

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

Ibrahim Ganjalipour<sup>1</sup>, Amir Hossein Refahi<sup>2\*</sup>, Sohrab Kordrostami<sup>3</sup>, Ali Asghar Hosseinzadeh<sup>4</sup>

1. PhD Student, Department of Applied Mathematics & Computer Science, Lahijan Branch, Islamic Azad University, Lahijan, Iran. [ibrahim.ganjali@gmail.com](mailto: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](mailto:ah_refahi@yahoo.com)
3. Full Professor, Department of Applied Mathematics & Computer Science, Lahijan Branch, Islamic Azad University, Lahijan, Iran. [sohrabkordrostami@gmail.com](mailto:sohrabkordrostami@gmail.com)
4. Assistant professor, Department of Applied Mathematics & Computer Science, Lahijan Branch, Islamic Azad University, Lahijan, Iran. [hosseinzadeh\\_ali@yahoo.com](mailto: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.

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

دوره چهارم، زمستان ۱۴۰۲  
شماره چهارم، صص: ۱۱-۲۳

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

ابراهیم گنجعلی پور<sup>۱</sup>، امیرحسین رفاهی شیخانی<sup>۲\*</sup>، سهراب کردروستی<sup>۳</sup>، علی اصغر حسین زاده<sup>۴</sup>

۱. دانشجوی دکتری، دانشکده ریاضی کاربردی و علوم کامپیوتر، واحد لاهیجان، دانشگاه آزاد اسلامی، لاهیجان، ایران.

[ibrahim.ganjali@gmail.com](mailto:ibrahim.ganjali@gmail.com)

۲. دانشیار، دانشکده ریاضی کاربردی و علوم کامپیوتر، واحد لاهیجان، دانشگاه آزاد اسلامی، لاهیجان، ایران.. (نویسنده مسئول)

[ah\\_refahi@yahoo.com](mailto:ah_refahi@yahoo.com)

۳. استاد، دانشکده ریاضی کاربردی و علوم کامپیوتر، واحد لاهیجان، دانشگاه آزاد اسلامی، لاهیجان، ایران.

[sohrabkordrostami@gmail.com](mailto:sohrabkordrostami@gmail.com)

۴. استادیار، دانشکده ریاضی کاربردی و علوم کامپیوتر، واحد لاهیجان، دانشگاه آزاد اسلامی، لاهیجان، ایران.

[hosseinzadeh\\_ali@yahoo.com](mailto: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 معرفی شده است. ابتدا یک مجموعه داده چندزبانه جامع با داده‌های موازی برای زبان‌های SOV و SVO را ایجاد کرده و تنوع زبانی گسترده‌ای را تضمین می‌کنیم. ما یک مدل خودتوجه بهبودیافته با رمزگذاری نسبی موقعیت وزن‌دار جدید غنی‌شده با تزریق اطلاعات هم‌رخدادی از طریق عوامل اطلاعات مشترک نقطه‌ای (PMI) معرفی می‌کنیم. علاوه بر این، ما از یک چارچوب یادگیری مشترک استفاده می‌کنیم که نمونه‌های مشترک بین زبان‌ها را به منظور بهبود STS بین زبانی استفاده می‌کند. با آموزش همزمان در چندین جفت زبان، مدل ما توانایی انتقال دانش را به دست می‌آورد و به طور مؤثر پل ارتباطی بین زبان‌های با ساختارهای متفاوت SOV و SVO ایجاد می‌کند. مدل پیشنهادی ما بر روی مجموعه داده‌های STS- Benchmarks فارسی-انگلیسی و فارسی-فارسی ارزیابی شد و به ترتیب به ضریب همبستگی پیرسون ۰.۸۸۲۹٪ و ۰.۹۱۶۵٪ دست یافت. آزمایش‌های انجام شده نشان می‌دهد که مدل پیشنهادی ما در مقایسه با مدل‌های دیگر عملکرد بهتری داشته است. مطالعه کاهشی نیز نشان می‌دهد که سیستم ما قادر به همگرایی سریعتر است و کمتر مستعد بیش‌برازش 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 الهام می‌گیرند و قابلیت تمرکز بر بخش‌های مهم و معنادار متون را دارا می‌باشند. شبکه‌های عصبی مبتنی بر توجه، الهام گرفته 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 ظاهر شده است. مدل‌های STS کاربردی در شناسایی تقلب، ترجمه ماشینی و بازیابی اطلاعات، به ما امکان می‌دهد تا شباهت معنایی بین دو تکه متن را اندازه‌گیری کنیم. در حالی که STS پیشرفت‌های قابل توجهی داشته است، اما در زبان‌ها با ساختارهای نحوی متفاوت و منابع دیجیتال محدود چالش جدی باقی مانده است. یکی از جنبه‌های تنوع زبانی در تفاوت ترتیب کلمات در جملات است. برخی از زبان‌ها به ترتیب کلمات فاعل-مفعول-فعل (SOV) پایبند هستند، در حالی که دیگران الگوهای فاعل-فعل-مفعول (SVO) را دنبال می‌کنند. این تفاوت‌های ساختاری، در کنار عواملی مانند ضمیر حذف شده، وظیفه اندازه‌گیری STS در این زبان‌ها را پیچیده می‌کند. در اینجا، زبان فارسی با ترتیب کلمات SOV و ویژگی‌های زبانی منحصر به فرد خود، به‌عنوان یک مورد برای مطالعه STS چندزبانه در نظر گرفته می‌شود.

محققان از مدل‌های مبتنی بر ویژگی‌های سنتی تا معماری‌های پیچیده شبکه‌های عصبی پیش‌آموزش شده مانند [1] BERT، [2] GPT و نسخه‌های چندزبانه مانند XLM-R را برای محاسبه تشابه معنایی بررسی کرده‌اند. نقطه تحول مهمی در این زمینه توسعه مدل‌های زبان پیش‌آموزش شده مانند BERT، GPT و نسخه‌های چندزبانه‌شان مانند XLM-R بوده است که روش تشخیص تشابه معنایی متن را تغییر دادند. علارغم انبوه تحقیقات در زمینه تشابه متنی در زبان انگلیسی، کمبود تحقیقات در تحلیل تشابه متنی در زبان فارسی قابل مشاهده است.

