Scientific Journal of Electronical & Cyber Defence
Vol. 10, No. 2, Summer 2022, Serial No. 38

Investigation of the Semantic Similarity of Persian Sentences Using Vector Space Adaptation And Deep Learning

S. S. Sadidpour*, M. Haji Gholamreza, M. R. Mohammadzadeh, . M. R. Mohammadi, M. A. Keivanrad

* Assistant Professor, Malik Ashtar University of Technology, Tehran, Iran

(Received: 03/05/2021, Accepted: 01/01/2022)

ABSTRACT

Nowadays, similar texts recognition is a subject with many applications and due to its significance, has been analyzed and studied in various languages by researchers. In the past, sentences were often used as a set of words to be understood by computer systems. But today, with the spread of technology and the use of deep neural networks, the main concept of sentences can be extracted from the sentences themselves. Therefore, achieving a model that can encode sentences and extract the main concept of the sentence as accurately as possible is one of the essential needs for this purpose. This paper intends to use deep learning methods to evaluate the degree of semantic similarity between sentences. As the deep learning methods need many data, this paper employs an inter-linguistic mapping idea. The proposed method maps an English word embedding vector space into Persian, and Persian sentence similarity is calculated by a trained model in English and finally the outcome is compared with human scores. The results of the proposed method show the accuracy of the proposed system to be 89%, which is superior to other deep learning models.

Keywords: Sentence Concept Extraction, Word Embedding, Deep Learning, Sentence Similarity Extraction, Vector Space Adaptation.

* Corresponding Author Email: Sadidpour@mut.ac.ir

This article is an open-access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

Publisher: Imam Hussein University © Authors

نشریه علمی "پدافند الکترونیکی و سایبری"
سال دهم، شماره ۲، تابستان ۱۴۰۱، ص ۵۶-۴۳

علمی – پژوهشی

شباهت معنایی جملات فارسی با استفاده از تطبیق فضای برداری و یادگیری عمیق

سعیده سادات سدیدپور[1]*، مینا حاجی غلامرضا[2]، محمدرضا محمدزاده[3]، سیدمحمدرضا محمدی[4]، محمدعلی کیوان‌راد[5]

۱- استادیار، ۳،۲و ٤- کارشناسی ارشد، ۵- استادیار، دانشگاه صنعتی مالک‌اشتر، تهران، ایران

(دریافت: ۱٤۰۰/۰۲/۱۳، پذیرش: ۱٤۰۰/۱۰/۱۱)

چکیده

امروزه، شناسایی متون مشابه، موضوعی با کاربردهای فراوان می‌باشد که با توجه به اهمیت آن، توسط پژوهشگران زبان‌های مختلف مورد تحلیل و بررسی قرار گرفته است. در گذشته اغلب برای درک جملات توسط سامانه‌های رایانه‌ای، جملات به‌صورت مجموعه کلمات مورد بررسی قرار می‌گرفتند. اما امروزه، با گسترش فناوری و استفاده از شبکه‌های عصبی عمیق، می‌توان از خود جملات، مفهوم اصلی را استخراج نمود. بنابراین، رسیدن به مدلی که بتواند جملات را کدگذاری کرده و مفهوم اصلی جمله را با دقت هر چه بیشتر استخراج نماید، یکی از نیازهای ضروری برای این هدف به شمار می‌رود. این مقاله قصد دارد تا میزان شباهت جملات را از نقطه نظر معنایی به‌دست آورد که از روش‌های یادگیری عمیق استفاده می‌کند. از آنجایی که روش‌های یادگیری عمیق نیاز به داده آموزشی زیادی دارند، این مقاله از ایده نگاشت بین زبانی بهره می‌برد. روش پیشنهادی، فضای برداری تعبیه کلمات انگلیسی را به فارسی نگاشت کرده و با کمک مدل آموزش داده شده در زبان انگلیسی، شباهت جملات فارسی به‌دست می‌آید. درنهایت، نتایج نهایی با امتیازات انسانی مورد مقایسه قرار گرفته است. نتایج حاصل از روش پیشنهادی، میزان دقت این سامانه پیشنهادی را ۸۹ درصد ارائه می‌دهد که نسبت به سایر مدل‌های یادگیری عمیق برتری دارد.

کلیدواژه‌ها: استخراج مفهوم جملات، تعبیه کلمات، یادگیری عمیق، مشابهت‌یابی، تطبیق فضای برداری

۱- مقدمه

بسیاری از روش‌های پردازش متن و NLP، نیاز به نمایش عددی کلمات و متون دارند تا بتوان از انواع روش‌های عددی حوزه یادگیری ماشین استفاده نمود. همچنین اولین گام جهت طراحی سامانه‌های شباهت معنایی و بازیابی اطلاعات، مدل و الگوریتمی برای توصیف و تعیین شباهت‌های موجود میان اطلاعاتی که در اختیار دارد می‌باشد. یکی از الگوریتم‌های مهمی که در این بحث، کاربرد بسیار دارد، الگوریتم فضای برداری است که تمام مدارک را در مجموعه و جستجوهای کاربر به‌صورت بردار نشان می‌دهد و ضریب تشابه میان بردارهای اسناد و بردار جستجو را جهت بازیابی سند مربوطه، محاسبه می‌نماید.

در این سامانه‌ها یکی از این انتخاب‌ها، تعیین کوچک‌ترین واحد در متن مورد نظر است که معمولاً در پردازش زبان، کوچک‌ترین واحد را کلمه درنظر می‌گیرند. اما یکی از واحدهای مورد استفاده می‌تواند کاراکتر باشد. ساختار تعبیه کاراکتر به شکل n-تایی[1] کاراکتری می‌باشد که در تمام واژگان به‌صورت مشترک استفاده می‌شود. در بعضی موارد این مدل‌ها بهتر از مدل‌های تعبیه کلمه عمل می‌کند زیرا برای کلمات خارج از واژگان هم می‌توانند بردار لازم را تولید کند. برای مثال کلمه "OOV" که نمی‌توان با استفاده از مدل‌های تعبیه کلمه بردار متناظر آن را استخراج نمود، با روش تعبیه کاراکتر تبدیل به بردار می‌شود.

پژوهشگران تحقیقات زیادی در حوزه‌های مختلف شباهت معنایی از قبیل شباهت بین دو کلمه، متن کوتاه یا دو سند داشته‌اند [۱] که برای مدت طولانی مورد توجه محققین در زمینه‌های علمی بوده است. شباهت بین جملات نیز مسئله‌ای مهم است که به حوزه‌های متعدد پردازش زبان طبیعی مانند رتبه‌بندی پاسخ کوتاه [۲]، استخراج اطلاعات، پاسخگویی به سؤال [۳]، استفاده مجدد از متن یا تشخیص تشابه میان یک جمله و یک پاراگراف، روابط معنایی کلمات در یک توییت مشابه [۴] یا خلاصه‌سازی [۵] ارتباط دارد.

با توجه به آنکه در کاربردهای بسیاری از معیارهای شباهت استفاده می‌شود، اما بیشتر این روش‌ها جملات و متون را فقط بر اساس سطح ظاهری مقایسه می‌کنند و مقایسه بر اساس معنا و مفهوم کمتر صورت گرفته است.

رویکردهای قبل نیز در مشابهت جملات به شکلی بوده است

* رایانامه نویسنده مسئول: Sadidpour@mut.ac.ir

[1] N-gram

* این مقاله یک مقاله با دسترسی آزاد است که تحت شرایط و ضوابط مجوز Creative Commons Attribution (CC BY) توزیع شده است.

ناشر: دانشگاه جامع امام حسین (ع) © نویسندگان

که مطابقت از یک جهت واحد و به شکل کلمه به کلمه یا جمله به جمله اعمال می‌شده است. در پژوهشی، Wang و همکاران [۶] برای هر دو جمله ورودی یک شبکه عصبی رمزگذار مشابه را مورد استفاده قرار داده‌اند. در این روش ابتدا دو بردار لغات در یک فضا تشکیل، سپس تصمیم انطباق دو جمله بر اساس بردار لغات آن‌ها گرفته خواهد شد. از نقاط ضعف این روش این است که هیچگونه تعامل واضحی میان دو بردار لغات وجود ندارد و این مورد می‌تواند باعث از دست دادن برخی از اطلاعات شود.

در روش دیگر ارائه شده توسط Wang و همکاران [۶]، ابتدا واحدهای کوچک‌تر (مانند یک کلمه) با یکدیگر مقایسه شده، سپس نتایج این مقایسه با استفاده از یک شبکه CNN یا LSTM با یکدیگر تجمیع شده و تشکیل یک بردار را می‌دهند. درنهایت این بردار برای تصمیم‌گیری نهایی و انتخاب برچسب مناسب استفاده می‌شود. اما این روش محدودیت‌هایی نیز خواهد داشت. به‌عنوان مثال برخی از رویکردها تنها کلمه به کلمه منطبق شده و تشخیص، تنها در یک جهت صورت می‌گیرد. یعنی تطبیق بین جمله اول و دوم صورت می‌گیرد اما بین جمله دوم با اول انجام نمی‌شود.

پژوهشگران برای برطرف کردن این محدودیت‌ها راهکارهایی اندیشیده‌اند. مدل تطبیقی دوجهته چند جانبه[1] (BiMPM) روش پیشنهادی آن‌ها بوده که ابتدا دو جمله به‌عنوان ورودی به یک شبکه BiLSTM برای رمزگذاری داده شده، سپس این دو جمله، در دو جهت با یکدیگر تطبیق داده می‌شوند [۶].

Mueller و Thyagarajan [۷]، یک معماری بر اساس شبکه‌های عصبی بازگشتی Siamese ارائه دادند که هم‌زمان با کار بر روی دو بردار ورودی مختلف، از وزن‌های یکسانی استفاده می‌کند تا بردارهای خروجی قابل مقایسه را محاسبه کند. غالباً یکی از بردارهای خروجی مقدم است، بنابراین پایه‌ای را تشکیل می‌دهد که بردار دیگر خروجی، با آن مقایسه می‌شود. این روش به‌طور معمول برای مقایسه موارد مشابه در مجموعه‌های مختلف استفاده می‌شود.

