سامانه‌های پردازشی و ارتباطی چندرسانه‌ای هوشمند
Intelligent Multimedia Processing and Communication Systems (IMPCS)

Winter, 4 (4), 11-23
DOR:

Received: 1 Nov 2023
Accepted: 11 Dec 2023

مقاله پژوهشی

Joint Learning Approach with Attention-Based Model for Semantic Textual Similarity

Ibrahim Ganjalipour[1], Amir Hossein Refahi[2*], Sohrab Kordrostami[3], Ali Asghar Hosseinzadeh[4]

1. PhD Student, Department of Applied Mathematics & Computer Science, Lahijan Branch, Islamic Azad University, Lahijan, Iran. ibrahim.ganjali@gmail.com
2. Associate Professor, Department of Applied Mathematics & Computer Science, Lahijan Branch, Islamic Azad University, Lahijan, Iran. * *Corresponding Author*, ah_refahi@yahoo.com
3. Full Professor, Department of Applied Mathematics & Computer Science, Lahijan Branch, Islamic Azad University, Lahijan, Iran. sohrabkordrostami@gmail.com
4. Assistant professor, Department of Applied Mathematics &Computer Science, Lahijan Branch, Islamic Azad University, Lahijan, Iran. hosseinzadeh_ali@yahoo.com

Abstract

Introduction: Semantic Textual Similarity (STS) across languages is a pivotal challenge in natural language processing, with applications ranging from plagiarism detection to machine translation. Despite significant strides in STS, it remains a formidable task in languages with distinct syntactic structures and limited digital resources. Linguistic diversity, especially in word order variation, poses unique challenges, exemplified by languages adhering to Subject-Object-Verb (SOV) or Subject-Verb-Object (SVO) patterns, compounded by complexities like pronoun-dropping. This paper addresses the intricate task of measuring STS in Persian, characterized by SOV word order and distinctive linguistic features.

Method: We propose a novel joint learning approach, harnessing an enhanced self-attention model, to tackle the STS challenge in both SOV and SVO language structures. Our methodology involves establishing a comprehensive multilingual corpus with parallel data for SOV and SVO languages, ensuring a diverse representation of linguistic structures. An improved self-attention model is introduced, featuring weighted relative positional encoding and enriched context representations infused with co-occurrence information through pointwise mutual information (PMI) factors. A joint learning framework leverages shared representations across languages, facilitating effective knowledge transfer and bridging the linguistic gap between SOV and SVO languages.

Results: Our model, trained on Persian-English and Persian-Persian language pairs simultaneously, successfully extracts informative features, explicitly considering differences in word order and pronoun-dropping. During the training, the batch is sampled from STS benchmark with English and Translated Persian Pair texts and fed into customized encoder to get attention matrix and output embeddings. Then, the similarity module predicts the STS score. We use the STS score to compute the Mean Square Error (MSE) loss. Evaluation on Persian-English and Persian-Persian STS-Benchmarks demonstrates impressive performance, achieving Pearson correlation coefficients of 89.51% and 92.47%, respectively. Comparative experiments reveal superior performance against existing models, emphasizing the effectiveness of our proposed approach.

Discussion: The ablation study further substantiates the robustness of our system, showcasing faster convergence and reduced susceptibility to overfitting. The results underscore the significance of our enhanced model in addressing the complexities of measuring semantic similarity in languages with diverse linguistic structures and limited digital resources. The approach not only advances cross-lingual STS capabilities but also provides insights into handling syntactic variations, such as SOV and SVO word orders, and pronoun-dropping. This research opens avenues for future investigations into enhancing STS in languages with unique structural characteristics.

Keywords: Joint Learning, English-Persian Semantic Similarity, Transformer, SOV Word Order Language, Pointwise Mutual Information.

سامانه‌های پردازشی و ارتباطی چندرسانه‌ای هوشمند
Intelligent Multimedia Processing and Communication Systems (IMPCS)
انجمن علمی تجارت الکترونیکی ایران
واحد لاهیجان

رویکرد یادگیری اشتراکی بر مبنای شبکه‌های عصبی مبتنی بر توجه برای مشابهت‌یابی متون

تاریخ دریافت: ۱۰/ ۰۸/ ۱۴۰۲
تاریخ پذیرش: ۲۰/ ۰۹/ ۱۴۰۲

دورهٔ چهارم، زمستان ۱۴۰۲
شمارهٔ چهارم، صص: ۱۱-۲۳

ابراهیم گنجعلی‌پور[۱]، امیرحسین رفاهی‌شیخانی[۲*]، سهراب کردرستمی[۳]، علی‌اصغر حسین‌زاده[۴]

۱. دانشجوی دکتری، دانشکده ریاضی کاربردی و علوم کامپیوتر، واحد لاهیجان، دانشگاه آزاد اسلامی، لاهیجان، ایران.
ibrahim.ganjali@gmail.com
۲. دانشیار، داانشکده ریاضی کاربردی و علوم کامپیوتر ، واحد لاهیجان، دانشگاه آزاد اسلامی، لاهیجان، ایران.. (نویسندهٔ مسئول)
ah_refahi@yahoo.com
۳. استاد، دانشکده ریاضی کاربردی و علوم کامپیوتر ، واحد لاهیجان، دانشگاه آزاد اسلامی، لاهیجان، ایران.
sohrabkordrostami@gmail.com
۴. استادیار، دانشکده ریاضی کاربردی و علوم کامپیوتر ، واحد لاهیجان، دانشگاه آزاد اسلامی، لاهیجان، ایران.
hosseinzadeh_ali@yahoo.com

چکیده: مشابهت‌یابی معنایی متون (STS) یک وظیفه چالش‌برانگیز در زبان‌های با منابع دیجیتالی محدود است، دشواری‌های اصلی ناشی از کمبود مجموعه‌های آموزشی دسته‌بندی‌شده و مشکلات مرتبط با آموزش مدل‌های مؤثر است. در اینجا یک رویکرد یادگیری مشترک با استفاده از مدل خودتوجه بهبودیافته برای مقابله با چالش STS در ساختارهای زبانی (فاعل، مفعول، فعل) SOV و (فاعل، فعل، مفعول) SVO معرفی‌شده‌است. ابتدا یک مجموعه داده چندزبانه جامع با داده‌های موازی برای زبان‌های SOV و SVO را ایجاد کرده و تنوع زبانی گسترده‌ای را تضمین‌می‌کنیم. ما یک مدل خودتوجه بهبودیافته با رمزگذاری نسبی موقعیت وزن‌دار جدید غنی‌شده با تزریق اطلاعات هم‌رخدادی از طریق عوامل اطلاعات مشترک نقطه‌ای (PMI) معرفی‌می‌کنیم. علاوه بر این، ما از یک چارچوب یادگیری مشترک استفاده‌می‌کنیم که نمونه‌های مشترک بین زبان‌ها را به‌منظور بهبود STS بین زبانی استفاده‌می‌کند. با آموزش همزمان در چندین جفت زبان، مدل ما توانایی انتقال دانش را به‌دست‌می‌آورد و به‌طور مؤثر پل ارتباطی بین زبان‌های با ساختارهای متفاوت SOV و SVO ایجادمی‌کند. مدل پیشنهادی ما بر روی مجموعه داده‌های STS- Benchmarks فارسی-انگلیسی و فارسی-فارسی ارزیابی‌شد و به‌ترتیب به ضریب همبستگی پیرسون ۸۸,۲۹٪ و ۹۱,۶۵٪ دست‌یافت. آزمایش‌های انجام‌شده نشان‌می‌دهد که مدل پیشنهادی ما در مقایسه با مدل‌های دیگر عملکرد بهتری داشته‌است. مطالعه کاهشی نیز نشان‌می‌دهد که سیستم ما قادر به همگرایی سریعتر است و کمتر مستعد بیش برازش است.

واژه‌های کلیدی: پردازش زبان‌های طبیعی، مشابهت‌یابی معنایی متون، شبکه‌های عصبی مبتنی بر توجه، ترنسفورمر، اطلاعات مشترک نقطه‌ای.

۱- مقدمه

توسعه تکنولوژی در زمینه پردازش زبان‌های طبیعی به‌عنوان یکی از حوزه‌های مهم در علوم کامپیوتر و هوش مصنوعی در دهه‌های اخیر تاثیر بسزایی بر صنایع مختلف و نیازهای انسانی داشته است، یکی از رویکردهای مؤثر در این زمینه، استفاده از شبکه‌های عصبی مبتنی بر توجه است. این شبکه‌ها از ایده‌های مشتق شده از مکانیزم توجه در مغز انسان الهام‌می‌گیرند و قابلیت تمرکز بر بخش‌های مهم و معنادار متون را دارا می‌باشند. شبکه‌های عصبی مبتنی بر توجه، الهام‌گرفته از مکانیزم توجه انسانی، قادر به تمرکز بر بخش‌های مهم و معنادار متون هستند. این شبکه‌ها از یک رویکرد توجه‌محور برای ترجمه ماشینی، تحلیل متن، تولید متن و وظایف متعدد دیگر در پردازش زبان‌های طبیعی بهره‌می‌برند.

تشابه متنی معنایی (STS) در زبان‌ها به‌عنوان یک وظیفه مهم در پردازش زبان طبیعی ظاهرشده‌است. مدل‌های STS کاربردی در شناسایی تقلب، ترجمه ماشینی و بازیابی اطلاعات، به ما امکان‌می‌دهد تا شباهت معنایی بین دو تکه متن را اندازه‌گیری‌کنیم. درحالی‌که STS پیشرفت‌های قابل‌توجهی داشته‌است، اما در زبان‌ها با ساختارهای نحوی متفاوت و منابع دیجیتال محدود چالش جدی باقی‌مانده‌است. یکی از جنبه‌های تنوع زبانی در تفاوت ترتیب کلمات درجملات است. برخی از زبان‌ها به‌ترتیب کلمات فاعل-مفعول-فعل (SOV) پایبند هستند، درحالی‌که دیگران الگوهای فاعل-فعل-مفعول (SVO) را دنبال‌می‌کنند. این تفاوت‌های ساختاری، در کنار عواملی مانند ضمیر حذف‌شده، وظیفه اندازه‌گیری STS در این زبان‌ها را پیچیده‌می‌کند. در اینجا، زبان فارسی با ترتیب کلمات SOV و ویژگی‌های زبانی منحصربه‌فرد خود، به‌عنوان یک مورد برای مطالعه STS چندزبانه در نظر گرفته می‌شود.

