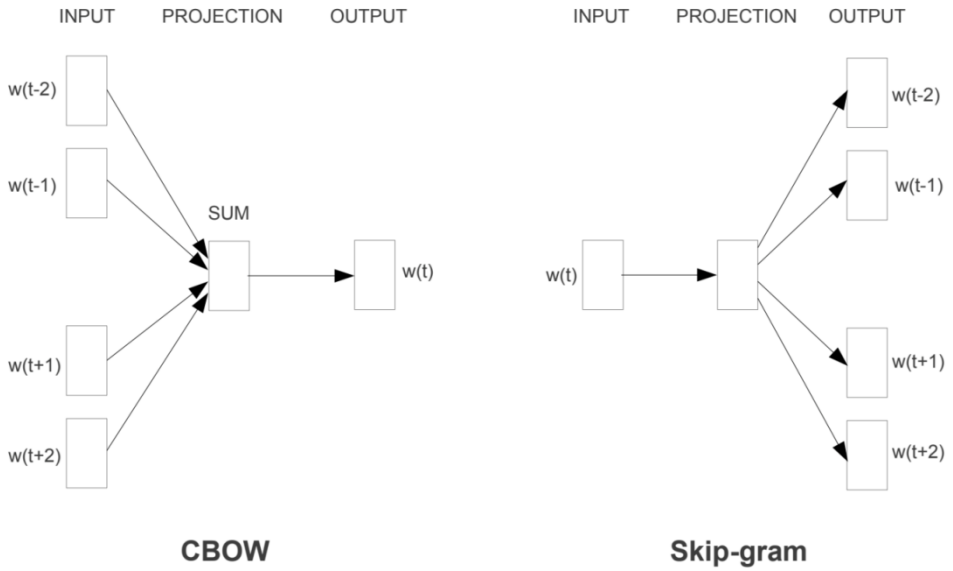
دو مدل یادگیری مختلف در شکل ۱.۲ ارائه‌شده است که می‌تواند به‌عنوان بخشی از رویکرد word2vec برای یادگیری کلمه تعبیه استفاده شود.

- بسته‌های کلمات مداوم یا مدل CBOW
- مدل پرش_گرم مداوم[2]

[1] **word2vec**

[2] skip-gram

مدل CBOW تعبیه را با پیش‌بینی کلمه فعلی بر اساس کلمات متن می‌آموزد. مدل پرش_گرم مداوم به پیش‌بینی کلمات همسایه با کمک کلمه فعلی می‌پردازد. مزیت اصلی این رویکرد این است که تعبیه‌های کلمه‌ای باکیفیت بالا می‌توانند به‌طور کارآمد آموخته شوند (فضای کم و پیچیدگی زمانی)، همچنین اجازه می‌دهد تا تعبیه‌های بزرگ‌تر (ابعاد بیشتر) از شرکت‌های بسیار بزرگ‌تر متن (میلیاردها کلمه) آموخته شود.



شکل ۱.۲: برآورد کارآمد از بازنمایی کلمات در فضای بردار

مدل CBOW مدل دیگری است که برای یادگیری بردارهای کلمات مورداستفاده قرار می‌گیرد. این مدل کلمه را از بین کلمات محتوایی منبع پیش‌بینی می‌کند. ازآنجایی‌که چند کلمه محتوایی وجود دارد، بردارهای کلمه متناظر با هر یک را که از طریق ضرب بردار ورودی با ماتریس حاصل‌شده‌اند، میانگین گرفته‌اند. از اینجایی‌که مرحله میانگین‌گیری بخش زیادی از اطلاعات توزیعی را مسطح می‌کند، بعضی افراد اعتقاددارند مدل CBOW برای دیتاست‌های کوچک مناسب‌تر است.

۸.۴.۲ مدل کیسه ای از کلمات[1]

مدل کیسه‌ای از کلمات نمایشی ساده است که در پردازش زبان طبیعی و بازیابی اطلاعات مورداستفاده قرار می‌گیرد. در این مدل، متنی (مانند یک جمله یا یک سند) به‌عنوان کیسه (چندصفحه‌ای) از کلمات آن نمایش داده می‌شود، بی‌توجه به دستور زبان و حتی دستور کلمه اما با حفظ فراوانی، مدل کیف‌دستی کلمات همچنین برای دید رایانه‌ای استفاده‌شده است[۴۸].

مدل کیف واژه‌ها معمولاً در روش‌های طبقه‌بندی اسناد استفاده می‌شود که در آن فرکانس وقوع

[1] **Bag of words**

هر کلمه به‌عنوان ویژگی‌ای برای آموزش طبقه‌بندی استفاده می‌شود. در عمل، کیسه‌ای از کلمات عمدتاً به‌عنوان ابزاری برای تولید ویژگی‌ها مورداستفاده قرار می‌گیرد. پس از تبدیل متن به کیسه‌ای از کلمات، می‌توان اقدامات مختلفی را برای توصیف متن محاسبه کرد. متداول‌ترین نوع ویژگی‌ها یا ویژگی‌های محاسبه‌شده از مدل کیسه‌ای از کلمات، فرکانس مدت است، یعنی تعداد دفعاتی که یک اصطلاح در متن ظاهر می‌شود. این نمایه لیست (بردار) ترتیب کلمات موجود در جملات اصلی را حفظ نمی‌کند. این فقط ویژگی اصلی مدل کیسه‌ای از کلمات است. همچنین به‌عنوان مدل فضای برداری شناخته می‌شود. مدل بسته کلمات برای بینایی کامپیوتر مدل بسته کلمات معمولاً در روش‌های دسته‌بندی اسناد مورداستفاده قرار می‌گیرد که در آن وقوع هر کلمه به‌عنوان یک ویژگی برای آموزش طبقه‌بندی آماری استفاده می‌شود.

۹.۴.۲ تخصیص پنهان دیریکلت[1]

تخصیص پنهان دیریکلت، یک مدل تولیدی بدون نظارت است که موضوع را به اسناد اختصاص می‌دهد. کلمات موجود در هر سند به تشخیص موضوع کمک می‌کنند. موضوعات ممکن است پیش‌بینی نشوند، اما تعداد موضوعات باید مشخص شود. ممکن است کلمات بین موضوعات هم‌پوشانی داشته باشند، بنابراین چندین موضوع ممکن است کلمات مشابهی داشته باشند. از کاربردهای LDA به موارد زیر می‌توان اشاره‌کرده درک موضوعات مختلف در یک مجموعه، شناخت بهتر از نوع اسناد کمی کردن بیشترین کاربرد و مهم‌ترین کلمات در یک مجموعه شباهت اسناد.

این موارد مدل متغیرهای نهفته را ایجاد می‌کند:

۱.توزیع موضوعات برای هر سند

۲.توزیع کلمات برای هر موضوع

برخلاف الگوریتم word2vec، که با جملات کاملاً ساختاریافته عملکرد بسیار خوبی دارد، LDA یک کلمه مدل است، به معنی ترتیب کلمات در یک سند، شمارش نمی‌شود. این همچنین بدان معنی است که کلمات متوقف و کلمات نادر باید کنار گذاشته شوند، به‌طوری‌که مدل برای کلمات بسیار مکرر و کلمات بسیار کمیاب، که هر دو به موضوعات کلی کمک نمی‌کنند، بیش‌ازحد جبران نمی‌کند.

LDA فرض می‌کند که هر سند توسط یک فرآیند آماری تولید می‌شود. یعنی هر سندی ترکیبی از موضوعات است و هر موضوع ترکیبی از کلمات است.

۱۰.۴.۲ فاصله جنسن شانون[2]

در تئوری و آمار احتمال، واگرایی جنسن شانون روشی برای اندازه‌گیری شباهت بین دو توزیع احتمال است. همچنین به شعاع اطلاعات یا واگرایی کل نسبت به میانگین معروف است. این

[1] Latent Dirichlet Allocation (LDA)

[2] Jensen-Shannon distance

تفاوت بر اساس واگرایی Kullback – Leibler است، با برخی تفاوت‌های قابل‌توجه ازجمله اینکه متقارن است و همیشه دارای ارزش محدود است و برخلاف واگرایی Kullback، این‌یک معیار را به روش مستقیم تعیین می‌کند. ریشه مربع واگرایی جنسن شانون معیاری است که اغلب به‌عنوان فاصله جنسن و شانون شناخته می‌شود. با استفاده از ماشین‌آلات آزمون‌های آماری، روش‌های مناسبی برای تصمیم‌گیری در مورد مشابه بودن دو سند داریم. نوع آزمون قابل‌استفاده برای شباهت آزمون در ادبیات آمار به‌عنوان آزمون همگنی شناخته‌شده است.

۱۱.۴.۲ ماشین بردار پشتیبان [1]

ماشین بردار پشتیبان در آزمایشگاه‌های A&T Bell توسط Vapnik و همکارانش توسعه‌یافته است و مبتنی بر چارچوب‌های آماری است. ماشین بردار پشتیبان یکی از روش‌های یادگیری بانظارت است که از آن برای طبقه‌بندی و رگرسیون استفاده می‌کنند. مبنای کاری دسته‌بندی کننده‌ی ماشین بردار پشتیبان دسته‌بندی خطی داده‌ها است و در تقسیم خطی داده‌ها سعی می‌کنیم خطی را انتخاب کنیم که حاشیه اطمینان بیشتری داشته باشد. حل معادله پیدا کردن خط بهینه برای داده‌ها به‌وسیله روش‌های QP که روش‌های شناخته‌شده‌ای در حل مسائل محدودیت دار هستند صورت می‌گیرد. قبل از تقسیم خطی برای اینکه ماشین بتواند داده‌های با پیچیدگی بالا را دسته‌بندی کند داده‌ها را به‌وسیله‌ی تابع phi به فضای با ابعاد خیلی بالاتر می‌بریم [۴۹]. برای اینکه بتوانیم مسئله ابعاد خیلی بالا را با استفاده از این روش‌ها حل کنیم از قضیه دوگانی لاگرانژ برای تبدیل مسئله‌ی مینیمم سازی موردنظر به فرم دوگانی آن‌که در آن به‌جای تابع پیچیده‌ی phi که ما را به فضایی با ابعاد بالا می‌برد، تابع ساده‌تری به نام تابع هسته که ضرب برداری تابع phi است ظاهر می‌شود استفاده می‌کنیم.

در یادگیري ماشین، ماشین‌هاي بردار پشتیبان مدل‌هاي یادگیري با الگوریتم‌هاي یادگیري مرتبط هستند که داده‌هاي مورداستفاده براي طبقه‌بندی و تحلیل رگرسیون را تحلیل می‌کنند. با توجه به مجموعه‌ای از مثال‌های آموزشی، که هر یک‌به‌دو دسته دیگر تعلق دارند، یک الگوریتم آموزش ماشین بردار پشتیبان مدلی را ایجاد می‌کند که نمونه‌های جدید را به یک دسته یا گروه دیگر اختصاص می‌دهد و آن را به‌عنوان طبقه‌بندی کننده خطی باینری غیرممکن تبدیل می‌کند. یک مدل ماشین بردار پشتیبان نمایانگر نمونه‌ها به‌عنوان نقاط در فضا است که نقشه‌برداری شده است به‌طوری‌که نمونه‌های دسته‌های جداگانه با یک شکاف واضح که تا حد ممکن گسترده است تقسیم می‌شود. نمونه‌های جدید سپس در همان فضا ترسیم می‌شوند و پیش‌بینی می‌شود که بر اساس طرف شکافی که روی آن قرار می‌گیرند، به یک دسته تعلق دارند.

[1] **Support vector machines(SVM)**

۱۲.۴.۲ تجزیه و تحلیل آماری[1]

تجزیه‌کننده زبان طبیعی برنامه‌ای است که ساختار گرامری جملات را انجام می‌دهد، به‌عنوان‌مثال کدام گروه از کلمات باهم جمع می‌شوند (به‌عنوان عبارات) و کدام کلمات موضوع یا هدف یک فعل هستند. تجزیه‌کنندگان احتمالی از دانش زبان به‌دست‌آمده از جملات تجزیه‌شده دستی استفاده می‌کنند تا سعی کنند محتمل‌ترین تحلیل جملات جدید را تولید کنند. این تجزیه‌وتحلیل‌های آماری هنوز هم برخی از اشتباهات را انجام می‌دهند، اما معمولاً بسیار خوب عمل می‌کنند. پیشرفت آن‌ها یکی از بزرگ‌ترین پیشرفت در پردازش زبان طبیعی در دهه ۱۹۹۰ بود. آن‌ها طراحی‌شده‌اند تا توسط افرادی که می‌خواهند روابط متنی را استخراج کنند، به‌راحتی درک و به کار گرفته شوند.

۱۳.۴.۲ حافظه کوتاه مدت طولانی [2]

حافظه کوتاه‌مدت طولانی یک معماری شبکه عصبی مکرر مصنوعی[3] است که در زمینه یادگیری عمیق مورداستفاده قرار می‌گیرد. حافظه کوتاه‌مدت طولانی برخلاف شبکه‌های عصبی استاندارد تغذیه‌شده، اتصالات بازخورد دارد. این نه‌تنها می‌تواند نقاط داده‌های واحد (مانند تصاویر) را پردازش کند، بلکه توالی‌های کل داده‌ها (مانند گفتار یا فیلم) را نیز پردازش می‌کند. به‌عنوان‌مثال، حافظه کوتاه‌مدت طولانی برای کارهایی از قبیل تشخیص نام‌نویسی دست‌نویس، متصل به آن، تشخیص گفتار و تشخیص ناهنجاری در ترافیک شبکه یا سامانه‌های تشخیص نفوذ[4] کاربرد دارد.

یک واحد مشترک حافظه کوتاه‌مدت طولانی از یک سلول، یک دروازه ورودی، یک دروازه خروجی و یک دروازه فراموشی تشکیل‌شده است. سلول مقادیر را در فواصل زمانی دلخواه به یاد می‌آورد و سه دروازه جریان اطلاعات به داخل و خارج از سلول را تنظیم می‌کند.

شبکه‌های حافظه کوتاه‌مدت طولانی به‌خوبی برای طبقه‌بندی، پردازش و پیش‌بینی‌های مبتنی بر داده‌های سری زمانی مناسب هستند، زیرا می‌توان مدت‌زمان ناشناخته‌ای رابین وقایع مهم در یک سری زمان ایجاد کرد. حافظه کوتاه‌مدت طولانی برای مقابله با مشکلات شیب منفجر و ناپدیدشده که می‌توانند هنگام آموزش با RNN سنتی روبرو شوند، ایجادشده‌اند. عدم حساسیت نسبی نسبت به طول فاصله، مزیت LSTM نسبت به RNN، مدل‌های پنهان مارکوف و سایر روش‌های یادگیری توالی در کاربردهای بی‌شمار است. چندین واحد معماری از واحدهای LSTM وجود دارد. یک معماری معمول از سلول (قسمت حافظه واحد LSTM) و سه تنظیم‌کننده تشکیل‌شده است، که معمولاً به‌عنوان دروازه نامیده می‌شوند، جریان اطلاعات در داخل واحد LSTM یک دروازه ورودی، یک دروازه خروجی و یک دروازه فراموشی. برخی از تغییرات واحد LSTM دارای یک یا چند مورد از این دروازه‌ها نیستند یا ممکن است دارای دروازه‌های دیگری باشند. به‌عنوان‌مثال، واحدهای

---

[1] **Statistical Analysis**

[2] **Long short-term memory(lstm)**

[3] Recurrent neural network(RNN)

[4] Intrusion detection systems(IDS)

عودکننده درج‌شده (GRU) دروازه خروجی ندارند.

به‌طور شهودی، سلول مسئول پیگیری وابستگی بین عناصر در دنباله ورودی است. دروازه ورودی میزان ورود مقدار جدید به داخل سلول را کنترل می‌کند، دروازه فراموشی تا چه حد مقداری در سلول باقی می‌ماند و دروازه خروجی میزان استفاده از مقدار سلول را برای محاسبه خروجی کنترل می‌کند. عملکرد فعال‌سازی دروازه LSTM اغلب عملکرد سیگموئید لجستیک است. اتصالات داخل و خارج از دروازه LSTM وجود دارد، که تعداد کمی از آن‌ها در حالت عود است. وزن این اتصالات، که باید در طول آموزش یاد بگیرند، نحوه عملکرد دروازه‌ها را تعیین می‌کند.

۵.۲ پیشینه تحقیق

شباهت معنایی مبتنی بر معنا که به معنی جملات بستگی دارد، رویکردهای مختلفی دارد. این روش‌ها فن‌های مختلفی را برای مقایسه دو جمله ازنظر معنایی به کار می‌گیرند. رویکرد اول، مبتنی بر پیکره است که شباهت کلمات را بر اساس تجزیه‌وتحلیل آماری پیکره می‌یابد. همچنین، از یادگیری عمیق برای تجزیه‌وتحلیل جملات استفاده می‌کند. رویکرد دوم، مبتنی بر دانش است که معنی کلمات و روابط بین کلمات در یک شبکه معنایی گنجانده‌شده است. رویکرد سوم مبتنی بر ساختار می‌باشد که از اطلاعات ساختار یک جمله برای به دست آوردن معنی آن جمله استفاده می‌کند.

۱.۵.۲ روش‌های مبتنی بر پیکره

دو روش مختلف برای تجزیه‌وتحلیل آماری یک متن وجود دارد. مورد اول استفاده از آنالیز آماری عادی مانند LSA و دوم استفاده از یادگیری عمیق است. در وهله اول، یک قشر بزرگ از طریق شمارش کلمات در پیکره و اسناد به‌صورت آماری تحلیل می‌شود. محاسبه Tf-idf که به‌منظور وزن دادن به کلمه استفاده می‌شود، یک هدف مهم در تجزیه‌وتحلیل بدنه است. علاوه بر این، استفاده از یادگیری عمیق نتایج خوبی را در نمایش کلمات ازنظر معنایی نشان داده است[۱].

۱.۱.۵.۲ روش‌هاي آماري

شباهت آماری آموخته‌شده از داده‌ها یعنی مجموعه نوشته‌ها، که مجموعه‌ای از متن نوشتاری یا گفتاری است. در این روش ابتدا یک مدل آماری ساخته‌شده و سپس شباهت تخمین زده می‌شود.

در ارزیابی [۵۰] چترجی(۲۰۰۱) نتایج اولیه یک طرح اندازه‌گیری شباهت را که مبتنی بر یک مدل خطی است و ضرایب آن با استفاده از فن رگرسیون چندگانه تعیین‌شده است، گزارش می‌دهد. داده‌های تجزیه‌وتحلیل از نظرسنجی تعدادی از پاسخ‌دهندگان جمع‌آوری‌شده است. سه جنبه عمده شباهت، یعنی عمل‌گرا، نحوی و معنایی در نظر گرفته‌شده است. از هر مخاطب خواسته‌شده است تا شباهت بین جفت‌های مختلف جملاتی را که با دقت طراحی‌شده‌اند تا یکی از انواع فوق شباهت را

ارزیابی کند. تجزیه‌وتحلیل آماری از این ارزیابی‌ها، درک عمومی انسان را در مورد شباهت اصولی نشان می‌دهد، که به طراحی یک برنامه بازیابی مناسب کمک می‌کند. ترجمه ماشین مبتنی بر مثال[1] مبتنی بر ایده انجام ترجمه با تقلید از مثال‌های ترجمه گذشته است. در این نوع سیستم ترجمه، مقدار زیادی از نمونه‌های ترجمه بین دو زبان در یک پایگاه داده متنی ذخیره می‌شوند. ترجمه جملات بازیابی شده پس‌ازآن تغییریافته برای به دست آوردن ترجمه جمله ورودی داده‌شده مناسب است. بدیهی است، این طرح بستگی به این دارد که طرح بازیابی چقدر خوب و مؤثر است. هرچه جمله بازیابی شده به آن نزدیک‌تر باشد، تطبیق آن با نیاز ترجمه فعلی آسان‌تر خواهد بود و به‌تبع آن، کیفیت کلی ترجمه بهبود می‌یابد. اولین کار در جهت هدف فوق، یافتن مؤلفه‌های کلیدی است که در اندازه‌گیری شباهت بین جملات نقش دارند. یک لیست جامع باید تعیین شود. برای هر یک از این مؤلفه‌ها مقادیر ممکن مشخص می‌شوند و مسافت نسبی آن‌ها تعیین می‌شود.

در ارزیابی [۴۵] شیراکاوا و همکاران(۲۰۱۵) یک روش اندازه‌گیری شباهت معنایی مبتنی بر ویکی‌پدیا ارائه داده‌اند که برای متون کوتاه در دنیای واقعی در نظر گرفته‌شده است. این مقاله روشی را ارائه می‌دهد که نوعی تحلیل معنایی صریح (ESA) است که یک کیف از اشخاص ویکی‌پدیا (صفحات ویکی‌پدیا) را به‌عنوان متن معنایی آن به متن اضافه می‌کند و از وکتور موجودات برای محاسبه شباهت معنایی استفاده می‌کند. اضافه کردن نهادهای مرتبط به یک متن، نه یک کلمه یا عبارت، یک مشکل عملی چالش‌برانگیز است زیرا معمولاً از چندین زیر مشخص تشکیل می‌شود، روش پیشنهادی در این کار با استفاده از نوی بیزین، یک مکانیزم وزن احتمالی مبتنی بر قضیه بیز، این مشکل جمع را حل می‌کند. همچنین در این مقاله به این نتیجه رسیده است که کاهش ابعاد بردار به اشخاص نماینده ویکی‌پدیا به‌ندرت عملکرد را تحت تأثیر قرار می‌دهد درحالی‌که اندازه بردار را کاهش می‌دهد و ازاین‌رو فضای ذخیره‌سازی و زمان پردازش محاسبه شباهت کسینوس را نشان می‌دهد.