شباهت همیشه واضح نیست و بنابر شرایط مسئله متغیر است. واژه‌ها می‌توانند به دو صورت لغوی[2] و معنایی[3] مشابه باشند. اگر دنباله‌ای از کاراکترهای مشابه موجود باشد، یعنی کلمات از لحاظ لغوی مشابه هستند و اگر آن‌ها به همان شیوه مورد استفاده قرار گیرند، در همان زمینه استفاده شوند و یک نوع دیگری از همان کلمه با مفهوم مشابه باشد کلمات از لحاظ معنایی مشابه هستند. شباهت لغوی با استفاده از الگوریتم‌های مختلف رشته محور[4] معرفی شده است و شباهت معنایی با استفاده از الگوریتم‌های دانش محور[5] و پیکره محور[6] معرفی می‌شود [۸].

Farouk [۹] رویکردهای استفاده شده برای اندازه‌گیری میزان شباهت جملات را نیز بررسی نمود. اندازه‌گیری شباهت بین جملات زبان طبیعی بر اساس روش اتخاذ شده او به سه دسته طبقه‌بندی می‌شود: دسته اول شباهت مبتنی بر کلمه[7] می‌باشد که برای محاسبه تشابه جملات فقط به شباهت مجموعه کلمات آن جمله بستگی دارد. دسته دوم در طبقه‌بندی، شباهت مبتنی بر ساختار[8] می‌باشد که از اطلاعات ساختار جملات در سنجش تشابه استفاده می‌شود. سه روش مختلف برای بهره‌برداری از اطلاعات نحوی یک جمله: استفاده از دستور زبان جملات، استفاده از POS و استفاده از ترتیب کلمات نیز می‌باشد. تشابه مبتنی بر بردار[9] دسته سوم در طبقه‌بندی پیشنهادی است. بسته به محاسبه شباهت بین بردارهای جملات، این نوع رویکرد شباهت آن‌ها را اندازه‌گیری می‌کند.

Wang و همکاران [۱۰] برای محاسبه تشابه معنایی بین جملات، مدل مبتنی بر توجه را نیز ارائه داده‌اند. در این روش با استفاده از شبکه‌های حافظه کوتاه مدت ماندگار دو سویه ویژگی-های معنایی یک جمله استخراج می‌شود و بازنمایی هر جمله با توجه به سطح کلمه یاد گرفته می‌شود. سپس با استفاده از یک سازوکار توجه[10]، اهمیت هر کلمه بررسی می‌شود و به هر کلمه با توجه به اهمیتشان وزن مربوطه اختصاص داده می‌شود. این میزان اهمیت در اصلاح وزن‌ها و بایاس مورد توجه قرار می‌گیرد. درنهایت با بردار رابطه وزن‌دهی شده به وسیله توجه در سطح جمله، رابطه بین دو جمله استخراج می‌شود.

انگیزه Wang و همکارانش از سازوکار توجه این است که کلمات مختلف باید در بازنمایی معنایی نهایی یک جمله نقش داشته باشند. این روش فقط به کلماتی توجه می‌کند که رابطه معنایی آن‌ها از طریق سازوکارهای توجه به سطح کلمه تأثیر زیادی در معنی جمله دارد. شبکه BiLSTM نیز می‌تواند بازنمایی این ویژگی را در معماری توجه در سطح کلمه استخراج کند. در لایه توجه، خروجی بازنمایی‌ها الحاق شده و از تابع تشابه کسینوسی برای محاسبه مقدار تشابه دو جمله استفاده می‌شود [۱۰].

ایده اصلی این پژوهش، یافتن میزان شباهت جملات با

[1] Bilateral Multi-perspective Matching
[2] Lexically
[3] Semantically

[4] String-Based Algorithms
[5] Knowledge-Based Algorithms
[6] Corpus-Based
[7] Word Based
[8] Structure Based
[9] Vector Based
[10] Attention Mechanism

استفاده از یادگیری عمیق و تطبیق تعبیه کلمات از یک زبان آماده به زبانی جدید می‌باشد. مبنای پیدایش این ایده، وجود داده آموزشی کافی در زبان انگلیسی و عدم وجود داده کافی در زبان فارسی می‌باشد. بنابراین، ابتدا فضای برداری تعبیه کلمات دو زبان انگلیسی و فارسی تطبیق داده می‌شود. سپس، شباهت جملات از لحاظ مفهومی و نه ظاهری با استفاده از شبکه‌های عمیق مورد مقایسه قرار گرفته و میزان شباهت آن‌ها تشخیص داده می‌شود. بنابراین، از نوآوری‌های این مقاله می‌توان به استفاده از ایده نگاشت به‌منظور جایگزینی عدم وجود داده کافی شباهت معنایی در زبان فارسی اشاره کرد.

در ادامه، نگاشت بین زبانی شرح داده می‌شود. در بخش بعد، ایده نگاشت و استفاده از آن برای یافتن شباهت معنایی جملات مورد بحث قرار داده شده است. سپس، نتایج این پژوهش ارائه و تفسیر شده است. درنهایت، نتیجه‌گیری و پیشنهادات بیان خواهد شد.

نگاشت بین زبانی

ایده نگاشت بر این اساس می‌باشد که Mikolov و همکاران [۱۱] با بررسی‌های انجام شده‌شان به نتایجی دست یافتند که فضاهای تعبیه کلمه[1] در زبان‌های مختلف، پیوسته می‌توانند ساختار مشابهی داشته باشند و بتوان با استفاده از این ساختار مشابه، نگاشت بین زبانی انجام داد تا در کاربردهای مختلف NLP از جمله مشابهت‌یابی، از آن استفاده نمود. تعبیه کلمه یکی از منابع پر استفاده در پردازش زبان طبیعی است که ثابت کرده‌اند اهمیت زیادی برای مدل‌سازی پدیده‌های زبانی در هر دو حالت با نظارت و بدون نظارت دارد.

یکی از دلایل اصلی اهمیت آن‌ها، نقش مهمی است که در انتقال دانش از یک زبان به دیگری، به ویژه در بازیابی اطلاعات و طبقه‌بندی متن دارند. در عین حال در مسائل NLP چند-زبانه مانند ترجمه ماشین بهبود قابل توجهی بخشیده است. به تازگی نشان داده شده است که می‌توان این تعبیه را توسط دو فضای برداری متفاوت با استفاده از تبدیلات خطی و یک فرهنگ لغت دو-زبانه کوچک با نظارت هم‌تراز نمود [۱۲].

روش‌های مختلفی برای به‌دست آوردن این پیوند-زبانی وجود دارد. پژوهشگران در سال ۲۰۱۳، ابتدا دریافتند که فضاهای تعبیه کلمه، پیوسته ساختارهای مشابهی را در سراسر زبان‌ها، حتی با توجه به زوج زبان‌های مختلف مانند انگلیسی و ویتنامی دارند. پیشنهاد آن‌ها این بود که با استفاده از یک نگاشت خطی از یک منبع به فضای تعبیه هدف، از این شباهت بهره‌برداری کنند. آن‌ها یک واژگان موازی از پنج هزار کلمه به‌عنوان نقاط پایه برای یادگیری این نگاشت مورد استفاده قرار داده و روش خود را در یک کار ترجمه ارزیابی نمودند [۱۳].

از آن زمان، چندین مطالعه در بهبود این پیوند-زبانی تعبیه کلمه صورت گرفته است. یکی از موفق‌ترین روش‌ها، یادگیری دو-زبانه از طریق یک فرآیند سه مرحله‌ای است: به این صورت که ابتدا به‌طور مستقل برای هر زبان، بردار تعبیه واژگان آموزش داده می‌شود. این بردار را می‌توان با مدل‌های از پیش-آموزش دیده کلمات، مانند Word2vec و GloVe یا FastText به‌دست آورد.

در مرحله بعد، یک راهبرد ترازبندی خطی نگاشت تک-زبانه برای ادغام آن‌ها در یک فضای برداری دو-زبانه مشترک اعمال می‌شود و در آخر، یک تغییر نهایی در تعویض ترازبندی اعمال می‌شود. بنابراین بردارهای کلمات از هر دو زبان اصلاح و با یکدیگر هم‌تراز می‌گردد.

فرضیه اصلی این است که چنین رویکردهایی می‌توانند بهبود بیشتری یابند، زیرا فرض می‌کنند که ساختار داخلی دو فضای تک-زبانه اولیه بدون تغییر می‌ماند. با این حال، این ساختار تحت تأثیر پدیده‌های خاص زبان قرار می‌گیرد. پیشنهاد کارشناسان در این باره این است که بعد از مرحله هم‌ترازی اولیه، تغییراتی اعمال شود تا مترادف‌های پیوند-زبانی به سمت نقطه میانی بین آن‌ها حرکت کند.

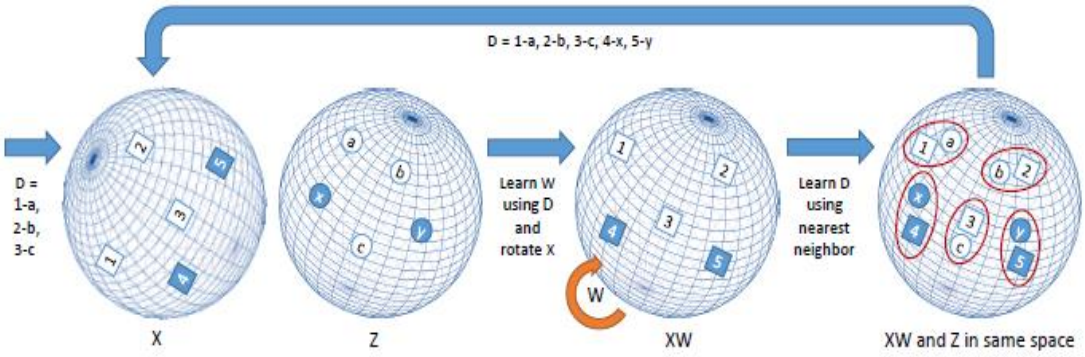
هدف این است با استفاده از این تغییر، یک پیوند-زبانی مناسب از این فضای برداری به‌دست آید و علاوه بر این، فضاهای تک-زبانه نیز با این تغییرات بهبود یابد. همچنین این مقایسه به‌طور معمول یاد می‌گیرد که ساختار فضاهای تک-زبانه نیز حفظ شود. آزمایش‌های انجام شده تأیید می‌کنند که نتیجه تعبیه‌های پیوند-زبانی از مدل‌های پیشرفته در هر دو کار یک و چند-زبانه دارای ارزیابی بهتری است [۱۲].