محققان از مدل‌های مبتنی بر ویژگی‌های سنتی تا معماری‌های پیچیده شبکه‌های عصبی پیش‌آموزش‌شده مانند BERT [1]، [۲] GPT و نسخه‌های چندزبانه مانند XLM-R را برای محاسبه تشابه معنایی بررسی‌کرده‌اند. نقطه تحول مهمی در این زمینه توسعه مدل‌های زبان پیش‌آموزش‌شده مانند BERT، GPT و نسخه‌های چندزبانه‌شان مانند XLM-R بوده است که روش تشخیص تشابه معنایی متن را تغییردادند. علارغم انبوه تحقیقات در زمینه تشابه متنی در زبان انگلیسی، کمبود تحقیقات در تحلیل تشابه متنی در زبان فارسی قابل‌مشاهده است.

در اینجا، یک رویکرد یادگیری ترکیبی نوین برای پرداختن به چالش‌های STS در هر دو ساختار زبان SOV و SVO معرفی‌می‌شود. رویکرد ما از یک مکانیزم خودتوجه بهبود یافته بهره‌می‌برد که از کدگذاری موقعیت‌های رابطه‌ای وزن‌دار بهرهمند است. ما به چالش‌های مربوط به STS چندزبانه می‌پردازیم، شامل جمع‌آوری داده، معماری مدل و استراتژی‌های آموزش. یک جنبه حیاتی از کار ما شامل ایجاد یک مجموعه داده چندزبانه جامع برای داده‌های موازی برای زبان‌های SOV و SVO است. این مجموعه داده نه‌تنها نمایندگی گسترده‌ای از تنوع زبانی را تضمین‌می‌کند، بلکه همچنین اساس قابلیت‌های چندزبانه مدل ما است. در اساس رویکرد ما، مدل خودتوجه پیشرفته‌ای قراردارد که طراحی‌شده‌است تا ارتباط‌های پیچیده داخلی و بین جملات و همچنین اطلاعات همزمانی را از طریق گراف مرجع نسبی سطح جمله عامل اطلاعات مشترک نقطه‌ای (PMI) درنظربگیرد. اضافه‌کردن کدگذاری موقعیت‌های رابطی وزن‌دار، مدل ما را قادر به درنظرگرفتن اهمیت موقعیت کلمات و روابط متنی آن‌ها می‌کند و اندازه‌گیری دقیق‌تری از STS را ممکن‌می‌سازد. برای آموزش موازی گونه‌های زبانی SOV و SVO از یک چارچوب یادگیری مشترک استفاده‌می‌کنیم. این چارچوب انتقال دانش را از طریق همزمان آموزش در چندین جفت نمونه از مجموعه داده فراهم‌می‌کند. این به مدل ما امکان‌می‌دهد تا با استفاده از تمثیل‌های مشترک در زبان‌ها، دید چندزبانه‌ای ارز شمندی از تشابه متنی ارائه‌دهد. مدل ما همچنین به ویژگی‌های ویژه ترتیب کلمات و قابلیت حذف ضمیر توجه خاصی می‌کند و اطلاعات متنی وابسته و وابستگی‌های دوربرد را به‌طور مؤثر دربرمی‌گیرد. با استخراج ویژگی‌های معنایی، رویکرد ما دقت اندازه‌گیری تشابه معنایی را افزایش‌می‌دهد. در ارزیابی‌های دقیق در مجموعه داده‌های STS فارسی-انگلیسی و STS فارسی-فارسی، مدل پیشنهادی ما عملکرد قابل‌توجهی را نشان‌می‌دهد و بهبودهای معناداری نسبت به روش‌های موجود کسب‌می‌کند. علاوه بر این، مطالعه موردی و مطالعه فرسایشی نشان‌می‌دهد که مدل ارائه شده کارایی، پایداری و قابلیت تعمیم بالایی دارد.

به‌طور خلاصه در این تحقیق به موارد زیر پرداخته‌ایم:

- ما یک مدل مبتنی بر ترانسفورمر جدید برای تشخیص موجودیت نام‌دار فارسی معرفی‌می‌کنیم که از مدل پیش‌آموزش‌شده XLM-R و گراف مرجع نسبی سطح جمله عامل اطلاعات مشترک نقطه‌ای (PMI) استفاده‌می‌کند
- ما یک مدل مبتنی بر ترانسفورمر جدید برای استخراج رابطه فارسی معرفی‌می‌کنیم که از مدل پیش‌آموزش‌شده XLM-R و گراف مرجع نسبی سطح جمله عامل اطلاعات مشترک نقطه‌ای (PMI) استفاده‌می‌کند.
- ما یک رویکرد یادگیری ترکیبی با مدل خودتوجه بهبودیافته برای پرداختن به چالش‌های STS در ساختارهای زبانی SOV و SVO معرفی‌می‌کنیم.
- ما کدگذار ترانسفورمر را با درنظرگرفتن ویژگی‌های زبانی قابلیت حذف ضمیر و ترتیب کلمات فاعل-مفعول-فعل در زبان فارسی، اصلاح‌می‌کنیم.
- نتایج آزمایش‌ها، عملکرد بهتر روش ارائه‌شده در مقایسه با کارهای پیشین در انجام وظایف پردازش زبانی مورد نظر را نشان‌می‌دهد.

۲- پیشینه پژوهش

در این بخش مروری بر کارهای پیشین در تشابه‌یابی معنایی متون ارائه‌می شود. در اینجا ما تلاش‌می‌کنیم که کارهای قبلی مرتبط با روش پیشنهادی را به‌منظور ایجاد ارتباط بین دانش موجود و ارتقاء های پژوهش حاضر ارائه‌کنیم.

تعیین دقیق مشابهت معنایی متون (STS)، به مدت سال‌ها یک شاخه مهم در تحقیقات پردازش زبان طبیعی بوده‌است. محققان از تکنیک‌ها و رویکردهای مختلفی برای انجام این وظیفه چالش‌برانگیز در ساختارهای زبانی متنوع استفاده‌کرده‌اند. در این بخش، ما به مرور تکامل‌ها و مشارکت‌های کلیدی در زمینه STS می‌پردازیم و تمرکز خود را بر چالش‌های موجود در زبان‌های SOV و SVO قرارمی‌دهیم. محاسبه مشابهت بین متون کوتاه برای اولین بار در سال ۲۰۰۶ گزارش شد[۳]. از آن زمان، از سال ۲۰۱۲ در کارگاه بین‌المللی ارزیابی معنایی (SemEval)، وظیفه مشابهت معنایی از مشابهت یا عدم مشابهت دودویی به محاسبه درجه مشابهت گسترده شده‌است[۴]. این محاسبات عمدتاً توسط مقدار عددی از ۰ تا ۵ نمایش‌داده‌می شود برای هر جفت متن یا جمله. ابتدایی‌ترین ایده‌ها برای شناسایی مشابهت معنایی بین دو جمله بر مبنای تطابق معنایی بین کلمات در جملات بوده‌است، که در نهایت معادل به جمع جبری مشابهت کلمات شده است. با این حال، بیشتر تحقیقات معاصر در این زمینه بر بازنمایی معنایی جملات با استفاده از تکنیک‌های یادگیری عمیق تمرکز دارند. از طریق این روش ها، جملات به بردار های عددی با ابعاد مختلف تبدیل‌می شوند. ایجاد بردارهای تعبیه معمولاً با ا ستفاده از متون بزرگ انجام‌می‌شود. در زبان انگلیسی به‌دلیل استفاده گسترده و دسترسی به متون بزرگ، تحقیقات بیشتری در این حوزه انجام شده‌ا ست. اما برای زبان‌های با منابع و متون محدودتر مانند فارسی، تحقیقات در این حوزه به نسبت کمتر بوده‌است.

در اینجا ما تلاش‌می‌کنیم که کارهای قبلی مرتبط با روش پیشنهادی را به‌منظور ایجاد ارتباط بین دانش موجود و ارتقاءهای ما خلا صه‌کنیم. ما تعدادی از مطالعات تحقیقاتی برجسته در زمینه مشابهت متنی چندزبانه و یکزبانه معرفی‌می‌کنیم. مدل Multilingual BERT [۱] یک مدل زبان مبتنی بر ترانسفورمر است که پیش‌آموزش بر روی یک مجموعه داده چندزبانه بزرگ انجام‌شده‌است. ثابت‌شده‌است که M-BERT عملکرد بسیار خوبی را در و ظایف مختلف پردازش ز بان چندزبانه، شامل م شابهت معنایی چندزبانه، دارد. توانایی درک و تولید بردار های بازنمایی برای چندین ز بان آن را به یک ابزار مهم برای کاربردهای چندزبانه می‌کند.

یک نسخه دیگر از BERT به نام DistilBert [۵] از دانش انلاین هنگام پیش‌آموزش بهره‌می‌برد و نشان‌می‌دهد که می‌توان اندازه یک مدل BERT را ۴۰٪ کاهش‌داد، در حالی که ۹۷٪ از توانایی‌های درک زبان آن حفظ‌شود و سرعت آن ۶۰٪ افزایش‌یابد.