در اینجا، یک رویکرد یادگیری ترکیبی نوین برای پرداختن به چالش‌های STS در هر دو ساختار زبان SOV و SVO معرفی می‌شود. 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 نه تنها نمایندگی گسترده‌ای از تنوع زبانی را تضمین می‌کند، بلکه همچنین اساس قابلیت‌های چندزبانه مدل ما است. در اساس رویکرد ما، مدل خودتوجه پیشرفته‌ای قرار دارد که طراحی شده است تا ارتباط‌های پیچیده داخلی و بین جملات و همچنین اطلاعات همزمانی را از طریق گراف مرجع نسبی سطح جمله عامل اطلاعات مشترک نقطه‌ای (PMI) در نظر بگیرد. اضافه کردن کدگذاری موقعیت‌های رابطی وزن‌دار، مدل ما را قادر به در نظر گرفتن اهمیت موقعیت کلمات و روابط متنی آن‌ها می‌کند و اندازه‌گیری دقیق‌تری از STS را ممکن می‌سازد. برای آموزش موازی گونه‌های زبانی SOV و SVO از یک چارچوب یادگیری مشترک استفاده می‌کنیم. این چارچوب انتقال دانش را از طریق همزمان آموزش در چندین جفت نمونه از مجموعه داده فراهم می‌کند. این به مدل ما امکان می‌دهد تا با استفاده از تمثیل‌های مشترک در زبان‌ها، دید چندزبانه‌ای ارزشمندی از تشابه متنی ارائه دهد. 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 در برمی‌گیرد. با استخراج ویژگی‌های معنایی، رویکرد ما دقت اندازه‌گیری تشابه معنایی را افزایش می‌دهد. در ارزیابی‌های دقیق در مجموعه داده‌های STS فارسی-انگلیسی و STS فارسی-فارسی، مدل پیشنهادی ما عملکرد قابل توجهی را نشان می‌دهد و بهبودهای معناداری نسبت به روش‌های موجود کسب می‌کند. علاوه بر این، مطالعه موردی و مطالعه فرسایشی نشان می‌دهد که مدل ارائه شده کارایی، پایداری و قابلیت تعمیم بالایی دارد.

به‌طور خلاصه در این تحقیق به موارد زیر پرداخته ایم:

- ما یک مدل مبتنی بر ترانسفورمر جدید برای تشخیص موجودیت نام‌دار فارسی معرفی می‌کنیم که از مدل پیش‌آموزش شده XLM-R و گراف مرجع نسبی سطح جمله عامل اطلاعات مشترک نقطه‌ای (PMI) استفاده می‌کند
- ما یک مدل مبتنی بر ترانسفورمر جدید برای استخراج رابطه فارسی معرفی می‌کنیم که از مدل پیش‌آموزش شده XLM-R و گراف مرجع نسبی سطح جمله عامل اطلاعات مشترک نقطه‌ای (PMI) استفاده می‌کند.
- ما یک رویکرد یادگیری ترکیبی با مدل خودتوجه بهبود یافته برای پرداختن به چالش‌های STS در ساختارهای زبانی SOV و SVO معرفی می‌کنیم.
- ما کدگذار ترانسفورمر را با در نظر گرفتن ویژگی‌های زبانی قابلیت حذف ضمیر و ترتیب کلمات فاعل-مفعول-فعل در زبان فارسی، اصلاح می‌کنیم.
- نتایج آزمایش‌ها، عملکرد بهتر روش ارائه شده در مقایسه با کارهای پیشین در انجام وظایف پردازش زبانی مورد نظر را نشان می‌دهد.

## ۲- پیشینه پژوهش

در این بخش مروری بر کارهای پیشین در تشابه‌یابی معنایی متون ارائه می‌شود. در اینجا ما تلاش می‌کنیم که کارهای قبلی مرتبط با روش پیشنهادی را به منظور ایجاد ارتباط بین دانش موجود و ارتقاءهای پژوهش حاضر ارائه کنیم.

تعیین دقیق مشابهت معنایی متون (STS)، به مدت سال‌ها یک شاخه مهم در تحقیقات پردازش زبان طبیعی بوده است. محققان از تکنیک‌ها و رویکردهای مختلفی برای انجام این وظیفه چالش‌برانگیز در ساختارهای زبانی متنوع استفاده کرده‌اند. 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 قرار می‌دهیم. 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 شده است [۴]. این محاسبات عمدتاً توسط مقدار عددی از ۰ تا ۵ نمایش داده می‌شود برای هر جفت متن یا جمله. ابتدایی‌ترین ایده‌ها برای شناسایی مشابهت معنایی بین دو جمله بر مبنای تطابق معنایی بین کلمات در جملات بوده است، 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 روش‌ها، جملات به بردارهای عددی با ابعاد مختلف تبدیل می‌شوند. ایجاد بردارهای تعبیه معمولاً با استفاده از متون بزرگ انجام می‌شود. در زبان انگلیسی به دلیل استفاده گسترده و دسترسی به متون بزرگ، تحقیقات بیشتری در این حوزه انجام شده است. اما برای زبان‌های با منابع و متون محدودتر مانند فارسی، تحقیقات در این حوزه به نسبت کمتر بوده است.

در اینجا ما تلاش می‌کنیم که کارهای قبلی مرتبط با روش پیشنهادی را به منظور ایجاد ارتباط بین دانش موجود و ارتقاءهای ما خلاصه کنیم. ما تعدادی از مطالعات تحقیقاتی برجسته در زمینه مشابهت معنایی چندزبانه و یکزبانه معرفی می‌کنیم. مدل Multilingual BERT [۱] یک مدل زبان مبتنی بر ترانسفورمر است که پیش‌آموزش بر روی یک مجموعه داده چندزبانه بزرگ انجام شده است. ثابت شده است که M-BERT عملکرد بسیار خوبی را در وظایف مختلف پردازش زبان چندزبانه، شامل مشابهت معنایی چندزبانه، دارد. توانایی درک و تولید بردارهای بازنمایی برای چندین زبان آن را به یک ابزار مهم برای کاربردهای چندزبانه می‌کند.