یکی از مدل‌های پیشرفته روش تبدیل خطی و نگاشت بین-زبانی، VecMap می‌باشد. در این روش ابتدا نگاشت خطی با استفاده از یک فرهنگ لغت دو-زبانه کوچک تخمین زده شده، سپس این فرهنگ لغت واژه‌های جدید را از زبان مبدأ با استفاده از تبدیل یاد گرفته شده می‌افزاید. درنهایت، روند تا زمانی که معیار همگرایی ایجاد شود، تکرار می‌شود [۱۲].

در شکل (۱)، X و Z ماتریس تعبیه کلمه را در دو زبان مختلف بیان می‌کنند، به‌طوری که *Xi مربوط به تعبیه کلمه زبان مبدأ و *Zj مربوط به تعبیه کلمه زبان هدف است. فرهنگ لغت مربوطه به‌عنوان ماتریس D دودویی درنظر گرفته شده است

[1] WordEmbedding

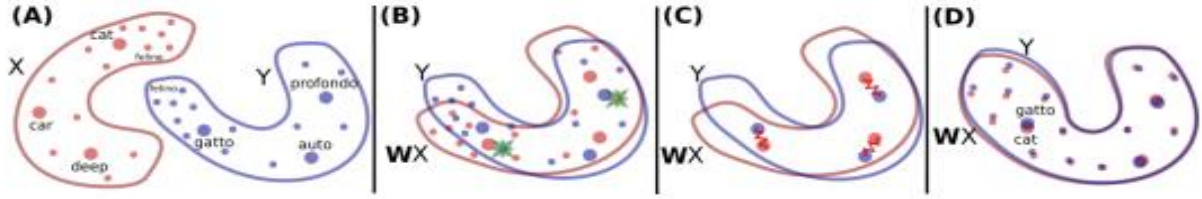
به‌طوری که اگر Dij = 1 برقرار باشد یعنی کلمه زبان هدف با کلمه زبان مقصد هم‌تراز می‌باشد. هدف به‌دست آوردن ماتریس نگاشت بهینه W* می‌باشد، به‌طوری که مجموع مربع فاصله اقلیدسی بین نگاشت تعبیه مبدأ Xi*W و جایگذاری هدف Zj* برای ورودی فرهنگ لغت Dij به حداقل برسد [۱۴].



شکل (۱): نگاشت دو فضای برداری به یک فضا با استفاده از فرهنگ لغت دو-زبانه و روش VecMap [۱۴]

MUSE روش دیگری است که با استفاده از ماتریس تبدیل می تواند از طریق یک فرهنگ لغت دو-زبانه با استفاده از هم‌ترازی اولیه عرضه شده یا از طریق مدل بدون نظارت به‌دست آید. این مدل بدون نظارت با استفاده از یادگیری مشارکتی برای به‌دست آوردن هم‌ترازی اولیه از دو فضای تک-زبانه آموزش داده می‌شود. این معیار اندازه‌گیری که نزدیک‌ترین همسایگان هر دو واژه منبع و هدف را درنظر می‌گیرد، نشان می‌دهد که بهتر از مناطق فضایی با تراکم بالا برخورد می‌کند، به‌طوری که باعث کاهش مشکل هاب‌نس[1] در مدل‌های تعبیه‌سازی کلمات می‌شود. این مشکل زمانی رخ می‌دهد که چندین نقطه (به نام هاب) تبدیل به نزدیک‌ترین همسایگان بسیاری از نقاط دیگر در فضای تعبیه می‌شوند [۱۲].

این روش که از آموزش مشارکتی[2] برای یادگیری نگاشت خطی از یک منبع به یک فضای هدف بهره می‌برد، در دو مرحله اجرا می‌گردد. در ابتدا، یک متمایز کننده[3] برای تفکیک بین نگاشت تعبیه منبع و تعبیه هدف آموزش داده می‌شود، در حالی که نگاشت به‌طور مشترک آموزش داده می‌شود تا فریب دهنده[4] را تمیز دهد. سپس، یک فرهنگ لغت ترکیبی از فضای تعبیه مشترک به‌دست می آید و نگاشت را با الگوریتم SVD، تنظیم دقیق می‌کند. در شکل (۲-A) دو توزیع از تعبیه کلمات می‌باشد که کلمات انگلیسی به رنگ قرمز توسط X و کلمات ایتالیایی به رنگ آبی توسط Y نشان داده شده که باید هم‌تراز و ترجمه شود. هر نقطه یک کلمه در آن فضا را نشان می‌دهد. اندازه نقطه متناسب با فرکانس کلمات در بخش آموزش آن زبان می‌باشد. در شکل (۲-B) با استفاده از یادگیری مشارکتی، یک ماتریس چرخش فرا گرفته می‌شود که تقریباً دو توزیع را هم‌تراز می‌کند. ستاره‌های سبز به‌صورت تصادفی کلمات انتخاب شده هستند که به متمایز کننده داده می‌شوند تا تعیین کنند که آیا دو کلمه جایگذاری از یک توزیع مشابه می‌آید یا نه. نگاشت W توسط الگوریتم SVD ترمیم می‌شود (شکل ۲-C). این روش با استفاده از هم‌ترازی کلمات تکرار شده با مرحله قبل به‌عنوان نقاط لنگر[5] استفاده می‌شود و یک تابع انرژی را که بین نقاط لنگر مربوط می‌شود به حداقل می‌رساند. این نگاشت سپس برای نقشه کلی کلمات در فرهنگ لغت استفاده می‌شود. درنهایت، با استفاده از نگاشت W و معیار اندازه‌گیری فاصله CSLS ترجمه می‌شود (شکل ۲-D) [۱۳].



شکل (۲): نگاشت خطی دو-زبانه با روش MUSE [۱۳]

[1] Hubness
[2] Adversarial
[3] Discriminator
[4] Fool
[5] Anchor

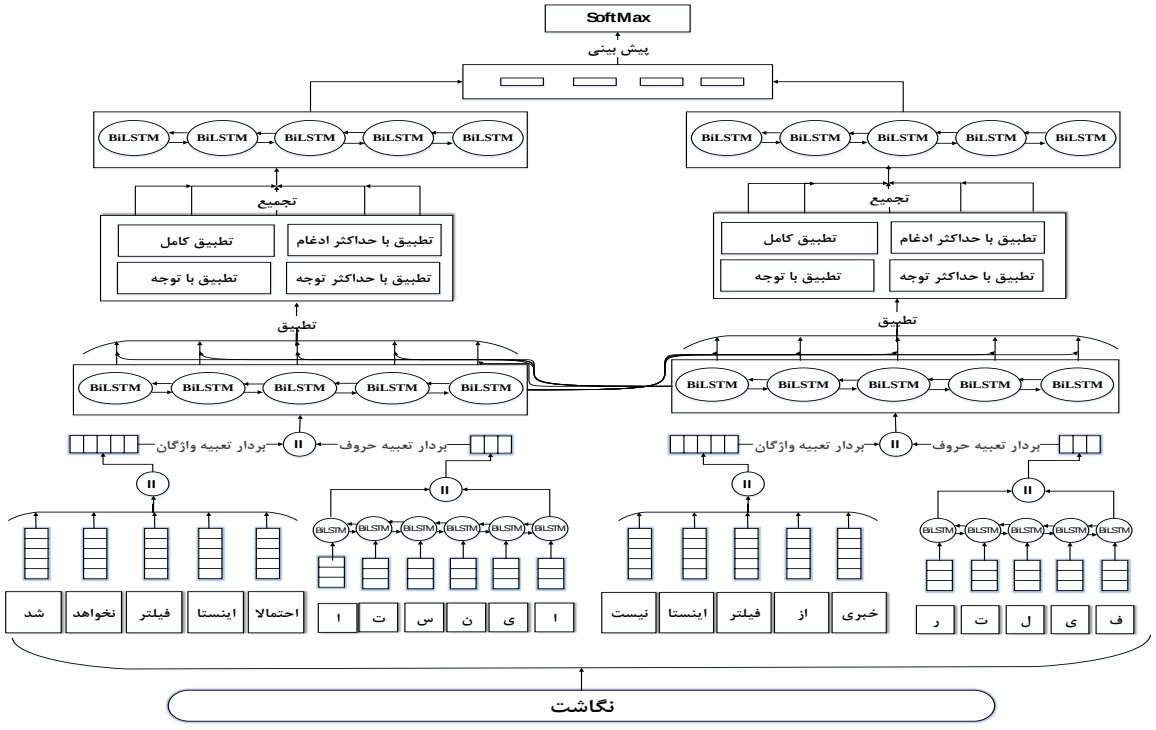
در سال ۲۰۱۸ شبکه BERT برای مسائل متنی، یا به‌طور دقیق‌تر NLP، توسط Devlin و همکاران ارائه شد. شبکه‌ای بسیار بزرگ که با دادگان زیادی (Wikipedia + BookCorpus) توسط مهندسان گوگل آموزش داده شد و در دسترس دیگر پژوهشگران قرار گرفت. بنابر بررسی‌های انجام شده، گرچه BERT در درک عبارات مثبت درست عمل می‌کند، در عبارات منفی نمی‌تواند کارآمد عمل کند. برای مثال وقتی کسی بنویسد، "کفتر چاهی یک ... است." خود الگوریتم درک کرده و جای خالی را با پرنده پر می‌کند اما اگر شخص دیگری بنویسد "کفتر چاهی یک ... نیست."نمی‌داند که باید با چه چیزی جای خالی را پر کند. بنابراین در مواردی که BERT نمونه‌ها یا زمینه‌های نفی را پیش از این ندیده است، درک این موضوع برایش سخت خواهد بود [۱۵].