هدف پینهیرو و همکاران (۲۰۱۷) در [۵۱] ارائه یک رویکرد جدید برای اندازه‌گیری شباهت بین جملات نوشته‌شده به زبان پرتغالی با استفاده از ویژگی‌های آماری و یادگیری عمیق برای غلبه بر مشکلات معنا بوده است. این مسئله با استفاده از یک رویکرد یادگیری عمیق برای اندازه‌گیری شباهت بین کلمات، بر مشکل معنا غلبه می‌کند. ایده اصلی این است که چهار ویژگی مختلف را بر اساس تطبیق مستقیم[2] Tf-idf، چارچوب کلمات به بردار و اندازه جملات ترکیب کند تا شباهت نهایی را به دست آورد. درواقع این چهار ویژگی مختلف را برای اندازه‌گیری شباهت جملات برای پرتغالی‌ها ترکیب می‌کند، همچنین از چارچوب کلمات به بردار و یک روش مبتنی بر ماتریس برای مقابله با مشکل معنی استفاده می‌کند. علاوه بر این، از یک سیستم رگرسیون برای به دست آوردن مقدار تشابه با ترکیب ویژگی‌های کار شباهت معنایی متن و طبقه‌بندی‌های مختلف یعنی ماشین بردار پشتیبان، نوی بیزین و شبکه عصبی به کار RTE استفاده‌شده است. این الگوریتم ویژگی‌های استخراج‌شده را به‌عنوان ورودی با شباهت پیشنهادی در مجموعه آموزش دریافت می‌کند و خروجی

[1] Example-Based Machine Translation(EBMT)

[2] Direct Matching

تابعی است برای ترکیب این ویژگی‌ها که باعث ایجاد شباهت نهایی در مجموعه آزمون می‌شود. برای تعیین کمیت شباهت نهایی از یک الگوریتم رگرسیون استفاده‌شده است.

در [۵۲] ماجومدر و همکاران (۲۰۱۸) روشی ارائه داده‌اند که قادر خواهد بود میزان تفسیر بین دو جمله را با درجه‌بندی میزان هم‌ارزی معنایی در مقیاس درجه‌بندی‌شده از ۰ (تراز نشده) به ۵ (معادل معنایی) اندازه‌گیری کند. هدف مقاله اضافه کردن یک‌لایه قابل تفسیر نیست بلکه روشی را تدوین کرده است که می‌تواند شباهت‌ها و تفاوت‌های بین این دو جمله را توضیح دهد. آن‌ها برای داده‌های استاندارد به‌دقت ۰/۶۴ برای نوع و نمره تراز رسیده‌اند.

برای توسعه این روش ابتدا تکه‌ها / عبارت را در بین جملات شناسایی کرده، سپس آن‌ها را در بین جملات قرار داده است. در طول تراز عبارت روابط میان آن‌ها را مشخص کرده و پس‌ازآن بر اساس رابطه، نمره شباهت بین عبارت اندازه‌گیری می‌شود. مسئله تشابه متن معنایی قابل تفسیر، به چهار وظیفه فرعی تقسیم‌شده است. اولین کار تولید درختان پارس برای جملات است (یعنی تکه‌ها را مشخص کرده است). وظیفه بعدی ترسیم تکه‌های جملات است و سوم اختصاص دادن استدلال تراز از فهرستی از برچسب‌های احتمالی. سرانجام، نمره همبستگی / شباهت بین عبارت تراز وسط قرار می‌گیرد. ابتدا جملات ورودی را از قبل پردازش کرده و سپس تکه‌های جملات مشخص‌شده‌اند. علاوه بر این، تکه‌ها وارد روند تراز شده‌اند، که می‌تواند آن‌ها را بر اساس نمره مرتبط بودن تنظیم کند. سرانجام، نمره کلی اندازه‌گیری شده است و بر اساس آن‌یک برچسب رابطه اختصاص داده‌شده است.

در [۲۶] فراندز و همکاران(۲۰۱۹)، یک روش تعبیه‌کننده ارائه داده‌اند، که هدف آن یادگیری بازنمایی جملات بدون نظارت از متن بدون برچسب است. در حقیقت یک روش بدون نظارت را پیشنهاد کرده که یک جمله را به‌عنوان یک سری وزن‌دار تعبیه کلمات مدل می‌کند. وزن این سریال با استفاده از اطلاعات متقابل شانون[1] در میان کلمات، جملات و پیکره متناسب است. در این بررسی به بازنمایی جمله‌ها پرداخته تا رفتار آماری معنای جمله را موردمطالعه قرار دهد. آن‌ها با استفاده از داده‌های SICK به همبستگی۰/۷۲۴ رسیده‌اند.

۲.۱.۵.۲ روش‌های مبتنی بر یادگیری عمیق

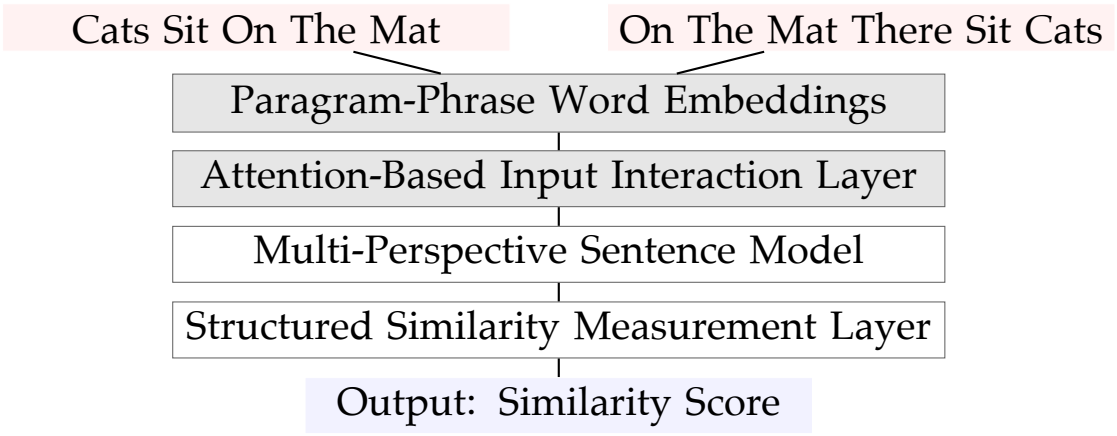
در [۵۳] هی و همکاران (۲۰۱۶) از یک شبکه عصبی کانولوشن مبتنی بر شباهت معنایی متن استفاده کرده‌اند که یک‌لایه تعامل ورودی مبتنی بر توجه را توسعه داده و آن را با شبکه عصبی کانولوشن چندرسانه‌ای که با استفاده از ضمائم کلمه پاراگرام_عبارت[2] آموزش داده‌شده در جفت‌های ترجمه ترکیب کرده است. در اینجا از شبکه عصبی حلقوی چندمنظوره[3] استفاده‌شده است که جملات ورودی را به‌صورت جداگانه تلقی می‌کند و عمدتاً از تعاملات حساس به متن چشم‌پوشی می‌کند. برای برطرف کردن این مشکل لایه تعامل ورودی مبتنی بر توجه را توسعه داده، دو جمله مستقل

[1] Mutual Information(MI)

[2] PARAGRAM-PHRASE

[3] multi-perspective convolutional neural network(MPCNN)

ورودی را به یک جفت جمله مرتبط می‌کند که به مدل کمک می‌کند تا کلمات ورودی مهم را برای اندازه‌گیری بیشتر تشابه انتخاب کند و همچنین از سازگاری کلمات قوی پاراگرام_ عبارت استفاده می‌کند. ساختار کلی مدل در شکل ۲.۲ نشان داده‌شده است.



شکل ۲.۲: نمای کلی مدل

مدل پایه: شبکه عصبی مصنوعی چندبعدی: از مدل چندبعدی شبکه عصبی کانولوشن استفاده کرده است که دو جزء اصلی دارد:

۱. یک مدل چندبعدی جمله برای تبدیل یک جمله به نمایندگی: یک شبکه عصبی کانولوشن اطلاعات مختلف در هر جمله را با استفاده از انواع مختلف فیلترهای کانوولاتور[1]، انواع توزیع و اندازه پنجره‌ها را در برمی‌گیرد.

۲. یک‌لایه اندازه‌گیری شباهت ساختاریافته: از این لایه برای مقایسه مناطق محلی بازنویسی جمله‌ها از متریک‌های شباهت چندگانه استفاده می‌کند. در این بررسی با استفاده از مجموعه داده‌های آزمون رقابت STS انگلیسی SemEval 2016 که شامل پنج مجموعه داده از حوزه‌های مختلف می‌باشد به ضریب پیرسون۸۰۴٪ رسیده است.

مولر(۲۰۱۶) از یک شبکه سیامس که درواقع شبکه طولانی‌مدت را برای داده‌های برچسب‌گذاری‌شده از جفت توالی‌های متغیر طول ارائه داده استفاده کرده است. برای این برنامه‌های کاربردی، ارائه دهنده‌های تعبیه کلمات[2] با اطلاعات مترادف به LSTMها ارائه داده که از یک بردار اندازه ثابت برای رمزگذاری معنی اساسی بیان‌شده در جمله (بدون در نظر گرفتن اصطلاح خاص / نحو خاص) استفاده می‌کند. با محدود کردن عملیات بعدی با تکیه‌بر متریک ساده منهتن، مجازات جمله را که توسط این مدل آموخته‌شده است را مجبور کرده تا یک فضای بسیار ساختاریافته را شکل دهد که هندسه آن روابط معنایی پیچیده را نشان می‌دهد. در این بررسی با استفاده از مجموعه داده‌های SICK به‌دقت ۸۴/۲ و ضریب همبستگی پیرسون ۸۸۲۲٪ دست‌یافته‌اند [۵۴].

در ارزیابی [۵۵] وایتینگ و همکاران (۲۰۱۶) شش معماری ترکیبی را باهم مقایسه می‌کنند،

[1] convolutional
[2] wordembedding

آن‌ها را در مجموعه داده‌های شباهت متنی حاشیه‌نویسی که هر دو از همان توزیع داده‌های آموزش و همچنین از طیف گسترده‌ای از حوزه‌های دیگر ترسیم‌شده است، ارزیابی می‌کند. پیچیده‌ترین معماری‌ها، مانند شبکه‌های عصبی مکرر حافظه کوتاه‌مدت، بهترین عملکرد را در داده‌های درون دامنه‌دارند. بااین‌حال، در سناریوهای خارج از حوزه، معماری‌های ساده مانند کلمه با میانگین بسیار زیاد از LSTMها بهتر است. ساده‌ترین مدل میانگین ما حتی با سامانه‌هایی که برای کارهای خاص تنظیم‌شده‌اند بسیار رقابتی است و درعین‌حال بسیار کارآمد و آسان برای استفاده است. در این بررسی بر روی سه کار تحت نظارت NLP آزمایش صورت گرفته است : شباهت جملات، بازپرداخت و طبقه‌بندی احساسات و با استفاده از مجموعه داده‌های SICK به ضریب همبستگی ۰/۷۰ درصد رسیده است.

زو و همکاران یک مدل D-LSTM مبتنی بر LSTM را پیشنهاد داده‌اند و بر روی‌داده‌های SICK به ضریب همبستگی پیرسون ۰/۸۳۰۵ و ضریب همبستگی اسپرمن ۰/۷۷۲۹ و خطای MSE، ۰/۳۴۴۲ دست‌یافته‌اند[۵۶].

پونتز و همکاران (۲۰۱۸) سامانه‌ای را برای اندازه‌گیری شباهت معنایی جملات، ترکیبی از شبکه عصبی راجعه و عودکننده را طراحی کردند. این سیستم از یک شبکه همخوانی برای در نظر گرفتن زمینه محلی کلمات و یک LSTM برای در نظر گرفتن زمینه جهانی جملات استفاده می‌کند. این ترکیبی از شبکه‌ها به حفظ اطلاعات مربوط به جملات کمک می‌کند و محاسبه شباهت بین جملات را بهبود می‌بخشد. در این بررسی از فن NNها استفاده‌شده است. ابتدا، از CNN سیامس استفاده کرده تا متن محلی کلمات را در یک جمله تحلیل کند و بازنمایی از اهمیت یک کلمه و همسایگی آن ایجاد نماید. سپس، از LSTM سیامس استفاده کرده تا کل جمله را بر اساس کلمات و متن‌های محلی آنالیز کند. در آخر، شباهت معنایی جفت جملات را با استفاده از فاصله منهتن پیش‌بینی نموده است. در انتها به ضریب همبستگی پیرسون ۰/۸۵۴۹ و ضریب اسپرمن ۰/۷۹۳۳ و خطای ۰/۲۸۹۸ دست‌یافته است[۵۷].

در [۵۸] راناسینگه و همکاران از معماری‌های مکرر سیامس برای اندازه‌گیری شباهت معنایی متن استفاده کرده‌اند. در این روش شبکه‌ها دارای بردارهای نمایشی با همان معنای مشابه هستند و باعث مقایسه آسان‌تر آن‌ها می‌شوند. با توجه به اینکه وزن‌ها در شبکه‌های فرعی به اشتراک گذاشته می‌شود، پارامترهای کمتری برای آموزش وجود دارد که به‌نوبه خود به معنای نیاز به داده‌های آموزشی کمتری و تمایل به بیش‌ازحد مناسب بودن است. با توجه به میزان نیروی انسانی موردنیاز برای تولید مجموعه داده برای STS شبکه‌های عصبی سیامی می‌توانند راه‌حل ایده آل برای کار شباهت معنایی متن باشند. راناسینگه و همکاران باکار بر روی مجموعه داده‌ی SICK به ضریب همبستگی ۰/۸۹۰۱ درروش GRU دست‌یافته‌اند. این روش از بقیه روش‌ها نتیجه بهتری داشته است.

۲.۵.۲ روش‌های مبتنی بر معنا

در [۵۹] لی و همکاران (۲۰۰۶) الگوریتمی را ارائه داده‌اند که اطلاعات معنایی و اطلاعات مربوط به ترتیب کلمه را در جملات ذکر کرده، سپس شباهت معنایی دو جمله با استفاده از اطلاعاتی که از یک پایگاه داده ساختاری واژگانی ساخت‌یافته به دست آورده و از آمار بدنه محاسبه‌شده است استفاده می‌کند. استفاده از دیتابیس واژگانی باعث می‌شود که از مفهوم دانش مشترک استفاده کند و درج آمار اطلاعات جسمی، باعث می‌شود تا با حوزه‌های مختلف سازگار باشد. متن به‌عنوان دنباله‌ای از کلمات در نظر گرفته می‌شود که هر یک از آن‌ها اطلاعات مفیدی دارند. کلمات به همراه ساختار ترکیبی آن‌ها باعث می‌شود که متن معنای خاصی را منتقل کند. این روش به‌صورت پویا یک کلمه مشترک را با استفاده از تمام کلمات مجزا در جفت جملات تشکیل می‌دهد. برای هر جمله، یک ویکتور معنایی خام با کمک یک بانک اطلاعاتی واژگانی به دست می‌آورد. برای هر جمله یک ویکتور مرتبه کلمه تشکیل می‌دهد و مجدداً از اطلاعات پایگاه داده واژگان استفاده می‌کند، ازآنجاکه هر کلمه در یک جمله متفاوت به معنای کل جمله است، اهمیت یک کلمه با استفاده از محتوای اطلاعاتی که از یک جسم به دست می‌آید، وزن می‌گیرد. با ترکیب بردار معنایی خام با محتوای اطلاعاتی از بدنه، یک ویکتور معنایی برای هر یک از دو جمله به دست می‌آید. شباهت معنایی بر اساس دو بردار معنایی محاسبه می‌شود. شباهت سفارش با استفاده از دو بردار مرتبه محاسبه می‌شود. سرانجام شباهت جملات با ترکیب شباهت معنایی و تشابه نظم حاصل می‌شود. اندازه‌گیری شباهت الگوریتم این بررسی ضریب همبستگی پیرسون به‌طور معقول خوب ۰٫۸۱۶ با رتبه‌های انسانی، در سطح ۰٫۰۱ معنی‌دار به دست آورد.

در [۴۰] امیر و همکاران (۲۰۱۶) روشی ارائه داده‌اند که روشی کارآمد برای تشخیص پاراگراف‌ها در متون است که با یک عبارت خاص مطابقت دارد. این مقاله از یک معیار شباهت بر اساس هسته‌های معنایی استفاده کرده است. ابتدا از Stanford Parser به‌عنوان یک ابزار NLP برای تفسیر جملات استفاده کرده است. سپس بر اساس روابط دستوری هسته‌های معنایی را تشکیل داده که شامل موضوعات سه‌گانه، فعل‌ها و اشیا است. درواقع هسته‌ها معنای عمومی جملات را بیان می‌کند که انعطاف‌پذیری خاصی برای مقابله با ساختارهای ناهمگن دارند و سرعت محاسبات شباهت معنایی را زیاد می‌کند. بر اساس هسته‌های استخراج‌شده، یک تجزیه‌وتحلیل برای استخراج اطلاعات مرتبط انجام‌شده است. برای انجام این کار از یک روش یادگیری ماشین برای یافتن بهترین راه برای جمع‌کردن ویژگی‌ها در یک مدل معنایی استفاده‌شده است. در این بررسی خطای استاندارد باقی‌مانده[1] ۰٫۱۶ تعیین شد.

ویکی‌پدیا یک مخزن دائره‌المعارف مستقل از دامنه و یک شبکه معنایی برای محاسبه شباهت معنایی بین مفاهیم را فراهم می‌کند. اقدامات مبتنی بر ویژگی‌های سنتی به ترکیب خطی از خصوصیات مختلف با دو محدودیت اصلی، اطلاعات ناکافی و از بین رفتن اطلاعات معنایی متکی است. در این بررسی، چندین روش اندازه‌گیری شباهت معنایی هیبریدی را با استفاده از محتوای اطلاعات و

[1] Residual standard error(RSE)

ویژگی‌های مفاهیم ارائه داده است، که از محدودیت‌های معرفی‌شده در بالا جلوگیری می‌کنند. با توجه به ادغام خصوصیات گسسته در یک جزء، دو مدل بازنمایی معنایی به نام‌های CORM و CARM را ارائه داده، سپس، شباهت معنایی را بر اساس این مدل‌ها محاسبه نموده است و IC مقوله‌ها را به‌عنوان مکمل اندازه‌گیری شباهت معنایی در نظر گرفته است. به‌طور خلاصه، رویکردهای مورداستفاده در این بررسی در تعیین شباهت معنایی بین مفاهیم کارآمدتر بوده و از همبستگی انسانی بهتری نسبت به روش‌های قبلی مانند word2vec و NASARI برخوردار است. این روش ابتدا از همسایگان به‌عنوان یک ویژگی اضافی برای تقویت نمایش مفاهیم استفاده می‌کند. ثانیاً، چند جفت دسته مشابه از دو مفهوم را انتخاب کرده و برای ایجاد ویژگی‌های معنایی اضافی با کمک اقدامات مبتنی بر IC، دو زیر گراف را بناکرده است. سوم، ویژگی‌های همسایگان و دسته‌ها را در یک بازنمایی مفهوم ادغام نموده است. سرانجام، چندین فرمول را برای اندازه‌گیری شباهت بین دو مفهوم بر اساس مدل‌های نمایش معنایی پیشنهاد کرده که باعث شده است، اطلاعات گمشده در یک ویژگی با استفاده از دیگر خصوصیات مرتبط تکمیل شود و اطلاعات معنایی نهفته در بین خصوصیات مختلف، استخراج و در اندازه‌گیری شباهت معنایی استفاده شود[۲۳].

شباهت معنایی متن می‌خواهد درجه همگرایی بین دو جمله را ارزیابی کند. بیشتر معنای سطحی کلمات را در نظر می‌گیرد ولی همبستگی و ارتباط مفهومی دو کلمه را حساب نمی‌کند. برای حل این مشکل از مفهوم انتقال فضا استفاده می‌شود که از مفاهیم اسم با روابط is-a در شبکه واژگان استفاده می‌کند که این مدل هم فقط اسم‌ها را در نظر می‌گیرد و فعل و صفت و عبارات دیگر را نادیده می‌گیرد درنتیجه اطلاعات از بین می‌رود. راه‌حل ارائه‌شده برای حل این مشکل باید از پیوندهای مشتق شده در شبکه واژگان برای مرتبط کردن فعل‌ها و صفت‌هاوغیره با مفاهیم مربوط به اسم آن‌ها استفاده کرد و همچنین برای بهبود دقت مدل مبتنی بر مفهوم انتقال فضا اهمیت انواع مختلف کلمات را با توجه به تعیین وزن‌های آن‌ها تعیین می‌کند.

برای بهره‌برداری از مفهوم جملات محتوای اطلاعات را معرفی کرده، که برای اندازه‌گیری مفاهیم اسم به کار می‌رود. محتوای اطلاعات در مقایسه با روش‌های دیگر دو تا مزیت دارد. ۱) طبقه‌بندی مبتنی بردانش یعنی رابطه‌ی is-a از مفاهیم در شبکه واژگان می‌تواند شباهت‌های ذهنی و ارزیابی شباهت معنایی را نشان دهد. ۲) تعیین مفهوم اطلاعات با روش آماری با شهود انسانی سازگار است مثلاً هسته کلمه شامل اطلاعات بیشتری نسبت به خود کلمه است.

بااین‌حال روش محتوای اطلاعات هم به دلیل مشکل محاسبه هم‌پوشانی مفهومی در میان مفاهیم چندگانه به‌ندرت استفاده می‌شود. به‌طور خاص اسامی در یک جمله متعلق به مفاهیم مختلف نقشه‌برداری می‌شوند. که یا به مفاهیمی که دارای ارتباط زیرین هستند یا مفاهیم پراهمیت هستند. برای رفع مشکل هم‌پوشانی محتوای اطلاعات پیشنهاد نمایشی ازجمله دانش مبتنی بر دانش یعنی مفهوم فضای انتقال داده‌شده است و برای رفع پیچیدگی محاسباتی متد فوق برای محتوای اطلاعات متنی روش مبتنی برافزایش اطلاعات برای کاهش پیچیدگی محاسبات محتوای اطلاعات پیشنهادشده است. سپس شباهت معنایی جملات را می‌توان با اندازه‌گیری نسبت محتوای اطلاعات مشترک بین آن‌ها محاسبه کرد. بااین‌حال بسیاری از روش‌های مبتنی بر دانش فقط مفاهیم

را در یک سطح در نظر می‌گیرند که باعث از بین رفتن اطلاعات زیادی در جملات می‌شود که دو دلیل اساسی دارد:

۱. فقدان محتوای اطلاعات کلمات نامنظم در شبکه واژگان. چون تنها مفاهیم اسم قادر به ایجاد یک ساختار سلسله مراتبی هستند و مفاهیم فعل و قید و صفت و غیره. در زیر شبکه‌های چندگانه وجود ندارند.