یک نسخه دیگر از BERT به نام DistilBert [۵] از دانش آنلاین هنگام پیش‌آموزش بهره‌می‌برد و نشان می‌دهد که می‌توان اندازه یک مدل BERT را ۴۰٪ کاهش داد، در حالی که ۹۷٪ از توانایی‌های درک زبان آن حفظ شود و سرعت آن ۶۰٪ افزایش یابد.

مدل XLM-R یک افزونه از Multilingual BERT است که مدل‌سازی چندزبانه را بهبود می‌بخشد. این مدل بر روی حجم زیادی از داده‌های ۱۰۰ زبان پیش‌آموزش دیده است و بهترین نتایج را در مجموعه گسترده‌ای از وظایف پردازش زبان چندزبانه، شامل مشابهت معنایی چندزبانه، به دست آورده است. کارایی این مدل در انجام وظایف زبان‌های کم‌منبع آن را یک گزینه برجسته برای تحقیقات چندزبانه می‌کند.

برای زبان‌های کم‌منبع مانند اسپانیایی، عربی، اندونزیایی و تایلندی، تانگ و همکاران در سال ۲۰۱۸ [۶] یک مدل چندزبانه معرفی کردند. آن‌ها یک چارچوب مدل مشابهت معنایی یکزبانه را به یک تنظیم چندزبانه گسترش دادند و نشان دادند که با استفاده از یک کدگذار چندزبانه مشترک، هر جمله می‌تواند نمایش‌های مختلفی را نشان دهد که به زبان هدف وابسته هستند.

بریکچین در [۷] ایده‌هایی معرفی کرد که فضاهای معنایی چندزبانه در یک فضای مشترک با استفاده از لغتنامه‌های دوزبانه تبدیل می‌شوند. آن‌ها از روش‌های بدون نظارت برای محاسبه مشابهت جملات تنها بر مبنای تعبیه‌های معنایی بهره‌بردند. آن‌ها نشان دادند که بهبود فضاهای مشترک معنایی از طریق وزن‌دهی به کلمات می‌تواند به نتایج کمک کند. یافته‌های آن‌ها نشان داد که ضریب همبستگی پیرسون ۰٫۸۶۱ درصد در جفت جملات عربی-انگلیسی دارد.

در [۸] تعاریف Wordnet به ۷ زبان مختلف برای ایجاد یک زمینه آزمون مشابهت معنایی متنی چندزبانه به کار می‌رود. یک وظیفه تطابق اسناد برای استفاده بین نمادهای توصیفی Wordnet در ۷ زبان مختلف ایجاد شده است. روش‌های مشابهت متنی بدون نظارت مانند فاصله واسترشتاین، فاصله سینکهورن و مشابهت کسینوس با یک مدل عمیق آموزش داده شده Siamese مقایسه شده‌اند. این وظیفه به عنوان یک وظیفه بازیابی و وظیفه تطابق مدل می‌شود تا اثر توابع مشابهت معنایی را بررسی کند و نتایج نشان دادند که مدنظر گرفتن مسأله به عنوان یک مسأله بازیابی و تطابق تأثیر مخربی بر روی نتایج دارد.

پیرس و همکاران [۹] مطالعاتی را در خصوص کیفیت Multilingual BERT برای وظایف چندزبانه انجام دادند. آن‌ها آزمایش‌های مختلفی را روی مجموعه داده‌های متنوع با استفاده از مدل Multilingual BERT انجام دادند و نتایج مطلوبی را به دست آوردند. در برخی از آزمایش‌های انجام شده 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 دنبال می‌کنند، که ساختار جمله‌ای معمولاً شامل ترتیب کلمات فاعل-فعل-مفعول است. در تضاد با این موضوع، زبان‌های مانند فارسی ساختار SOV (فاعل-مفعول-فعل) را دنبال می‌کنند، 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 منحصر به فردی اقدام می‌کند. با بهره‌گیری از این نوآوری‌ها، ما قصد داریم برای STS چندزبانه با پیچیدگی‌های خاصی که تنوع زبانی و تغییرات ترتیب کلمات ایجاد می‌کنند یک راهکار نوین ارائه کنیم.

### ۳- روش پیشنهادی

#### ۳-۱- ویژگی‌های مؤثر 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 حذف پذیری 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 به عنوان ورودی نپذیرد و آن را قطع کند (فاعل و فعل در یک قالب پردازشی نیستند)، بردار تعبیه خروجی مفهوم کلی دقیق را شامل نخواهد شد.

#### ۳-۲- توکن سازی

برای توکن سازی، ما از مدل پیش‌آموز XLM-R [۵] استفاده می‌کنیم و همچنین از بردارهای تعبیه کلمات تولید شده برای بردارهای ورودی مدل سفارشی خود نیز استفاده می‌کنیم. XLM-R در مقایسه با BERT چندزبانه (M-BERT) [۴] در مجموعه متنوعی از معیارهای میان‌زبانی بهتر عمل می‌کند، 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) آموزش داده شده است. این مدل بهبود قابل توجهی نسبت به مدل‌های چندزبانه منتشر شده قبلی مانند Multilingual BERT را در وظایف پسینی مانند طبقه‌بندی، برچسب‌گذاری دنباله‌ها و پاسخ به سوالات ارائه می‌دهد. XLM-R از روش Sentence-piece [۱۰] برای توکن سازی زیرکلمه‌ای استفاده می‌کند که به خصوص در زبان‌های با منابع کم مانند فارسی عملکرد خوبی دارد.