مدل XLM-R یک افزو نه از Multilingual BERT است که مدل‌سازی چندزبانه را بهبودمی‌بخشد. این مدل بر روی حجم زیادی از داده های ۱۰۰ ز بان پیش‌آموزش‌د یده است و بهترین ن تایج را در مجموعه گ سترده‌ای از وظایف پردازش زبان چندزبانه، شامل م شابهت معنایی چندزبانه، به‌د ست‌آورده‌ا ست. کارآیی این مدل در انجام وظایف زبان‌های کم منبع آن را یک گزینه برجسته برای تحقیقات چندزبانه می‌کند.

برای زبان‌های کم‌منبع مانند ا سپانیایی، عربی، اندونزیایی و تایلندی، تانگ و همکاران در سال ۲۰۱۸ [۶] یک مدل چندزبانه معرفی‌کردند. آن‌ها یک چارچوب مدل مشابهت معنایی یک‌زبانه را به یک تنظیم چندزبانه گسترش‌دادند و نشان‌دادند که با استفاده از یک کدگذار چندزبانه م شترک، هر جمله می‌تواند نمایش‌های مختلفی را ن شان‌دهد که به زبان هدف وابسته هستند.

بریکچین در [۷] ایده‌هایی معرفی کرد که ف ضاهای معنایی چندزبانه در یک فضای مشترک با استفاده از لغتنامه‌های دوزبانه تبدیل‌می‌شوند. آن‌ها از روش‌های بدون نظارت برای محاسبه مشابهت جملات تنها بر مبنای تعبیه‌های معنایی بهره‌بردند. آن‌ها ن شان‌دادند که بهبود فضاهای مشترک مع نایی از طریق وزن‌دهی به کل مات می‌توا ند به ن تایج کمک‌کند. یافته‌های آن‌ها نشان‌داد که ضریب همبستگی پیرسون ۶۱,۸ درصد در جفت جملات عربی-انگلیسی دارد.

در [۸] تعاریف Wordnet به ۷ زبان مختلف برای ایجاد یک زمینه آزمون مشابهت معنایی متنی چندزبانه به‌کارمی‌رود. یک وظیفه تطابق اسناد برای استفاده بین نمادهای توصیفی Wordnet در ۷ زبان مختلف ایجادشده‌است. روش‌های مشابهت متنی بدون نظارت مانند فاصله واسترشتاین، فاصله سینکهورن و مشابهت کسینوس با یک مدل عمیق آموزش‌داده‌شده Siamese مقایسه‌شده‌اند. این وظیفه به‌عنوان یک وظیفه بازیابی و وظیفه تطابق مدل می‌شود تا اثر توابع مشابهت معنایی را برر سی‌کند و نتایج ن شان‌دادند که مدّنظر گرفتن م سأله به‌عنوان یک مسأله بازیابی و تطابق تأثیر مخربی بر روی نتایج دارد.

پیرس و همکاران [۹] مطالعاتی را درخصوص کیفیت Multilingual BERT برای وظایف چندزبانه انجام‌دادند. آن‌ها آزمایش‌های مختلفی را روی مجموعه‌داده‌های متنوع با استفاده از مدل Multilingual BERT انجام دادند و نتایج مطلوبی را به‌دست‌آوردند. در برخی از آزمایش‌های انجام‌شده در دو زبان مختلف، تعبیه‌های چندزبانه برای جفت جملات در برخی زبان‌ها، مانند انگلیسی و ژاپنی، دقت نسبتاً پایینی داشتند. کاهش دقت می‌تواند به تفاوت‌های در ساختارهای زبانی بین دو زبان برگردد. زبان‌های مانند انگلیسی ساختار SVO را دنبال‌می‌کنند، که ساختار جمله‌ای معمولاً شامل ترتیب کلمات فاعل-فعل-مفعول ا ست. در ت ضاد با این مو ضوع، زبان‌های مانند فار سی ساختار SOV (فاعل-مفعول-فعل) را دنبال‌می‌کنند، که معمولاً شامل قرار گرفتن فاعل یا نهاد در ابتدا، سپس مفعول و در آخر فعل است. با توجه به تحقیقات موجود، کار ما با ترکیب عناصری از مکانیزم‌های خودتوجه پیشرفته،

کدگذاری موقعیت تطابقی با وزن و تکنیک‌های یادگیری مشترک برای مقابله با چالش‌های STS چندزبانه در هر دو زبان SOV و SVO به روش منحصربه‌فردی اقدام‌می‌کند. با بهره‌گیری از این نوآوری‌ها، ما قصدداریم برای STS چندزبانه با پیچیدگی‌های خاصی که تنوع زبانی و تغییرات ترتیب کلمات ایجادمی‌کنند یک راهکارنوین ارائه‌کنیم.

۳- روش پیشنهادی

۳-۱- ویژگی‌های موثر در مکانیزم خود-توجه در زبان فارسی

فاعل و فعل بخش‌های کلیدی در بیان معنی جمله هستند و توجه میان آن‌ها در ایجاد نمایش متنی زمینه‌ای بهتر بسیار مؤثر است. فارسی دارای ترتیب کلماتی فاعل-مفعول-فعل (SOV) و مشخصه حذف‌پذیری ضمیر است. در فارسی، فعل جمله اغلب تا انتهای جمله آشکار می‌شود. بر خلاف زبان انگلیسی، در فارسی فاعل و فعل فاصله مکانی از یکدیگر دارند. اگر لایه توجه موقعیت کل دنباله را به‌عنوان ورودی نپذیرد و آن را قطع‌کند (فاعل و فعل در یک قالب پردازشی نیستند)، بردار تعبیه خروجی مفهوم کلی دقیق را شامل نخواهدشد.

۳-۲- توکن سازی

برای توکن‌سازی، ما از مدل پیش‌آموز XLM-R [۵] استفاده‌می‌کنیم و همچنین از بردارهای تعبیه کلمات تولید شده برای بردارهای ورودی مدل سفارشی خود نیز استفاده‌می‌کنیم. XLM-R در مقایسه با BERT چندزبانه (M-BERT) [۴] در مجموعه متنوعی از معیارهای میان‌زبانی بهتر عمل‌می‌کند، از جمله دقت میانگین ۱۴٫۶+ در استنباط زبان طبیعی میان‌زبانی، امتیاز F1 میانگین ۱۳+ در پاسخ‌دهی به سوالات چندزبانی و امتیاز F1 +2.4 در تشخیص نام واژه‌های موجود در زبان. XLM-R بر روی داده‌های زبانی با حجم ۲٫۵ ترابایت از داده‌های متنی Common Crawl در ۱۰۰ زبان (شامل فارسی) آموزش‌داده‌شده‌است. این مدل بهبود قابل‌توجهی نسبت به مدل‌های چندزبانه منتشرشده قبلی مانند Multilingual BERT را در وظایف پسینی مانند طبقه‌بندی، برچسب‌گذاری دنباله‌ها و پاسخ به سوالات ارائه‌می‌دهد. XLM-R از روش Sentence-piece [۱۰] برای توکن‌سازی زیرکلمه‌ای استفاده‌می‌کند که به‌خصوص در زبان‌های با منابع کم مانند فارسی عملکرد خوبی دارد.

پس از فرآیند توکن سازی، ما یک گراف ناهمگون در سطح جمله از توکن ها ایجادمی‌کنیم. خروجی فرآیند توکن‌سازی مجموعه‌ای از توکن‌های زیرکلمه‌ای با شناسه‌های متناظر است. برای درنظرگرفتن هم رخدادی توکن‌ها در متن یا مجموعه داده، ابتدا توکن‌های با فراوانی بالا و کلمات توقف (مثل ".", "!", "آن", "از", "به", ...) حذف‌می‌کنیم تا مقدار PMI آن‌ها صفر شود. ما یک گراف متنی ناهمگون G=(V,E) می‌سازیم. گراف متن شامل گره‌های توکن (V) است که تمام توکن‌های موجود در لغتنامه متنی را نمایندگی‌می‌کنند. گراف متن همچنین شامل یال‌های توکن-توکن (E) است که براساس هم‌رخدادی محلی توکن‌ها در پنجره‌های کشویی (قالب‌های فرضی که شامل چند توکن هستند) در متن ایجادمی شود، و وزن یال‌ها توسط اطلاعات مشترک نقطه‌ای (PMI) اندازه‌گیری‌می‌شود. همان‌طور که در بخش ۳,۱ توضیح‌داده شده‌است، فاعل (در ابتدای جمله) و فعل (در انتهای جمله) در زبان با ترتیب SOV معمولاً دور از یکدیگر قراردارند. ما می‌خواهیم هم‌رخدادی آن‌ها را با عامل PMI درنظربگیریم. بنابراین، هنگامی که پنجره کشویی در ابتدای جمله است، همچنین هم‌رخدادی با آخرین کلمات و برعکس را اندازه‌گیری‌می‌کنیم. به‌عنوان مثال، اگر طول پنجره ۴ باشد، برای کلمه اول، هم رخدادی با سه کلمه آخر نیز اندازه‌گیری‌می‌شود، و برای کلمه آخر، هم‌رخدادی با سه کلمه اول اندازه‌گیری‌می شود. برای اندازه‌گیری هم‌رخدادی کلمات، ما از اطلاعات مشترک نقطه‌ای استفاده‌می‌کنیم. PMI یک معیار ارتباطی بین یک جفت نتایج گسسته x و y است، که به صورت زیر تعریف‌می‌شود:

$$PMI(x, y) = log\big(P(x, y) / P(x)P(y)\big) \quad (1)$$

ما یک مجموعه کلمات $w \in VW$ و متناظر با آن‌ها متن $c \in VC$ را فرض‌می‌کنیم، جایی که VW و VC واژگان کلمه و متن هستند. کلمات از یک مجموعه متنی از کلمات w_1، w_2، ...، w_n می‌آیند و متون مرتبط با کلمه w_i کلماتی هستند که در اطراف آن در یک پنجره L اندازه‌گیری‌می‌شوند: w_{i-L}، ...، w_{i-1}، w_{i+1}، ...، w_{i+L}. ما مجموعه جفت‌های کلمات و متن مشاهده‌شده را با D نمایش‌می‌دهیم. ما #(w, c) را برای تعداد باری که جفت (w, c) در D ظاهرمی‌شود، استفاده‌می‌کنیم. به همین ترتیب، #(w) و #(c) تعداد دفعاتی هستند که w و c در D ظاهرشده‌اند.