مدل‌های توجه، و به‌طور خاص BERT، در پردازش زبان طبیعی، نتایج امیدوار کننده‌ای کسب کرده‌اند. Facebook AI، XLM را به‌عنوان نسخه بهبود یافته BERT برای دستیابی به نتایج پیشرفته در رده‌بندی و ترجمه ارائه داده است [۱۶]. XLM برای یادگیری روابط بین کلمات در زبان‌های مختلف، از روش شناخته شده پیش‌پردازش (BPE) و یک سازوکار آموزش دو زبانه با BERT استفاده می‌کند. این مدل در یک کار طبقه‌بندی بین زبانی از سایر مدل‌ها پیشی می‌گیرد (در ۱۵ زبان) و هنگامی که از یک مدل از قبل آموزش دیده برای مقداردهی اولیه مدل ترجمه استفاده می‌شود، ترجمه ماشین را به‌طور قابل توجهی بهبود می‌بخشد [۱۷].

همچنین، Multilingual-BERT (mBERT) نسخه‌ای از BERT است که در سال ۲۰۱۹ به ۱۰۴ زبان مختلف با استفاده از منبع غنی ویکی‌پدیا آموزش و منتشر شده است. کارهای اخیر نشان می‌دهد که برای کارهای مختلف علی‌رغم آموزش بدون نظارت موازی، بسیار عالی عمل می‌کند. mBERT نمایش‌های بین زبانی غنی را می‌آموزد که امکان انتقال بین زبان‌ها را فراهم می‌کند. نتایج نشان می‌دهد که بیشتر اطلاعات به‌صورت غیر خطی رمزگذاری می‌شوند، در حالی که برخی از آن‌ها را می‌توان با ابزارهای کاملاً خطی نیز بازیابی کرد. اما در این مدل فقط می‌توان بر نمایش آخرین لایه و ماتریس تعبیه شده که بین لایه‌های ورودی و خروجی تقسیم می‌شود تمرکز نمود. در این مدل امکان گسست آسان بین هویت زبان و معنای واژگانی در نمایش‌های mBERT افزایش یافته است [۱۸].

مدل انتخابی

مدل تطبیقی دو جهته چند جانبه، یکی از روش‌های بررسی شده برای محاسبه شباهت معنایی می‌باشد که در این روش دو جمله به اصطلاح P و Q در دو جهت ($P \rightarrow Q$ و $P \leftarrow Q$) مطابقت داده می‌شود. در هر جهت مدل، دو جمله از دیدگاه‌های مختلف تطبیق می‌یابند. در صورتی که ورودی به‌صورت یک جفت جمله P و Q باشد، این مدل توزیع احتمال $Pr(y|P,Q)$ را از طریق پنج لایه بازنمایی کلمه، بازنمایی متن، تطبیق، تجمیع و پیش‌بینی کننده محاسبه می‌کند. این مدل در شکل () نمایش داده شده است که اجزای این مدل پیشنهادی در قسمت‌های بعدی پرداخته می‌شود.



شکل (۳): روش پیشنهادی و مدل

۱-۱- بردار تعبیه

روش پیشنهادی از دو سطح بردار تعبیه بهره می‌برد. در یک بخش، برای هر واژه‌ای که وارد شبکه می‌شود، یک بردار بازنمایش با استفاده از بردار تعبیه واژگان از پیش-آموزش دیده نگاشت شده Word2vec ساخته می‌شود. بنابراین هر واژه با بردار بازنمایشی که حاوی یکسری اطلاعات معنایی و ارتباطی از واژه است، نمایش داده می‌شود.

در بخش دیگر، بردار بازنمایش واژگان با استفاده از استخراج ویژگی از سطح حروف هر واژه به‌دست می‌آید. برای استخراج اطلاعات از واژگانی که ناشناخته هستند و پیش از این بازنمایشی برای این واژگان، توسط Word2Vec، درنظر گرفته نشده است، از بردار تعبیه حروف استفاده می‌شود.

از آنجایی که شبکه عصبی عمیق، عمل استخراج ویژگی را به‌صورت خودکار انجام می‌دهد و شبکه‌های LSTM اطلاعات دنباله‌ای از داده‌ها را حفظ می‌کند و همچنین به دلیل آنکه بتوان اطلاعات را از دو جهت دنباله استخراج و حفظ نمود، از شبکه BiLSTM استفاده گردیده است. به این صورت که کاراکترهای هر واژه به‌عنوان ورودی به شبکه وارد شده و بردار خروجی آخرین واحد LSTM از شبکه LSTM پیش‌خور، به بردار خروجی آخرین واحد LSTM از شبکه LSTM پس‌خور متصل می‌شود و بردار بازنمایش واژه در سطح حروف ساخته می‌شود.

در بخشی دیگر بازنمایشی از واژگان با توجه به زمینه متنی که واژگان در آن قرار دارند، به‌دست می‌آید. از آنجایی که شبکه‌های LSTM نسبت به RNN وابستگی‌های طولانی‌تری را در یک دنباله از کلمات می‌توانند استخراج کنند، برای استخراج ویژگی با توجه به حوزه متنی که واژگان در آن قرار گرفته‌اند، از شبکه‌های LSTM استفاده می‌شود. برای اینکه بتوان بازنمایشی برای هر واژه با توجه به کلمات قبل و بعد از آن به‌دست آورد از LSTM دو جهته BiLSTM استفاده شده است.

۱-۲- تطبیق

لایه اصلی در این مدل، لایه تطبیق به‌صورت چند جانبه می‌باشد. هدف این لایه مقایسه هر یک از بردارهای ساخته شده از جملات همراه با مفاهیم متنی، با دیگر جملات است. در این بخش دو بازنمایش از بردار تعبیه متن جمله اول و دوم در هر گام زمانی[1] از دو جهت مطابقت داده خواهد شد. در هر گام زمانی جمله اول در برابر تمام گام‌های زمانی جمله دوم و برعکس تطبیق داده می‌شوند. عملیات تطبیق چند جانبه بر اساس مراحل زیر تعریف می‌شود:

- تطبیق کامل

تطبیق در این حالت به این شکل است که هر تعبیه متن پیش‌خور یا پس‌خور با آخرین گام زمانی جمله دیگر مطابقت داده خواهد شد.

- تطبیق با حداکثر ادغام

تطبیق در این حالت به این شکل است که هر تعبیه متن رو به جلو یا رو به عقب یک جمله را با همه تعبیه متن رو به جلو یا رو به عقب جمله دیگر برای ژهای عضو مجموعه ۱ تا N مطابقت داده می‌شود و بیشترین مقدار استخراج خواهد شد.

- تطبیق با توجه

در این بخش، ابتدا میزان شباهت کسینوسی بین هر تعبیه متن رو به جلو یا رو به عقب با همه تعبیه متن‌های رو به جلو یا رو به عقب جمله دیگر محاسبه می‌شود. سپس $\vec{\alpha}_{ij}$ یا $\overleftarrow{\alpha}_{ij}$ به‌عنوان وزن $\vec{h}_j^q$ یا $\overleftarrow{h}_j^q$ انتخاب می‌شود و یک بردار دقت برای کل جمله Q با جمع وزن تمام تعبیه‌های متنی Q به‌دست می‌آید. درنهایت، هر تعبیه متن رو به جلو یا رو به عقب $\vec{h}_i^p$ یا $\overleftarrow{h}_i^p$ با یک بردار دقت متناظر مطابقت داده می‌شود.

- تطبیق با حداکثر توجه

این استراتژی شبیه به استراتژی تطبیق با دقت است. با این حال، به جای اینکه مجموع وزن‌های همه تعبیه‌های متنی به‌عنوان بردار دقت انتخاب شود، تعبیه متنی با بیشترین شباهت کسینوسی به‌عنوان بردار دقت انتخاب شده است. سپس، هر تعبیه متنی از جمله P را با بردار دقت جدید مطابقت داده می‌شود.

تمام این چهار راهبرد تطبیق برای هر گام زمانی از جمله P اعمال می‌شود و هشت بردار تولید شده به‌عنوان بردار تطبیق نهایی برای هر گام زمانی از جمله P پیوند داده می‌شود. همچنین همین فرآیند در جهت معکوس نیز انجام می‌شود. با توجه به آزمایش‌های انجام شده، تمام این چهار راهبرد باید به‌صورت هم‌زمان برای به‌دست آوردن دقت بالاتر استفاده شود. اگر هر کدام از آن‌ها حذف شود کارایی سامانه کاهش پیدا می‌کند.

۱-۳- تجمیع

این لایه برای جمع کردن دو توالی بردارهای تطبیق به یک بردار

[1] Time Step

تطبیق با طول ثابت استفاده می‌شود. یک مدل BiLSTM دیگر نیز بر روی دو بردار تطبیق به‌صورت جداگانه اعمال می‌شود تا نتایج ۴ حالت تطبیق مرحله قبل را جمع کند. در انتهای این مرحله یک بردار با طول ثابت، (آخرین مرحله زمانی از شبکه BiLSTM) ساخته می‌شود.

۱-۴- پیش‌بینی

در انتها لایه کاملاً متصل با تابع فعالیت Softmax قرار داده شده است تا برچسب بین دو عبارات پیش‌بینی شود. هدف از این لایه، ارزیابی توزیع احتمال $Pr(y \mid P,Q)$ است. در این مسئله y به دو صورت (y=۱) دو جمله مشابه و (y=۰) دو جمله غیر مشابه می باشد.

۲- روش پیشنهادی

روش پیشنهادی این مقاله از دو بخش نگاشت بین زبانی و مـدل تشکیل شده که در ادامه تشریح می‌گردد.