۲. فقدان واحدهای نامشخص (oovNEs): تعداد نهادهای نام‌گذاری شده در شبکه واژگان کاملاً محدود است. ازاین‌رو مقادیر محتوای اطلاعات تعداد زیادی از کلمات را نمی‌توان مستقیماً به دست آورد.

به‌منظور حل چالش‌های فوق مدل انتقال فضا را مبتنی بر این جنبه‌ها توسعه داده است. ۱) لینک‌های مشتق شده در شبکه واژگان برای مرتبط کردن فعل‌ها صفت‌ها و واژگان با مفاهیم اسم مربوط به هرکدام استفاده می‌شود. ۲) علاوه بر این ovvNEها توسط نرم‌افزارهای مانند stanford corenlp کشف‌شده است. که برای تشخیص دو کلمه شبیه به هم ولی بامعنای متفاوت مثل کیف و کیف به کار می‌رود که افزایش اطلاعات جملات توسط ovvNEها بر اساس گرایش یاد گرفته‌شده از NEهای شناخته‌شده برای کاهش بیشتر اطلاعات ازدست‌رفته استفاده می‌کند. علاوه بر این برای بهبود دقت مدل مبتنی بر مفهوم انتقال فضا اهمیت انواع مختلف کلمات را برای ارزیابی شباهت معنایی با توجه به تعیین وزن‌های به آن‌ها اندازه می‌گیرد. روش CTS بر روی‌داده‌های لی و همکاران به همبستگی ۰/۸۶۷۲ رسیده است [۵].

در ارزیابی [۶۰] پاوار(۲۰۱۸)روشی را ارائه داده است که با استفاده از یک الگوریتم قبلاً ایجادشده و ادغام آن با قشر دامنه با استفاده از آمار دامنه می‌پردازد. دامنه تفکیک‌شده به‌عنوان داده یادگیری نظارت‌شده برای الگوریتم عمل می‌کند. همچنین برای محاسبه شباهت بین افعال عمل در اهداف یادگیری با اشاره به طبقه‌بندی بلوم، شاخص Bloom را معرفی کرده است. طبقه‌بندی بلوم دستور العمل‌های کلیدی کلمه کلیدی را ارائه می‌دهد و یک ساختار سلسله مراتبی را هنگام تعیین نتایج یادگیری مورداستفاده قرار می‌دهد. این روش در ابتدا تشابه معنایی بین کلمات را با توجه به معنی کلمه در متن جمله محاسبه می‌کند. سپس برای شکل دادن یک بردار معنایی برای هر دو جمله به‌طور جداگانه استفاده می‌شود. بردارهای معنایی شکل‌گرفته برای محاسبه شباهت معنایی استفاده می‌شوند. وکتور کلمه سفارش با در نظر گرفتن ساختار نحوی جملات، یعنی وقایع کلمات مربوط به یکدیگر ساخته می‌شود. طبقه‌بندی بلوم یک مدل ساختاری مشهور، تثبیت‌شده و سلسله مراتبی است که شامل افعال اکشن در سطوح مختلف سلسله‌مراتب است. هرچه سلسله‌مراتب را بالا می‌بریم، سطح فعل عمل افزایش می‌یابد. سه لایه برتر، آنالیز، سنتز و ارزیابی افعال را با تفکر انتقادی نشان می‌دهد. در اینجا طبقه‌بندی بلوم را به‌عنوان بخش جداگانه‌ای از روش‌شناسی پیاده‌سازی کرده است و آن را در استفاده از متدلوژی شباهت جمله اصلی محدود نموده است.

در ارزیابی [۲۴] تی و همکاران(۲۰۱۹)هر متن کوتاه را به‌عنوان دو بردار متراکم نشان داده‌اند: اولی با استفاده از شباهت کلمه به کلمه مبتنی بر بردارهای کلمه از قبل آموزش‌دیده، دومی با استفاده

از تشابه کلمه به کلمه مبتنی بر منابع خارجی دانش ساخته‌شده است. همچنین یک الگوریتم پیش‌پردازش ایجاد کرده، که زنجیره‌ای از نام‌های اصلی را باهم متصل کرده و تقسیم کلمه را برای حفظ معنای افعال اصطلاحات و عبارات انجام داده است. روش پیشنهادی از شباهت بین متن‌های کلمه ساخته‌شده با استفاده از کلمات جاسازی‌شده و ارتباط معنایی بین مفاهیم مبتنی بر منابع خارجی دانش بهره‌برداری می‌کند. همچنین این دو قطعنامه هسته‌ای اشخاص نام‌گذاری شده را در دو متن کوتاه انجام می‌دهد تا موجودات را به‌هم‌پیوسته و تقسیم‌بندی کلمات را برای حفظ معنای افعال اصطلاحات و عبارات حفظ کند. درنتیجه ضریب همبستگی پیرسون ۸۳۷۸٪ را به دست آورده است.

۳.۵.۲ روش‌های مبتنی بر رشته

سیستم ukp از مدل رگرسیون خطی ورودی برای آموزش داده‌ها استفاده می‌کند تا چندین متن را با پیچیدگی‌های مختلف ترکیب کند. این سیستم با استفاده از N ـ گرم کلمات به تحلیل مقادیر بردار تجزیه‌وتحلیل صریح معناشناختی و تجمیع شباهت کلمه بر اساس منابع واژگانی ـ معنایی می‌پردازد. از سیستم جایگزینی واژگان و ترجمه ماشین آماری برای اضافه کردن لکسیم‌های اضافی استفاده کرده است تا شکاف‌های واژگانی را کاهش دهد. به‌منظور مقایسه‌ی شباهت کلمات دوبعدی در سطح متن، از استراتژی ارائه‌شده توسط میهالکا (Mihalcea) استفاده‌شده است که با استفاده از مجموع نمرات idf وزن‌دار شباهت بین هر کلمه با همتای خودش در متن دیگر را به دست آورده است. از جایگزینی واژگانی و ترجمه ماشین آماری و سپس با استفاده از همین ویژگی‌ها اما با تعداد تصادفی برای جفت جملات استفاده کرده است [۶۱].

در بررسی [۶۲] آتوم و همکاران (۲۰۱۶) اقدامات شباهت کلمه قبلی در مجموعه داده‌های معیار موردبررسی قرارگرفته و سپس از یک شباهت کلمه ترکیبی در یک اندازه‌گیری تشابه متن رمان (TSM) استفاده‌شده است. TSM پیشنهادی مبتنی بر محتوای اطلاعات و روابط معنایی شبکه واژگان است. TSM شامل تطابق دقیق کلمات، طول هر دو جمله در یک جفت و حداکثر شباهت بین یک کلمه و متن مقایسه شده است. این مقاله یک اندازه‌گیری تشابه متن جدید را بر اساس اندازه‌گیری شباهت کلمه از فاصله مشترک با محتوای مشترک و همچنین محتوای اطلاعات کلمات مقایسه شده ارائه کرده است. شباهت متن پیشنهادی بر اساس طول متن مقایسه شده و کل کلمات دقیق وزن می‌شود. اندازه‌گیری شباهت بیشتر از اقدامات شباهت مقایسه شده را دارد و در سطح ۰٫۰۵ معنی‌دار است. دلیل دستیابی زیاد به این روش ناشی از به‌کارگیری اطلاعات تکمیلی (کالبد و محتوای اطلاعات) و اثربخشی اندازه‌گیری شباهت کلمه وام شده است.

در [۶۳] بابا و همکاران (۲۰۱۷) از نمایندگی توزیع کلمات برای تعریف یک سند مورد ارزیابی، استفاده کرده‌اند. این تحقیق یک روش تشخیص سرقت ادبی را بر اساس حداکثر مقدار طول جمله یعنی طولانی‌ترین توالی مشترک با وزن تعریف‌شده توسط نمایندگی توزیع پیشنهاد کرده است. که این روش در برنامه‌های کاربردی که نیاز به شناسایی دقیق سرقت‌های پیچیده دارند مفید است.

نمایندگی توزیع‌شده به‌عنوان یک تابع در نظر گرفته می‌شود که یک کلمه را به یک بردار با یک بعد کوچک ترسیم می‌کند و فاصله بین بردارها یک شباهت بین کلمات است که با بردارها مطابقت دارد. نمایش ساده توزیع‌شده با کاهش ابعاد یک نمایشگر بردار ساده بر اساس فرکانس یک کلمه در دسترس است. در این روش به‌دقت ۸۳/۷۰ در صد برای داده‌های wllcs رسیده است.

شباهت متنی معنایی و استنباط زبان طبیعی دو وظیفه رایج در زمینه زبان طبیعی است که برای محک زدن مدل‌های بیان جمله درجایی که دو جمله در یک زوج قرار دارند استفاده می‌شود. در چنین وظایفی جملات به‌عنوان کیسه کلمات، توالی‌ها، درختان یا حلقه‌ها ارائه می‌شوند، اما مدل موردنظر به زوج‌های کلمه‌ای مبتنی است. نتایج به‌دست‌آمده در این بررسی در پنج مجموعه داده با توجه به مدل توجه به کلمه محور، خطا را تا ٪۴۱ کاهش داده است [۹].

۴.۵.۲ روش‌های ترکیبی

در [۶۴] اسلام و همکاران (۲۰۰۸)یک روش برای اندازه‌گیری شباهت معنایی متون با استفاده از یک اندازه‌گیری مبتنی بر بدنه از شباهت کلمه معنایی و یک نسخه عادی و اصلاح‌شده از طولانی‌ترین نتایج مشترک الگوریتم رشته ارائه داده است. همچنین بر محاسبه شباهت بین دو جمله یا دو پاراگراف کوتاه متمرکز می‌باشد. روش پیشنهادی این مقاله شباهت دو متن را از اطلاعات معنایی و نحوی (ازنظر نظم کلمه مشترک) که در آن وجود دارد، تعیین می‌کند. درواقع سه کارکرد شباهت را در نظر گرفته است، ابتدا شباهت رشته و شباهت کلمه معنایی محاسبه‌شده و سپس از یک تابع تشابه ترتیب کلمه اختیاری استفاده کرده است تا بتواند اطلاعات نحوی را درروش بیان‌شده بگنجاند. در انتها، شباهت متن با ترکیب شباهت رشته، شباهت معنایی و تشابه نظم کلمه مشترک با عادی‌سازی حاصل می‌شود. آن‌ها به ضریب همبستگی پیرسون۰٫۸۵۳ و دقت۷۲/۴۲ بر روی مجموعه داده‌ی STS-65 دست‌یافته‌اند.

روستایی و همکاران (۲۰۱۴) از پردازش زبان طبیعی و فن‌های یادگیری ماشین برای استفاده از یک طرح شباهت معنایی جمله استفاده کرده‌اند. در بخش اول این طرح مجموعه‌های مختلف ویژگی‌های زبانی شامل ویژگی‌های مبتنی بر رشته، مبتنی بر معنا، مبتنی بر اسم و مبتنی بر نحو استخراج‌شده است. علاوه بر این، شبکه واژگان برای استخراج اطلاعات معنایی استفاده می‌شود. در بخش دوم الگوریتم‌های یادگیری ماشین‌های مانند k ـ نزدیک‌ترین همسایه[1]، مدل رگرسیون لجستیک و ماشین‌های بردار پشتیبان برای ساخت مدل‌های طبقه‌بندی بر روی مجموعه‌ای از ویژگی‌های استخراج‌شده استفاده می‌شوند. نتایج تجربی در بخش اول نشان می‌دهد که ویژگی‌های استخراج‌شده برای شباهت معنایی جمله معتبر است. علاوه براین با مقایسه عملکرد الگوریتم‌های مختلف طبقه‌بندی در بخش دوم الگوریتم KNN موفق‌تر است. این طرح ازنظر دقت برای محاسبه تشابه معنایی بهتر عمل می‌کند. برای ارزیابی عملکرد روش پیشنهادی، سه مجموعه داده‌ی استاندارد متفاوت استفاده می‌شود. مجموعه داده‌های MSRP، RTE3، TREC9 هستند. MSRP شامل

[1] K-Nearest neighbor (KNN)

۴۰۷۶ جفت آموزش و ۱۷۲۵ جفت آزمون است. RTE3 شامل ۸۰۰ جمله جفت است. TREC9 حاوی ۱۹۳ سؤال پارافرز ساخته‌شده توسط ارزیابی‌های انسانی با استفاده از سؤالات اصلی استخراج‌شده از پرسش پرس‌وجو از سؤالات ارسال‌شده توسط کاربر است. در این بررسی نشان داده‌شده است که الگوریتم‌های یادگیری ماشین به‌ویژه KNN و رگرسیون لجستیک ازنظر دقت از روش‌های دیگر بهتر هستند[۶۵].

در مقاله [۸] ولی و همکاران(۲۰۱۶) کار را با پیش‌پردازش جفت‌های جمله شروع می‌کنند؛ سپس آن را با انتساب نمرات شباهت، نمرات واژگانی، معنایی و نحوی و درنهایت با نمره کل محاسبه با استفاده از یادگیری تحت نظارت به پایان می‌رسانند. آن‌ها بر روی مجموعه داده‌های عربی به ضریب همبستگی۰/۹۲ رسیدند.

در مقاله[۷] ریچالسکاو همکاران (۲۰۱۶) برای شناسایی تشابه از ترکیبی از رمزگذار خودکار بازگشتی با یک سیستم جریمه‌ـ‌جایزه[1] که بر اساس شبکه واژگان برای ارتباط معنایی و یک طبقه‌بندی کننده ماشین بردار پشتیبان، که نتیجه نهایی را از ماتریس‌های تشابه تولید می‌کند استفاده کرده‌اند. این راه‌حل بیشتر توسط یک طبقه‌بندی گروه پشتیبانی می‌شود، تراز کننده را با یک شبکه عصبی دوطرفه گیت مکرر و ویژگی‌های اضافی ترکیب نموده‌اند، سپس از رگرسیون بردار خطی پشتیبانی را برای تعیین مجموعه دیگری از امتیازات استفاده کرده‌اند. اولاً رمزگذار خودکار بازگشتی برای انجام یادگیری بر روی درختان تجزیه استفاده می‌شود. سپس شبکه واژگان فاصله بردارهای رمزگذار خودکار بازگشتی را با استفاده از سیستم جریمه‌ـ‌جایزه بر اساس شباهت‌های معنایی کلمات تنظیم می‌کند.

سیستم MayoNLP دو روش برای ارزیابی شباهت معنایی جملات به‌کاربرده است: اولین روش تطبیق متن بر اساس معادلات معنایی با استفاده از یک شبکه معنایی و پیکره آماری است و دومین روش یک مدل یادگیری ماشین است که از یک مدل معنایی یادگیری عمیق استفاده کرده است. این دو روش را به‌صورت خطی ترکیب کرده‌اند. یک بردار لغت مشترک JWV با جمع‌آوری کلمات منحصربه‌فرد که در یک جفت جمله وجود دارد تشکیل‌شده است. شباهت معنایی با مقایسه معنای پنهان در جملات محاسبه می‌شود و شباهت نحوی بر اساس ترکیب کلمه در دو جمله محاسبه می‌شود. این بررسی به ضریب همبستگی۰/۸۰۱۵ در بهترین اجرا دست‌یافته است[۶۶].

در [۳۳] دو اجرا برای شباهت معنایی متن تک‌زبانه ارسال‌شده است، یکی مبتنی بر یادگیری نظارت‌شده و دیگری سیستم بدون ناظر می‌باشد.

- **UWB sup**: سیستم نظارت‌شده بر پایه رگرسیون SVM با هسته rbf که از تمام فن‌های استفاده‌شده در sts به‌عنوان ویژگی‌های رگرسیون استفاده کرده است تا وابستگی بین ویژگی‌های یکسان را بهتر مدل کند.
- **UWB unsup**: سیستم بدون نظارت بر اساس همگرایی کلمات وزن‌دار با استفاده از انجام کار متقابل زبانی شباهت معنایی متن را با دو زبان اسپانیایی و انگلیسی در دو مرحله انجام داده است. اول جملات اسپانیایی را از طریق مترجم گوگل به زبان انگلیسی ترجمه کرده است. جملات انگلیسی

[1] award-penalty

ثابت باقی‌مانده، سپس از سیستم شباهت معنایی متن برای کار با تک‌زبانه‌ها استفاده کرده است. این بررسی بر روی مجموعه داده‌های STS2016 به ضریب همبستگی پیرسون ۰/۸۱۱۶ دست‌یافته است.

در [۳] کای و همکاران(۲۰۱۷) یک رویکرد ترکیبی مبتنی بر شبکه واژگان CSSM-ICSP برای اندازه‌گیری شباهت معنایی مفهوم پیشنهاد داده‌اند که از مفهوم اطلاعات مفاهیم برای وزن کمترین فاصله بین مفاهیم استفاده کرده است. این مقادیر مفهوم اطلاعات مشترک را در محاسبه کمترین فاصله مسیر بین مفاهیم مقایسه شده ادغام می‌کند. برای تعیین کمیت مفهوم اطلاعات مفاهیم، در درجه اول با در نظر گرفتن خصوصیات معنایی و ساختارهای مختلف شبکه واژگان، یک مدل مفهوم اطلاعات ذاتی ایجادشده است. در این مقاله، یک روش ترکیبی را برای محاسبه شباهت معنایی بین جفت کلمات بر اساس شبکه واژگان پیشنهاد کرده است. معیار پیشنهادی غیرخطی فاصله مسیر بین مفاهیم را برای اندازه‌گیری شباهت معنایی آن‌ها تغییر داده است. در این روش، وزن مفهوم اطلاعات مسیر به‌عنوان يك عامل محاسبه در نظر گرفته‌شده است که در آن اختلاف مقادیر مفهوم اطلاعات بین مفاهیم مجاور براي وزن کردن لبه پیونددهنده مفاهیم به‌کاررفته است. با توجه به روش‌های مختلف شناسایی فاصله معنایی بین مفاهیم، این رویکرد به‌عنوان دو استراتژی ارائه می‌شود: استراتژی ۱ از کوتاه‌ترین مسیری عادی استفاده می‌کند و استراتژی ۲ از همپوشانی مسیر بین دو مفهوم مقایسه شده و کمترین مشترک آن‌ها استفاده می‌کند.

رویکردهای مبتنی بر محتوای اطلاعات به مدل‌های آماری ارزش مفهوم اطلاعات مفاهیم به‌عنوان فاکتورهای محاسبه تکیه کرده‌اند، که برای ارزیابی چگونگی یک مفهوم خاص و آموزنده استفاده‌شده است. بر اساس بردارهای ویژگی، رویکردهای مبتنی بر ویژگی روی ویژگی‌های مشترک یا ویژگی‌های یکسان مفاهیم تطبیقی متمرکز می‌شوند. در کنار ظهور معیارهای مختلف شباهت، ادغام روش‌های مختلف برای غلبه بر محدودیت یک رویکرد واحد امکان‌پذیر است. در اینجا یک روش ترکیبی مبتنی بر شبکه واژگان را برای ارزیابی شباهت معنایی بین مفاهیم ایجاد کرده است. برخلاف رویکردهایی که نیاز به یک مقیاس بزرگ همراه با اصطلاحات حوزه گرا دارند، این رویکرد فقط از شبکه واژگان استفاده می‌کند و به‌جای تراز هستی‌شناسی بر شباهت معنایی مفاهیم در شبکه واژگان تمرکز می‌کنند. این به‌طور عمده به: اولاً تجزیه‌وتحلیل مقایسه‌ای ویژگی‌های فنی رویکردهای شباهت موجود در مورد عوامل محاسبه ، مزایا و اشکالات علاوه بر دقت پرداخته است. ثانیاً ارائه یک مدل جدید از IC ذاتی مفاهیم، که دو ویژگی معنایی شبکه واژگان را در نظر می‌گیرد. ثالثاً به‌صورت غیرخطی فاصله مسیری با وزن IC را به شباهت بین مفاهیم مقایسه شده توسط یک عملکرد نمایی تبدیل کنید. این مدل بر روی‌داده‌های M&C به ضریب همبستگی پیرسون ۰/۸۹۲ دست‌یافته است.

در بررسی [۶] ورداسبی و همکاران(۲۰۱۷) با به کار بردن سه ویژگی بر روی روش هم‌ترازی به نتایج خوبی در تخمین شباهت معنایی متون رسیده‌اند. دو مورد از ویژگی‌های پیشنهادی این مقاله برای تجزیه‌وتحلیل ویژه‌ای هستند، که برای شباهت یابی واژه‌های معنایی بهتر عمل می‌کنند. با محاسبه شباهت یک جفت جمله، نمره سطح همبستگی بین آن دو جمله به دست می‌آید. برای این کار نیاز است تا فراتر از تشابه واژگانی پرداخته و شباهت معنایی جملات را در نظر بگیرند. یکی از

ابزارهایی که به‌طور گسترده برای محاسبه شباهت کلمات به کار می‌رود شبکه واژگان است. مخزن دانش به‌عنوان یک شبکه معنایی روابط نحو را ذخیره می‌کند. علاوه بر شبکه واژگان یکی دیگر از روش‌های مناسب برای تخمین شباهت معنایی کلمه، روش متداول شبکه مبتنی بر word2vec است. در حقیقت word2vec یک کلمه تعبیه‌شده را در فضای بردار ایجاد می‌کند و شباهت معنایی بین دو کلمه به‌سادگی با محاسبه شباهت کوسین از کلمات بازتاب‌های بردار محاسبه می‌کند. روش مشخص‌شده از نمایش word2vec استفاده می‌کند و بردارهای کلمات داخل یک جمله را برای تولید ماتریس جمله تکرار می‌کند. به‌این‌ترتیب وظیفه اندازه‌گیری شباهت جمله، به محاسبه شباهت ماتریس‌ها کاهش می‌یابد. یکی از ویژگی‌های به‌کاربرده شده برای ارزیابی شباهت جمله، خطای تخمین خطی[1] است که نتیجه خوبی در برآورد خطی یک ماتریس دارد. همچنین از یک ترازنامه تک‌زبانه استفاده کرده تا به‌طور صحیح کلمه مشابهی از جفت جملات را باهم جمع کند و برای هر شباهت تخمینی ارزشی ارائه دهد.