$$\#(w) = \sum_{c' \in V_c} \#(w, c') \quad (2)$$

$$\#(c) = \sum_{w' \in V_w} \#(w', c) \quad (3)$$

PMI(w, c) ارتباط بین یک کلمه w و یک متن c را با محاسبه لگاریتم نسبت بین احتمال مشترک آن‌ها (تعداد بارهایی که به همراه هم ظاهر می‌شوند) و احتمال‌های حاشیه‌ای آن‌ها (تعداد بارهایی که به‌تنهایی ظاهر می‌شوند) اندازه‌گیری‌می‌کند. مقدار PMI می‌تواند به صورت تجربی با درنظرگرفتن تعداد واقعی مشاهدات در یک متن تخمین زده‌شود:

$$PMI(w, c) = log\big(\#(w, c) \cdot |D| / \#(w) \cdot \#(c)\big) \quad (4)$$

استفاده از PMI به‌عنوان یک اندازه ارتباطی در پردازش زبان طبیعی توسط چرچ و هنکس [۱۱] معرفی‌شد و به‌عنوان وظیفه‌های شباهت واژه‌ها به‌طور گسترده مورد قبول قرارگرفت. کار با ماتریس PMI چالش‌های محاسباتی متعددی را ایجادمی‌کند. ردیف‌های ماتریس PMI حاوی بسیاری از ورودی‌های جفت کلمه-متن (w, c) هستند که هرگز در متن مشاهده نشده‌اند و بنابراین PMI(w, c) = log 0 منفی بی‌نهایت است. این ماتریس نه تنها خوش تعریف نیست، بلکه اسپارس نیز نیست، که یک مسأله عملی اساسی به دلیل ابعاد بزرگ آن است. برای حل این مسأله، ما در موارد #(w, c) = 0 مقدار PMI(w, c) = 0 تعیین می‌کنیم و ماتریس پراکنده به‌دست‌می‌آوریم. توجه‌می‌کنیم که

این ماتریس ناسازگار است، به معنای آن‌که جفت‌های کلمه-متن مشاهده‌شده ولی بی‌ارتباط دارای یک ورودی منفی در ماتریس هستند، درحالی‌که جفت‌های مشاهده‌نشده در سلول متناظر با آن‌ها مقدار ۰ دارند. به‌عنوان مثال، یک جفت کلمات نسبتاً متداول (با احتمال بالا برای w و c) که تنها یک بار با هم ظاهرمی‌شوند، شواهد قوی این موضوع را نشان‌می‌دهد که این کلمات تمایل‌دارند که با هم ظاهرنشوند، که منجر به مقدار منفی PMI می‌شود و به همین دلیل مقدار ورودی منفی در ماتریس دارد. از سوی دیگر، جفت کلماتی با احتمال‌های یکسان برای w و c که هرگز در متن با هم ظاهرنمی‌شوند، مقدار ۰ خواهندداشت. جایگزینی همه مقادیر منفی توسط ۰: (PPMI) مقدار PMI مثبت به معنای هم‌بستری زیاد کلمات در یک متن است، درحالی‌که مقدار PMI منفی به عدم هم‌بستری یا کم هم‌بستری در متن اشاره دارد. بنابراین، ما تنها یال‌ها را بین جفت‌های کلمه با مقادیر مثبت PMI اضافه‌می‌کنیم.

$$PPMI(w,c) = max(PMI\ (w,c), 0) \quad (5)$$

هنگام بازنمایی کلمات، برخی از استدلال‌ها وجوددارد که مقادیر منفی را نادیده بگیرند: انسان‌ها به راحتی می‌توانند ارتباطات مثبت را در نظر بگیرند (مثلاً دریا و ماهی) اما بسیار سخت‌تر است که ارتباطات منفی (مثلاً دریا و بیابان) استخراج‌کنند. این نشان‌می‌دهد که تشابه معنایی دو کلمه بیشتر تحت‌تأثیر محیط‌های مثبتی که به اشتراک می‌گذارند است تا محیط‌های منفی که به اشتراک می‌گذارند. به همین دلیل از محیط های منفی صرف‌نظرکردن و آن ها را به‌عنوان بی‌اطلاع (۰) نشان‌دادن نیز معقول به‌نظرمی‌آید. به‌واقع، نشان‌داده‌شده‌است که معیار PPMI عملکرد خوبی در وظایف تشابه معنایی ارائه‌می‌دهد. مقایسه معیارهای ارتباط کلمه-متن نشان‌می‌دهد که PMI و به‌خصوص PPMI برترین نتایج را برای یک مجموعه گسترده از وظایف تشابه کلمه ارائه‌می‌دهند. برای بهره‌برداری از اطلاعات کلی هم‌رخدادی کلمات در مجموعه داده، ما از یک پنجره کشویی با اندازه ثابت بر روی تمام اسناد برای هر جمله از مجموعه داده‌ها در زبان فارسی استفاده‌می‌کنیم تا آمار تطابق را جمع‌آوری‌کنیم. بعد از محاسبه PMI در تمام مجموعه داده، ماتریس مجاورت را برای هر جمله ایجادمی‌کنیم [۱۲, ۱۳] .

$$M_{ij} = \begin{cases} \text{PMI}(w_i, w_j) & w_i, w_j \text{ are words}, \text{PMI}(w_i, w_j) > 0 \\ 1 & i = j \\ 0 & \text{otherwise} \end{cases} \quad (6)$$

M ماتریس مجاورت کلمات هر جمله است و (wi، wj) شناسه مختص کلمات مجموعه داده هستند. با توجه به عامل تطابق PMI، ما در فرآیند زبان فارسی به‌عنوان یک زبان با ترتیب SOV و امکان حذف ضمیر، سه تغییر را درنظرمی‌گیریم. اول، با درنظرگرفتن تطابق کلان ضمیر فاعل و فعل، شبکه آموزش‌می‌بیند تا نمایش فعل را از جملاتی که ضمیر حذف-نشده‌است برای جملاتی که ضمیر حذف‌شده است، بیاموزد. ما متوجه-می‌شویم که با واردکردن تطابق کلان ضمیر و فعل با عامل PMI، روش ما تشخیص جنسیت و جمع یا مفرد بودن فعل (تطابق فاعل و فعل) را در صورت حذف ضمیر بهبودمی‌بخشد. دوم، با بهره‌گیری از تطابق کلمات کلی واردشده، مدل ارائه‌شده به دقت بالایی دست‌یافت. سوم، ما تنها تطابق کلمات در جمله را درنظرنگرفته‌ایم، بلکه تطابق در کل مجموعه داده را نیز درنظرگرفته‌ایم. بنابراین، مدل، توانایی تعمیم بالاتری برای به‌دست‌آوردن عملکرد پایدارتر دارد. در بخش نتایج و آزمایش ها، ما عملکرد تعمیم مدل و اثر واردکردن تطابق کلمات کلی PMI در تشخیص تشابه معنایی و تطابق فاعل و فعل را تحلیل‌می‌کنیم. مقاله برای حل کاستی‌های موجود در پژوهش‌های قبلی ذکرشود. از ذکر مراجع به‌صورت گروهی و بدون تحلیل محتوای آن‌ها خودداری‌شود.

۳-۳- روش پیشنهادی برای کدگذاری موقعیت نسبی در بازنمایی متن

مدل سفارشی ما از بردار تعبیه مدل توکن‌سازی XLM-R [14] با کدگذاری موقعیت مناسب برای زبان فارسی استفاده‌می‌کند. در این بخش خواص ترانسفورمر را تجزیه و تحلیل می‌کنیم و دو بهبود خاص برای محاسبه بردار بازنمایی پیشنهادمی‌کنیم.

اولین بهبود این است که با توجه به مشخصات زبان فارسی (توضیح داده‌شده در بخش ۳-۱)، ما تصمیم‌گرفتیم تا حداکثر مکان نسبی موجود را در اساس طول دنباله ها در مجموعه داده آموزشی درنظربگیریم. کدگذاری موقعیت نسبی [۱۵] فرض‌کرده‌است که اطلاعات دقیق موقعیت نسبی برای فاصله دور بی‌فایده است ، در حالیکه در زبان فارسی به‌عنوان یک زبان با حذف ضمیر، فاعل و فعل فاصله دارند (فعل در انتهای جمله ظاهر می‌شود) و همچنین آن‌ها قسمت‌های معنوی کلیدی یک جمله برای تولید مفهوم کلی هستند. بنابراین، ما در مدل خود از برش‌زنی استفاده‌نمی‌کنیم و وزن موقعیت نسبی بین تمام توکن‌ها را درنظرمی‌گیریم.

بهبود دوم مرتبط با واردکردن اطلاعات متقابل نقطه‌ای بین عناصر دنباله در کل متن است. ما گراف کامل متصل بین توکن‌های جملات را در مرحله پیش‌پردازش توضیح‌داده‌شده در بخش ۲-۳ ایجاد می‌کنیم و عامل PMI را در لایه توجه استفاده‌می‌کنیم. وزن‌های یال‌های ارتباطی بین توکن‌ها از ماتریس مجاورت تولید شده از معادله (۶) به‌دست‌می‌آید. در ادامه بهبود ما برای کدگذاری ترانسفورمر معمولی [15] را مطرح‌می‌کنیم.