۲-۱- چارچوب روش پیشنهادی نگاشت

در این پژوهش از ایده نگاشت بین زبانی در زبان‌های انگلیسی و فارسی استفاده شده است تا از طریق نگاشت خطی از زبان مبدأ به زبان هدف بتوان شباهت برداری کلمات مشترک در دو زبان را استخراج نمود. این کلمات مشترک بین دو زبان، با استفاده از الگوریتم SVD هم‌تراز می‌شود، تا فضای برداری آن کلمات به هم نزدیک‌تر شود.

در اصل با این روش می‌توان تا حدود زیادی فضای برداری آن‌ها را یکی نمود. در انتها با محاسبه فاصله کسینوسی بین بردار تعبیه واژگان مبدأ با بردار تعبیه واژگان نگاشت شده، و استفاده در مدل‌های مورد نظر، می‌توان میزان کاهش فاصله را دید و این میزان تغییر را در خروجی مدل‌ها متوجه شد.

نگاشت مورد بررسی بر اساس یک فرآیند سه مرحله‌ای انجام شده است: در ابتدا دو ماتریس از دو بردار تعبیه واژگان مربوطه ساخته می‌شود. در این پژوهش، از سه بردار تعبیه واژگان مختلف به دو زبان فارسی و انگلیسی به شکل بردار تعبیه واژگان فارسی با ۳۵،۶۱۰ کلمه و انگلیسی با ۲،۰۰۰،۰۰۰ کلمه در ۳۰۰ بعد برای نگاشت اولیه استفاده شده است. در ابتدا ماتریسی از هر یک از این دو بردار واژگان ساخته می‌شود تا عملیات بعدی روی آن‌ها انجام شود.

سپس هم‌پوشانی مربوط به دو ماتریس استخراج می‌گردد. این به معنی است که تمام کلماتی که در هر دو ماتریس ساخته شده وجود دارد، استخراج گردد و در لیست جداگانه‌ای قرار داده شود. در این آزمایش به تعداد ۱۴،۶۸۵ کلمه مشابه در دو ماتریس بردار واژگان به‌صورت مشابه ساخته شده است.

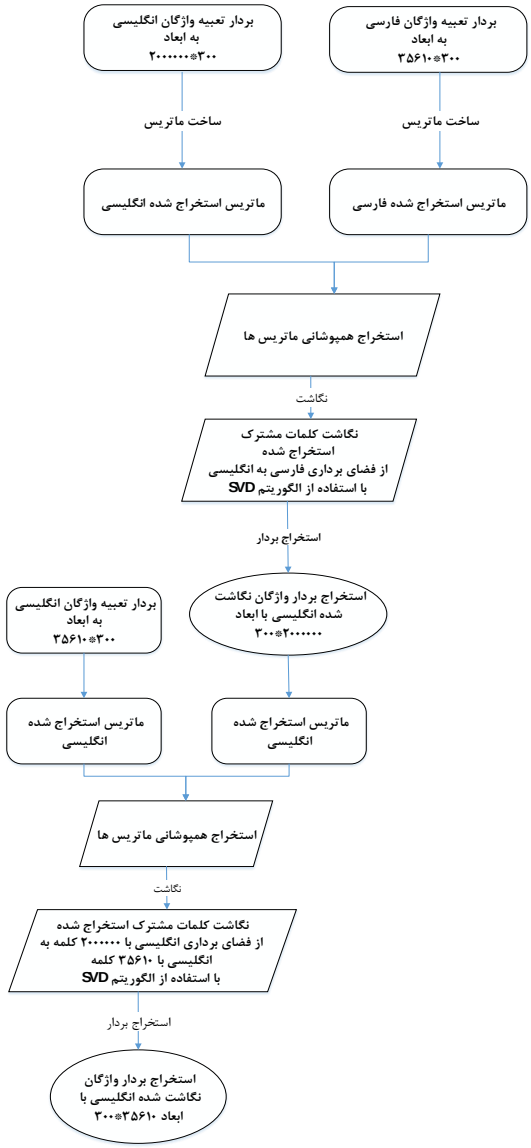
بدین طریق به احتمال بالا می‌توان تمام کلمات مشترکی که در فارسی با انگلیسی مشابه هستند را بهتر استخراج نمود. در حقیقت عمل نگاشت دو بار صورت می‌گیرد. یک بار نگاشت بین دو بردار تعبیه به زبان‌های فارسی و انگلیسی به ترتیب ۳۵،۶۱۰ کلمه‌ای و ۲،۰۰۰،۰۰۰ کلمه‌ای انجام شده، سپس نگاشت دوم بین دو بردار تعبیه انگلیسی ۲،۰۰۰،۰۰۰ کلمه‌ای و ۳۵،۶۱۰ کلمه‌ای انجام شده است که ابعاد تمام بردارهای استفاده شده ۳۰۰ می‌باشد.

اکثر روش‌ها برای نگاشت بین دو زبان نیاز به ساختارهای بزرگ موازی مانند فرهنگ لغت دو-زبانه داشته که این ساختار بزرگ برای بیشتر زبان‌ها موجود نمی‌باشد. برای مثال در این پژوهش فرهنگ لغت دو-زبانه فارسی-انگلیسی مناسبی برای استفاده در دسترس نبوده است. برای مثال برای کلمه "در" فارسی چند ترجمه انگلیسی وجود دارد که از جمله آن‌ها می‌توان به in، door، at، unto و wicket اشاره کرد.

با توجه به این مسئله، این پژوهش بدون فرهنگ لغت دو-زبانه فقط فضای برداری دو بردار تعبیه واژگان مورد استفاده را یکی نموده و نتایج خوبی به دنبال داشته است. ایده این روش از آنجایی به نظر رسید که مدل‌های مورد بررسی یادگیری عمیق، نیازمند دادگان بسیار زیاد بوده تا مدل به خوبی آموزش داده شود. متأسفانه دادگان فارسی جمع‌آوری شده با حجم بالا و به اندازه کافی برای این مدل‌ها موجود نمی‌باشد. در شکل (۴)، چارچوب این روش نمایش داده شده است.

۲-۲- مدل

این پژوهش برای رسیدن به بهترین خروجی مدل‌های مختلفی را مورد بررسی قرار داده است. یکی از مدل‌های متناسب با هدف پژوهش مدل BiMPM بوده است که به‌عنوان مدلی انتخابی قرار گرفته است. اجزای این مدل در بخش ۳، به‌طور کامل تفسیر شده است. روش پیشنهادی و مدل مورد بررسی در **شکل** () نمایش داده شده است.



شکل (۴): چارچوب روش نگاشت

۲-۳- دادگان

در زبان انگلیسی مجموعه‌های آماده متعددی با طول‌های متفاوت و نیز با موضوعات مختلف از قبل تهیه و در دسترس پژوهشگران قرار می‌گیرد؛ اما در زبان فارسی، به دلیل عدم وجود مجموعه آماده، هر پژوهشگر در کار خود نیازمند ساخت مجموعه متفاوتی است که توسط خود فرد تهیه می‌گردد.

در زبان انگلیسی پیکره‌های بسیاری برای این منظور توسعه یافته است که از آن میان می‌توان به پیکره جملات SNLI[1] دانشگاه استنفورد [۱۹] اشاره نمود.

این پیکره یکی از معروف‌ترین و پرکاربردترین پیکره‌های موجود است که شامل ۵۷۰ هزار جفت جمله می‌باشد که به‌صورت دستی به سه حالت مشابه، خنثی و غیر مشابه برچسب‌گذاری[2] شده است. این دادگان یکی از بزرگ‌ترین منابع با کیفیت با نشانه برای تفسیر معنایی جملات است.

ماهیت معناشناختی استنتاج زبان طبیعی، آن را یک کاندید خوب برای یادگیری تعبیه جملات عمومی در روش‌های با نظارت می‌داند. به دلایل اشاره شده در بالا، در پردازش زبان طبیعی فارسی نیاز به وجود دادگان با برچسب و با کیفیت بالا به شدت احساس
می‌شود. از آنجا که دسترسی به دادگان فارسی با برچسب محدود و دارای مشکلاتی است که قابل استفاده نبوده‌اند، تصمیم بر آن شد که برای این منظور از وب گاه‌های خبری نسبتاً با کیفیت، دادگان مورد نیاز جمع‌آوری شود.

در این پژوهش، ابتدا بیش از ۲۰ وب گاه و خبرگزاری مورد ارزیابی قرار گرفتند، که ۶ مورد با بالاترین کیفیت برای استخراج دادگان مناسب انتخاب شدند.

نمونه‌های استخراج شده بعد از نشانه‌گذاری، جهت ساخت پیکره نهایی استفاده می‌شوند. به‌منظور اعمال ویژگی قابل حمل بودن در پیکره، از قالب XML استفاده شده است تا سازگاری بهتری با ابزارهای مختلف برنامه‌نویسی داشته باشد. دادگان ساخته شده دارای یک شناسه برای هر نمونه می‌باشد که توسط عنصر PairId مشخص شده است. عناصر Sentence1 و Sentence2 حاوی مقدار اصلی هر یک از عبارت‌های نمونه هستند و Labels که حاوی برچسب تشخیص داده شده توسط فرد متخصص در دو حالت مشابه و غیر مشابه می‌باشد.

سپس به‌منظور بهبود کیفیت دادگان جمع‌آوری شده، مورد پردازش و مرتب‌سازی قرار گرفته شد. همچنین برای اطمینان از کیفیت دادگان، توسط چند نیروی انسانی مورد بررسی قرار گرفت. در انتها ۵۰۰۰ جفت جمله استاندارد تولید و مورد استفاده قرار گرفت.