در این روش از دو منبع شباهت کلمات به‌عنوان ورودی ترازنامه استفاده‌شده است:

۱. شباهت کوسین از نمایش برداری کلمات حاصل از word2vec

۲. عناصر متناظر ماتریس شباهت کلمه لغات معنی لغوی

از ترکیب سه ویژگی خطای تخمین خطی، w2vta و spata با یک سیستم Tf-idf نتیجه برآورد شباهت معنایی به‌دست‌آمده است.

مناسب‌ترین سیستم برای مقایسه فن‌های مختلف بر روی پنج زبان مختلف(انگلیسی، فارسی، آلمانی، ایتالیایی و اسپانیایی) Luminoso بود که یک سیستم ترکیبی است که به‌طور مؤثر کلمات و اطلاعات از منابع دانش را ادغام می‌کند[۶۷].

در مقاله[۶۸] پاگلیاردینی و همکاران(۲۰۱۸) یک روش ریاضی، محاسباتی کارآمد الهام گرفته از CBOW را برای آموزش جمله به‌جای تعبیه کلمات پیشنهاد داده‌اند. این مدل، یک مدل ساده بدون نظارت است که اجازه می‌دهد تعبیه کلمات را با استفاده از بردارهای کلمه‌ای همراه با تعبیه N_گرم‌ها هم‌زمان با ترکیب آموزش و خود بردارهای تعبیه‌شده انجام دهد. پیچیدگی محاسباتی تعبیه‌های این مدل در عملیات برداری، در هر کلمه پردازش‌شده، هم در حین آموزش و هم در استنتاج تعبیه‌های جمله فقط O(۱) است. آن‌ها بر روی مجموعه داده‌های SICK به ضریب پیرسون ۰/۷۱۵ رسیده‌اند. در این پژوهش از مدل Sent2Vec نیز بر روی‌داده‌های SICK استفاده کرده‌اند.

در[۶۹] پاوار(۲۰۱۸) متدلوژی ارائه داده‌شده است که با درج تشابه معنایی و آمار پیکره، به n محاسبه شباهت معنایی بین جملات می‌پردازد. برای محاسبه شباهت معنایی بین کلمات و جملات، روش پیشنهادی از یک رویکرد مبتنی بر لبه با استفاده از بانک اطلاعاتی واژگان پیروی می‌کند. این روش در هر دو معیار و میانگین مجموعه داده‌های شباهت انسان آزمایش‌شده است. روش پیشنهادی از پایگاه داده شبکه واژگان برای مقایسه معنای مناسب کلمه استفاده می‌کند. یک وکتور معنایی برای هر جمله تشکیل‌شده است که شامل وزن اختصاص داده‌شده به هر کلمه برای هر کلمه دیگر ازجمله دوم در مقایسه است. این مرحله همچنین محتوای اطلاعات کلمه را در نظر

[1] Linear estimation error(LEE)

می‌گیرد، به‌عنوان‌مثال فرکانس کلمه از یک کالای استاندارد. شباهت معنایی بر اساس دو بردار معنایی محاسبه می‌شود. برای هر جمله یک بردار مرتبه تشکیل می‌شود که شباهت نحوی بین جملات را در نظر می‌گیرد. سرانجام، شباهت معنایی بر اساس بردارهای معنایی و بردارهای ترتیب محاسبه می‌شود. در این ارزیابی برای تشابه کلمه به ضریب همبستگی پیرسون ۸۷۵۳٪ رسیده‌اند و برای شباهت جمله، همبستگی به‌دست‌آمده ۸۷۹۴٪ است.

این الگوریتم در ابتدا هر دو جمله را متمایز می‌کند و در بخش‌های گفتار آن‌ها را برچسب‌گذاری می‌کند. این رویکرد، تفکیک معنای صحیح کلمه مقایسه را تضمین می‌کند. شباهت بین کلمات بر اساس یک رویکرد مبتنی بر لبه قبلاً محاسبه می‌شود. محتوای اطلاعاتی از یک قشر می‌تواند برای تأثیر شباهت در حوزه خاص استفاده شود. بردارهای معنایی حاوی شباهت‌ها بین کلمات برای جملات تشکیل‌شده و بیشتر برای محاسبه تشابه جمله استفاده می‌شود. بردارهای ترتیب کلمه نیز برای محاسبه تأثیر ساختار نحوی جملات تشکیل‌شده‌اند. ازآنجاکه نظم کلمه بر شباهت کلی کمتر از شباهت معنایی تأثیر می‌گذارد، شباهت ترتیب کلمات به میزان کمتری وزن می‌شود. این متدلوژی بر روی مجموعه داده‌های قبلاً ایجادشده آزمایش‌شده است که حاوی نتایج استاندارد و همچنین میانگین نتایج انسانی است.

۵.۵.۲ سایر روش‌ها

در [۷۶] به ازای هر جمله یک بردار شباهت معنایی و یک بردار شباهت چیدمان کلمات تشکیل می‌شود و شباهت بین دو جمله با ترکیب وزن‌دار این دو بردار را به دست آورده است. الگوریتم این مقاله جهت محاسبه میزان شباهت میان کلمات به دلیل نبودن پیکره کامل در زبان فارسی، از پیکره اینترنتی استفاده کرده است . سپس با کمک این مقادیر و ترکیب دو بردار شباهت معنا و چیدمان به ترکیب نقش نحو معنا پرداخته است. درنهایت نتایج به‌دست‌آمده را در سه سطح فازی دسته‌بندی کرده است این روش بر روی ۵۰ جمله از اخبار برگزیده فارسی پیاده‌سازی شده است .

در[۷۷] با استخراج ویژگی متن کلمات کلیدی را استخراج می‌کند و با استفاده از تجزیه‌کننده‌های متنی درخت‌های تجزیه مربوط به اسناد را به دست آورده و با استفاده از وزن کلمات کلیدی به‌دست‌آمده میزان شباهت بین درخت‌ها را تخمین میزند. برای بررسی میزان شباهت بین درختان از الگوریتم به دست آوردن تعداد زیر درختان مشابه استفاده کرده‌اند . در انتها با استفاده از الگوریتم خوشه‌بندی سلسله مراتبی k ـ نزدیک‌ترین همسایه گروه‌بندی سندهای متنی را انجام داده است.

در[۷۰] جان ویتینگ و همکاران (۲۰۱۶) با استفاده از تعبیه‌های CHARAGRAM، یک رویکرد ساده برای یادگیری مدل‌های ترکیبی شخصیت پایه را برای تعبیه توالی متن ارائه دادند. یک کلمه یا جمله با استفاده از یک بردار شمارش N ـ گرم کاراکتر نمایش داده می‌شود و به دنبال آن‌یک دگرگونی غیرخطی برای انجام تعبیه کم بعدی انجام می‌شود. برای ارزیابی از سه روش استفاده کرده است: تشابه کلمه ، تشابه جمله و برچسب زدن بخشی از گفتار. در این بررسی به این نتیجه رسیده‌اند که تعبیه‌های CHARAGRAM از معماری‌های پیچیده‌تر بر اساس شبکه‌های عصبی مکرر و

دستیابی به عملکرد جدیدترین روش در چندین کار مشابهت، بهتر است. آن‌ها به ضریب همبستگی پیرسون۶۳/۳۳ بر روی مجموعه داده‌ی sl999 و۵۳/۸۷ روی مجموعه داده‌ی ws353 دست‌یافته‌اند.

رکیب و همکاران(۲۰۱۷) به ترتیب ۲ اقدام مرتبط با کلمه موجود را مبتنی بر Google N-gram و Global Vectors برای نمایندگی کلمه اتخاذ نموده‌اند و آن‌ها را به روش متفاوتی با یک روش مرتبط بودن جمله مبتنی بر عبارت N ـ گرم مبتنی بر Google برای محاسبه مرتبط بودن متن ترکیب کرده‌اند. ترکیب ارتباط Google با کلمه گرام و عبارت Google به‌تنهایی با دستیابی به میانگین وزنی بالاتر پیرسون، یعنی ۰/۶۳۹ و ۰/۶۱۹ منجر شده است. به‌طور مشابه، ترکیبی از ارتباط کلمه مبتنی بر GloVe و مرتبط بودن با عبارت N ـ گرم مبتنی بر Google عملکرد بهتری نسبت به کلمه مبتنی بر GloVe دارد و با دستیابی به میانگین وزنی بالاتر پیرسون، یعنی به ترتیب، برابر ۰/۶۱۹ و ۰/۶۰۵ در همان ۱۴ مجموعه داده رسیده است[۷۱].

در مقاله [۷۲] یک روش بدون ناظر به نام رویکرد منبع نور[1] برای اندازه‌گیری شباهت معنایی بین متون به زبان‌های مختلف پیشنهادشده است. این رویکرد برای کار در فضای دوزبانه (یا چندزبانه)، بردارهای مداوم کلمه (یعنی تعبیه کلمه) را از یک‌زبان به فضای بردار زبان دیگر از طریق مدل ترجمه خطی طراحی کرده است. سپس کلمات را با توجه به شباهت بردارهای آن‌ها در فضای تعبیه دوزبانه تراز کرده و اقدامات مختلف نظارت‌نشده از شباهت معنایی با بهره‌گیری از تعبیه‌های دوزبانه و تراز کلمات را بررسی کرده‌اند. این روش ساده عملکردی نزدیک به روش‌های نظارت‌شده دارد و ثبات در زوج‌های مختلف زبان را نشان می‌دهد. این روش برای استخراج جملات موازی از مجموعه‌ای از نوشتجات[2] و تشخیص سرقت ادبی قابل‌مقایسه است. در این مدل به ابزار و منبع مشخصی نیاز نیست و صرفاً برپایه در دسترس بودن مجموعه‌ای نوشتجات بزرگ برای هر دو زبان ورودی استوار است. همچنین به یک مجموعه نسبتاً کوچک از جفت ترجمه کلمات احتیاج دارد.

در [۷۳] از تعبیه جمله BERT استفاده می‌کند. ابتدا ارتباط نظری بین هدف از پیش آموزش داده‌شده مدل زبان پوشیده و وظیفه شباهت معنایی را ازلحاظ نظری آشکارکرده و سپس تعبیه‌های جمله BERT را به‌صورت تجربی تجزیه‌وتحلیل نموده است. این تعبیه جمله همیشه یک فضای معنایی ناهمسانگرد ناهموار برای جملات ایجاد می‌کند، که به عملکرد شباهت معنایی آن آسیب می‌رساند. برای پرداختن به این موضوع در این مقاله توزیع تعبیه‌شده جمله ناهمسانگرد را به یک توزیع گاوسی صاف و همسانگرد از طریق جریان‌های نرمال‌سازی که با یک روش بدون ناظر صورت می‌گیرد تبدیل کرده است. در این روش به نمره همبستگی ۶۵/۶۲ دست‌یافته‌اند.

۶.۲ جمع‌بندی

در این فصل ویژگی‌های مورداستفاده در شباهت معنایی متن را موردبررسی قراردادیم و به بررسی تعدادی از این مقالات در این حوزه پرداختیم. در زیر تعدادی از این مقالات را باهم مقایسه نموده و

[1] cross-lingua
[2] Corpor

نتایج آن‌ها را موردبحث قرار داده‌ایم. در فصل بعدی روش پیشنهادی خود را برای تخمین شباهت معنایی متون ارائه خواهیم داد.

با توجه به جدول ۱.۲ بر روی‌داده‌های SICK، راناسینگه و همکاران با روش GRU به ضریب همبستگی بهتری نسبت به بقیه با خطای کمتری در این مجموعه داده رسیده‌اند. بر روی مجموعه داده‌های SemEval نیز افضل و همکاران به ضریب همبستگی بهتری دست پیداکرده‌اند. پاور و همکاران بر روی مجموعه داده‌های sts-۶۵ که حاوی ۶۵ جفت جمله می‌باشد به همبستگی بهتری دست‌یافته‌اند.

جدول ۱.۲: مقایسه برخی کارهای صورت گرفته در زمینه شباهت معنایی متن

روش(سال)	مجموعه داده	ضریب همبستگی پیرسون	ضریب همبستگی اسپرمن	مجموع مربع خطا	دقت
ISLAM and D (2008)	STS۶۵	۰/۸۵۳		۲۷/۵۸	۷۲/۴۲
GRU (2014)	SICK	۰/۸۹۰۱	۰/۸۳۹۶	۰/۲۱۱۲	۷۸/۸۸
MaLSTM (2016)	SICK	۰/۸۸۲۲	۰/۸۳۴۵	۰/۲۲۸۶	۸۴/۲
Afzal et al (2016)	SemEval-۲۰۱۶	۰/۸۰۱۵			
IC model (2017)	J،C	۰/۸۹۲	۰/۶۷		
Proposed Measure (2017)	AssIN	۰/۷۱		۰/۳۷	۸۵/۳۵
Siamese # local context: 5 + Siamese LSTM(2018)	SICK	۰/۸۵۴۹	۰/۷۹۳۳	۰/۲۸۹۸	۷۸/۰۲
Pawar et al(2018)	STS۶۵	۰/۸۷۹۴			
CTS(۲۰۱۸)	SemEval-۲۰۱۶	۰/۷۷۹۹			
Gazpio et al(2019)	SICK	۰/۸۶۰		۰/۳۸	۶۲

فصل ۳

روش تحقیق

۱.۳ مقدمه

در فصل قبل مروری بر روش‌های تخمین شباهت معنایی متن و کاربرد آن‌ها در شباهت یابی و همچنین کارهای صورت گرفته در این زمینه و نتایج به‌دست‌آمده از آن‌ها داشتیم. در این فصل به رویکرد پیشنهادی خود که مبتنی بر پیش‌پردازش داده‌های متنی و استفاده از رویکرد تخصیص پنهان دیریکلت به همراه فاصله جنسن شانون برای به دست آوردن نتیجه نهایی شباهت معنایی متون مورد آزمایش خود می‌باشد خواهیم پرداخت. سپس با استفاده از رویکرد ماشین بردار پشتیبان دقت دسته‌بندی را به دست خواهیم آورد . ابتدا به تعریف دقیق مسئله پرداخته و سپس با بررسی دقیق هر یک از ویژگی‌های روش پیشنهادی، مدل خود را به تفضیل بیان خواهیم کرد.

۲.۳ طرح مسئله

در فصل قبل مبانی نظری و روش‌های ارزیابی شباهت معنایی متن موردمطالعه قرار گرفت. با توجه به کاربردی بودن شباهت یابی در زمینه‌های بسیار ازجمله غلط‌یابی املایی، تشخیص سرقت ادبی، تشخیص تقلب، بازیابی اطلاعات، طبقه‌بندی اسناد و غیره درصدد هستیم تا به‌دقت بهتری برای شباهت بین جملات برسیم. همچنین میزان همبستگی را افزایش داده و از مجموع مربعات خطای نتایج به‌دست‌آمده کم کنیم.

امروزه شیوه‌های بسیاری برای کارهای پردازش زبان طبیعی صورت می‌گیرد. یکی از این شیوه‌ها استفاده از روش‌های آماری در ترکیب با سایر روش‌ها می‌باشد و همچنین نیاز به یک طبقه‌بند سریع و قابل‌اعتماد برای دسته‌بندی داده‌ها داریم. یکی از روش‌های مورداستفاده در این زمینه استفاده از روش LDA می‌باشد که ما از آن برای به دست آوردن شباهت جملات استفاده می‌کنیم. همچنین از ترکیب این فن با روش آماری جنسن و شانون برای اندازه‌گیری فاصله بین جملات استفاده خواهیم کرد. درنهایت از ماشین بردار پشتیبان برای تخمین دقت فاصله به‌دست‌آمده استفاده می‌نماییم.

۳.۳ روش پیشنهادی

برای به دست آوردن شباهت معنایی بین جملات ابتدا باید جملات تبدیل به زبان قابل‌فهم برای رایانه شوند. به همین منظور باید تبدیل به بردارهای عددی شوند. قبل از تبدیل به بردار عددی به پیش‌پردازش جملات خواهیم پرداخت. به این صورت که علائم نگارشی اضافه و کلماتی که تأثیر زیادی درمعنای متن ندارند را ازجمله حذف می‌کنیم و ریشه کلمات را جایگزین خواهیم کرد. سپس به استخراج ویژگی آن‌ها که همان تبدیل به بردار عددی می‌باشد خواهیم پرداخت و فرکانس کلمات

پرکاربرد را برای آموزش LDA به دست خواهیم آورد. در انتها فاصله بین بردارها را محاسبه کرده که شباهت نهایی را به ما خواهد داد. درواقع شباهت معنایی، اندازه‌گیری فاصله مفهومی بین دو موضوع است، هر چه این فاصله به صفر نزدیک‌تر باشد شباهت جفت جمله بیشتر است و هر چه به یک نزدیک‌تر باشد تناقض بین جملات بیشتر می‌باشد. پس از به دست آوردن نمره شباهت هر جفت جمله به دسته‌بندی آن‌ها با استفاده از برچسب نمره شباهت خواهیم پرداخت و دقت دسته‌بندی همچنین همبستگی پیرسون و اسپرمن را به دست خواهیم آورد.

در ادامه به شرح موارد گفته‌شده می‌پردازیم:

۴.۳ ارائه‌ی مدل کلی

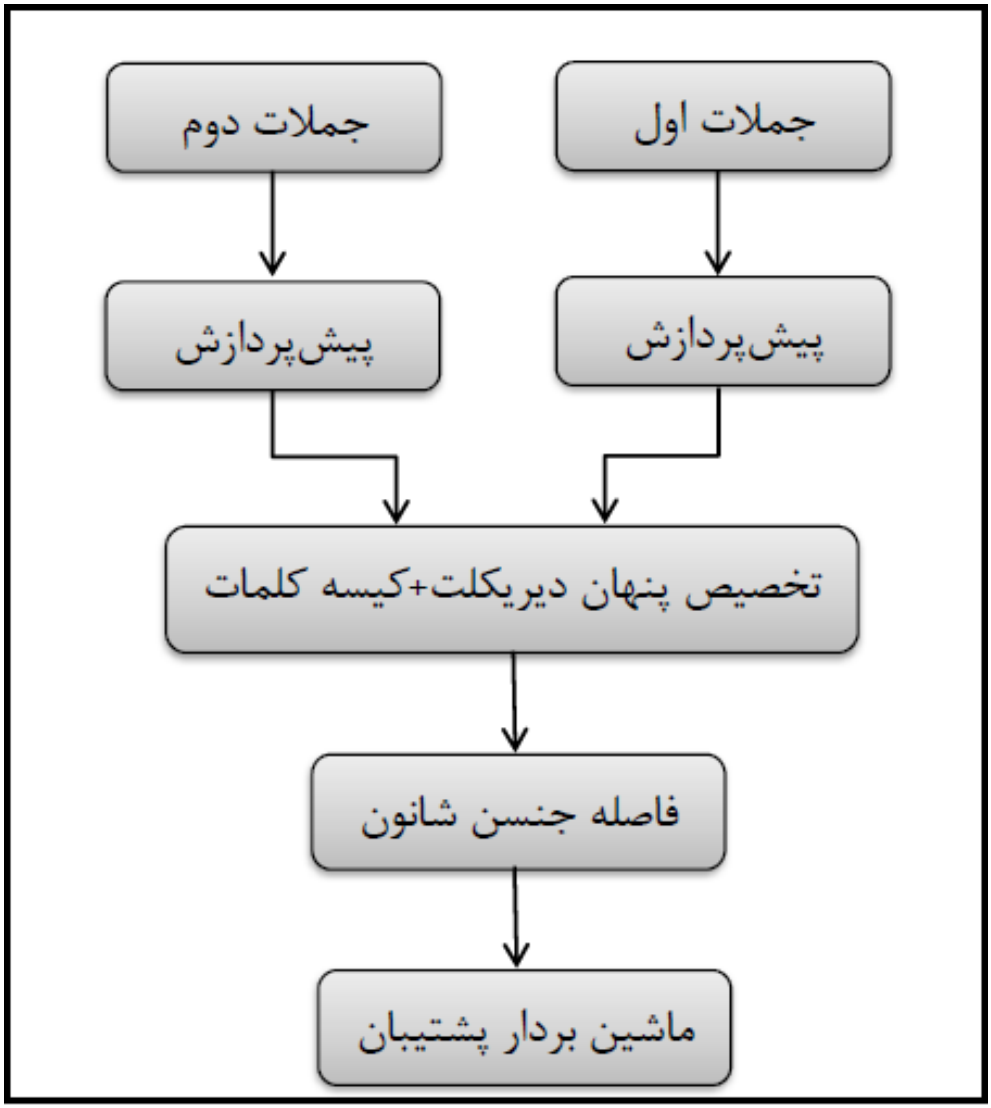
همان‌طور که می‌دانیم شباهت متن باید تعیین کند که نزدیک بودن دو قطعه متن هم ازنظر نزدیک بودن سطح (تشابه واژگانی) و هم‌معنی (تشابه معنایی) وجود دارد. به‌جای مقایسه کلمه به کلمه، برای گرفتن بیشتر ارتباط معنایی باید به زمینه نیز توجه کنیم. برای در نظر گرفتن شباهت معنایی، باید روی معنای کلی جمله متمرکز شویم. که در آن‌یک قطعه متن قبل از محاسبه شباهت به یک گروه مرتبط از کلمات مرتبط تقسیم‌شده است. ما با پردازش جملات علائم اضافی و کلمات بی‌تأثیر را حذف می‌نماییم و جملات را توکن‌بندی می‌کنیم و نقش هر کلمه در جمله را به دست می‌آوریم. ایده اصلی تهیه اسناد به‌عنوان ورودی و دریافت بردارهای ویژگی به‌عنوان خروجی است و درنهایت با اندازه‌گیری فاصله بین این ویژگی‌ها، جملات را مقایسه کنیم.