کدگذار ترانسفورمر ورودی $X \in R_{n\times d}$ را می‌پذیرد، جایی که n طول دنباله و d ابعاد بردار تعبیه ورودی است. ماتریس ورودی از فرآیند توکن سازی بخش ۳-۲ می‌آید. سپس سه ماتریس یادگیری‌پذیر Wq، Wk، Wv برای تبدیل‌کردن X به فضاهای مختلف استفاده‌می‌شوند. معمولاً ابعاد سه ماتریس همه $R_d\times d_k$ هستند، جایی که d_k پارامتر فرضی است. پس از آن، می‌توان درایه‌های ماتریس توجه را با معادلات

زیر محاسبه‌کرد:

$$Q,K,V = XW^{Q}, XW^{K}, XW^{V},$$
$$A_{t,j} = Q_t K_j^{T},$$
$$Z = Attention(K,Q,V) = softmax(\frac{A}{\sqrt{d_k}})V,$$

$$z_i = \sum_{j=1}^{n} \alpha_{ij} (x_j W^{V} + a_{ij}^{V}) \quad (8)$$

که Q_t بردار پرسش برای توکن t امین است، j توکنی است که توکن t امین به آن توجه می‌کند. K_j نمایش بردار کلیدی از توکن j امین است. عملگر softmax در طول بعد آخر اعمال می‌شود.

$$e_{ij} = \frac{x_i W^{Q} (x_j W^{K} + a_{ij}^{K})^{T}}{\sqrt{d_z}}, \quad (9)$$

$$a_{ij}^{v} = b_{j-i}^{v} sigmoid(M_{ij}), \quad (10)$$

$$a_{ij}^{k} = b_{j-i}^{k} sigmoid(M_{ij}), \quad (11)$$

ما اطلاعات موقعیت نسبی ویژگی‌های مدل را در دو سطح تغییر داده‌ایم: مقادیر و کلیدها. این در معادلات موجود خودتوجه تغییر یافته نشان داده شده است. اطلاعات موقعیت نسبی سفارشی به عنوان یک مؤلفه اضافی به کلیدها ارائه می‌شود. ما معادله (۹) را برای منتقل کردن وزن‌های یال موقعیت نسبی که اطلاعات کلان تطابق کلی کلمات را شامل می شود پیشنهاد می‌کنیم. b_{j-i} (برای k و v) موقعیت نسبی قابل یادگیری است، که توسط تابع سیگموئید اعمال شده بر M_{ij} (ماتریس M در مرحله پیش‌پردازش تولید شده، توضیح داده شده در بخش ۲-۳) وزن‌دهی می شود. با استفاده از معادله (۱۰-۱۱) ما تطابق و موقعیت نسبی را به لایه خود-توجه وارد می‌کنیم به شرح زیر:

$$\alpha_{ij} = \frac{\exp e_{ij}}{\sum_{k=1}^{n} \exp e_{ij}} \quad (12)$$

برای پیادهسازی بهینه تر از معادله ۱۳ استفاده می کنیم.

$$e_{ij} = \frac{x_i W^{Q} (x_j W^{K})^{T} + x_i W^{Q} (a_{ij}^{K})^{T}}{\sqrt{d_z}} \quad (13)$$

$$head^{(h)} = Attention(Q^{(h)}, K^{(h)}, V^{(h)}), \quad (14)$$

جایی که x_i عناصر دنباله ورودی هستند و α بردارهای ماتریس توجه هستند. b_{i-j} وزن‌های موقعیت نسبی را نمایان می‌کنند. اگر فاصله بین دو عنصر جمله ۳ باشد، b_3 کدگذاری موقعیت نسبی ۳ را نمایان می‌کند و وزن‌های بردار در فرآیند یادگیری در موقعیت نسبی ۳ (برای $a_{1,4}$) به روز می شوند. همانطور که در (۸-۱۰) مشاهده می شود، اضافه کردن وزن‌های موقعیت جدید به بردار کلید و ضرب با بردار پرسش این معنی را می‌دهد که توجه بیشتری به عناصر توالی متناظر دارد و همچنین با استفاده از معادله ۱۰ ضرب b_{j-i} و سیگموئید($M_{i,j}$) اطلاعات کلان تطابق کلی بین عناصر i و j را در کدگذاری موقعیت نسبی وارد می‌کند. به عبارت دیگر با این تغییر، بیشترین تطابق عناصر دنباله باعث توجه بیشتر به موقعیت نسبی آن‌ها می‌شود. عملیات softmax در فرایند خود-توجه معمول تغییر نمی‌کند. برای محاسبه ماتریس توجه، از معادله (۱۱) استفاده می‌شود. به جای استفاده از یک گروه W_q، W_k، W_v، استفاده از چندین گروه توانایی خودتوجه را افزایش می‌دهد. هنگام استفاده از چندین گروه، به آن خودتوجه چندسر می‌گویند، و محاسبه می‌تواند به صورت زیر فرموله شود:

$$Multihead(H) = [head^{(1)}; ...; head^{(n_h)}]W_O \quad (15)$$

که n_h تعداد سرهاست، *head* نمایانگر شاخص سر است و الحاق آنها در بعد آخر است. به طور معمول $d_k \times n_h = d$ است، که به این معناست که خروجی $[head(1); ...; head(n)]$ از اندازه R n×d خواهد بود. W_o یک پارامتر یادگیری‌پذیر متناسب با خروجی مد نظر مدل می باشد.

۳-۴- مدل پیشنهادی برای مشابهت یابی معنایی متون

بعد از به دست آوردن بردارهای بازنمایی برای هر جمله با استفاده از روش پیشنهادی، تشابه بین آن‌ها توسط اندازه‌گیری تشابه یا معکوس فاصله در فضای بردار محاسبه می‌شود. معیارهای تشابه معیارهای فاصله هستند که فاصله یا نزدیکی بین دو بردار را تعیین می‌کنند. آشکار است که معیارهای تشابه به صورت معکوس با معیارهای فاصله مرتبط هستند، به این معنی که هرچه تشابه بیشتر باشد، فاصله بین دو بردار کمتر است. انواع مختلفی برای محاسبه فاصله وجود دارد از جمله فاصله اقلیدسی، فاصله منهتن و فاصله مینکوفسکی و غیره [۱۶]. تشابه کسینوسی یکی از پراستفاده‌ترین معیارها برای اندازه‌گیری تشابه معنایی بین بردارها است. در برخی از مقالات مرتبط با تشخیص تشابه معنایی، تشابه کسینوسی به فاصله زاویه‌ای تبدیل‌می شود. می‌توان از تابع آرک‌کسینوس به این منظور استفاده‌کرد. آرک‌کسینوس تشابه کسینوسی را به فاصله زاویه‌ای تبدیل‌می‌کند که به فاصله مثلثی پایبند

اســـت. طبق این رویه، عدم وجود زاویه عملکرد بهتری در تشـــخیص تشـــابه معنایی بین جملات نســـبت به تشـــابه کســـینوســـی دارد [۱۷]. معادله ۱۶ نحوه محاســـبه تشـــابه بین دو بردار u و v با اســـتفاده از آرک‌کسینونس را توضیح‌می‌دهد.

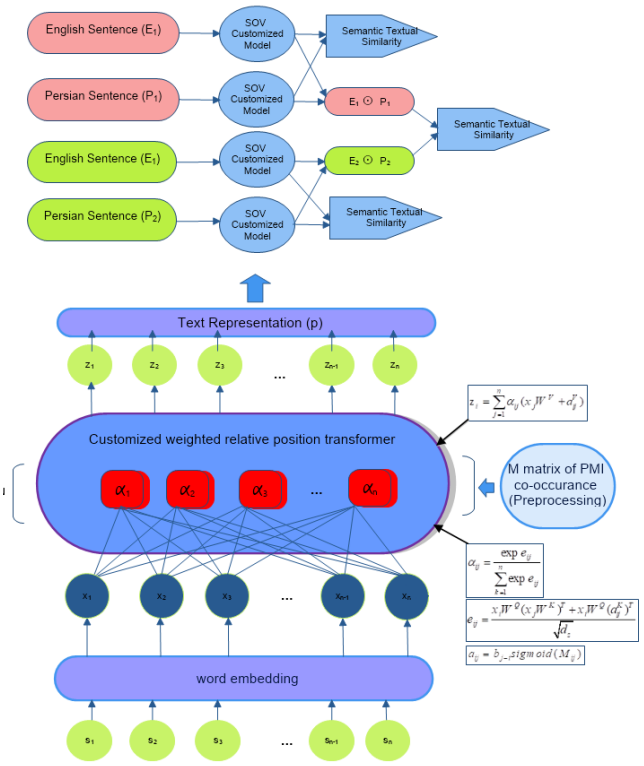
$$Similarity(U,V) = -arccos(\frac{U \cdot V}{||U|| ||V||}) \quad (16)$$

با استفاده از معیارهای مبتنی بر فاصله مانند فاصله اقلیدسی و فاصله منهتن، ما می‌توانیم تشـــابه بین دو بردار را با انتقال معکوس فاصـــله تعیین‌کنیم [۱۸]. همان‌طور که در معادله ۱۷ آمده‌اســـت، فاصـــله اقلید سی کوتاهترین فا صله بین دو بردار را بر ا ساس ق ضیه پیتاگورس محا سبه‌می‌کند. اگر x و y دو بردار تعبیه با بعد p برای جملات با شند، فاصله اقلیدسی و فاصله منهتن بین این دو جمله به‌صورت معادلات ۱۷ و ۱۸ بیان‌می‌شود.

$$D_{euc} = \sqrt{(\sum_{i=1}^{p}(x_i - y_i)^2)} \quad (17)$$

$$D_{man} = \sum_{i=1}^{p}|x_i - y_i| \quad (18)$$

چارچوب مدل ما در شکل 3-1 نشان‌داده‌شده‌است. مدل ما شامل یک ماژول پیش‌پردازش برای م حاســـبه فاکتور PMI، یک لا یه توکن‌ســـازی، یک ماژول کدگذاری ســـفارشـــی با وزن‌دهی PMI به موقعیت نسبتی و یک ماژول تشابه متنی معنایی با یادگیری مشترک است.