پس از فرآیند توکن سازی، ما یک گراف ناهمگون در سطح جمله از توکن‌ها ایجاد می‌کنیم. خروجی فرآیند توکن سازی مجموعه‌ای از توکن‌های زیرکلمه‌ای با شناسه‌های متناظر است. برای در نظر گرفتن هم‌رخدادی توکن‌ها در متن یا مجموعه داده، ابتدا توکن‌های با فراوانی بالا و کلمات توقف (مثل "!", "!", "ان", "از", "به", "...") حذف می‌کنیم 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 لغت نامه متنی را نمایندگی می‌کنند. 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 ایجاد می‌شود، و وزن یال‌ها توسط اطلاعات مشترک نقطه‌ای (PMI) اندازه‌گیری می‌شود. همان‌طور که در بخش ۳٫۱ توضیح داده شده است، فاعل (در ابتدای جمله) و فعل (در انتهای جمله) در زبان با ترتیب SOV معمولاً دور از یکدیگر قرار دارند. ما می‌خواهیم هم‌رخدادی آن‌ها را با عامل PMI در نظر بگیریم. بنابراین، هنگامی که پنجره کشویی در ابتدای جمله است، همچنین هم‌رخدادی با آخرین کلمات و برعکس را اندازه‌گیری می‌کنیم. به عنوان مثال، اگر طول پنجره ۴ باشد، برای کلمه اول، هم‌رخدادی با سه کلمه آخر نیز اندازه‌گیری می‌شود، و برای کلمه آخر، هم‌رخدادی با سه کلمه اول اندازه‌گیری می‌شود. برای اندازه‌گیری هم‌رخدادی کلمات، ما از اطلاعات مشترک نقطه‌ای استفاده می‌کنیم. PMI یک معیار ارتباطی بین یک جفت نتایج گسسته  $x$  و  $y$  است، که به صورت زیر تعریف می‌شود:

$$PMI(x, y) = \log \left( \frac{P(x, y)}{P(x)P(y)} \right) \quad (1)$$

ما یک مجموعه کلمات  $w \in VW$  و متناظر با آن‌ها متن  $c \in VC$  را فرض می‌کنیم، جایی که  $VW$  و  $VC$  واژگان کلمه و متن هستند. کلمات از یک مجموعه متنی از کلمات  $w_1, w_2, \dots, w_n$  می‌آیند و متون مرتبط با کلمه  $w_i$  کلماتی هستند که در اطراف آن در یک پنجره  $L$  اندازه‌گیری می‌شوند:  $w_{i-L}, \dots, w_{i-1}, w_{i+1}, \dots, w_{i+L}$ . ما مجموعه جفت‌های کلمات و متن مشاهده شده را با  $D$  نمایش می‌دهیم.  $(w, c) \in D$  را برای تعداد باری که جفت  $(w, c)$  در  $D$  ظاهر می‌شود، استفاده می‌کنیم. به همین ترتیب،  $\#(w)$  و  $\#(c)$  تعداد دفعاتی هستند که  $w$  و  $c$  در  $D$  ظاهر شده‌اند.

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

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

$PMI(w, c)$  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 به تنهایی ظاهر می‌شوند) اندازه‌گیری می‌کند. مقدار PMI می‌تواند به صورت تجربی با در نظر گرفتن تعداد واقعی مشاهدات در یک متن تخمین زده شود:

$$PMI(w, c) = \log \left( \frac{\#(w, c) \cdot |D|}{\#(w) \cdot \#(c)} \right) \quad (4)$$

استفاده از PMI به عنوان یک اندازه ارتباطی در پردازش زبان طبیعی توسط چرچ و هنکس [۱۱] معرفی شد و به عنوان وظیفه‌های شباهت واژه‌ها به طور گسترده مورد قبول قرار گرفت. کار با ماتریس PMI چالش‌های محاسباتی متعددی را ایجاد می‌کند. 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 مسئله عملی اساسی به دلیل ابعاد بزرگ آن است. برای حل این مسئله، ما در موارد  $\#(w, c) = 0$  مقدار  $PMI(w, c) = 0$  تعیین می‌کنیم و ماتریس پراکنده به دست می‌آوریم. توجه می‌کنیم که

این ماتریس ناسازگار است، به معنای آن که 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 ظاهر می‌شوند، شواهد قوی این موضوع را نشان می‌دهد که این کلمات تمایل دارند که با هم ظاهر نشوند، 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 ظاهر نمی‌شوند، مقدار ۰ خواهند داشت. جایگزینی همه مقادیر منفی توسط ۰: (PPMI) مقدار PMI مثبت به معنای هم‌بستگی زیاد کلمات در یک متن است، درحالی‌که مقدار PMI منفی به عدم هم‌بستگی یا کم هم‌بستگی در متن اشاره دارد. بنابراین، ما تنها یال‌ها را بین جفت‌های کلمه با مقادیر مثبت PMI اضافه می‌کنیم.

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

هنگام بازنمایی کلمات، برخی از استدلال‌ها وجود دارد 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) استخراج کنند. این نشان می‌دهد که تشابه معنایی دو کلمه بیشتر تحت تأثیر محیط‌های مثبتی که به اشتراک می‌گذارند است تا محیط‌های منفی که به اشتراک می‌گذارند. به همین دلیل از محیط‌های منفی صرف‌نظر کردن و آن‌ها را به‌عنوان بی‌اطلاع (۰) نشان دادن نیز معقول به نظر می‌آید. به‌واقع، نشان داده شده است که معیار PPMI عملکرد خوبی در وظایف تشابه معنایی ارائه می‌دهد. مقایسه معیارهای ارتباط کلمه-متن نشان می‌دهد که PMI و به‌خصوص PPMI برترین نتایج را برای یک مجموعه گسترده از وظایف تشابه کلمه ارائه می‌دهند. برای بهره‌برداری از اطلاعات کلی هم‌خدادی 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 استفاده می‌کنیم تا آمار تطابق را جمع‌آوری کنیم. بعد از محاسبه PMI در تمام مجموعه داده، ماتریس مجاورت را برای هر جمله ایجاد می‌کنیم [۱۲، ۱۳].