در شکل ۱.۳ مدل کلی را رسم نموده‌ایم. همان‌طور که در این شکل مشاهده می‌کنیم، مدل کلی دارای ۴ مرحله می‌باشد. در بخش اول، پیش‌پردازش روی‌داده‌ها انجام می‌گیرد. در مرحله بعد به کمک LDA توزیع بین کلمات برای هر دسته مشخص می‌شود و برای هر جمله بردار عددی را به دست می‌آوریم، در حقیقت این مرحله، مرحله‌ی استخراج ویژگی می‌باشد. سپس این بردارهای ایجادشده و ماتریس به‌دست‌آمده از LDA را به ماژول تخمین شباهت ارائه می‌دهیم تا هر جمله را به یک دسته اختصاص دهد و خروجی را برگرداند و پس‌ازآن خروجی ایجادشده را به مرحله بعدی می‌فرستیم تا نتایج به‌دست‌آمده طبقه‌بندی و بررسی شوند و دقت انجام‌شده به دست آید.

۱.۴.۳ پیش پردازش متن

همان‌طور که می‌دانیم داده‌های متنی غیر ساخت‌یافته، بزرگ‌ترین منبع داده‌های تولیدشده به‌وسیله انسان محسوب می‌شوند. این داده‌ها باید از زبان انسان به قالب قابل‌خواندن توسط ماشین برای پردازش بیشتر تبدیل شوند.

اولین مرحله در دسته‌بندی متن تبدیل اسناد به‌صورت رشته‌ای از کاراکترها با فرمت‌های مختلف می‌باشد که برای روش‌های یادگیری و طبقه‌بندی نمایش داده می‌شود. روش‌های پیش‌پردازش متن ازآن‌جهت حائز اهمیت هستند که ابزارهای لازم را برای تبدیل متن از زبان طبیعی به فرمت



شکل ۱.۳: ساختار کلی روش پیشنهادی

قابل‌خواندن توسط ماشین را فراهم می‌کنند. هر زمان که داده‌های متنی داشته باشیم، باید چندین مرحله پیش‌پردازش را برای داده‌ها انجام دهیم تا کلمات را به ویژگی‌های عددی تبدیل کنیم که با الگوریتم‌های یادگیری ماشین کار می‌کنند. مراحل پیش‌پردازش یک مسئله عمدتاً به دامنه و خود مسئله بستگی دارد، بنابراین نیازی نیست که همه مراحل را برای هر مشکلی اعمال کنیم.

پس از به دست آوردن متن، ما با نرمال‌سازی متن شروع می‌کنیم. نرمال‌سازی متن شامل موارد زیر است:

- **تبدیل متن به حروف کوچک**

اولین مرحله در پیش‌پردازش متن تبدیل داده‌ها به فرم یکنواخت می‌باشد. بهتر است داده‌های متنی را به حروف کوچک تبدیل کنیم تا مطمئن شویم به‌طور مثال کلمه "Nlp" با "nlp" یکسان است.

- **توکن‌بندی**[1]

توکن‌بندی فرآیند تقسیم متن داده‌شده به قطعات کوچک‌تر به نام توکن است که کلمات، اعداد، علائم نگارشی مانند(’’،’’ ، ’’.’’ ، ’’!’’) و سایر موارد را می‌توان به‌عنوان نشانه در نظر گرفت. همچنین شامل تقسیم قطعات بزرگ متن به جملات و عبارات و کلمه‌ها، نشانه‌ها یا عناصر معنی‌دار می‌باشد. این مرحله همچنین به‌عنوان تقسیم‌بندی یا تجزیه‌وتحلیل واژگانی شناخته می‌شود که برای انجام پردازش بیشتر لازم است.

- **حذف ایست واژه‌ها**[2]

درحالی‌که با داده‌های متنی کار می‌کنیم، با موارد مختلفی روبرو می‌شویم که برای تجزیه‌وتحلیل ما کاربرد چندانی ندارند زیرا هیچ معنی به داده‌های ما نمی‌دهند. این‌ها می‌توانند ضمایر باشند (مانند من، تو و غیره) و یا متداول‌ترین کلمات در یک‌زبان مانند ”the“، ”a“، ”on“، ”is“، ”all“ هستند. این کلمات در معنی یک جمله سهیم نیستند و معمولاً از متن حذف می‌شوند. لیست ایست واژه‌ها برای اکثر زبان‌ها باید استخراج شود که با استفاده از این لیست می‌توان دید که اگر این کلمه‌ها در داخل اسناد متنی وجود داشته باشد از اسناد حذف می‌شود و اگر کلمه در لیست نبود حذف نمی‌شود و با این کار از تعداد ویژگی‌ها کم می‌شود.

- **ریشه‌یابی**[3]

کلمات کلیدی، مرتبط‌ترین لغات موجود در یک داده متنی محسوب می‌شوند و می‌توانند برای خلاصه‌سازی محتویات آن‌ها مورداستفاده قرار بگیرند. استفاده از روش‌های استخراج کلمات کلیدی به مدل متن‌کاوی اجازه می‌دهند تا داده‌هایی که قرار است جستجو شوند را شاخص گذاری، محتویات متون را خلاصه‌سازی و متون را برچسب‌گذاری کند.

ریشه‌یابی یک مرحله اساسی در پیش‌پردازش متن برای NLP است. Lemmatization فرآیند الگوریتمی یافتن ریشه یک کلمه است که به معنی آن بستگی دارد. ریشه‌یابی معمولاً به تجزیه‌وتحلیل واژگان اشاره دارد که هدف آن حذف انتهای عطف است. الگوریتم ساقه زنی[4] با بریدن پسوند از کلمه کار می‌کند. به معنای گسترده‌تر، ابتدا یا انتهای کلمه را می‌برد. برعکس، ریشه‌یابی عملی قدرتمندتر است و تجزیه‌وتحلیل ریخت‌شناسی کلمات را در نظر می‌گیرد. لمی را که فرم پایه همه اشکال عطف آن است برمی‌گرداند. برای ایجاد فرهنگ لغت و جستجوی فرم مناسب کلمه، دانش عمیق زبانی لازم است. ساقه زنی یک عمل کلی است درحالی‌که ریشه‌یابی یک عمل هوشمندانه است که در آن فرهنگ لغت به دنبال فرم مناسب آن می‌گردد. ازاین‌رو، ریشه‌یابی به شکل‌گیری ویژگی‌های بهتر یادگیری ماشین کمک می‌کند. تفاوت در اینجاست که کلمه ریشه همواره ازنظر فرهنگ واژگانی یک کلمه صحیح است؛ اما

[1] Tokenization
[2] stopwords
[3] Lemmatization
[4] Stemming

stem ریشه ممکن است چنین نباشد. ازاین‌رو کلمه ریشه به نام lemma نیز نامیده می‌شود که همواره در دیکشنری وجود دارد.

- **برچسب‌گذاری کلمات**[1]

برچسب‌گذاری کلمات دسته‌بندی واژگانی خاصی هستند که در آن واژه‌ها بر اساس زمینه و نقش معنایشان به بخش‌های مختلف انتساب می‌یابند. به‌طورمعمول واژه‌ها می‌توانند دریکی از دسته‌های اصلی زیر قرار گیرند:

اسم[2]: این جزء معمولاً واژه‌هایی را شامل می‌شود که یک شیء یا نهاد را معرفی می‌کنند که می‌تواند زنده یا غیرزنده باشد. برخی نمونه‌ها شامل روباه، سگ، کتاب و غیره هستند. نماد تگ POS برای اسم‌ها به‌صورت N است.

فعل[3]: افعال همان واژه‌هایی هستند که اعمال، وضعیت‌ها یا رخدادهای خاصی را توصیف می‌کنند. طیف متنوعی از زیر دسته‌های افعال مانند افعال کمکی، افعال انعکاسی و افعال گذرا و موارد بسیار دیگر هستند. برخی نمونه‌های معمول افعال به‌صورت دویدن، پریدن، خواندن و نوشتن هستند. نماد تگ POS برای افعال به‌صورت V است.

صفت[4]: صفت‌ها، اسم‌هایی هستند که برای توصیف یا تعیین کیفیت واژه‌های دیگر که معمولاً اسم یا عبارت‌های اسمی هستند، به کار می‌روند. عبارت گل زیبا یک اسم (N) دارد که با استفاده از یک صفت (ADJ) به‌صورت زیبا توصیف‌شده است. نماد تگ POS برای صفت‌ها به‌صورت ADJ است.

قید[5]: قیدها معمولاً به‌صورت اصلاح‌کننده[6] برای واژه‌های دیگر شامل اسم، صفت، فعل و یا دیگر قیدها استفاده می‌شوند. عبارت گل بسیار زیبا دارای قید (ADJ) «بسیار» است که صفت زیبا را تعدیل کرده و میزان زیبایی آن را تعیین می‌کند. نماد تگ POS برای قیدها به‌صورت ADJ است.

علاوه بر این چهار دسته اصلی اجزای گفتار، دسته‌های دیگری نیز وجود دارند که به‌طور مکرر در زبان انگلیسی ظاهر می‌شوند. این موارد شامل ضمایر، حروف اضافه، حروف ندا، حروف ربط، ضمایر اشاره و موارد بسیار دیگر هستند. به‌علاوه، هر تگ POS مانند اسم (N) را می‌توان به زیر دسته‌هایی مانند اسامی مفرد (NN)، اسامی خاص مفرد (NNP) و اسامی جمع (NNS) تقسیم کرد.

این فرایند طبقه‌بندی و برچسب‌گذاری تگ‌های POS برای کلمات به نام تگ گذاری اجزای گفتار نامیده می‌شود. تگ‌های POS برای حاشیه‌نویسی کلمات و تعیین POS آن‌ها استفاده

[1] part-of-speech tagging(pos)
[2] Noun
[3] Verb
[4] Adjective
[5] Adverb
[6] Modifier

می‌شوند که در جهت اجرای آنالیزهای خاص مانند خلاصه‌سازی اسامی و تعیین رایج‌ترین اسم‌ها، ابهام‌زدایی از معنی کلمه و آنالیز گرامری کاملاً مفید هستند. درروش پیشنهادی ما این مراحل انجام‌شده‌اند. در شکل ۲.۳ مراحل این فرآیند نشان داده‌شده است.

شکل ۲.۳: مراحل پیش‌پردازش متن

۲.۴.۳ استخراج ویژگی از متن

همان‌طور که قبلاً بیان کردیم تخمین شباهت معنایی متن در حوزه‌های بسیاری مورداستفاده قرارگرفته است. مشکلی که در دسته‌بندی متن وجود دارد، حجم زیاد ویژگی‌ها است که باعث کاهش دقت نتایج دسته‌بندی می‌شود. برای انتخاب و برای حل این مشکل و کاهش ابعاد ویژگی‌ها از متدهای انتخاب ویژگی استفاده می‌کنند.

انتخاب ویژگی متن یک موضوع اساسی و مهم برای استخراج متن و بازیابی اطلاعات است، زیرا در این مرحله واژه‌های کلیدی انتخاب می‌شوند تا به‌عنوان بهترین نمایش‌دهنده برای سند متنی مورداستفاده قرار بگیرند. اگر تعداد واژه‌های کلیدی انتخاب‌شده کم باشد صحت و کارایی سیستم تحت تأثیر قرار می‌گیرد و کاهش می‌یابد و در مقابل اگر تعداد واژه‌های کلیدی انتخاب‌شده زیاد باشد باعث کاهش کارایی سیستم در بعد زمان خواهد شد و سرعت آموزش در فاز آموزش پایین می‌آید. در طول استخراج ویژگی، ویژگی‌های نامربوط یا اضافی حذف می‌شوند. محاسبه وزن را استخراج ویژگی متن می‌نامند. هم‌زمان با استخراج ویژگی با استفاده از تخصیص پنهان دیریکلت به تشخیص اینکه هر کلمه در چه دسته‌ای قرار دارد می‌پردازیم. این دسته‌ها بر اساس نمره شباهت

موجود به‌دست‌آمده‌اند. درواقع کلمات را بر اساس نمره شباهت در یک دسته قرار می‌دهیم و این فرآیند را تکرار می‌کنیم تا کل داده‌ها به دسته درست تعلق گیرند. تعداد دسته‌ها را با این فرآیند مشخص می‌کنیم. ممکن است کلمات در میان دسته‌ها هم‌پوشانی داشته باشند، بنابراین چندین دسته کلمات مشابهی دارند.

مدل با اختصاص دادن هر کلمه در هر جمله به یک دسته تصادفی، کار خود را شروع می‌کند. سپس، این مرحله را برای هر کلمه تکرار می‌کند. دسته فعلی را لغو کرده و کلمه را بر اساس احتمال به دسته بعدی اختصاص می‌دهیم. یعنی دسته به جمله فعلی اختصاص دارد و کلمه به دسته فعلی مختص می‌شود. مدل LDA طبقه‌بندی دسته‌ها را انجام نمی‌دهد، بلکه بررسی می‌کند که آیا این کلمه برای این دسته مناسب هست یا نه، یعنی توزیع گسسته بین کلمات برای یک دسته را انجام می‌دهد. می‌توان دسته‌ها را بعد از آموزش نام برد. وقتی هر کلمه را در هر جمله تکرار می‌کنیم دسته فعلی آن را لغو می‌کنیم و کلمه را به دسته جدید اختصاص می‌دهیم. دسته‌ای که کلمه را مجدداً به آن اختصاص می‌دهیم بر اساس احتمالات زیر است:

$$P(document" Like" thetopic) * (topic" Like" thewordw\prime) \quad (۱.۳)$$

$$\Rightarrow \frac{n_{i,k} + \alpha}{N_i - ۱ + K_\alpha} * \frac{m_{w\prime,k} + \eta}{\sum_{w \in v} m_{w,k} + V_\eta} \quad (۲.۳)$$

در این فرمول به ترتیب:

$n_{i,k}$: تعداد واژه‌های اختصاص یافته به عنوان k در سند i

α: پارامتر هموارسازی

N_i: تعداد کلمات در سند i

K: تعداد موضوعات

$m_{w,k}$: تعداد واگذاری‌ها، کل مجموعه، کلمه w به موضوع k

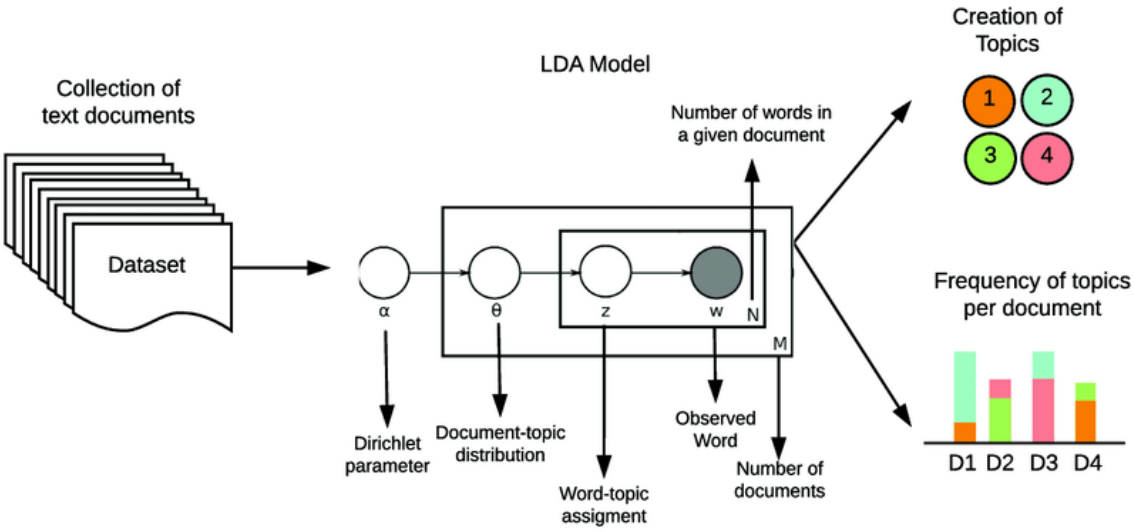
η: پارامتر هموارسازی

V: اندازه واژگان یعنی تعداد کلمات مجزا

LDA دارای ۲ ابر پارامتر است: α و η مقدار کم برای α به این معنی است که اسناد فقط تعداد کمی از عناوین را دارند. مقدار زیاد آن نیز به این معنی که اسناد در یک مجموعه شبیه یکدیگر هستند. مقدار کم برای η به این معنی است که موضوعات دارای تعداد کم کلمات کمک‌کننده هستند. مقدار زیادی از η، به این معنی که موضوعات باهم تداخل دارند و بیشتر شبیه یکدیگر می‌شوند.

الگوریتم LDA به ترتیب کلمات در جمله اهمیت نمی‌دهد. بهتر است از کلماتی که زیاد در پایگاه داده وجود ندارد در این روش حذف شوند. همان‌طور که در مرحله اول ما ایست واژه‌ها را نیز حذف کردیم تا بتوانیم در این روش به نتیجه بهتری برسیم.

در شکل ۳.۳ روند کلی LDA نشان داده‌شده است.



شکل ۳.۳: مدل تخصیص پنهان دیریکلت [۷۴]

در این مرحله هر خروجی از مرحله قبل را به کیسه کلمات تبدیل می‌کنیم. برای به دست آوردن BOW برای هر جمله‌ای که تاکنون آموزش ندیده است فقط کلمات موجود در فرهنگ لغت موجود برای آموزش مدل استفاده می‌کنیم. بنابراین اگر کلمات جدیدی درون جمله وجود داشته باشد هنگام استنباط از توزیع دسته، به آن‌ها توجه نمی‌شود و هیچ خطایی برای کلمات جدید دیده نمی‌شود. اما یک مشکل دارد آن‌هم این است که برخی کلمات ممکن است حذف شوند. برای حل این مشکل مجموعه آموزش را افزایش می‌دهیم و سعی در آموزش بیشتر جملات داریم.

۳.۴.۳ جزئیات روش جنسن شانون و دسته‌بندی

پس از به دست آوردن BOW هر جمله فرکانس‌های جملات را به‌روز می‌کنیم و شناسه‌هایی برای کلمات جدید ایجاد می‌کنیم. هم‌زمان فرکانس‌های جمله را نیز به‌روز می‌کنیم، برای هر کلمه‌ای که در این سند ظاهر می‌شود، فرکانس سند آن را یکی افزایش می‌دهیم.

درنهایت ما یک توزیع دسته برای هر جمله جدید داریم. اگر بخواهیم مشابه‌ترین جملات را در مجموعه بیابیم می‌توانیم این کار را با مقایسه توزیع دسته جمله جدید با تمام توزیع‌های موضوعی جملات در مجموعه انجام دهیم. ما از معیار فاصله جنسن شانون برای یافتن مشابه‌ترین اسناد استفاده می‌کنیم.

واگرایی جنسن شانون یک نسخه متقارن و صاف از واگرایی موازی است. این واگرایی بین دو بردار احتمال p و q به این صورت تعریف‌شده است:

$$JSD(P \parallel Q) = \frac{۱}{۲} D(P \parallel M) + \frac{۱}{۲} D(Q \parallel M) \qquad (۳.۳)$$

که M در آن:

$$M = \frac{۱}{۲}(P + Q)$$

و D واگرایی $Kullback - Leibler$ است:

$$D(P \parallel Q) = \sum_i P(i) \log \frac{P_i}{Q_i} \quad (۴.۳)$$

$$JSD(P \parallel Q) = \frac{۱}{۲} \sum_i [P_i \log(\frac{P_i}{\frac{۱}{۲}(P_i) + (Q_i)}) + Q_i \log(\frac{Q_i}{\frac{۱}{۲}(Q_i) + (P_i)})] \quad (۵.۳)$$

ریشه مربع واگرایی جنسن شانون فاصله جنسن شانون است:

$$\sqrt{JSD(P \parallel Q)}$$

واگرایی هندسی جنسن شانون با در نظر گرفتن میانگین هندسی فرمول بسته‌ای را برای واگرایی بین دو توزیع گوسی ارائه می‌دهد. یک تعریف کلی‌تر، که امکان مقایسه بیش از دو توزیع احتمال را فراهم می‌کند، این است:

$$JSD_{\pi ۱,\ldots,\pi n}(P_۱, P_۲, \ldots, P_n) = H(\sum_{i=۱}^{n} - \sum_{i=۱}^{n} \pi_i H(P_i)) \quad (۶.۳)$$

که در آن :

$$p۱ = p_۱, p۲ = P_۲, \pi_۱ = \pi_۲ = \frac{۱}{۲}$$

واگرایی جنسن شانون روشی برای اندازه‌گیری شباهت بین دو توزیع احتمال است که ما در این پژوهش برای اندازه‌گیری شباهت بین جملات استفاده می‌کنیم. برای یافتن مشابه‌ترین جملات در مجموعه، توزیع موضوع سند جدید را با همه توزیع‌های موضوعی اسناد در مجموعه مقایسه می‌کنیم و از معیار فاصله جنسن شانون برای یافتن مشابه‌ترین اسناد استفاده می‌کنیم. آنچه مسافت جنسن و شانون به ما می‌گوید، این است که با مقایسه واگرایی توزیع آن‌ها، کدام جملات ازنظر آماری ’’نزدیك‌تر’’ و درنتیجه شباهت بیشتری دارند.

هرچه فاصله جنسن شانون کوچک‌تر باشد، دو توزیع مشابه‌تر است و دو جمله شباهت بیشتری به هم دارند. با استفاده از میانگین توزیع احتمالات آن‌ها یا ادغام محتویات هر دو جمله، میانگین سند M رابین دو سند ۱ و ۲ ایجاد می‌کنیم و اندازه هر یک از اسناد ۱ و ۲ با متوسط سند M از طریق $KL(P||M)$ و $KL(Q||M)$ را به دست می‌آوریم. در آخر اختلافات را از نقطه ۲ به‌طور متوسط انجام می‌دهیم.