شکل 3-1: چارچوب مدل شامل یک ماژول پیش‌پردازش برای محاسبه عامل PMI، یک لایه توکن‌سازی، یک ماژول کدگذاری سفارشی با وزن‌دهی PMI به موقعیت نسبتی و ماژول تشابه متنی معنایی

Sentence 1		Sentence 2		Score
E_1	She won the gold medal in the swimming competition.	E_2	The swimming champion received a gold medal.	0.88
P_1	او مدال طلای مسابقات شنا را به دست آورد.	P_2	قهرمان شنا مدال طلا گرفت.	

جدول 3-1: نمونه‌ها از مجموعه داده E_1 و P_1 به تعبیری معنای یکسانی دارند و همچنین E_2 و P_2. امتیاز تشابه بین $<E_1 , P_1>$ و $<E_2 , P_2>$ برابر با ۰,۸۸ است.

در مجموعه داده ارزیابی تشابه متنی معنایی (STS)، جایی که متون جفت‌شده با یک امتیاز تشابه همراه هستند، یک بردار متناسب برای هر جمله در هر دو زبان اول (فارسی) و دوم (انگلیسی) ایجاد می‌شود. به‌عنوان مثال، در جدول 3-1 ، <E1 , P1> به تعبیری نمونه جملات معادل در انگلیسی و فارسی را در STS-B نشان‌می‌دهد، در حالی که <E2 , P2> نمایانگر نمونه‌ها در سوی دیگر است. ازآنجاکه هر دوی این‌ها از نظر معنایی معادل هستند، انتظارمی‌رود دارای یک امتیاز تشابه بیشینه باشند که برابر با یک است. E_1، P_1، E_2 و P_2 توسط کدگذار

سفارشی پیشنهادی ایجادمی‌شوند. چارچوب یادگیری مشترک ما با استفاده از معادلات (۱۹، ۲۰) بردارهای متناظر را به هم متصل‌می‌کند و آن‌ها را به ماژول تشابه منتقل‌می‌کند.

$$U = E_1 \square P_1 \ , \ V = E_2 \square P_2 \quad (19)$$

$$Similarity(U,V) = -arccos(\frac{U \cdot V}{||U|| ||V||}) \quad (20)$$

بنابراین، با درنظرگرفتن L_1 برای $<P_1,E_1>$، L_2 برای $<P_2,E_2>$ و L_3 برای $<U, V>$ به‌عنوان مؤلفه‌های توابع از دست دادن تشابه معنایی، از معادله (۲۱) برای محاسبه تابع هزینه مشترک استفاده‌می‌شود.

$$L_{joint} = L_1 + L_2 + L_3 \quad (21)$$

۳-۵- الگوریتم یادگیری برای مشابهت‌یابی متون

به منظور آموزش مدل بهینه شـــده از کتابخانه‌های transformers که توســط Hugging Face تحت فریم ورک Pytorch ارائه‌شـــده‌اســت اســتفاده‌می‌کنیم. مدل پیشـــنهادی به شـــیوه مینی-دســـته‌ای (mini-batch) آموزش می‌بیند. الگوریتم آموزش را در اینجا ارائه‌می‌دهیم. ابعاد تعبیه و تعداد سرانه‌ها ورودی‌های مورد نیاز ه ستند. ماتریس مجاورت هم ر خدادی در مرح له پیش‌پردازش م حاســـ به‌می‌شـــود. عمل یات توکن‌گـذاری و مقـداردهی پـارامترهـای یـادگیری پیش از آموزش انجام‌می شوند. در طول آموزش، د سته‌ای از متن‌های جفت انگلی سی و فار سی ترجمه شده از مجموعه داده STS انتخاب شده و وارد رمزگذار ترانسفورمر با موقعیت نسبی وزنی سفارشی می‌شوند تا ماتریس توجه و نماینده‌های خروجی را به‌دست‌آورند. سپس ماژول تشابه امتیاز STS را پیش‌بینی‌می‌کند. ما از امتیاز STS برای محا سبه تابع از هزینه میانگین مربعات (MSE) اســـتفاده‌می‌کنیم. در نهایت، الگوریتم پارامترهای مدل را براساس گرادیان‌های تابع هرینه به‌روزرسانی می‌کند.

Algorithm۱ : Training algorithm for presented model.

Require: preprocessed adjacency matrix of global word co-occurrence M using (1-6)
Require: training sentences set $\boldsymbol{S}$ from STS benchmark contains sentence pairs
Require: embedding dimension d
Require: number of head n_h
Require: initialize embeddings and learnable parameters

for $t = 1, 2, 3, \dots n_{epoch}$ do
  sample a train set $\boldsymbol{S}_{batch}$ of size k
  $Loss \leftarrow 0$
  for $(\boldsymbol{s_1}, \boldsymbol{s_2}, \dots, \boldsymbol{s_k})$ in $\boldsymbol{S}_{batch}$ do
    P_1 , $E_1 \leftarrow$ compute representation of text (9-15)
    P_2 , $E_2 \leftarrow$ compute representation of text (9-15)
    U , $V \leftarrow$ compute concatenated vectors using (19)
    $Score_1 \leftarrow$ compute similarity score of P_1 and E_1 (16)
    $Score_2 \leftarrow$ compute similarity score of P_2 and E_2 (16)
    $Score_3 \leftarrow$ compute similarity score of U and V (16)
    $L(y) \leftarrow$ compute MSE loss L_1 , L_2 and L_3 using (21)
    $Loss \leftarrow Loss + L(y)$
    Update embeddings and learnable parameters
  end
  w.r.t the gradients using $\nabla Loss$
end

Output Similarity Score and Accuracy

۴- یافته‌های پژوهش

۴-۱- داده‌ها برای مشابهت‌یابی معنایی متون

در این مطالعه، ما از مجموعه‌داده مرجع ارزیابی تشابه معنایی متون در زبان فارسی استفاده‌کرده‌ایم. ما این مجموعه‌داده را با ترجمه مجموعه‌داده STS Benchmark انگلیسی (STS-B) از طریق رابط برنامه‌نویسی Google Cloud Translation API ایجاد و نتایج مرجع برای فارسی و انگلیسی فراهم‌کرده‌ایم. برای ارزیابی کیفیت مدل تولیدشده، نیاز به داده‌های برچسب‌گذاری‌شده توسط منبع انسانی داریم. به‌علاوه برای یکدست شدن متن و سادگی پردازش، لازم است حروف ا ضافی حذف‌شوند. علاوه بر آن وجود حروف اضافی در ادامه راه به ما کمکی-نمی‌کند. حروف اضافی شامل » ! * + = - @] [< > / . «است [۱۹]. ازآنجاکه مجموعه‌داده مرجعی برای اندازه‌گیری مشابهت معنایی بین زبان‌های فارسی و انگلیسی وجودندارد، ما از مجموعه‌داده STS Benchmark انگلیسی استفاده‌کردیم. ما یکی از جفت جمله‌های این مجموعه‌داده را به زبان فارسی ترجمه‌کردیم تا امکان ارزیابی مدل ایجادشود. این مجموعه‌داده شامل ۸،۶۲۸ جفت جمله به همراه امتیاز مشابهت معنایی از ۰ (کمترین مشابهت) تا ۵ (بیشترین مشابهت) می‌باشد. این مجموعه به سه بخش تقسیم‌شده‌است: داده‌های آموزش (۷۰٪)، اعتبارسنجی (۱۵٪) و آزمون (۱۵٪)

۴-۲- ارزیابی و مقایسه نتایج در مشابهت یابی معنایی متون

در این زیربخش، ما تجزیه و تحلیل دقیقی از عملکرد مدل ارائه‌شده در وظایف STS تک‌زبانه و چندزبانه ارائه‌می‌دهیم. ما نتایج را با نتایج مدل‌های برجسته دیگر مقایسه‌می‌کنیم تا برتری رویکردمان را نشان-دهیم. ما مدل خود را بر روی مجموعه داده‌های STS-B فارسی-فارسی و انگلیسی-فارسی آموزش‌داده‌ایم و آن را آزمایش‌کرده‌ایم. هر مجموعه داده را به ۵ زیرمجموعه مساوی تقسیم‌کرده و از اعتبارسنجی ۵ تایی استفاده‌کرده‌ایم. مرحله آموزش را ۵ بار جداگانه تکرارکرده‌ایم، هر بار یکی از ۵ زیرمجموعه به‌عنوان مجموعه آزمون استفاده‌شده و ۴ زیرمجموعه باقی‌مانده به‌عنوان مجموعه آموزش ترکیب‌شده‌اند. در تمام آزمایش‌ها، همبستگی پیرسون با شباهت کسینوسی، فاصله اقلیدسی و فاصله منهتن به‌عنوان معیاری برای ارزیابی عملکرد مدل محاسبه‌شده-است.