مدل مورد بررسی، بر روی مجموعه دادگان ساخته شده فارسی که شامل ۵۰۰۰ هزار جفت جمله با برچسب مشابه و غیر مشابه می‌باشد، مورد آزمایش قرار گرفته است. به‌طور تصادفی ۶۰۰ عبارت مرتبط و ۴۰۰ عبارت نامرتبط به‌عنوان مجموعه ارزیابی و ۵۰۰ عبارت مرتبط و ۵۰۰ عبارت نامرتبط به‌عنوان مجموعه آزمایش انتخاب شده است. مقدار باقیمانده به‌عنوان مجموعه آموزشی استفاده شده است.

[1] Stanford Natural Language Inference

[2] Label

در جدول (۱) نمونه‌های ساخته شده از چند موضوع نمـایش داده شده است.

جدول (۱): نمونه‌ای از دادگان جمع‌آوری شده

موضوع	نمونه‌ای از جملات انتخابی هر موضوع
بازگشت آمریکا به برجام	● وقت آن است آمریکا به برجام بازگردد. ● آمریکا دیر یا زود روزی به برجام بازخواهد گشت.
پایبندی ایران به برجام	● ایران به تمام تعهدات خود در برجام پایبند بوده است. ● آژانس انرژی اتمی نیز بارها تأکید کرده که ایران به تعهداتش پایبند بوده است.
مخالفت کشورها علیه خروج آمریکا از برجام	● هر مذاکره‌ای باید در تداوم برجام باشد. ● نخست وزیر ایتالیا نیز حمایت خود را از برجام اعلام کرد.

۳- معیارهای ارزیابی

به‌طور کلی، روش‌ها و ایده‌های پیشنهادی در این پژوهش مربوط به شباهت معنایی جملات می‌باشد. با توجه به اینکه تاکنون کمتر کاری در حوزه شباهت معنایی جملات فارسی انجام شده و کمتر به این موضوع در این حوزه پرداخته شده است، مجموعه محکی مناسبی برای ارزیابی وجود ندارد؛ از این رو ۱۰۰۰ جفت جمله از اخبار و نیز جملات کوتاه فارسی با برچسب متناسب برگزیده و روش پیشنهادی بر روی این جملات، پیاده‌سازی شده است. برچسب‌ها به‌صورت دو کلاس «شبیه=۱» و «غیر شبیه=۰» درنظر گرفته شده‌اند. برای برچسب‌گذاری داده‌ها، جملات در راستای یک موضوع شبیه و جملات در دسته‌های موضوعی متفاوت، غیر شبیه برچسب‌گذاری شده‌اند که نمونه‌ای از آن در جدول (۲) نشان داده شده است.

در انتهای نتایج این سامانه مشابهت‌یاب، با حاشیه‌نویسی‌های انسانی مورد ارزیابی قرار گرفته شد. همچنین یکی از روش‌های پایه در ارزیابی شباهت‌یابی جملات، معیارهای دقت و بازخوانی می‌باشد. از این رو عملکرد پیشنهادی با استفاده از این معیارها نیز مورد مقایسه قرار گرفت.

یکی از مهم‌ترین معیارهای تعین کارایی یک الگوریتم دسته-بندی دقت یا نرخ دسته‌بندی[1] می‌باشد که این معیار، دقت کل یک دسته‌بند را محاسبه می‌کند. در واقع این معیار مشهورترین و عمومی‌ترین معیار محاسبه کارایی الگوریتم‌های دسته‌بندی است که نشان می‌دهد، دسته‌بند طراحی شده چند درصد از کل مجموعه رکوردهای آزمایشی را به درستی دسته‌بندی کرده است که این معیار نیز مورد ارزیابی قرار گرفت و نتایج در جدول (۳) نمایش داده شده است.

جدول (۲): نمونه‌ای از دادگان برچسب‌گذاری شده

جملات	برچسب
تصویری که در خواب می‌بینیم رؤیا است.	۱
تصویر سازی در خواب را رؤیا گویند.	
به نظر می‌رسد اختصاص ۳۰ لیتر بنزین ماهانه برای هر نفر کافی است.	۱
سهمیه طرح‌های پیشنهادی بین ۲۰ تا ۳۰ لیتر در یک ماه است.	
به نظر می‌رسد اختصاص ۳۰ لیتر بنزین ماهانه برای هر نفر کافی است.	۰
این ۰/۸ لیتر در روز که ۲۴ لیتر در ماه است، سهمیه بنزین به قیمت ۱۰۰۰ تومان است.	

۳-۱- روش بهینه‌سازی

نرخ یادگیری[2] بیانگر سرعت (گام) به‌روزرسانی وزن‌ها است که می‌تواند مقداری ثابت داشته باشد یا به‌صورت سازگار شونده تغییر کند. به‌منظور بهینه‌سازی، روش‌های مختلفی از جمله کاهش در امتداد گرادیان[3] و Adam مورد استفاده قرار گرفتند. با توجه به اینکه نتایج حاصل از روش Adam بهتر بود، در تمامی معماری‌های استفاده شده از این روش به‌عنوان روش بهینه‌سازی در جهت یادگیری شبکه استفاده گردید.

همچنین، نرخ یادگیری ۰/۰۱ درنظر گرفته شده است. این پارامترها قابل تغییر هستند و اعداد انتخابی با انجام چندین آزمایش و به‌صورت تجربی انتخاب شده‌اند.

۳-۲- ارزیابی مدل

در این بخش نتایج همه روش‌ها مورد بررسی، در جدول (۳) نمایش داده شده است و مقایسه‌ای بین روش با نگاشت و بدون نگاشت صورت می‌گیرد. علاوه بر این سه معیار دقت، بازخوانی و F1 مورد بررسی قرار گرفته که نتایج این بررسی در شکل (۳) نمایش داده شده است.

[1] Classification Accuracy - Rate

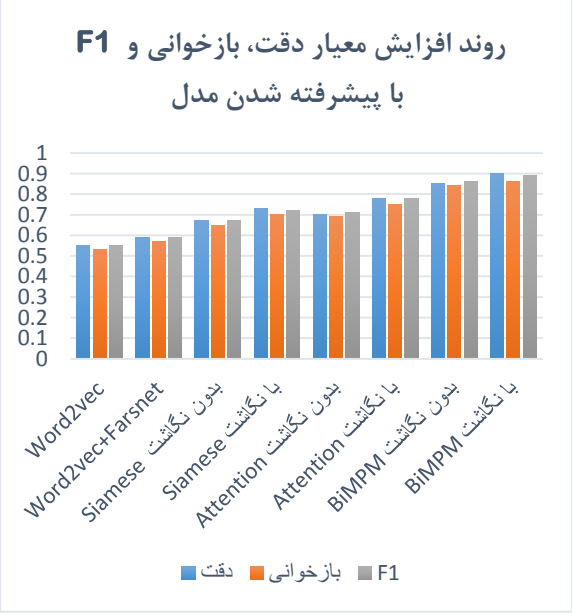
[2] Learning Rate

[3] Stochastic Gradient Descent (SGD)

جدول (۳): نتایج روش‌ها (٪)

معماری	دقت	بازخوانی	F1
Word2vec	۰/۵۵	۰/۵۳	۰/۵۵
Word2vec+Farsnet	۰/۵۹	۰/۵۷	۰/۵۹
Siamese بدون نگاشت	۰/۶۷	۰/۶۵	۰/۶۷
Siamese با نگاشت	۰/۷۳	۰/۷۰	۰/۷۲
Attention بدون‌نگاشت	۰/۷۰	۰/۶۹	۰/۷۱
Attention با نگاشت	۰/۷۸	۰/۷۵	۰/۷۸
BiMPM بدون نگاشت	۰/۸۵	۰/۸۴	۰/۸۶
BiMPM با نگاشت	۰/۹۰	۰/۸۶	۰/۸۹

در این قسمت، بخشی از آزمایش‌هایی که در این پژوهش انجام گرفته است، بیان گردید. آزمایش‌های زیادی برای تعیین هایپر پارامترها انجام شده است که تنها بهترین نتایج ذکر می‌گردد. در واقع آزمایش‌ها از یک معماری ساده‌تر شروع می‌شود و در هر مرحله با افزودن اجزایی به شبکه بهترین نتیجه گزارش شده است. در تمامی آزمایش‌ها از بردار Word2Vec با ۳۰۰ بعد استفاده شده است.



شکل (۳): نمودار دقت در مدل‌های مختلف مورد بررسی

اولین روش بررسی شده تشخیص شباهت بین دو جمله بر اساس محاسبه فاصله کسینوسی از طریق بردار تعبیه واژگان می‌باشد که در سطر اول جدول نتایج ذکر شده است. در این روش ابتدا پیش‌پردازش لازم بر روی دادگان صورت گرفته، سپس دو جمله به بردار متناظر از طریق word2vec پیش-آموزش دیده تبدیل شده و میانگین هر بردار به‌دست آمده، سپس فاصله کسینوسی بین دو بردار محاسبه شده است.

به‌عنوان مثالی برای این روش اگر سه جمله زیر را درنظر گرفته، در صورتی که مقایسه‌ای مفهومی و معنایی بین دو جمله انجام شود، به تشخیص انسانی، جمله اول و دوم شبیه‌تر از دو جمله اول و سوم می‌باشد. اما با توجه به نتایج به‌دست آمده از این روش، به دلیل توجه بیشتر به میانگین کلمات و توجه به تعداد کلمات مشابه‌تر در دو جمله، جمله اول و سوم را شبیه‌تر از جمله اول و دوم درنظر می‌گیرد.

۱. پایتخت ایران یکی از آلوده‌ترین شهرهای کشور است.
۲. تهران یکی از غبارآلوده‌ترین شهرهای جهان می‌باشد.
۳. پایتخت ایران یکی از بزرگ‌ترین‌های شهرهای کشور است.

خروجی این مدل، فاصله کسینوسی بین دو بردار می‌باشد که به‌صورت مقداری بین ۰ و ۱ می‌باشد. برای به‌دست آوردن خروجی مناسب چند مقدار آستانه درنظر گرفته شده و خروجی هر کدام نمایش داده شده است که خروجی حاصل از مقدار آستانه ۰/۸ بهترین مقدار خروجی بوده است. نتایج حاصل از آستانه‌های مختلف در جدول (۴) نمایش شده است.