پس‌ازاین مرحله از ماشین بردار پشتیبان برای به دست آوردن دقت نتایج به‌دست‌آمده استفاده می‌شود. علاوه بر انجام طبقه‌بندی خطی، ماشین‌های بردار پشتیبان می‌توانند با استفاده ازآنچه

ترفند هسته نامیده می‌شود، یک طبقه‌بندی غیرخطی را انجام دهند. به‌طور ضمنی ورودی‌های آن‌ها را در فضاهای دارای ویژگی‌های بسیار بالا نقشه‌برداری می‌کنند. هنگامی‌که داده‌ها بدون مجوز هستند، یادگیری نظارت‌شده امکان‌پذیر نیست و یک رویکرد یادگیری بدون نظارت موردنیاز است، که سعی می‌کند خوشه‌بندی طبیعی داده‌ها را به گروه‌ها پیدا کند و سپس داده‌های جدید را به این گروه‌های تشکیل‌شده نقشه ببرد. ماشین بردار پشتیبان با استفاده از برچسب آموزش برای هر گروه، متن جدید را دسته‌بندی کنند.

یک الگوریتم یادگیری ماشین معمولی سعی می‌کند مرزی پیدا کند که داده‌ها را به‌گونه‌ای تقسیم کند که خطای طبقه‌بندی غلط به حداقل برسد. SVM بهینه‌ترین مرز تصمیم‌گیری را پیدا می‌کند. بهینه‌ترین مرز تصمیم‌گیری محدوده‌ای است که حداکثر حاشیه را از نزدیک‌ترین نقاط همه کلاس‌ها داشته باشد.

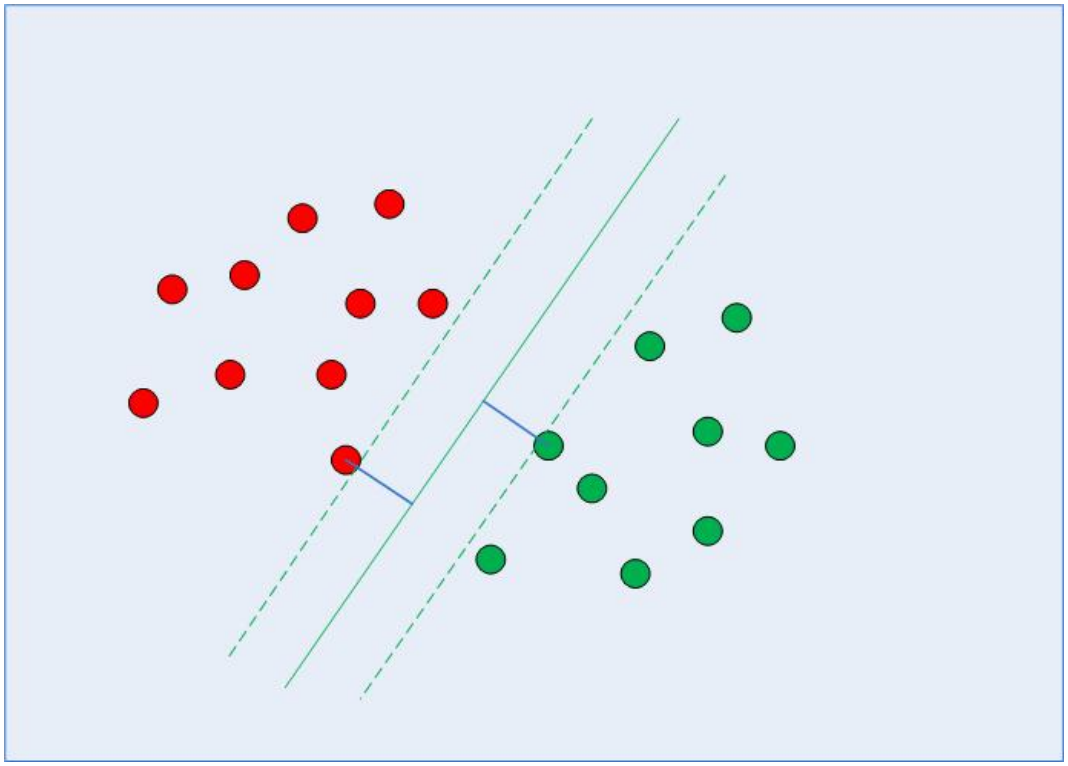
در الگوریتم‌های SVM عملکرد هسته این است که داده‌ها را به‌عنوان ورودی گرفته و آن‌ها را به شکل موردنیاز تبدیل کند. الگوریتم‌های مختلف SVM از انواع مختلف توابع هسته استفاده می‌کنند. این توابع می‌توانند انواع مختلفی داشته باشند. به‌عنوان‌مثال خطی[۱]، غیرخطی[۲]، چندجمله‌ای[۳]، تابع پایه شعاعی[۴] و سیگموئید[۵].

هسته چندجمله‌ای یک عملکرد هسته است که با ماشین‌های بردار پشتیبان و سایر مدل‌های هسته‌دار استفاده می‌شود و نشان‌دهنده شباهت بردارها در یک فضای مشخصه نسبت به چند جمله‌ای‌های متغیرهای اصلی است. در شکل۴.۳ مرز تصمیم‌گیری برای بردار پشتیبان خطی و در شکل ۵.۳ مرز تصمیم‌گیری برای داده‌های غیرخطی قابل‌تفکیک مشاهده می‌کنیم.

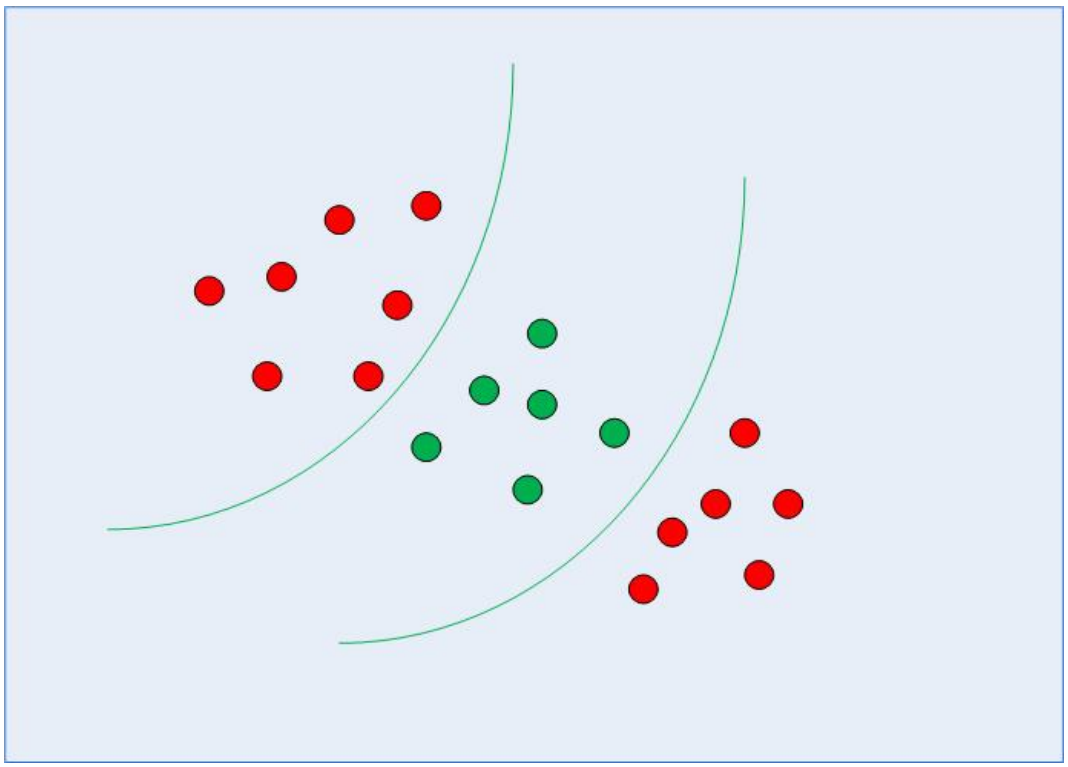
۵.۳ جمع‌بندی

در این فصل روش پیشنهادی خود را به تفضیل و با جزئیات بیان نمودیم. در فصل بعد این روش را با استفاده از سناریوهای ارائه‌شده می‌آزماییم و نتایج را برای هر مرحله ارائه می‌دهیم.

[1] linear
[2] Nonlinear
[3] Polynomial
[4] Radial base function(RBF)
[5] Sigmoid



شکل ۴.۳: مرز تصمیم برای بردار پشتیبان خطی[۷۵]



شکل ۵.۳: مرز تصمیم برای داده‌های غیرخطی[۷۵]

فصل ۴

نتایج و تحلیل آن‌ها

۱.۴ مقدمه

در فصل گذشته به معرفی روش پیشنهادی خودپرداختیم، در این فصل نتایج عملی این روش را شرح خواهیم داد. ابتدا مجموعه دادگان مورداستفاده خود در این پژوهش را معرفی می‌کنیم و پس از آن به جزئیات پیاده‌سازی انجام‌گرفته می‌پردازیم. در ادامه نتایج رویکرد خود را بررسی کرده و جداول و نمودارهای به‌دست‌آمده از تحقیق را توصیف می‌نماییم. در انتها خلاصه‌ای از کارهای صورت گرفته در این فصل را بیان می‌کنیم.

۲.۴ مجموعه دادگان مورد استفاده

بانک اطلاعاتی SICK[۱]: جملات داده شامل جملات دانش ترکیبی در وظیفه مشترک SemEval2014 استفاده می‌شود. مجموعه دادگان ما شامل ۱۰۰۰۰ جفت جمله است. ارتباط میان جملات با مقدار بین ۱ تا ۵ نشان داده‌شده است. این مجموعه داده به‌عنوان معیار برای ارزیابی فن تشابه جملات استفاده می‌شود. این مجموعه داده حاوی متون نوشته‌شده به زبان انگلیسی توسط کاربران اینترنت می‌باشد. بخش مهمی از مجموعه داده‌ها از مخازنی مانند اخبار روزانه، مجموعه داده‌های ویکی‌پدیا، توییت‌ها، متن‌ها، عناوین خبری، شرح تصاویر و فیلم‌ها و دیگر موارد است. این مجموعه داده شامل جفت‌های جملات معنایی است که به‌صورت خودکار انتخاب و دستکاری‌شده‌اند تا پدیده‌هایی مانند جملات غنی ازنظر لغوی، مترادف و متضاد را شامل شوند. هدف از استفاده از این دادگان ارزیابی مدل‌های معنایی توزیع ترکیبی در جملات کامل از طریق ارتباط معنایی و پیام متنی می‌باشد.

مجموعه داده‌های SICK از دو مجموعه پارافراس موجود ساخته‌شده است: مجموعه داده‌های Image FLickr 8k و مجموعه داده‌های تشابه متن معنایی SemEval 2012 STS MSR-Video Description بین دو عنصر. مجموعه داده‌های قبل از آموزش شامل داده‌های جداگانه جمله‌ای است که برای کارهای قبلی SemEval2013 معنایی متن معنایی ارائه‌شده است. مجموعه داده‌های قبل از آموزش شامل تقریباً ۱۱۰۰۰ جفت جمله است که دارای برچسب [۱،۵] است.

در مجموعه نهایی همان‌طور که در جدول ۱.۴ نشان داده‌شده است، نمرات ارتباط به شرح زیر توزیع شد: امتیازات مربوط به ۹۲۳ جفت در محدوده [۱،۲]، ۱۳۷۳ جفت در محدوده [۲،۳]، ۳۸۷۲ جفت در دامنه [۴،۳] و ۳۶۷۲ جفت در محدوده [۴،۵]. در فرآیند برچسب‌گذاری، جملات را به دسته تناقض[۲]، مترادف[۳] و خنثی[۴] تقسیم می‌کنند، که منجر به ۵۵۹۵ جفت خنثی، ۱۴۲۴ جفت

[1] Sentences Involving Compositional Knowledge
[2] contradiction
[3] entailment
[4] neutral

تناقض و ۲۸۲۱ جفت مترادف دلالت شد(طبق جدول ۲.۴)

جدول ۱.۴: طبقه‌بندی مجموعه دادگان مورد استفاده براساس امتیازات

محدوده داده‌ها	تعداد	درصد
محدوده [۱،۲]	۹۲۳	۹/۳
محدوده [۲،۳]	۱۳۷۳	۱۳/۹۵
محدوده [۳،۴]	۳۸۷۲	۳۹/۳۴
محدوده [۴،۵]	۳۶۷۲	۳۷/۳۱

جدول ۲.۴: طبقه‌بندی مجموعه دادگان مورد استفاده براساس برچسب گذاری موجود

برچسب موجود	تعداد
خنثی	۵۵۹۵
متناقض	۱۴۲۴
مترداف	۲۸۲۱

در جدول ۳.۴ نمونه‌ای از مجموعه داده‌ها همراه با ضریب همبستگی موجود مشاهده می‌کنید.

جدول ۳.۴: نمونه‌ای از مجموعه داده‌ها همراه با ضریب همبستگی

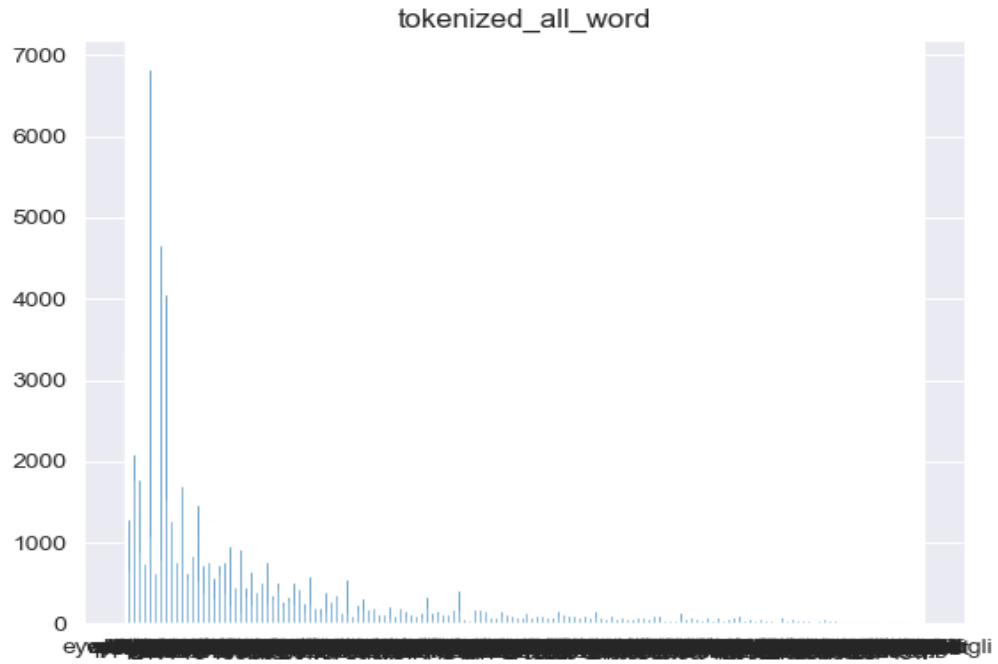
جمله دوم	جمله اول	نمره ارتباط
Two women are sparring in a kickboxing matc	Two young women are sparring in a kickboxing fight	۴.۹
There is no man dunking the ball at a basketball game	A man with a jersey is dunking the ball at a basketball game	۴

علاوه بر داده‌های SICK، مجموعه دادگان SIS65 نیز جهت ارزیابی شباهت معنایی مورداستفاده قرار می‌گیرد. این مجموعه داده شامل ۶۵ جفت جمله می‌باشد. به دلیل بیشتر بودن جملات پایگاه داده SICK و همچنین به دلیل دسترسی رایگان به این مجموعه و استفاده مقالات بسیاری از این پایگاه داده به‌عنوان یک پایگاه داده استاندارد، ما جملات این داده‌ها را مورد ارزیابی قرار می‌دهیم.

۳.۴ پیاده‌سازی

در این بخش به جزئیات پیاده‌سازی انجام‌شده می‌پردازیم.

مجموعه دادگان با فرمت متنی هستند و طبق فرآیندی که گفته شد طبقه‌بندی‌شده‌اند. ابتدا ایست واژه‌ها ازجمله حذف می‌شوند. زیرا ایست واژه‌ها تعداد رخداد بالا و بار معنایی کمی دارند، بنابراین حذف این واژه‌ها از پایگاه داده حجم محاسبات را کاهش و کارایی را بهبود می‌دهد. همچنین به کمک یک تابع به توکن‌بندی جملات می‌پردازیم. سپس ریشه کلمات رو با ریشه‌یابی lemma به دست می‌آوریم و pos کلمات را در تابعی دیگر به دست می‌آوریم. نمودار کلمات پیش‌پردازش شده در شکل ۱.۴ نشان داده‌شده است.



شکل ۱.۴: نمودار داده‌های پیش پردازش شده

در ادامه فرکانس کلمات را با استفاده از کتابخانه nltk به دست می‌آوریم. دادگان را به مجموعه آموزش و آزمون تقسیم می‌کنیم. مجموعه آموزش برای آموزش مدل LDA استفاده می‌شود، درحالی‌که مجموعه آزمون برای بازیابی جملات جدید ورودی به مدل ما استفاده می‌شود. برای اجرای LDA روش‌های متعددی وجود دارد که ما از نمونه‌برداری gibbs استفاده خواهیم کرد، زیرا روشی ساده برای درک موضوع است. کتابخانه‌هایی که برای اجرای LDA مورداستفاده قرا می‌گیرند Genism و Scikit-learn هستند. در میان این دو بسته Genism بهتر عمل می‌کند زیرا مدل‌سازی موضوع با Genism مرجع خوبی برای یادگیری LDA است. ماهم از همین کتابخانه برای پیاده‌سازی LDA استفاده کردیم.

برای دستیابی به نتایج دقیق بهتر است مدل را مجدداً آموزش دهیم. در این مرحله جملات را با استفاده از ماژول استخراج‌کننده ویژگی‌ها به کیسه کلمات تبدیل می‌کنیم. برای به دست آوردن BOW از کتابخانه Genism استفاده کردیم و از کلمات موجود در فرهنگ لغت برای آموزش استفاده

کردیم. تمام توزیع‌های دسته‌های به‌دست‌آمده از LDA را در یک ماتریس قرار می‌دهیم. این کار امکان محاسبه سریع و کارآمد را فراهم می‌کند.

هدف ما از استفاده کردن از LDA یافتن میزان شباهت جملات باهم است. برای این منظور ما از پیاده‌سازی آنتروپی استفاده کردیم.

ما فاصله ماتریس به‌دست‌آمده از دسته‌های LDA و ماتریس موجود از کل مجموعه موجود را با کمک فن جنسن شانون که کمترین فواصل مربوط به آن‌ها را در آرایه به دست می‌آورد محاسبه می‌کنیم با این کار می‌توانیم مشابه‌ترین جملات را به دست آوریم.

فاصله جنسن شانون یک روش آماری برای به دست آوردن شباهت بین دو توزیع احتمال می‌باشد. پیاده‌سازی این روش با کتابخانه Scipy انجام گرفت.

۱.۳.۴ طبقه‌بندی نمرات شباهت به‌دست‌آمده

برای طبقه‌بندی داده‌ها نیز از ماشین بردار پشتیبان خطی، هسته چندجمله‌ای و هسته تابع پایه شعاعی و دسته‌بندی‌کننده بیز ساده استفاده کردیم. به‌طور ساده روش بیز روشی برای دسته‌بندی پدیده‌ها، بر پایه احتمال وقوع یا عدم وقوع یک پدیده است. شیوه یادگیری در روش بیز ساده از نوع یادگیری با نظارت است. مرحله آخر هر الگوریتم یادگیری ماشین ایجاد ارزیابی برای هسته چندجمله‌ای است. برای این منظور نیز از کتابخانه sklearn کمک گرفته‌ایم. درست است محاسبات پشت SVM پیچیده است، ولی ما با کمک این کتابخانه توانسته‌ایم به‌راحتی به پیاده‌سازی این طبقه‌بندی بپردازیم. روش بیز ساده دقت کمتری نسبت به ماشین بردار پشتیبان دارد. بنابراین، ماشین بردار پشتیبان خطی روشی مؤثر در این زمینه می‌باشد.

۴.۴ آزمون و پارامترهای مورد نظر

در این بخش به مقایسه روش پیشنهادی خود با سایر روش‌های انجام‌شده در این زمینه می‌پردازیم. جملات موجود در پایگاه داده خود را با روش‌های ازجمله فاصله کوسین، فاصله اقلیدسی، فاصله منتهن بررسی می‌کنیم. همچنین روش خود را با بعضی از مقالات فصل دوم مقایسه خواهیم نمود. ازجمله روش‌های استفاده‌شده در فصل دوم که از پایگاه داده موردنظر ما استفاده کرده‌اند:

- روش Sent2Vec
- روش PARAGRAM-PHRASE
- روش MaLSTM features با دسته‌بند SVM
- روش DLSTM features با دسته‌بند SVM

همچنین در این فصل مشاهده نمودیم که برای استخراج ویژگی‌ها روش‌های N ـ گرم، Tf-idf، مدل کیسه‌ای از کلمات و word2vec مورداستفاده قرارگرفته‌اند. برای تخمین شباهت به کمک این ویژگی‌ها از روش‌های شبکه‌های عصبی CNN و LSTM، رگرسیون خطی، بیزین، طبقه‌بندی Bloom و K ـ نزدیک‌ترین همسایه و ترکیبی از این روش‌ها استفاده‌شده است. هدف ما از انجام این پژوهش مقایسه این روش با روش‌هایی که از پایگاه داده SICK استفاده کرده‌اند می‌باشد. همچنین به مقایسه روش خود با روش‌های زیر پرداختیم.