۴-۳- STS تک‌زبانه (فارسی - فارسی)

نتایج مدل ما بر روی وظیفه STS-B فارسی به فارسی در جدول 4-1 آمده‌است. در STS-B فارسی -فارسی، ما به‌ترتیب ۸۹،۰۹٪ هم‌بستگی پیرسون با شباهت کسینوسی، ۹۱،۵۲٪ هم‌بستگی پیرسون با فاصله

اقلیدسی و ۹۱٫۶۵٪ هم‌بستگی با فاصله منهتن به‌دست‌آوردیم. مدل سفارشی‌سازی‌شده با SOV برترین هم‌بستگی را با فاصله منهتن داشت و این عملکرد از مدل‌های XML-R، DistilBert و M-BERT آموزش‌دیده‌شده بهتر بود. همه مدل‌های مبتنی بر توجه نتایج با امتیاز شباهت کسینوسی بالا را به‌دست‌آوردند، اما در مواردی که باید شباهت بالایی بین جملات وجودنداشته‌باشد، پیش‌بینی‌های ضعیف‌تری داشتند و همبستگی کمتری با امتیازهای واقعی نشان‌دادند. هنگامی که بازنمایی جملات در XML-R را به صورت مستقیم برای شباهت متنی معنایی به‌کاربردیم، تقریباً تمام جفت جملات نمره شباهتی بین ۰٫۶ تا ۱٫۰ کسب‌کردند، حتی اگر برخی از جفت‌ها توسط مترجم به‌عنوان کاملاً مرتبط محسوب‌نشوند. به عبارت دیگر، نمایش‌های جملات مبتنی بر ترنسفورمر تمامی جملات را به‌نحوی فشرده‌کرده‌اند که تقریباً همه جملات به فضای برداری کوچکتری نگاشت می‌شوند و درنتیجه شباهت بالا ایجادمی‌کنند. بنابراین، مناسب نیست که به‌طور مستقیم بازنمایی به‌دست‌آمده از XML-R را برای تطابق معنایی یا بازیابی متن به‌کاربریم. مدل سفارشی پیشنهادی با رمزگذاری موقعیت نسبی وزن‌دار نمایش‌های دقیق‌تری از جملات ایجادکرد و در نتیجه شباهت پیش‌بینی‌شده به شباهت واقعی متناسب بود. روش ما شباهت کسینوسی پیش‌بینی‌شده کمتری را در شباهت واقعی کم پیش‌بینی‌کرد و شباهت پیش‌بینی‌شده بالاتری را برای شباهت واقعی بیشتر ایجادکرد.

جدول 4-1 : نتایج مدل ما در مقایسه با مدل‌های دیگر برای STS

Method	Pearson Correlation with Cosine Similarity	Pearson Correlation with Euclidean distance	Pearson Correlation with Manhattan distance
M-BERT	65.06	63.66	63.65
M-BERT (Fine-tuned model)	73.77	75.34	75.37
DistilBert	66.98	67.31	67.21
DistilBert (Fine-tuned model)	72.63	74.50	75.75
XLM-R	76.57	75.31	78.37
XLM-R (Fine-tuned model)	84.57	84.11	85.68
SOV Customized	89.09	91.52	91.65

تک‌زبانه

۴-۴- نتایج بین‌زبانی (فارسی - انگلیسی)

در زمینه تطابق متنی چندزبانه، نتایج ما بر روی مجموعه داده STS-B انگلیسی به فارسی، که درجدول 4-2 نشان‌داده‌شده‌است، کارآیی مدل ما را نشان‌می‌دهد. مدل سفارشی‌سازی‌شده با SOV بهبودهای قابل‌توجهی را نسبت به مدل‌های دیگر کسب‌کرده‌است و به‌ترتیب ۸۶٫۹۸٪ هم‌بستگی پیرسون با شباهت کسینوسی، ۸۷٫۶۲٪ هم‌بستگی پیرسون با فاصله اقلیدسی و همچنین ۸۸٫۲۹٪ هم‌بستگی پیرسون با فاصله منهتن را فراهم‌کرده‌است. در اینجا نتایج مدل‌های XML-R، DistilBert و M-BERT نشان داده شده و موقعیت مدل ما به عنوان یک راه‌حل مدرن و برتر تثبیت شده است.

کارایی مدل ما در آزمون چندزبانه نشان می‌دهد که این مدل ، بین زبان‌های SOV و SVO مانند فارسی و انگلیسی راه حل مشترک برای شباهت یابی ایجاد می کند. این نتایج نشان دهنده قوت رویکرد ما در گرفتن روابط معنایی پیچیده در میان زبان‌ها است. در مورد چندزبانه فارسی-انگلیسی هنگامی که از بهینه‌سازی با استفاده از مدل BERT چندزبانه (بدون بهینه‌سازی با استفاده از متن موازی) استفاده‌نمی‌کنیم، ضریب هم‌بستگی به ۶۴٫۱۱٪ می‌رسد. با این حال، هنگام استفاده از متن موازی، ضریب هم‌بستگی افزایش‌می‌یابد و با افزایش تعداد جفت جملات فارسی-انگلیسی در متن موازی، ضریب همبستگی پیرسون نیز افزایش‌می‌یابد. به‌عنوان مثال، هنگام استفاده از جفت جملات فارسی-انگلیسی از متن موازی برای مدل BERT چندزبانه بهینه، ضریب هم‌بستگی به ۷۳٫۱۹٪ می‌رسد. با توجه به ویژگی‌های زبانی فارسی، مدل ما فاصله بین بردارهای متنی انگلیسی و فارسی جملات مشابه را کاهش-داده‌است. با کسب ۸۸٫۲۹٪ هم‌بستگی در آزمون چندزبانه، روش ما نسبت به مدل‌های XML-R، DistilBert و M-BERT برتری دارد و نتایج مدرنی را به‌دست‌آورده‌است. علاوه بر این، روش ما با چارچوب یادگیری مشترک، با کسب ۸۹٫۵۱٪ هم‌بستگی در آزمون چندزبانه، بهتر از مدل سفارشی‌سازی SOV بدون یادگیری مشترک، مدل XML-R، DistilBert و M-BERT عمل‌کرده و نتایج برتری را به‌دست آورده-است.

جدول 4-2 : نتایج مدل ما در مقایسه با مدل‌های دیگر برای STS چند زبانه

Method	Pearson Correlation with Cosine Similarity	Pearson Correlation with Euclidean distance	Pearson Correlation with Manhattan distance
M-BERT	63.88	64.03	64.11
M-BERT (Fine-tuned model)	72.61	72.39	73.19
DistilBert	65.74	65.82	66.12
DistilBert (Fine-tuned model)	69.25	69.73	70.08
XLM-R	72.02	72.85	72.28
XLM-R (Fine-tuned model)	82.39	82.35	83.47
SOV Customized	86.98	87.62	88.29

۴-۵- مطالعه فرسایشی برای مشابهت‌یابی معنایی متون

برای اعتبارسنجی بیشتر کارایی مدل ارائه‌شده برای مشابهت‌یابی متون، ما یک مطالعه فرسایشی اجراکردیم. این مطالعه تأثیر اجزای مختلف مدل ما بر عملکرد کلی را بررسی‌می‌کند. همان‌طور که در جدول 4-3 آمده‌است، ما آزمایش‌ها را در دو شرط انجام‌دادیم. در اینجا، مدل پایه

RPE-Transformer نشان‌دهنده مدل ترانسفورمر (مدل XML-R) است که برای هر موقعیت نسبی در داخل یک فاصله قطعی بازنمایی را یادمی‌گیرد. SOV Customized (مناسب برای زبان‌های با ترتیب کلمات SOV) به مدل خودتوجهی اشاره‌دارد که نمایش را برای هر موقعیت نسبی در کل جمله بدون قطع‌کردن یادمی‌گیرد. PMI weighted SOV Customized نشان‌دهنده مدل کامل است که شامل سفارشی‌سازی برای زبان‌های با ترتیب کلمات SOV و تزریق عامل PMI برای همرخدادی کلمات جهانی است. همان‌طور که از نتایج مشخص است، سفارشی‌سازی SOV و همرخدادی کلمات PMI هر دو به نتایج کلی کمک‌می‌کنند و نسبت به مدل پایه، دقت را به‌ترتیب در مجموعه داده STS-B فارسی-فارسی ۵٫۹۷٪ و در مجموعه داده STS-B فارسی-انگلیسی ۴٫۸۲٪ افزایش‌می‌دهند. هنگامی که ساختار مدل تغییرمی‌کند، امتیازهای مشابهت نیز متفاوت هستند. روش ما بهبود قابل‌توجهی در آزمون‌های چندزبانه و تک‌زبانه ایجادمی‌کند. توجیه منطقی این موضوع این است که روش ما، علاوه‌بر فاصله بین فاعل و فعل، بر ارتباط میان این دو تمرکز دارد و PMI عامل وزن‌دار هم رخدادی کلمات سرتاسری را در بازنمایی کلمات درون جمله اعمال-

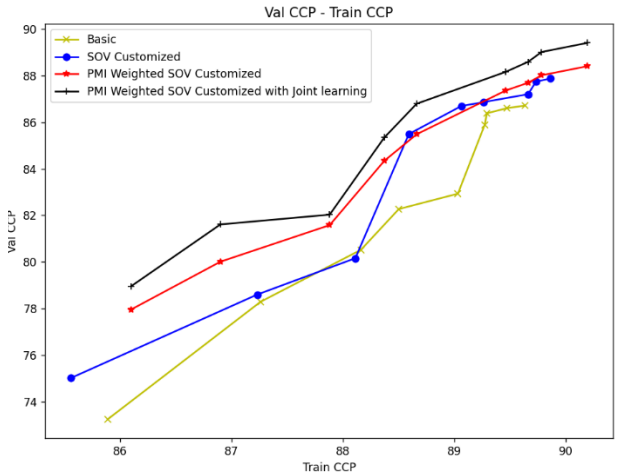
Model	STS-B Persian-English	STS-B Persian-Persian
Relative position encoding with clipping (Basic XML-R model)	83.47	85.68
Relative position encoding without clipping (SOV customized)	87.91	89.01
PMI weighted relative position encoding without clipping (PMI weighted SOV Customized)	88.29	91.65
PMI weighted relative position encoding without clipping with joint learning (PMI weighted SOV Customized with Joint Learning)	89.51	92.47

می‌کند. بر اساس این، ما می‌فهمیم که مدل پیشنهادی بازنمایی برداری بهتری را ایجادکرده و به عبارتی دقت STS بیشتری به‌دست‌آورده‌است.