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

$M$  ماتریس مجاورت کلمات هر جمله است و  $(w_j, w_i)$  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 در نظر می‌گیریم. اول، با در نظر گرفتن تطابق کلان ضمیر فاعل و فعل، شبکه آموزش می‌بیند تا نمایش فعل را از جملاتی که ضمیر حذف-

نشده است برای جملاتی که ضمیر حذف شده است، بیاموزد. ما متوجه می‌شویم که با وارد کردن 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 بهبود می‌بخشد. دوم، با بهره‌گیری از تطابق کلمات کلی وارد شده، مدل ارائه شده به دقت بالایی دست‌یافت. سوم، ما تنها تطابق کلمات در جمله را در نظر گرفته‌ایم، بلکه تطابق در کل مجموعه داده را نیز در نظر گرفته‌ایم. بنابراین، مدل، توانایی تعمیم بالاتری برای به‌دست آوردن عملکرد پایدارتر دارد. در بخش نتایج و آزمایش‌ها، ما عملکرد تعمیم مدل و اثر وارد کردن تطابق کلمات کلی PMI در تشخیص تشابه معنایی و تطابق فاعل و فعل را تحلیل می‌کنیم. مقاله برای حل کاستی‌های موجود در پژوهش‌های قبلی ذکر شود. از ذکر مراجع به‌صورت گروهی و بدون تحلیل محتوای آن‌ها خودداری شود.

### 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 استفاده می‌کند. در این بخش خواص ترانسفورمر را تجزیه و تحلیل می‌کنیم و دو بهبود خاص برای محاسبه بردار بازنمایی پیشنهاد می‌کنیم.

اولین بهبود این است که با توجه به مشخصات زبان فارسی (توضیح داده شده در بخش ۳-۱)، ما تصمیم گرفتیم تا حداکثر مکان نسبی موجود را در اساس طول دنباله‌ها در مجموعه داده آموزشی در نظر بگیریم. کدگذاری موقعیت نسبی [۱۵] فرض کرده است 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 استفاده نمی‌کنیم و وزن موقعیت نسبی بین تمام توکن‌ها را در نظر می‌گیریم.

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

کدگذار ترانسفورمر ورودی  $X \in \mathbb{R}^{n \times d}$  را می‌پذیرد، جایی که  $n$  طول دنباله و  $d$  ابعاد بردار تعبیه ورودی است. ماتریس ورودی از فرآیند توکن‌سازی بخش ۲-۳ می‌آید. سپس سه ماتریس یادگیری‌پذیر  $W_q$ ،  $W_v$ ،  $W_k$  برای تبدیل کردن  $X$  به فضاهای مختلف استفاده می‌شوند. معمولاً ابعاد سه ماتریس همه  $d_k \times d_k$  هستند، جایی که  $d_k$  پارامتر فرضی است. پس از آن، می‌توان درایه‌های ماتریس توجه را با معادلات

زیر محاسبه کرد:

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

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

جایی که  $x_i$  عناصر دنباله ورودی هستند و  $\alpha$  بردارهای ماتریس توجه هستند.  $b_{ij}$  وزن های موقعیت نسبی را نمایان می کنند. اگر فاصله بین دو عنصر جمله ۳ باشد،  $b_{ij}$  کدگذاری موقعیت نسبی ۳ را نمایان می کند و وزن های 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 وزن های موقعیت جدید به بردار کلید و ضرب با بردار پرسش این معنی را می دهد که توجه بیشتری به عناصر توالی متناظر دارد و همچنین با استفاده از معادله ۱۰ ضرب  $b_{ij}$  و سیگموئید ( $M_{ij}$ ) اطلاعات کلان تطابق کلی بین عناصر  $i$  و  $j$  را در کدگذاری موقعیت نسبی وارد می کند. به عبارت دیگر با این تغییر، بیشترین تطابق عناصر دنباله باعث توجه بیشتر به موقعیت نسبی آن ها می شود. عملیات softmax در فرایند خود-توجه معمول تغییر نمی کند. برای محاسبه ماتریس توجه، از معادله (۱۱) استفاده می شود. به جای استفاده از یک گروه  $W_k, W_v$ ، استفاده از چندین گروه توانایی خودتوجه را افزایش می دهد. هنگام استفاده از چندین گروه، به آن خودتوجه چندسره می گویند، و محاسبه می تواند به صورت زیر فرموله شود:

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

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  $[head^{(1)}; \dots; head^{(n)}]$  از اندازه  $R^{n \times d}$  خواهد بود.  $W_O$  یک پارامتر یادگیری پذیر 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 آن ها توسط اندازه گیری تشابه یا معکوس فاصله در فضای بردار محاسبه می شود. معیارهای تشابه معیارهای فاصله هستند که فاصله یا نزدیکی بین دو بردار را تعیین می کنند. 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 منهن و فاصله مینکوفسکی و غیره [۱۶]. تشابه کسینوسی یکی از پر استفاده ترین معیارها برای اندازه گیری تشابه معنایی بین بردارها است. در برخی از مقالات مرتبط با تشخیص تشابه معنایی، تشابه کسینوسی به فاصله زاویه ای تبدیل می شود. می توان از تابع آرک کسینوس به این منظور استفاده کرد. آرک کسینوس تشابه کسینوسی را به فاصله زاویه ای تبدیل می کند که به فاصله مثلثی پایبند

$$Q, K, V = XW^Q, XW^K, XW^V,$$

$$A_{i,j} = Q_i K_j^T,$$

$$Z = Attention(K, Q, V) = softmax\left(\frac{A}{\sqrt{d_k}}\right)V,$$

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

که  $Q_i$  بردار پرسش برای توکن  $t$  امین است،

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

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

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

توکن  $j$

است که توکن  $t$  امین به آن توجه می کند.  $K_j$  نمایش بردار کلیدی از توکن  $j$  امین است. عملگر softmax در طول بعد آخر اعمال می شود.

ما اطلاعات موقعیت نسبی ویژگی های مدل را در دو سطح تغییر داده ایم: 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 می شود. ما معادله (۹) را برای منتقل کردن وزن های یال موقعیت نسبی که اطلاعات کلان تطابق کلی کلمات را شامل می شود پیشنهاد می کنیم.  $b_{j-i}$  (برای  $k$  و  $v$ ) موقعیت نسبی قابل یادگیری است، که توسط تابع سیگموئید اعمال شده بر  $M_{ij}$  (ماتریس  $M$  در مرحله پیش پردازش تولید شده، توضیح داده شده در بخش ۲-۳) وزن دهی می شود. با استفاده از معادله (۱۰-۱۱) ما تطابق و موقعیت نسبی را به لایه خود-توجه وارد می کنیم به شرح زیر:

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

برای پیاده سازی بهینه تر از معادله ۱۳ استفاده می کنیم.