- ویژگی word2vec بافاصله کوسین
- ویژگی Tf-idf بافاصله منتهن
- ویژگی Tf-idf بافاصله اقلیدسی

دو جفت جمله انتخاب‌شده را به‌عنوان نمونه‌ای از پایگاه داده خود انتخاب کردیم، دقت به‌دست‌آمده از هر یک از روش‌های بیان‌شده را بررسی نمودیم. به این منظور دو جفت جمله ۴۱۴ و ۷۳۰ را انتخاب کرده و روش‌های ذکرشده را روی آن‌ها اعمال کردیم. در جدول ۴.۴ داده مورد آزمایش و در جدول ۵.۴ نتایج به‌دست‌آمده از این روش‌ها را مشاهده می‌نماییم. روش اول از فاصله کوسین استفاده کرده‌ایم و نتیجه بین ـ۱ و ۱ می‌باشد. هر چه دو جمله شباهت بیشتری داشته باشند نتیجه به یک نزدیک‌تر است. درروش دوم فاصله منتهن فاصله بین ۰ و ۱ به‌دست‌آمده است . فاصله جنسن شانون نیز نتیجه‌ای بین ۰ و۱ درخواهد گرداند. ما تمام نتایج را به معیار فاصله موردنظر یعنی بین ۱ تا ۵ تبدیل کردیم. همان‌طور که مشاهده می‌کنید روش ما نتایج صحیح‌تری نسبت به بقیه روش‌ها داشته است. دقت به‌دست‌آمده فقط برای همین دو جمله آزمایشی به‌دست‌آمده است.

جدول ۴.۴: جملات مورد مقایسه ۴۱۴ و ۷۳۰ مجموعه دادگان

	جمله A	جمله B	نمره ارتباط
DATA ۴۱۴	The man in the purple hat isn't operating a camera that smakes videos	The man in the purple hat is operating a camera that	۴/۷
DATA ۷۳۰	Two men are taking a break from a trip on a snowy road	The man in the purple hat is operating a camera that	۳/۴

طبق آنچه که در فصل گذشته توضیح دادیم روش پیشنهادی ما برای تخمین شباهت جفت جمله پس از بررسی ساختارهای متداول موجود در پردازش زبان طبیعی، به‌ویژه تخمین شباهت متون و جملات، از روش آماری جنسن شانون با ترکیب تخصیص پنهان دیریکلت استفاده کرده است و به‌وسیله بردار ماشین پشتیبان به طبقه‌بندی نمرات شباهت به‌دست‌آمده پرداختیم تا دقت مدل خود را به دست آوریم.

جدول ۵.۴: روش‌های مقایسه شده با روش پژوهش

روش مورد استفاده	نمره شباهت روی جملات ۴۱۴	نمره شباهت روی جملات ۷۳۰	دقت
word2vec-featurse + cosian similarity	۴/۰۸	۲/۷۲	۸۳/۴
Tfidf-featurse + manhatten-distance	۳/۷۳	۳/۲۸	۸۷/۹۱
Tfidf-featurse + euclidean-distance	۳/۴۲	۲/۷۱	۷۶/۲۳
LDA + jensen-shanon	۴/۷	۳/۳	۹۸/۵

داده‌های ورودی را به آموزش و تست تقسیم نمودیم. مجموعه آموزش برای آموزش مدل LDA استفاده می‌شود و مجموعه تست برای بازیابی جملات مشابه استفاده شدند.

طول کوتاه‌ترین جمله در مجموعه دادگان ما ۴ و طول بیشترین جمله ۳۰ می‌باشد. ما ۹۸۴۰ جمله داریم که به‌صورت تصادفی به دادگان آموزش و تست تقسیم کرده‌ایم. در اکثر اجراها تعداد دادگان آموزش بیشتر از ۶۰۰۰ و تقریباً دو برابر داده‌های تست می‌باشد. ما توانسته‌ایم با این روش زمان پیش‌پردازش داده‌ها را کاهش داده و به ۰/۱۲ ثانیه برسانیم. وقتی داده‌ها را با LDA آموزش می‌دهیم در۰/۰۷۸۹ ثانیه آموزش انجام می‌شود.

۵.۴ معیار ارزیابی

قدرت ارتباط بین دو متغیر به‌عنوان آزمون همبستگی شناخته می‌شود. برای ارزیابی روش پیشنهاد شده و مقایسه با سایر روش‌ها، از معیار ضریب همبستگی پیرسون، ضریب همبستگی اسپیرمن و دقت طبقه‌بندی استفاده می‌کنیم.

همبستگی پارامتری ـ همبستگی پیرسون[1](r): این‌یک وابستگی خطی بین دو متغیر (x,y) را اندازه‌گیری می‌کند که به‌عنوان یک آزمون همبستگی پارامتری شناخته می‌شود زیرا به توزیع داده‌ها بستگی دارد. هنگامی‌که پیرسون به حوزه تشابه معنایی اعمال می‌شود، درجه همبستگی بین نمرات آزمون‌های جفت مجموعه داده‌ها، انسان و نمرات شباهت مشاهده‌شده را نشان می‌دهد. بااین‌حال، همبستگی پیرسون نسبت به افراد خارج از مجموعه داده‌های معیار حساس است. اگرچه بسیاری از کارها مشکل اصلی را برطرف کرده‌اند، اما تحقیقات کمی در مورد توزیع داخلی سطل‌های مجموعه داده‌های معیار متمرکزشده است. مجموعه داده‌های معیار متن نماینده و توزیع‌شده،

[1] pearson rank correlation

طیف گسترده‌ای از مقادیر نمرات شباهت را در خود گنجانده است. بنابراین، مجموعه داده‌های معیار می‌تواند یک مجموعه داده مقطعی در نظر گرفته شود. اگرچه یک روش شباهت متن کامل می‌تواند یک همبستگی پیرسون را گزارش کند، اما همبستگی استاندارد پیرسون از مقادیر همبسته جداگانه در سطح مقطع داده آگاه نیست و دلیل آن این است که از محیط خارج می‌شود. بنابراین، با استفاده از روش خطای میانگین مربع مقیاس یافته عادی، که از خطای مقیاس استاندارد استنباط می‌شود، برای از بین بردن محیط‌های دور استفاده می‌کند.

در مباحث آماری، ضریب همبستگی پیرسون یا ضریب همبستگی حاصل‌ضرب‌ـ گشتاور پیرسون میزان همبستگی خطی بین دو متغیر تصادفی را می‌سنجد. مقدار این ضریب بین ۱ـ تا ۱ تغییر می‌کند که «۱» به معنای همبستگی مثبت کامل، «۰» به معنی نبود همبستگی، و «۱ـ» به معنی همبستگی منفی کامل است. این ضریب که کاربرد فراوانی در آمار دارد، توسط کارل پیرسون بر اساس ایده‌ی اولیه‌ی فرانسیس گالتون تدوین شد و با فرمول زیر محاسبه می‌گردد:

$$r_{xy} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\sqrt{n\sum y_i^2 - (\sum y_i)^2}} \qquad (۱.۴)$$

همبستگی غیر پارامتری ـ کندال (tau) و اسپیرمن[1](rho): آن‌ها ضرایب همبستگی مبتنی بر رتبه هستند که به‌عنوان همبستگی غیر پارامتری شناخته می‌شوند. در حقیقت برای سنجش ضریب همبستگی بین دو متغیر تصادفی است. مقدار این ضریب نشان‌دهنده‌ی قابلیت بیان یک متغیر به‌صورت تابعی یکنوا از متغیر دیگر است. همبستگی کامل پیرسون (۱+ یا ۱ـ) زمانی اتفاق می‌افتد که متغیری تابعی یکنوا از متغیر دیگر باشد. ضریب همبستگی اسپیرمن به‌صورت ضریب همبستگی پیرسون بین داده‌های رتبه‌بندی شده تعریف می‌شود. به‌عنوان‌مثال، اگر n زوج داده به‌صورت داده‌شده، ابتدا رتبهٔ هر داده را به‌صورت حساب کرده و سپس ضریب همبستگی اسپیرمن را به‌صورت زیر حساب می‌کنیم:

$$r_s = ۱ - \frac{۶\sum d_i^۲}{n(n^۲ - ۱)} \qquad (۲.۴)$$

معیارهای صحت[2]، حساسیت[3]، معیارـF[4] و دقت[5] از مهم‌ترین معیارهای مورداستفاده برای ارزیابی هستند. دقت معیاری برای نزدیک بودن اندازه‌گیری‌ها به یک مقدار خاص است، درحالی‌که صحت نزدیکی اندازه‌گیری‌ها به یکدیگر است، یعنی لزوماً با یک مقدار خاص نیست. به‌عبارت‌دیگر: اگر مجموعه‌ای از نقاط داده را از اندازه‌گیری‌های مکرر همان مقدار داشته باشیم، گفته می‌شود مجموعه دقیق است اگر میانگین آن‌ها نزدیک به مقدار واقعی اندازه‌گیری شده باشد. از طرف دیگر، اگر مقادیر نزدیک به هم باشند، مجموعه صحت دارد. این دو مفهوم مستقل از یکدیگر هستند، به

[1] spearman rank correlation
[2] Precision
[3] Recall
[4] F-measure
[5] Accuracy

این معنی که مجموعه داده‌ها می‌توانند دقیق باشند یا صحت داشته باشند یا هر دو باشند یا هیچ‌یک. هرکدام از آن‌ها طبق فرمول‌های زیر محاسبه می‌شوند:

$$Percision = \frac{t_p}{t_p + f_p} \tag{۳.۴}$$

$$Recall = \frac{t_p}{t_p + f_n} \tag{۴.۴}$$

$$F - measure = ۲ * \frac{Percision * Recall}{Percision + Recall} \tag{۵.۴}$$

$$Accuracy = \frac{t_p + t_n}{t_p + t_n + f_p + f_n} \tag{۶.۴}$$

در این روابط، tp تعداد پاسخ مثبت درست، tn تعداد پاسخ منفی درست، fp تعداد پاسخ مثبت غلط و fn تعداد پاسخ منفی غلط می‌باشد.

۶.۴ نتایج

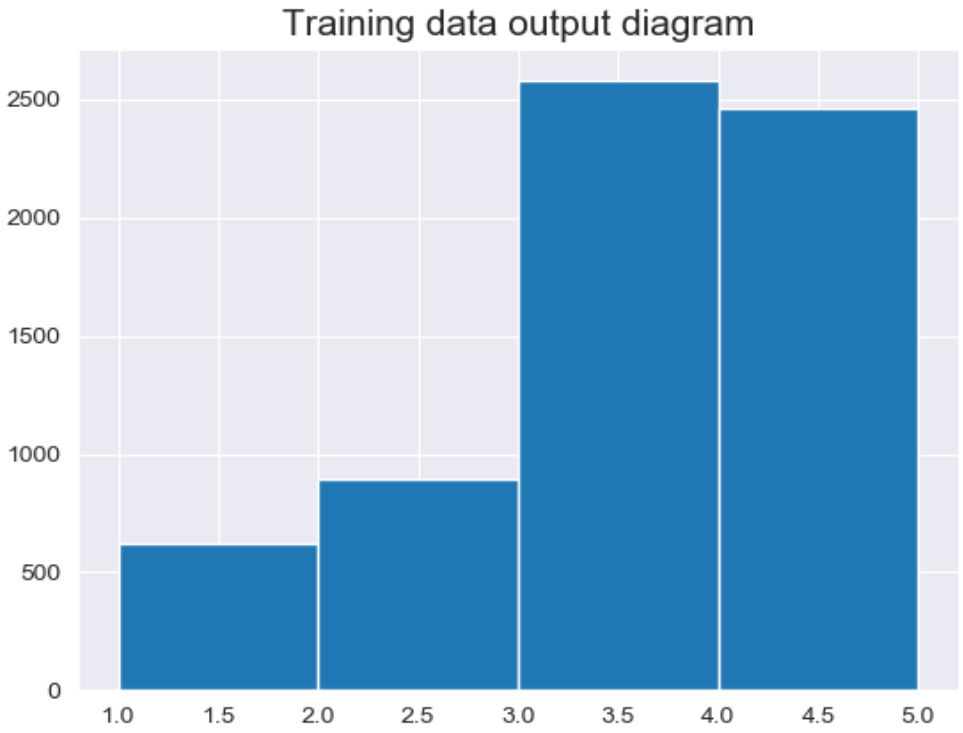
ما آموزش داده‌ها را چندین بار تکرار کردیم و توانستیم مدل خود را به‌خوبی آموزش دهیم. در جدول ۶.۴ نتایج به‌دست‌آمده بر اساس امتیاز ارتباط بین جملات روی‌داده‌های آموزش، به درصد بیان‌شده است. همان‌طور که مشاهده می‌کنید ما توانسته‌ایم با روش خود مدل را به‌خوبی آموزش داده و به خطای بسیار کمی برای به دست آوردن نمره ارتباط بین جملات دست پیدا کنیم. بعدازآن با آزمایش مدل نتیجه کار خود را بر روی‌داده‌های این مجموعه مشاهده می‌کنیم.

جدول ۶.۴: نتایج طبقه‌بندی مجموعه دادگان آموزش براساس امتیازات

محدوده داده‌ها	درصد
محدوده [۱،۲]	۸/۸۹
محدوده [۲،۳]	۱۵/۳۳
محدوده [۳،۴]	۳۹/۰۷
محدوده [۴،۵]	۳۶/۷۱

در شکل ۲.۴ مشاهده می‌کنیم که داده‌های آموزش بر اساس نمرات به ۴ دسته بین ۱ تا ۲ و دسته بعدی بین ۲ تا ۳ و به همین صورت دسته‌بندی‌شده‌اند.

شکل ۲.۴: نموداردسته‌بندی داده‌های آموزش



در انتها با ماشین بردار پشتیبان خطی، کرنل چندجمله‌ای و کرنل تابع پایه شعاعی به دسته‌بندی فاصله‌های به‌دست‌آمده پرداختیم. در جدول ۷.۴ با ماشین بردار پشتیبان خطی به‌دقت ۸۹ درصد بر روی داده‌ها دست‌یافته‌ایم. همچنین معیار صحت و حساسیت و معیار_F درهر دسته مشخص‌شده است.

جدول ۷.۴: ماشین بردار پشتیبان خطی

	Percision	**recall**	**f۱-score**	**Support**
۱	۸۹.۰	۷۸.۰	۸۳.۰	۹۱۲
۲	۸۹.۰	۷۸.۰	۸۳.۰	۱۳۵۹
۳	۸۹.۰	۷۸.۰	۸۳.۰	۳۸۳۲
۴	۸۹.۰	۷۸.۰	۸۳.۰	۳۳۵۷
۵	۰۰.۰	۰۰.۰	۰۰.۰	۲۸۲
Accuracy				
avg	۸۹.۰	۸۹.۰	۸۹.۰	۹۷۴۲

در جدول ۸.۴ نیز با ماشین بردار پشتیبان کرنل چندجمله‌ای به دقت ۷۶ درصد بر روی داده‌ها دست یافته‌ایم.

جدول ۸.۴: ماشین بردار پشتیبان کرنل چند جمله‌ای

	Percision	recall	f۱-score	Support
۱	۹۰.۰	۹۳.۰	۹۲.۰	۹۱۲
۲	۶۵.۰	۳۰.۰	۴۱.۰	۱۳۵۹
۳	۶۸.۰	۹۵.۰	۸۰.۰	۳۸۳۲
۴	۸۹.۰	۷۵.۰	۸۲.۰	۳۳۵۷
۵	۰۰.۰	۰۰.۰	۰۰.۰	۲۸۲
Accuracy				
avg	۷۶.۰	۷۶.۰	۷۶.۰	۹۷۴۲

جدول ۹.۴ هم با استفاده از ماشین بردار پشتیبان مدل کرنل تابع پایه شعاعی به دقت دسته‌بندی ۸۷ درصد دست یافته‌ایم .

جدول ۹.۴: ماشین بردار پشتیبان کرنل تابع پایه شعاعی

	Percision	recall	f۱-score	Support
۱	۹۰.۰	۸۸.۰	۸۹.۰	۹۱۲
۲	۸۰.۰	۹۲.۰	۸۶.۰	۱۳۵۹
۳	۸۷.۰	۹۵.۰	۹۱.۰	۳۸۳۲
۴	۹۰.۰	۸۴.۰	۸۷.۰	۳۳۵۷
۵	۰۰.۰	۰۰.۰	۰۰.۰	۲۸۲
Accuracy				
avg	۸۷.۰	۸۷.۰	۸۷.۰	۹۷۴۲

همانطور که در جدول ۱۰.۴ مشاهده می‌نمایید روش ماشین بردار پشتیبان خطی نسبت به سایر روش‌ها توانسته به دقت بالاتری دست یابد. به همین جهت برای دسته‌بندی از این روش استفاده کردیم.

در زیر ضریب همبستگی پیرسون و اسپرمن را نیز مشاهده می‌کنیم:

$$r_s = ۰/۹۲۵ \quad r_{xy} = ۰/۸۹۱$$

جدول ۱۰.۴: مقایسه دقت بدست آمده روش‌های دسته‌بندی

روش‌های دسته‌بندی	دقت
$LinerSVM$	۰/۸۹
$PolynomialSVM$	۰/۷۶
$RBFSVM$	۰/۸
$NaiveBayesclassifier$	۰/۷۱

همچنین ما روش خود را با روش‌های موجود در فصل دوم که از مجموعه دادگان SICK استفاده کرده‌اند را مورد مقایسه قرار دادیم. نتایج این مقایسه را در جدول۱۱.۴ مشاهده می‌نماییم. همانطور که در جدول می‌بینیم روش ما توانسته‌است نسبت MaLSTM، به بهبود دقت ۳/۸ درصدی و نسبت به روش DLSTM، به بهبود ضریب همبستگی پیرسون دست یابد.

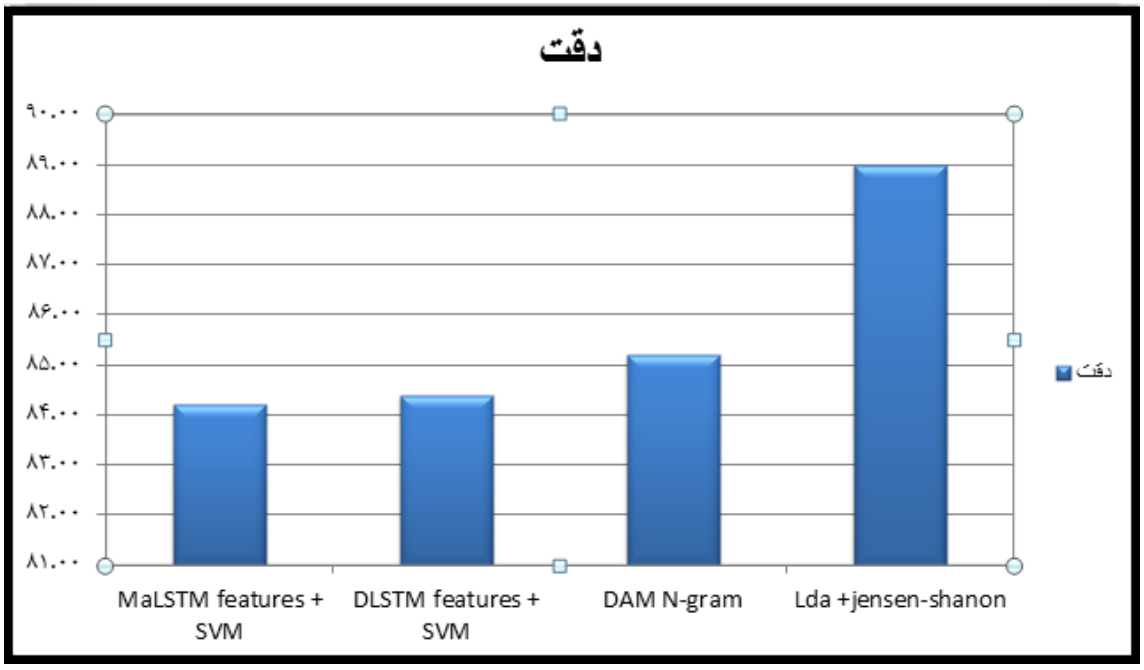
جدول ۱۱.۴: مقایسه روش ارائه شده با سایر روش‌ها

روش مورد استفاده	ضریب همبستگی	دقت
$PARAGRAM-PHRASE$	۰/۷۲	
$Sent۲Vec$	۰/۷۰	۸۱/۴
$MaLSTM features+SVM$	۰/۸۸۲۲	۸۵/۲
$DLSTM features+SVM$	۰/۸۳۰۵	۸۴/۲
$LDA+jensen-shanon$	۰/۸۹۱	۸۹

در شکل ۳.۴ نمودار مقایسه دقت روش پیشنهادی ما با سایر روش‌هایی که از مجموعه دادگان SICK استفاده کرده‌اند نشان داده شده است . همانطور که می‌بینید دقت این پژوهش نسبت به دیگر روش‌ها بهتر عمل کرده است.

۷.۴ جمع‌بندی

در این فصل به بررسی نحوه پیاده‌سازی روش پیشنهادی خود پرداخته، نتایج به‌دست‌آمده را نیز بیان کردیم. همچنین روش خود را با سایر روش‌ها مقایسه نمودیم. دیدیم که ازلحاظ دقت و ضریب همبستگی روش مایکی از بهترین روش‌های موجود میان روش‌های ذکرشده می‌باشد. این روش در به دست آوردن نمره جفت جملات و برچسب کلاس‌ها می‌تواند مؤثر باشد و در کارهای آینده مورداستفاده قرار گیرد.



شکل ۳.۴: نمودار دقت روش‌های بیان شده

فصل ۵

نتیجه‌گیری

۱.۵ نتیجه‌گیری

در این پژوهش دقت دسته‌بندی داده‌ها اهمیت داشت. در فصل اول به تعریف موضوع پرداختیم. در مورد تاریخچه این مطلب و همچنین کاربردهای شباهت معنایی و لزوم استفاده از آن‌ها صحبت کردیم.