جدول 4-3 : مطالعه فرسایشی برای STS

۴-۶- قابلیت تعمیم در مشابهت‌یابی معنایی متون

با توجه به مجموعه داده STS-B فارسی-انگلیسی، شکل 4-1 نمودار هم‌بستگی اعتبار-آموزش را تحت سه شرط نشان‌می‌دهد: مدل پایه (نمودار زرد)، سفارشی‌سازی SOV (نمودار آبی)، سفارشی‌سازی SOV با وزن PMI (نمودار قرمز) و سفارشی‌سازی SOV با یادگیری مشترک وزن‌دار PMI (نمودار سیاه).CCP معیار پیرسون را نشان‌می‌دهد و CCP آموزش نشان‌دهنده عملکرد مدل در مجموعه اعتباردهی و مجموعه آموزشی در طول فرآیند آموزش است.

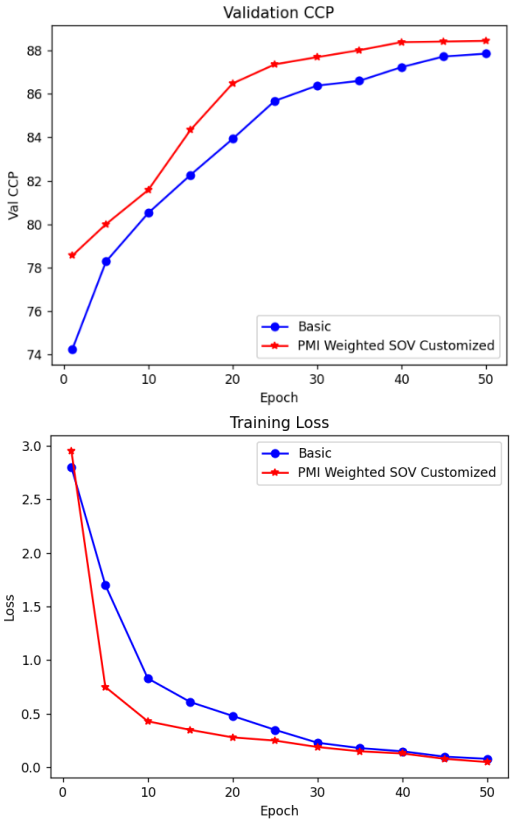


شکل 4-1 : نمودار همبستگی Validation-Train در یادگیری اشتراکی و شروط دیگر

نمودارهای مدل پایه (زرد) و سفارشی‌سازی SOV (آبی) در شکل 4-1 نسبتاً نزدیک به یکدیگر هستند، که نشان‌می‌دهد دارای توانایی تعمیم مشابهی هستند. موقعیت نمودار سیاه در بالایی نمودارهای دیگر است، که نشان‌می‌دهد که CCP داده‌های تست با یادگیری مشترک بالاتر از دقت به‌دست‌آمده در آموزشی یکسان است. بنابراین، می‌توانیم نتیجه-بگیریم که مدل ارائه‌شده دارای توانایی تعمیم بهتری است. علاوه بر این، با توجه به گوشه بالا و راست ، واضح است که مدل پایه، سفارشی‌سازی SOV و سفارشی‌سازی SOV با وزن PMI به موقعیت‌های بالاتری دست‌یافته‌اند. این نشان‌می‌دهد که سطح آموزش مدل در این چهار حالت عمیق‌تر شده‌است.

۴-۷- تاثیر مدل پیشنهادی در فرایند یادگیری برای مشابهت-یابی متون

اکنون تأثیر کدگذاری موقعیت نسبی با وزن PMI بدون کوتاه‌نمودن را با ارائه شاخص‌ها در فرآیند آموزش مدل مشابهت‌یابی، بررسی‌می‌کنیم. شکل 4-2 تغییرات از دست‌رفتن آموزش و هم‌بستگی (دقت) مجموعه اعتبار یا تست (دقت) با افزایش دوره آموزش را نشان‌می‌دهد. ما ۵۰ دوره اول را ثبت‌کرده‌ایم تا وضعیت در طول آموزش مشاهده‌شود. نمودار آبی شرایط پایه را نمایش‌می‌دهد و نمودار قرمز مدل ما را نشان‌می‌دهد. شکل 4-2 (a) نشان‌می‌دهد که خطای آموزش ما پایین‌تر است و سرعت همگرایی در طول آموزش سریع‌تر است، به ویژه در ۳۰ دوره اول. و مقادیر نهایی خطا هر دو نزدیک به ۰٫۰۳ هستند زیرا در آن زمان هر دو به بیش برازش دچار شده‌اند. از شکل 4-2 (b) مشخص است که هم‌بستگی مجموعه اعتبار آموزش ما سریع‌تر افزایش یافته‌است و مقدار نهایی آن بالاتر است. این نشان‌می‌دهد که کدگذاری موقعیت نسبی با وزن PMI بدون برش اثر مهاری بر روی بیش برازش تابع هزینه دارد.



شکل 4-2 : (a) همبستگی مجموعه اعتبارسنجی (b) هزینه آموزش. شاخص‌های فرآیند آموزش با یا بدون سفارشی‌سازی SOV با وزن PMI با یادگیری مشترک

۵- نتیجه‌گیری

ما یک رویکرد یادگیری مشترک با استفاده از مدل‌های خودتوجه بهبودیافته برای مقابله با چالش‌های STS در ساختارهای زبانی SOV و SVO معرفی‌کردیم. ابتدا مجموعه داده چندزبانه جامعی با داده‌های موازی برای زبان‌های SOV و SVO ایجاد کردیم تا تنوع زبانی را تضمین‌کنیم. از رمزگذاری موقعیت نسبی وزن‌دار و غنی‌شده با تزریق اطلاعات هم‌رخدادی از طریق اندازه‌گیری اطلاعات مشترک نقطه‌ای (PMI) برای تقویت بازنمایی متنی استفاده‌کردیم. علاوه‌بر این، از یک چارچوب یادگیری مشترک استفاده‌کردیم که نمونه‌های مشترک بین زبان‌ها را برای بهبود STS بین زبانی به‌کارمی‌برد. با آموزش همزمان بر روی چندین جفت زبان، مدل ما توانایی انتقال دانش را به‌دست‌آورد. مدل پیشنهادی ما بر مجموعه داده‌های STS-Benchmarks فارسی-انگلیسی و فارسی-فارسی ارزیابی‌شد و به‌ترتیب به ضریب همبستگی پیرسون به مقدار ۸۸,۲۹٪ و ۹۱,۶۵٪ دست‌یافت.

مراجع

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," presented at the Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019.

[2] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," 2018.

[3] E. Agirre, D. Cer, M. Diab, and A. Gonzalez-Agirre, "Semeval-2012 task 6: A pilot on semantic textual similarity," in ** SEM 2012: 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 (SemEval 2012)*, 2012, pp. 385-393.

[4] A. Islam and D. Inkpen, "Semantic text similarity using corpus-based word similarity and string similarity," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 2, no. 2, pp. 1-25, 2008.

[5] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," *arXiv preprint arXiv:1910.01108*, 2019.

[6] X. Tang *et al.*, "Improving multilingual semantic textual similarity with shared sentence encoder for low-resource languages," *arXiv preprint arXiv:1810.08740*, 2018.

[7] T. Brychcín, "Linear transformations for cross-lingual semantic textual similarity," *Knowledge-Based Systems*, vol. 187, p. 104819, 2020.

[8] Y. Sever and G. Ercan, "Evaluating cross-lingual textual similarity on dictionary alignment problem," *Language Resources and Evaluation*, vol. 54, pp. 1059-1078, 2020.

[9] T. Pires, E. Schlinger, and D. Garrette, "How multilingual is multilingual BERT?," *arXiv preprint arXiv:1906.01502*, 2019.

[10] T. Kudo and J. Richardson, "Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing," *arXiv preprint arXiv:1808.06226*, 2018.

[11] K. Church and P. Hanks, "Word association norms, mutual information, and lexicography," *Computational linguistics*, vol. 16, no. 1, pp. 22-29, 1990.

[12] J. A. Bullinaria and J. P. Levy, "Extracting semantic representations from word co-occurrence statistics: A computational study," *Behavior research methods*, vol. 39, no. 3, pp. 510-526, 2007.

[13] D. Kiela and S. Clark, "A systematic study of semantic vector space model parameters," presented at the Proceedings of the 2nd Workshop on Continuous Vector Space Models and their Compositionality (CVSC), 2014.

[14] Y. Liu *et al.*, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[15] P. Shaw, J. Uszkoreit, and A. Vaswani, "Self-attention with relative position representations," *arXiv preprint arXiv:1803.02155*, 2018.

[16] A. Singh, A. Yadav, and A. Rana, "K-means with Three different Distance Metrics," *International Journal of Computer Applications*, vol. 67, no. 10, 2013.

[17] D. Cer *et al.*, "Universal sentence encoder for English," in *Proceedings of the 2018 conference on empirical*

methods in natural language processing: system demonstrations, 2018, pp. 169-174.

[18] l. naderloo and M. Tahghighi Sharabyan, "Presenting a model for Multi-layer Dynamic Social Networks to discover Influential Groups based on a combination of Developing Frog-Leaping Algorithm and C-means Clustering," *Intelligent Multimedia Processing and Communication Systems (IMPCS)*, vol. 3, no. 3, pp. 29-39, 2022.

[19] L. Gonbadi and N. Ranjbar, "Sentiment Analysis of People's opinion about Iranian National Cars with BERT," *Intelligent Multimedia Processing and Communication Systems (IMPCS)*, vol. 3, no. 4, pp. 51-60, 2022.