است. طبق این رویه، عدم وجود زاویه عملکرد بهتری در تشخیص تشابه معنایی بین جملات نسبت به تشابه کسینوسی دارد [۱۷]. معادله ۱۶ نحوه محاسبه تشابه بین دو بردار  $u$  و  $v$  با استفاده از آرک کسینوس را توضیح می‌دهد.

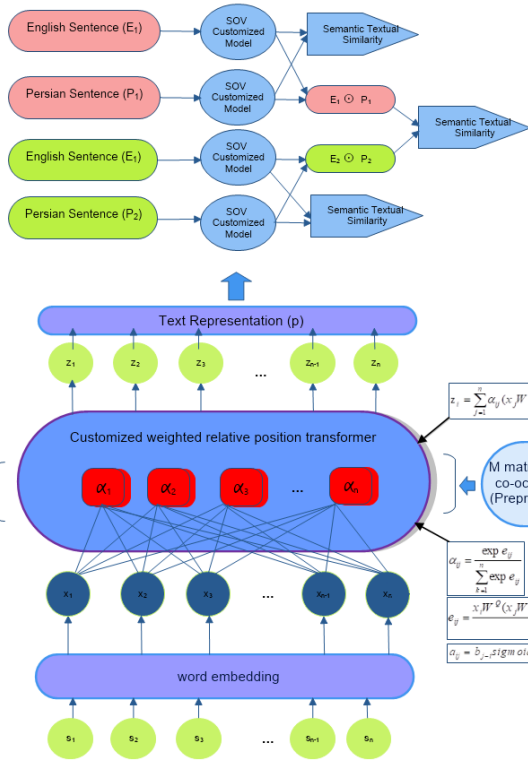
$$Similarity(U, V) = -\arccos\left(\frac{U \cdot V}{\|U\| \|V\|}\right) \quad (16)$$

با استفاده از معیارهای مبتنی بر فاصله مانند فاصله اقلیدسی و فاصله منهتن، ما می‌توانیم تشابه بین دو بردار را با انتقال معکوس فاصله تعیین کنیم [۱۸]. همان‌طور که در معادله ۱۷ آمده است، فاصله اقلیدسی کوتاهترین فاصله بین دو بردار را بر اساس قضیه پیتاگورس محاسبه می‌کند. اگر  $x$  و  $y$  دو بردار تعبیه با بعد  $p$  برای جملات باشند، فاصله اقلیدسی و فاصله منهتن بین این دو جمله به صورت معادلات ۱۷ و ۱۸ بیان می‌شود.

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

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

چارچوب مدل ما در شکل 1-3 نشان داده شده است. مدل ما شامل یک ماژول پیش پردازش برای محاسبه فاکتور PMI، یک لایه توکن سازی، یک ماژول کدگذاری موقعیت نسبی و یک ماژول تشابه متنی معنایی با یادگیری مشترک است.



شکل 1-3: چارچوب مدل شامل یک ماژول پیش پردازش برای محاسبه عامل PMI، یک لایه توکن سازی، یک ماژول کدگذاری سفارشی با وزن دهی PMI به موقعیت نسبی و ماژول تشابه متنی معنایی

Sentence 1		Sentence 2		Score
E <sub>1</sub>	She won the gold medal in the swimming competition.	E <sub>2</sub>	The swimming champion received a gold medal.	
P <sub>1</sub>	او مدال طلای مسابقات شنا را به دست آورد.	P <sub>2</sub>	قهرمان شنا مدال طلا گرفت.	

جدول 1-3: نمونه‌ها از مجموعه داده E<sub>1</sub> و P<sub>1</sub> به تعبیری معنایی یکسانی دارند و همچنین E<sub>2</sub> و P<sub>2</sub> امتیاز تشابه بین  $\langle E_1, P_1 \rangle$  و  $\langle E_2, P_2 \rangle$  برابر با ۰,۸۸ است.