در فصل دوم پیشینه‌ی تحقیق موردبررسی قرار گرفت. مبانی نظری استفاده‌شده در پژوهش‌های پیشین تعریف گردید. روش‌های زیادی برای استخراج ویژگی‌ها بیان شد ازجمله N ـ گرم‌ها، فرکانس مدت ـ فرکانس اسناد معکوس، چارچوب کلمه به بردار و دیگر موارد. سپس روش‌های طبقه‌بندی متن و اندازه‌گیری شباهت بیان گردید ازجمله روش‌های فاصله کسینوسی، فاصله اقلیدسی، شباهت ژاکارد و فاصله همینگ بودند. همچنین به توضیح روش‌های طبقه‌بندی ازجمله روش‌های آماری، شبکه‌های عصبی و روش‌های یادگیری عمیق پرداختیم و دیدیم روش‌های بسیاری برای تخمین شباهت متن وجود دارند ازجمله روش‌های آماری، روش‌های معنایی، مبتنی بر رشته و روش‌های ترکیبی که به مواردی از هرکدام اشاره شد و باهم مقایسه شدند.

در فصل سوم به توضیح کار خود پرداخته که ابتدا داده‌های متنی خود را پیش‌پردازش کردیم و از روش آماری تخصیص پنهان دیریکلت و استخراج ویژگی داده‌های پردازش‌شده از فن doc2bow استفاده کردیم که هر جمله را به مجموعه عددی تبدیل می‌کند. سپس از فاصله جنسن شانون برای تخمین شباهت جملات استفاده کردیم که یکی از روش‌های مؤثر در این زمینه می‌باشد. در انتها به دسته‌بندی نتایج با ۳ نوع ماشین بردار پشتیبان خطی، چندجمله‌ای و تابع پایه شعاعی پرداختیم.

در فصل چهارم دادگان مورداستفاده در پژوهش و پیاده‌سازی انجام‌شده را توضیح دادیم. در انتها نتایج به‌دست‌آمده یعنی نمره شباهت برای هر ۱۰۰۰۰ جمله را طبقه‌بندی نمودیم تا دقت راهکار خود را بررسی کنیم و به مقایسه روش خود با سایر روش‌هایی که از مجموعه دادگان موردنظر ما استفاده کرده بودند پرداختیم. دقت اندازه‌گیری یکی از مسائل مهم در این پژوهش است.

در دنیای امروزی داده‌های اینترنتی بیش‌ازپیش مورداستفاده قرارگرفته‌اند و استفاده از کتابخانه‌های دیجیتالی گسترش پیداکرده است. ممکن است داده‌های مشابه به دلایل مختلفی شکل‌گرفته باشند. برای بهره‌برداری صحیح از این داده‌ها استفاده از فن‌های یافتن مقدار این شباهت مهم قلمداد می‌شود.

همچنین یکی از روش‌هایی که برای تخمین شباهت معنایی متون مورداستفاده قرارگرفته است روش‌های آماری در ترکیب با سایر روش‌ها می‌باشد. ما نیز از این چارچوب برای پژوهش خود استفاده کرده‌ایم.

مورد دیگری که در این پژوهش وجود دارد، استفاده از پایگاه داده استاندارد می‌باشد که در پژوهش‌های مشابه دیگری نیز مورداستفاده قرارگرفته‌اند و ما به مقایسه روش خود با این روش‌ها پرداخته‌ایم.

۲.۵ کارهای آینده

پژوهش‌های بسیاری می‌تواند در رابطه با این موضوع صورت بگیرد که در زیر تعدادی از این روش‌ها را پیشنهاد می‌دهیم:

- استفاده از داده‌های فارسی به‌جای دادگان انگلیسی و ارائه روشی مناسب برای اینکه دقت تخمین شباهت معنایی داده‌های فارسی را بالا ببریم.
- استفاده از روش‌های یادگیری عمیق برای استخراج ویژگی‌ها ازجمله شبکه RBM و شبکه باور عمیق یعنی DBN که می‌تواند نتیجه مطلوبی داشته باشد و یکی از فن‌های مناسب استخراج ویژگی است.
- استفاده از روش‌های یادگیری عمیق و شبکه‌های پیچیده‌تر برای طبقه‌بندی به‌طور مثال شبکه‌های هیبرید عمیق و شبکه باور عمیق و دیگر شبکه‌های عمیق که نتایج مناسبی را برای این کار در پی‌دارند.
- استفاده از روش‌های ترکیبی برای اندازه‌گیری فاصله بین بردارها مثلاً ترکیب فاصله جنسن شانون و فاصله همینگ .
- استفاده از روش fasttext و glove برای استخراج ویژگی جملات پردازش‌شده.

مراجع

[1] Farouk, M. (2019). Measuring sentences similarity: a survey. arXiv preprint arXiv:1910.03940.

[2] Agirre, E., Banea, C., Cer, D., Diab, M., Gonzalez Agirre, A., Mihalcea, R., Wiebe, J. (2016). Semeval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In SemEval-2016. 10th International Workshop on Semantic Evaluation; 2016 Jun 16-17; San Diego, CA. Stroudsburg (PA): ACL; 2016. p. 497-511.. ACL (Association for Computational Linguistics).

[3] Cai, Y., Zhang, Q., Lu, W., Che, X. (2018). A hybrid approach for measuring semantic similarity based on IC-weighted path distance in WordNet.Journal of Intelligent Information Systems,51(1),23-47.

[4] Majumder, G., Pakray, P., Gelbukh, A., Pinto, D. (2016). Semantic textual similarity methods, tools, and applications: A survey. Computación y Sistemas,20(4),647-665.

[5] Huang, H., Wu, H., Wei, X., Gao, Y., Shi, S. (2019). Mapping sentences to concept transferred space for semantic textual similarity. Knowledge and Information Systems,60(3),1353-1376.

[6] Vardasbi, A., Faili, H., Asadpour, M. (2017, October). Eigenvalue based features for semantic sentence similarity. In 2017 Artificial Intelligence and Signal Processing Conference (AISP)(pp. 184-189).IEEE.

[7] Rychalska, B., Pakulska, K., Chodorowska, K., Walczak, W., Andruszkiewicz, P. (2016, June). Samsung Poland NLP Team at SemEval-2016 Task 1: Necessity for diversity; combining recursive autoencoders, WordNet and ensemble methods to measure semantic similarity. In Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)(pp.602-608).

[8] Wali, W., Gargouri, B., Hamadou, A. B. (2017). Enhancing the sentence similarity measure by semantic and syntactico-semantic knowledge. Vietnam Journal of Computer Science,4(1),51-60.

[9] Lopez-Gazpio, I., Maritxalar, M., Lapata, M., Agirre, E. (2019). Word n-gram attention models for sentence similarity and inference. Expert Systems with Applications, 132, .1-11

[10] Leacock, C., Chodorow, M. (1998). Combining local context and WordNet similarity for word sense identification. WordNet: An electronic lexical database, 49(2), 265-283.

[11] Rada, R., Mili, H., Bicknell, E., Blettner, M. (1989). Development and application of a metric on semantic nets. IEEE transactions on systems, man, and cybernetics, 19(1), 17-30.

[12] Wu, Z., Palmer, M. (1994, June). Verbs semantics and lexical selection. In Proceedings of the 32nd annual meeting on Association for Computational Linguistics (pp. 133-138). Association for Computational Linguistics

[13] Jiang, Y., Bai, W., Zhang, X., Hu, J. (2017). Wikipedia-based information content and semantic similarity computation. Information Processing Management, 53(1), 248-265.

[14] Meng, L., Gu, J., Zhou, Z. (2012). A new model of information content based on concept's topology for measuring semantic similarity in WordNet. International Journal of Grid and Distributed Computing, 5(3), 81-94.

[15] Sánchez, D., Batet, M., Isern, D. (2011). Ontology-based information content computation. Knowledge-Based Systems, 24(2), 297-303.

[16] Taieb, M. A. H., Aouicha, M. B., Hamadou, A. B. (2013). Computing semantic relatedness using Wikipedia features. Knowledge-Based Systems,50,260-278.

[17] Taieb, M. A. H., Aouicha, M. B., Tmar, M., Hamadou, A. B. (2012, November). Wikipedia category graph and new intrinsic information content metric for word semantic relatedness measuring. In International Conference on Data and Knowledge Engineering (pp. 128-140). Springer, Berlin, Heidelberg.

[18] Jiang, Y., Zhang, X., Tang, Y., Nie, R. (2015). Feature-based approaches to semantic similarity assessment of concepts using Wikipedia. Information Processing Management,51(3),215-234.

[19] Petrakis, E. G., Varelas, G., Hliaoutakis, A., Raftopoulou, P. (2006). X-similarity: Computing semantic similarity between concepts from different ontologies. Journal of Digital Information Management,4(4).

[20] Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. arXiv preprint cmp-lg/9511007

[21] Rodríguez, M. A., Egenhofer, M. J. (2003). Determining semantic similarity among entity classes from different ontologies. IEEE transactions on knowledge and data engineering,15(2),442-456.

[22] Tversky, A. (1977). Features of similarity. Psychological review,84(4),327.

[23] Qu, R., Fang, Y., Bai, W., Jiang, Y. (2018). Computing semantic similarity based on novel models of semantic representation using Wikipedia. Information Processing Management,54(6)1002-.1021

[24] Nguyen, H. T., Duong, P. H.,Cambria, E. (2019). Learning short-text semantic similarity with word embeddings and external knowledge sources. Knowledge-Based Systems, 182, 104842.

[25] Li, Y., McLean, D., Bandar, Z. A., O'shea, J. D., Crockett, K. (2006). Sentence similarity based on semantic nets and corpus statistics. IEEE transactions on knowledge and data engineering, 18(8), .1138-1150

[26] Arroyo-Fernández, I., Méndez-Cruz, C. F., Sierra, G., Torres-Moreno, J. M., Sidorov, G. (2019). Unsupervised sentence representations as word information series: Revisiting TF–IDF. Computer Speech Language,56,107-129.

[27] Li, Y., McLean, D., Bandar, Z. A., O'shea, J. D., Crockett, K. (2006). Sentence similarity based on semantic nets and corpus statistics. IEEE transactions on knowledge and data engineering,18(8),1138-1150.

[28] Mihalcea, R., Corley, C., Strapparava, C. (2006, July). Corpus-based and knowledge-based measures of text semantic similarity. In Aaai (Vol. 6, No.2006,pp.775-780)

[29] Agirre, E., Cer, D., Diab, M., Gonzalez Agirre, A., Guo, W. (2013). Semantic textual similarity, including a pilot on typed similarity. In The Second Joint Conference on Lexical and Computational Semantics.

[30] Agirre, E., Diab, M., Cer, D., Gonzalez-Agirre, A. (2012, June). Semeval-2012 task 6: A pilot on semantic textual similarity. In Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (pp. 385-393). Association for Computational Linguistics.

[31] Agirre, E., Cer, D., Diab, M., Gonzalez-Agirre, A., Guo, W. (2013, June). * SEM 2013 shared task: Semantic textual similarity. In Second Joint Conference on Lexical and Computational Semantics (SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity(pp32-43).

[32] Neculoiu, P., Versteegh, M., Rotaru, M. (2016, August). Learning text similarity with siamese recurrent networks. In Proceedings of the 1st Workshop on Representation Learning for NLP(pp.148-.(157

[33] Brychcín, T., Svoboda, L. (2016, June). UWB at SemEval-2016 Task 1: Semantic textual similarity using lexical, syntactic, and semantic information. In Proceedings of the 10th International Workshop on Semantic Evaluation(SemEval-2016)(pp.588-594).

[34] Sultan, M. A., Bethard, S.,Sumner, T. (2014). Back to basics for monolingual alignment: Exploiting word similarity and contextual evidence. Transactions of the Association for Computational Linguistics,2,219-230.

[35] Sultan, M. A., Bethard, S., Sumner, T. (2014, August). Dls@ cu: Sentence similarity from word alignment. In Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)(pp.241-246).

[36] Sultan, M. A., Bethard, S., Sumner, T. (2015, June). Dls@ cu: Sentence similarity from word alignment and semantic vector composition. In Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)(pp.148-153).

[37] Li Y, McLean D, Bandar ZA, O'shea JD, Crockett K (2006) Sentence similarity based on semantic nets and corpus statistics. IEEE Trans Knowl Data Eng 18(8):1138–1150.

[38] Mihalcea R, Corley C, Strapparava C (2006) Corpus-based and knowledge-based measures of text semantic similarity. In: Proceedings of the twenty-first national conference on artificial intelligence and the eighteenth innovative applications of artificial intelligence conference, Boston, Massachusetts, USA,pp 775–780.

[39] Sultan, M. A., Bethard, S.,Sumner, T. (2015, June). Dls@ cu: Sentence similarity from word alignment and semantic vector composition. In Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015) (pp. .(148-153

[40] Amir, S., Tanasescu, A., Zighed, D. A. (2017). Sentence similarity based on semantic kernels for intelligent text retrieval. Journal of Intelligent Information Systems,48(3),675-689.

[41] Crystal D. , The Cambridge Encyclopedia of Language, 2nd edition, Cambridge University Press, 1996. ISBN 0-521-55967-7 Manning,C. D.

[42] Schutze, H. , Foundations of Statistical Natural Language Processing, 5th edition, The MIT Press, 2002.ISBN 0-262-13360-1

[43] Borda, Monica (2011). Fundamentals in Information Theory and Coding. Springer. ISBN 978-3-642-20346-6.

[44] Han, Te Sun Kobayashi, Kingo (2002). Mathematics of Information and Coding. American Mathematical Society. ISBN .978-0-8218-4256-0

[45] Shirakawa, M., Nakayama, K., Hara, T., Nishio, S. (2015). Wikipedia-based semantic similarity measurements for noisy short texts using extended naive bayes. IEEE Transactions on Emerging Topics in Computing,3(2),205-219.

[46] Broder, A. Z. (1997). Syntactic clustering of the Web, J of Computer Networks and ISDN Systems, Volume 29,Issue .8-13

[47] Greenberg, R. I. (2003). Bounds on the number of longest common subsequences. arXiv preprint cs/0301030.

[48] Sivic, J., Zisserman, A. (2008). Efficient visual search of videos cast as text retrieval. IEEE transactions on pattern analysis and machine intelligence, 31(4), 591-606.

[49] Schölkopf, B., Smola, A. J. (2002). Learning with kernels. MIT Press, Cambridge, MA. Learning with kernels. MIT Press, Cambridge, MA.

[50] Chatterjee, N. (2001, February). A statistical approach for similarity measurement between sentences for EBMT. In Proceedings of Symposium on Translation Support Systems STRANS-2001.

[51] Cavalcanti, A. P., de Mello, R. F. L., Ferreira, M. A. D., Rolim, V. B., Tenório, J. V. S. (2017, October). Statistical and semantic features to measure sentence similarity in portuguese. In 2017 Brazilian Conference on Intelligent Systems (BRACIS) (pp. 342-347).IEEE.

[52] Majumder, G., Pakray, P., Avendaño, D. E. P. (2018, January). Semantic Textual Similarity Using Lexical and Cosine Similarity. In Annual Convention of the Computer Society of India (pp. 717-732). Springer, Singapore.

[53] He, H., Wieting, J., Gimpel, K., Rao, J., Lin, J. (2016, June). UMD-TTIC-UW at SemEval-2016 Task 1: Attention-based multi-perspective convolutional neural networks for textual similarity measurement. In Proceedings of the 10th International Workshop on Semantic Evaluation(SemEval-2016)(pp.1103-1108).

[54] Mueller, J.,Thyagarajan, A. (2016, March). Siamese recurrent architectures for learning sentence similarity. In Thirtieth AAAI Conference on Artificial Intelligence.

[55] Wieting, J., Bansal, M., Gimpel, K., Livescu, K. (2016). Charagram: Embedding words and sentences via character n-grams.arXiv preprint arXiv:1607.02789.

[56] Zhu, W., Yao, T., Ni, J., Wei, B., Lu, Z. (2018). Dependency-based Siamese long short-term memory network for learning sentence representations. PloS one, 13(3),e0193919.

[57] Pontes, E. L., Huet, S., Linhares, A. C.,Torres-Moreno, J. M. (2018). Predicting the semantic textual similarity with siamese CNN and LSTM. arXiv preprint arXiv:1810.10641.

[58] Orasan, C., Mitkov, R., Ranasinghe, T. (2019, September). Semantic textual similarity with siamese neural networks.RANLP.

[59] Li Y, McLean D, Bandar ZA, O'shea JD, Crockett K (2006) Sentence similarity based on semantic nets and corpus statistics. IEEE Trans Knowl Data Eng 1150–1138:(8)18

[60] Pawar, A., Mago, V. (2018). Similarity between learning outcomes from course objectives using semantic analysis, blooms taxonomy and corpus statistics.arXiv preprint arXiv:1804.06333.

[61] Bär, D., Biemann, C., Gurevych, I.,Zesch, T. (2012, June). Ukp: Computing semantic textual similarity by combining multiple content similarity measures. In Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (pp. 435-440). Ass ociation for Computational Linguistics.

[62] Atoum, I., Otoom, A. (2016). Efficient hybrid semantic text similarity using wordnet and a corpus. Int. J. Adv. Comput. Sci. Appl,7(9),124-130.

[63] Baba, K., Nakatoh, T., Minami, T. (2017). Plagiarism detection using document similarity based on distributed representation. Procedia computer science, 111, .382-387

[64] Islam, A.,Inkpen, D. (2008). Semantic text similarity using corpus-based word similarity and string similarity. ACM Transactions on Knowledge Discovery from Data (TKDD), 2(2), .10

[65] Roostaee, M., Fakhrahmad, S. M., Sadreddini, M. H., Khalili, A. (2014). Efficient calculation of sentence semantic similarity: a proposed scheme based on machine learning approaches and NLP techniques. Scientific Journal of Review,3(3),94-106.

[66] Afzal, N., Wang, Y., Liu, H. (2016). Mayonlp at semeval-2016 task 1: Semantic textual similarity based on lexical semantic net and deep learning semantic model. In Proceedings of the 10th International Workshop on Semantic Evaluation(SemEval-2016)(pp.674-679).

[67] Camacho-Collados, J., Pilehvar, M. T., Collier, N., Navigli, R. (2017, August). Semeval-2017 task 2: Multilingual and cross-lingual semantic word similarity. In Proceedings of the 11th International Workshop on Semantic Evaluation(SemEval-2017)(pp.15-26).

[68] Pagliardini, M., Gupta, P.,Jaggi, M. (2017). Unsupervised learning of sentence embeddings using compositional n-gram features. arXiv preprint arXiv:1703.02507.

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

[70] Wieting, J., Bansal, M., Gimpel, K.,Livescu, K. (2016). Charagram: Embedding words and sentences via character n-grams. arXiv preprint arXiv:1607.02789.

[71] Rakib, R. H., Islam, A., Milios, E. (2018). Improving text relatedness by incorporating phrase relatedness with word relatedness. Computational Intelligence,34(3),939-966.

[72] Glavaš, G., Franco-Salvador, M., Ponzetto, S. P., Rosso, P. (2018). A resource-light method for cross-lingual semantic textual similarity. Knowledge-Based Systems,143,1-9.

[73] Plaza-del-Arco, F. M., Molina-González, M. D., Ureña-López, L. A., Martín-Valdivia, M. T. (2021). Comparing pre-trained language models for Spanish hate speech detection. Expert Systems with Applications, 166,114120.

[74] Buenaño-Fernandez, D., González, M., Gil, D., Luján-Mora, S. (2020). Text Mining of Open-Ended Questions in Self-Assessment of University Teachers: An LDA Topic Modeling Approach. IEEE Access, 8,35318-35330.

[75] Schölkopf, B., Smola, A. J. (2002). Learning with kernels. MIT Press, Cambridge, MA. Learning with kernels. MIT Press, Cambridge, MA

[۷۶] شیرازی حسین،(۱۳۹۱)، پنجمین کنفرانس داده کاوی ایران، ''ارائه رویکرد جدید در محاسبه شباهت معنایی جملات فارسی''،ص۵۷ ، تهران

[۷۷] ملکوتی ن،(۱۳۹۱)، ''ارائه یک روش جدید برای بررسی میزان شباهت اسناد متنی''، دانشکده مهندسی برق و کامپیوتر، دانشگاه شیراز

Aabstract

In today's world, due to the increasing volume of internet information and the multiplicity of different digital cultures, the existence of similarities in textual data has increased for various reasons. For this reason, estimating the degree of similarity between them is necessary. Finding similarities between textual data is also used in areas such as information retrieval systems, plagiarism, data mining, document classification, and much more.

By entering a sentence or document or text into the system, its similarity with existing documents is checked and Used in the field in question. So far, various methods of natural language processing and machine learning have been introduced to calculate the similarity of textual data.These methods have achieved different degrees of accuracy, and better research is needed to improve this criterion. Among the statistical methods, which is one of the corpus-based methods, and by combining this method with other methods, interesting results can be achieved.

In this research, data is applied to the preprocessing machine before converting the text to readable format to reduce the features and increase the measurement accuracy. Each word is then assigned a correlation score using Latent Dirichlet Allocation, and this process is repeated based on the probabilities to improve the assignment of the correlation score to the words and the sentence is placed in the desired category. Doc2bow is also used to representation sentencesis. Then, to calculate the most similar sentences, the Jensen Shannon distance is used, which is obtained by comparing the divergence of the desired label distribution. Finally, the machine uses the linear support vector and the Radial base function to classify similar sentences. The proposed method achieved a classification accuracy of 89 percent and a correlation score of 0.92 percent on the SICK database and was able to be 3.8 percent better accurate than the MaLSTM research.

Keywords: text semantic similarity, Latent Dirichlet Allocation, document to Bag of words, Jensen-Shannon distance, Support vector machines

Shahrood University of Technology

Faculty Of Computer Engineering

MSc Thesis in: Artificial Intelligence Engineering

Estimating semantic based text similarity by using statistical methods

By: Fereshteh Riahi

Supervisor

Morteza Zahedi

Advisor

AliReza Tajari

February 2021