جدول (۴): نتایج حاصل از اولین معماری با آستانه‌های مختلف (٪)

مقدار آستانه	دقت	بازخوانی	F1
۰/۳	۴۱/۰۳	۴۰/۷۶	۴۱/۳۹
۰/۵	۴۹/۲۶	۴۷/۷۸	۴۹/۵۲
۰/۸	۵۵/۷۶	۵۳/۳۰	۵۵/۰۲

در مجموعه شبکه واژگانی، کلمات واحدی که در کاربردهای مختلف و با معانی مختلف ظاهر می‌شود، در اصطلاح دارای sense‌های مختلف می‌باشند. یکی از روابط مهم بین sense‌ها، رابطه Synonym یا هم‌معنی بودن است. دو sense متعلق به دو کلمه، زمانی هم‌معنی هستند که حامل مفهوم یکسانی بوده و قابلیت جایگزینی با یکدیگر را داشته باشد. Synset به مجموعه از sense‌ها گفته می‌شود که دارای رابطه هم‌معنی بودن باشند؛ به‌عنوان مثال، کلمات "والد-بابا-پدر" یک Synset محسوب می‌شوند [۵].

روش دوم بر این اساس مورد بررسی قرار گرفته تا بتوان کلماتی که از نظر معنایی در دو جمله مشابه هستند یا به عبارتی

مترادف‌های متفاوت در دو جمله به‌کار گرفته شده است (Synset)، با استفاده از فرهنگ واژگان فارسی به یکی از این کلمات در هر دو جمله تبدیل شود و بتوان با استفاده از میانگین‌گیری بردار واژگان نتایج بهتری حاصل نمود.

به‌عنوان مثال اگر سه جمله مثال زده شده در روش قبل درنظر گرفته شود، در جمله اول و دوم (غبارآلوده‌ترین و آلوده‌ترین) و (تهران و پایتخت ایران) که در فرهنگ واژگان فارس‌نت هر کدام از این دو کلمه به‌عنوان مترادف همدیگر درنظر گرفته شده است، می‌توان یکی از این دو کلمه را در هر دو جمله، جایگزین نمود تا در هر دو جمله کلمات مشابه‌تری موجود باشد. در حقیقت با این روش می‌توان "پایتخت ایران" به "تهران" و "غبارآلودترین" به "آلوده‌ترین" یا برعکس تبدیل شود که می‌توان نتایج بهتری نسبت به روش قبل حاصل نمود. خروجی به‌دست آمده با استفاده از این روش در سطر دوم جدول (۳) ذکر شده است.

تفاوت روش اول با روش دوم به این صورت است که محاسبه فاصله کسینوسی فقط بر اساس مدل‌های پیش‌آموزش داده شده انجام می‌شود که فقط می‌تواند کلمات نزدیک‌تر به هم و کلماتی که در دو جمله عیناً تکرار شده‌اند را تشخیص دهد. به‌عنوان مثال ایران و آمریکا شاید کلمات متضادی نسبت به هم باشند اما چون در اکثر جملات مربوطه این دو کنار هم قرار گرفته‌اند نزدیک به هم تشخیص داده و این کلمات در فضای برداری بهم نزدیک‌تر می‌باشد، در نتیجه فاصله کسینوسی کمتری نسبت به هم دارند.

اما در روش دوم با استفاده از فرهنگ واژگان فارس‌نت می‌توان هم‌معنی‌های مربوطه در دو جمله را پیدا نمود و آن‌ها را به یک کلمه تبدیل کرد تا فاصله کسینوسی کمتری نسبت به هم داشته باشند. در جدول (۵) نمونه‌هایی از این روش و میزان تشخیص سامانه نسبت به هر دو جمله، نمایش داده شده است که امتیازات به شرح زیر می‌باشد:

- امتیاز ۱: برچسب واقعی داده
- امتیاز ۲ : محاسبه فاصله کسینوسی با حذف ایست‌واژه‌ها و تبدیل جملات به بردار مربوطه با استفاده از مدل آموزش دیده word2vec
- امتیاز ۳: محاسبه فاصله کسینوسی با حذف تک‌واژه‌ها و با استفاده از فارس‌نت و تبدیل جملات به بردار مربوطه با استفاده از مدل آموزش دیده word2ve

جدول (۵): میزان شباهت جملات با استفاده از دو معماری مذکور

جمله ۱	جمله ۲	امتیاز ۱	امتیاز ۲	امتیاز ۳
فرد درون‌گرا کارهای خود را به فردی انجام می‌دهد	فرد درون‌گرا وقت خود را به عزلت می‌گذراند	۰	۰/۷۱	۰/۷۱
لاکچری به معنای عرض وجود کردن است	خودنمایی از بارزه لاکچری‌ها است	۱	۰/۷۰	۰/۸۵

در این مدل نیز خروجی حاصل از مقدار آستانه ۰/۸ بهترین نتیجه را حاصل شده است. نتایج حاصل از آستانه‌های مختلف در جدول (۶) نیز نمایش داده شده است.

جدول (۶): نتایج حاصل از دومین معماری با آستانه‌های مختلف (٪)

مقدار آستانه	دقت	بازخوانی	F1
۰/۳	۴۸/۰۳	۴۶/۷۶	۴۸/۳۹
۰/۵	۵۴/۲۶	۵۲/۷۸	۵۳/۵۲
۰/۸	۵۹/۷۶	۵۷/۳۰	۵۹/۰۲

نتایج حاصل از دو روش قبل که به‌عنوان روش‌های غیر یادگیری عمیق مورد بررسی قرار گرفت، نتوانستند انتظار این پژوهش در حوزه تشخیص شباهت دو جمله را به خوبی برآورده کنند. در حقیقت یک مدل استاندارد نباید به انواع حالت‌های بیان یک کلمه و نحوه نوشتن حساس باشد.

در این راستا، ایده استفاده از شبکه‌های عصبی در پردازش زبان مطرح شده است که متون مختلف را به یک بردار با طول ثابت نگاشت می‌کند که در کاربردهای مختلف مورد بررسی قرار گرفته است. در حقیقت در این روش به جای توجه به خود لغت، به کل متن با ابعاد بزرگ‌تر توجه خواهد شد. بنابراین، این مقاله نیز، چندین روش پیشنهادی در مقالات دیگر را به‌عنوان روش‌های پایه کار خود مورد بررسی قرار داده است که در ادامه بر روی نتایج آن‌ها بحث می‌شود.

یکی از مدل‌های مورد بررسی یادگیری عمیق، استفاده از شبکه‌های Siamese می‌باشد. در این مدل ورودی‌ها به بردارهای لغتی از قبل آموزش دیده تبدیل شده سپس، این بردارهای

بازنمایش وارد یک شبکه با ۴ لایه BiLSTM می‌شوند تا ویژگی-های معنایی از جملات استخراج شود.

در هر لایه LSTM، ۶۴ واحد یا به عبارتی ۱۲۸ واحد در BiLSTM تشکیل شده است. سپس میانگین این ۱۲۸ واحد در هر گام زمانی محاسبه می‌شود. این روش از طریق ورودی‌های تصادفی گوسی، مقداردهی اولیه شده است، سپس روی دادگان مورد نظر آموزش داده می‌شود.

در این معماری با استفاده از محاسبه فاصله منهتن دو خروجی شبکه BiLSTM میزان شباهت بین دو جمله تشخیص داده شده است. سپس دقت خروجی با برچسب واقعی محاسبه می‌شود. نتایج این روش نشان داده است که اگر به اندازه کافی دادگان به‌صورت جفت جمله موجود باشد، توانایی یادگیری معنایی غنی‌ای از جملات دارد. با توجه به دادگان محدود جمع-آوری شده فارسی، این روش در کمال سادگی نسبت به روش‌های دیگر برای تشخیص شباهت بین دو جمله به نسبت نتایج خوبی داشته است. خروجی به‌دست آمده از این مدل در سطر سوم جدول (۳) نمایش داده شده است.

مدل دیگر مورد بررسی از مدل‌های یادگیری عمیقی در این پژوهش، محاسبه شباهت معنایی با استفاده از سازوکار توجه می-باشد. در این مدل ورودی‌ها به بردارهای لغتی که از قبل آموزش دیده‌اند تبدیل می‌شوند. سپس از یک شبکه BiLSTM برای استخراج ویژگی‌های معنایی از جملات استفاده می‌شود. سازوکار توجه برای وزن‌دهی کلمات با اهمیت خروجی این بردار واژگان، اعمال می‌شود.

خروجی دو جمله از شبکه BiLSTM به لایه توجه با فرستاده شده و بردار وزن مربوط به این دو جمله استخراج می‌شود. خروجی لایه توجه با بردار تعبیه واژگان جمله اول به ۴ لایه با تابع فعالیت relu با اندازه ۲۰۰ و روش droupout بر روی واحدهای بازگشتی (با احتمال ۰/۱) استفاده شده تا از سرریز شدن بیش از حد جلوگیری شود. در انتها نیز خروجی این لایه‌ها به یک تابع sigmoid فرستاده می‌شود تا برچسب مورد نظر توسط سامانه پیش‌بینی شود.

خروجی این مدل یک لایه sigmoid قرار داده شده است که نشان دهنده میزان مقدار ارسال ورودی به خروجی ارسال است. مقدار صفر یعنی هیچ اطلاعاتی نباید به خروجی ارسال شود، در حالی که مقدار یک یعنی تمام ورودی به خروجی ارسال شود. سپس یک مقدار آستانه برای خروجی این مدل درنظر گرفته می‌شود که اگر بالاتر از حد آستانه (۰/۵) باشد، مشابه و اگر پایین‌تر باشد غیر مشابه درنظر گرفته می‌شود. با توجه به خروجی به‌دست آمده از مدل و خروجی واقعی دقت مدل به‌دست خواهد آمد.