در مجموعه داده ارزیابی تشابه متنی معنایی (STS)، جایی که متون جفت شده با یک امتیاز تشابه همراه هستند، یک بردار متناسب برای هر جمله در هر دو زبان اول (فارسی) و دوم (انگلیسی) ایجاد می‌شود. به عنوان مثال، در جدول 1-3،  $\langle E_1, P_1 \rangle$  به تعبیری نمونه جملات معادل در انگلیسی و فارسی را در STS-B نشان می‌دهد، در حالی که  $\langle E_2, P_2 \rangle$  نمایانگر نمونه‌ها در سوی دیگر است. از آنجاکه هر دوی این‌ها از نظر معنایی معادل هستند، انتظار می‌رود دارای یک امتیاز تشابه بیشینه باشند که برابر با یک است. E<sub>1</sub>, P<sub>1</sub>, E<sub>2</sub> و P<sub>2</sub> توسط کدگذار



```

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 استفاده کرده‌ایم. 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 فراهم کرده‌ایم. برای ارزیابی کیفیت مدل تولیدشده، نیاز به داده‌های برچسب‌گذاری شده توسط منبع انسانی داریم. به علاوه برای یکدست شدن متن و سادگی پردازش، لازم است حروف اضافی حذف شوند. علاوه بر آن وجود حروف اضافی در ادامه راه به ما کمکی نمی‌کند. حروف اضافی شامل «! \* + = - @ [ < > . / » است [۱۹]. از آنجاکه مجموعه‌داده مرجعی برای اندازه‌گیری مشابهت معنایی بین زبان‌های فارسی و انگلیسی وجود ندارد، ما از مجموعه‌داده STS Benchmark انگلیسی استفاده کردیم. ما یکی از جفت جمله‌های این مجموعه‌داده را به زبان فارسی ترجمه کردیم تا امکان ارزیابی مدل ایجاد شود. این مجموعه‌داده شامل ۸۶۲۸ جفت جمله به همراه امتیاز مشابهت معنایی از ۰ (کمترین مشابهت) تا ۵ (بیشترین مشابهت) می‌باشد. این مجموعه به سه بخش تقسیم شده است: داده‌های آموزش (۷۰٪)، اعتبارسنجی (۱۵٪) و آزمون (۱۵٪).

##### ۴-۲- ارزیابی و مقایسه نتایج در مشابهت‌یابی معنایی متون

در این زیربخش، ما تجزیه و تحلیل دقیقی از عملکرد مدل ارائه شده در وظایف STS تک‌زبانه و چندزبانه ارائه می‌دهیم. ما نتایج را با نتایج مدل‌های برجسته دیگر مقایسه می‌کنیم تا برتری رویکردمان را نشان دهیم. ما مدل خود را بر روی مجموعه داده‌های STS-B فارسی-فارسی و انگلیسی-فارسی آموزش داده‌ایم و آن را آزمایش کرده‌ایم. هر مجموعه داده را به ۵ زیرمجموعه مساوی تقسیم کرده و از اعتبارسنجی ۵ تایی استفاده کرده‌ایم. مرحله آموزش را ۵ بار جداگانه تکرار کرده‌ایم، هر بار یکی از ۵ زیرمجموعه به‌عنوان مجموعه آزمون استفاده شده و ۴ زیرمجموعه باقی‌مانده به‌عنوان مجموعه آموزش ترکیب شده‌اند. در تمام آزمایش‌ها، همبستگی پیرسون با شباهت کسینوسی، فاصله اقلیدسی و فاصله منتهن به‌عنوان معیاری برای ارزیابی عملکرد مدل محاسبه شده است.

##### ۴-۳- STS تک‌زبانه (فارسی - فارسی)

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

سفارشی پیشنهادی ایجاد می‌شوند. چارچوب یادگیری مشترک ما با استفاده از معادلات (۱۹، ۲۰) بردارهای متناظر را به هم متصل می‌کند و آن‌ها را به ماژول تشابه منتقل می‌کند.

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

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

بنابراین، با در نظر گرفتن  $L_1$  برای  $\langle P_1, E_1 \rangle$ ،  $L_2$  برای  $\langle P_2, E_2 \rangle$  و  $L_3$  برای  $\langle U, V \rangle$  به‌عنوان مؤلفه‌های توابع از دست دادن تشابه معنایی، از معادله (۲۱) برای محاسبه تابع هزینه مشترک استفاده می‌شود.

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

##### ۳-۵- الگوریتم یادگیری برای مشابهت‌یابی متون

به منظور آموزش مدل بهینه شده از کتابخانه‌های transformers که توسط Hugging Face تحت فریم ورک Pytorch ارائه شده است استفاده می‌کنیم. مدل پیشنهادی به شیوه مینی-دسته‌ای (mini-batch) آموزش می‌بیند. الگوریتم آموزش را در اینجا ارائه می‌دهیم. ابعاد تعبیه و تعداد سرانه‌ها ورودی‌های مورد نیاز هستند. ماتریس مجاورت هم‌رخدادی در مرحله پیش‌پردازش محاسبه می‌شود. عملیات توکن‌گذاری و مقداردهی پارامترهای یادگیری پیش از آموزش انجام می‌شوند. در طول آموزش، دسته‌ای از متن‌های جفت انگلیسی و فارسی 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 به دست آورند. سپس ماژول تشابه امتیاز STS را پیش‌بینی می‌کند. ما از امتیاز STS برای محاسبه تابع از هزینه میانگین مربعات (MSE) استفاده می‌کنیم. در نهایت، الگوریتم پارامترهای مدل را بر اساس گرادیان‌های تابع هزینه به‌روزرسانی می‌کند.

Algorithm 1: Training algorithm for presented model.

Require: preprocessed adjacency matrix of global word co-occurrence  $M$  using (1-6)

Require: training sentences set  $S$  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  $S_{batch}$  of size  $k$

$Loss \leftarrow 0$

for  $(s_1, s_2, \dots, s_k)$  in  $S_{batch}$  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)

$Loss \leftarrow Loss + L(y)$

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 شباهت‌یابی ایجاد می‌کند. 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) استفاده نمی‌کنیم، ضریب همبستگی به ۶۴٫۱۱٪ می‌رسد. با این حال، هنگام استفاده از متن موازی، ضریب همبستگی افزایش می‌یابد و با افزایش تعداد جفت جملات فارسی-انگلیسی در متن موازی، ضریب همبستگی پیروسون نیز افزایش می‌یابد. به عنوان 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 داده است. 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 به دست آورده است. 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 عمل کرده و نتایج برتری را به دست آورده-