با مقایسه سه معیار دقت، صحت و بازخوانی شاخص سایر روش‌ها با روش مورد بررسی، این روش تا حدودی به نتایج قابل قبول‌تری دست پیدا کرده است. دلیل این امر آن است که ترکیب BiLSTM و سازوکار توجه در سطح کلمه که برای بهبود وزن کلمات معنی‌دار در یک جمله اتخاذ شده است، به‌طور جامعی، اطلاعات محلی و خلاصه اطلاعات را درنظر می‌گیرد.

به‌منظور تأیید این نکته که این روش قادر به انتخاب کلمات مهم یک جمله است، نتیجه آزمایش‌ها نشان می‌دهد که مدل ترکیبی شبکه BiLSTM و مبتنی بر توجه کارآمد بوده و نسبت به روش‌های قبلی، عملکرد به مراتب بهتری برای اندازه‌گیری شباهت بین دو جمله دارد. خروجی به‌دست آمده از این مدل در سطر پنجم جدول (۳) نمایش داده شده است.

روش سوم از مدل‌های یادگیری عمیقی مورد بررسی، محاسبه شباهت معنایی با استفاده از مدل تطبیق دو جهته چند جانبه می‌باشد. در این مدل نیز ابتدا ورودی‌ها توسط بردار تعبیه واژگان از پیش آموزش دیده تبدیل و تعبیه واژه‌های خارج از لیست واژگان به‌صورت تصادفی تولید می‌شود.

به این صورت که هر کاراکتر به‌عنوان یک بردار ۲۰ بعدی و هر کلمه به یک بردار ۵۰ بعدی تبدیل می‌شود و وارد یک شبکه BiLSTM برای استخراج ویژگی‌های معنایی از جملات می‌شود. سازوکار توجه برای وزن‌دهی کلمات با اهمیت خروجی این بردار واژگان، اعمال می‌شود. سپس این دو بردار نیز با یک شبکه BiLSTM دیگر با هم ترکیب می‌شود.

در این روش تطبیق چند جانبه انجام می‌شود، سپس خروجی لایه تطبیق به یک بردار با طول ثابت تجمیع می‌شود. در انتها یک لایه کاملاً متصل که دارای تابع فعالیت Softmax است، برای تبدیل بردارهای بازنمایش به توزیع احتمالاتی، قرار داده شده است، تا مشخص شود که هر بردار دو جمله متعلق به کدام کلاس مشابه و غیر مشابه می‌باشد. خروجی به‌دست آمده از این مدل در سطر هفتم جدول (۳) نمایش داده شده است.

۳-۳- ارزیابی روش نگاشت

برای بهترکردن نتایج می‌توان از روش نگاشت و دادگان بیشتر انگلیسی استفاده نمود. با توجه به توضیحات داده شده در بخش قبل که نحوه استفاده از روش نگاشت توضیح داده شده، ابتدا مدل با بردار از پیش-آموزش-دیده نگاشت شده است و دادگان انگلیسی آموزش داده می‌شود، سپس از تنظیم دقیق آن برای آموزش فارسی استفاده می‌شود. با توجه به آزمایش‌های انجام

شده، می‌توان دقت مدل را با استفاده از این ایده و استفاده هم‌زمان از دادگان بسیار زیاد انگلیسی و سپس فارسی نتایج را بهبود بخشیده و دقت مدل در تشخیص شباهت معنایی بالاتر رود. خروجی به‌دست آمده با استفاده از روش نگاشت بر روی سه مدل یادگیری عمیق در سطر چهارم، ششم و هشتم جدول (۳) نمایش داده شده است.

با توجه به ارزیابی نتایج، مدل BiMPM نتایج بهتری را بعد از عمل نگاشت به‌دست آورد. این مدل از بردار واژگان به علاوه بردار تعبیه حروف با هم استفاده می‌نماید. دوم اینکه این مدل از چند لایه شبکه BiLSTM استفاده نموده است که می‌تواند اطلاعات قبل و بعد از آن گام را نگه‌داری کند. این مدل، تطابق بین دو جمله را از چند جهت مورد بررسی قرار داده و تجمیع این چند تطبیق را با هم استفاده می‌نماید. یکی از حالت‌های تطبیق استفاده از سازوکار توجه می‌باشد که این سازوکار اطلاعات معنایی با اهمیت بیشتر را استخراج می‌کند و به خروجی می‌فرستد.

۴- نتیجه‌گیری

در سال‌های اخیر، به‌کارگیری یادگیری عمیق در حوزه‌های مختلف پردازش زبان طبیعی توسعه چشمگیری داشته است. یادگیری عمیق این امکان را فراهم می‌کند که در هر کاربردی از پردازش زبان طبیعی از جمله تشخیص شباهت معنایی جملات، بدون نیاز به مهندسی ویژگی‌ها به‌صورت سنتی و وجود فرد خبره، مدل به‌صورت سراسری آموزش ببینند.

به تازگی از مدل‌های یادگیری عمیق برای داده‌هایی با ماهیت دنباله‌ای استفاده می‌گردد. متن، جزو داده‌هایی است که ماهیت دنباله‌ای دارد و برای این منظور شبکه‌های LSTM پیش‌بینی و دقت بهتری را به نسبت شبکه‌های RNN نتیجه می‌دهند و وابستگی‌های طولانی‌تری را در دنباله داده‌ها به‌دست می‌آورند. در کاربردهایی از زبان طبیعی همچون سرویس ترجمه ماشینی گوگل، از این شبکه‌ها استفاده شده است و نتایج بسیار خوبی به‌دست آمده است.

از نتایج حاصل شده که در بالا ذکر شد، می‌توان نتیجه گرفت که استفاده از هر دو بردار تعبیه واژگان و حروف نتایج را بهبود می‌بخشد. بنابراین ایجاد معماری خاصی که قادر به استخراج ویژگی‌های غنی‌تر از محتوای متن آموزشی باشد، یادگیری را افزایش می‌دهد.

برای مثال در سرویس ترجمه ماشینی گوگل از چندین لایه LSTM پس از BiLSTM استفاده شده است ولی در این پژوهش بعد از هر حالت یک BiLSTM درنظر گرفته شده است. چرا که افزایش تعداد لایه‌ها با توجه به کم بودن حجم پیکره، موجب پیچیده شدن مدل گشته و عملکرد سامانه امکان افت دارد. نگاشت بین دو زبان نیز نتایج خوبی به دنبال داشته است که می‌توان از هر دو زبان به‌صورت مشترک استفاده نمود و دقت تشخیص شباهت را بالا برد.

۵- مراجع

[1] R., Mihalcea, C. Corley, and C. Strapparava, "Corpus-based and Knowledge-based Measures of Text Semantic Similarity," In Aaai, Vol. 6, pp.775-780, 2006.

[2] D. T. Tolciu, C. Sacarea, and C. Matei, "Analysis of Patterns and Similarities in Service Tickets using Natural Language Processing," Journal of Communications Software and Systems, vol. 17, no. 1, pp. 29-35, 2021.

[3] 3.V. Bahel and A. Thomas, "Text Similarity Analysis for Evaluation of Descriptive Answers," ArXiv Preprint ArXiv, 2105.02935, 2021.

[4] S. Mizzaro, M. Pavan, and I. Scagnetto, "Content-based Similarity of Twitter Users," In European conference on information retrieval, Springer, 2015.

[5] Z. Sepehrian, S. S. Sadidpour, and H. Shirazi, "An Approach Based on Semantic Similarity in Persian Query-Based Summarization," Scientific Journal of Electronic and Cyber Defense, vol. 2, no. 3, pp. 51-63, 2014 (in Persian).

[6] Z. Wang, W. Hamza, and R. Florian, "Bilateral Multi-perspective Matching for Natural Language Sentences," ArXiv Preprint ArXiv, 1702.03814, 2017.

[7] J. Mueller and A. Thyagarajan, "Siamese Recurrent Architectures for Learning Sentence Similarity," In Thirtieth AAAI Conference on Artificial Intelligence, 2016.

[8] W. H. Gomaa and A. A. Fahmy, "A Survey of Text Similarity Approaches," International Journal of Computer Applications, vol. 68, no. 13, pp. 13-18, 2013.

[9] M. Farouk, "Measuring Sentences Similarity: A Survey," arXiv:1910.03940v1, July 2019.

[10] Y. Wang, X. Di, J. Li, H. Yang, and L. Bi, "Sentence Similarity Learning Method based on Attention Hybrid Model," In Journal of Physics: Conference Series, IOP Publishing, 2018.

[11] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space." ArXiv Preprint ArXiv, 1301.3781, 2013.

[12] Y. Doval, J. Camacho-Collados, L. Espinosa-Anke, and S. Schockaert, "Improving Cross-lingual Word Embeddings by Meeting in the Middle," ArXiv Preprint ArXiv, 1808.08780, 2018.

[13] A. Conneau, G. Lample, M. A. Ranzato, L. Denoyer, and H. Jégou, "Word Translation Without Parallel Data," ArXiv Preprint ArXiv, 1710.04087, 2017.

[14] M. Artetxe, G. Labaka, and E. Agirre, "Learning Bilingual Word Embeddings with (almost) no Bilingual Data," In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2017.

[15] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding," ArXiv Preprint ArXiv,1810.04805, 2018.

[16] H., Huang, Y. Liang, N. Duan, M. Gong, L. Shou, D. Jiang, and M. Zhou, "Unicoder: A Universal Language Encoder by Pre-training with Multiple Cross-lingual Tasks," ArXiv Preprint ArXiv, 1909.00964, 2019.

[17] G. Lample and A. Conneau, "Cross-lingual Language Model Pretraining," ArXiv Preprint ArXiv, 1901.07291, 2019.

[18] H. Gonen, S. Ravfogel, Y. Elazar, and Y. Goldberg, "It's not Greek to mBERT: Inducing Word-Level Translations from Multilingual BERT," ArXiv Preprint ArXiv, 2010.08275, 2020.

[19] R. Samuel, G. A. Bowman, and C. Potts, "The Stanford Natural Language Inference (SNLI) Corpus, arXiv preprint arXiv:1508.05326, 2015.