اقلیدسی و ۹۱٫۶۵٪ همبستگی با فاصله منهن به دست آوردیم. مدل سفارشی‌سازی شده با SOV برترین همبستگی را با فاصله منهن داشت و این عملکرد از مدل‌های XML-R، DistilBert و M-BERT آموزش دیده شده بهتر بود. همه مدل‌های مبتنی بر توجه نتایج با امتیاز شباهت کسینوسی بالا را به دست آوردند، اما در مواردی که باید شباهت بالایی بین جملات وجود داشته باشد، پیش‌بینی‌های ضعیف‌تری داشتند و همبستگی کمتری با امتیازهای واقعی نشان دادند. هنگامی که بازنمایی جملات در XML-R را به صورت مستقیم برای شباهت متنی معنایی به کار بردیم، تقریباً تمام جفت جملات نمره شباهتی بین ۰٫۶ تا ۱٫۰ کسب کردند، حتی اگر برخی از جفت‌ها توسط مترجم به عنوان کاملاً مرتبط محسوب نشوند. به عبارت دیگر، نمایش‌های جملات مبتنی بر ترنسفورمر تمامی جملات را به نحوی فشرده کرده‌اند که تقریباً همه جملات به فضای برداری کوچکتری نگاشت می‌شوند و در نتیجه شباهت بالا ایجاد می‌کنند. بنابراین، مناسب نیست که به طور مستقیم بازنمایی به دست آمده از XML-R را برای تطابق معنایی یا بازیابی متن به کار ببریم. مدل سفارشی پیشنهادی با رمزگذاری موقعیت نسبی وزن‌دار نمایش‌های دقیق‌تری از جملات ایجاد کرد و در نتیجه شباهت پیش‌بینی شده به شباهت واقعی متناسب بود. روش ما شباهت کسینوسی پیش‌بینی شده کمتری را در شباهت واقعی کم پیش‌بینی کرد و شباهت پیش‌بینی شده بالاتری را برای شباهت واقعی بیشتر ایجاد کرد.

جدول 1-4: 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	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

است.

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

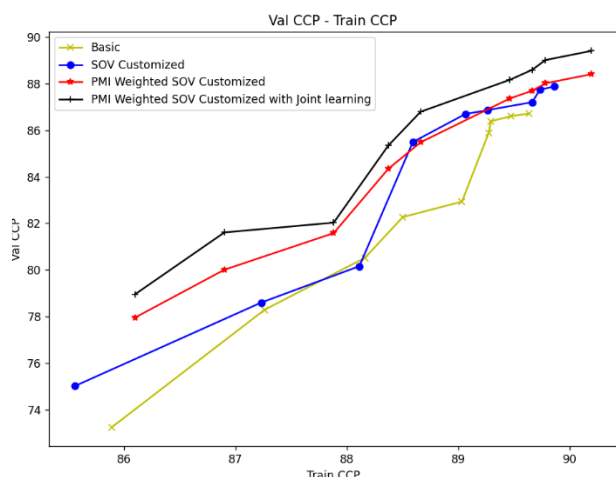
تک‌زبانه

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

۴-۵- مطالعه فرسایشی برای مشابهت‌یابی معنایی متون، برای اعتبارسنجی بیشتر کارایی مدل ارائه شده برای مشابهت‌یابی متون، ما یک مطالعه فرسایشی اجرا کردیم. این مطالعه تأثیر اجزای مختلف مدل ما بر عملکرد کلی را بررسی می‌کند. همان‌طور که در جدول 3-4 آمده است، ما آزمایش‌ها را در دو شرط انجام دادیم. در اینجا، مدل پایه

#### ۴-۴- نتایج بین‌زبانی (فارسی - انگلیسی)

در زمینه تطابق متنی چندزبانه، نتایج ما بر روی مجموعه داده STS-B انگلیسی به فارسی، که در جدول 2-4 نشان داده شده است، کارایی مدل ما را نشان می‌دهد. مدل سفارشی‌سازی شده با SOV بهبودهای قابل توجهی را نسبت به مدل‌های دیگر کسب کرده است و به ترتیب ۸۶٫۹۸٪ همبستگی پیروسون با شباهت کسینوسی، ۸۷٫۶۲٪ همبستگی پیروسون با فاصله اقلیدسی و همچنین ۸۸٫۲۹٪ همبستگی پیروسون با فاصله منهن را فراهم کرده است. در اینجا نتایج مدل‌های XML-R،



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

نمودارهای مدل پایه (زرد) و سفارشی سازی SOV (آبی) در شکل 1-4 نسبتاً نزدیک به یکدیگر هستند، که نشان می‌دهد دارای توانایی تعمیم مشابهی هستند. موقعیت نمودار سیاه در بالایی نمودارهای دیگر است، که نشان می‌دهد که CCP داده‌های تست با یادگیری مشترک بالاتر از دقت به دست آمده در آموزشی یکسان است. بنابراین، می‌توانیم نتیجه بگیریم که مدل ارائه شده دارای توانایی تعمیم بهتری است. علاوه بر این، با توجه به گوشه بالا و راست، واضح است که مدل پایه، سفارشی سازی SOV و سفارشی سازی SOV با وزن PMI به موقعیت‌های بالاتری دست یافته‌اند. این نشان می‌دهد که سطح آموزش مدل در این چهار حالت عمیق تر شده است.

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

اکنون تاثیر کدگذاری موقعیت نسبی با وزن PMI بدون کوتاه نمودن را با ارائه شاخص‌ها در فرایند آموزش مدل مشابهت‌یابی، بررسی می‌کنیم. شکل 2-4 تغییرات از دست رفتن آموزش و هم بستگی (دقت) مجموعه اعتبار یا تست (دقت) با افزایش دوره آموزش را نشان می‌دهد. ما ۵۰ دوره اول را ثبت کرده‌ایم تا وضعیت در طول آموزش مشاهده شود. نمودار آبی شرایط پایه را نمایش می‌دهد و نمودار قرمز مدل ما را نشان می‌دهد. شکل 2-4 (a) نشان می‌دهد که خطای آموزش ما پایین تر است و سرعت همگرایی در طول آموزش سریع تر 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 برآزش دچار شده‌اند. از شکل 2-4 (b) مشخص است که هم بستگی مجموعه اعتبار آموزش ما سریع تر افزایش یافته است و مقدار نهایی آن بالاتر است. این نشان می‌دهد که کدگذاری موقعیت نسبی با وزن PMI بدون برش اثر مهاری بر روی بیش برآزش تابع هزینه دارد.

RPE-Transformer نشان دهنده مدل ترانسفورمر (مدل XML-R) است که برای هر موقعیت نسبی در داخل یک فاصله قطعی بازنمایی را یاد می‌گیرد. SOV Customized (مناسب برای زبان‌های با ترتیب کلمات SOV) به مدل خودتوجهی اشاره دارد که نمایش را برای هر موقعیت نسبی در کل جمله بدون قطع کردن یاد می‌گیرد. PMI weighted SOV Customized نشان دهنده مدل کامل است که شامل سفارشی سازی برای زبان‌های با ترتیب کلمات SOV و تزریق عامل PMI برای هم‌خدادی کلمات جهانی است. همان طور که از نتایج مشخص است، سفارشی سازی SOV و هم‌خدادی کلمات PMI هر دو به نتایج کلی کمک می‌کنند و نسبت به مدل پایه، دقت را به ترتیب در مجموعه داده STS-B فارسی-فارسی ۵.۹۷٪ و در مجموعه داده STS-B فارسی-انگلیسی ۴.۸۲٪ افزایش می‌دهند. هنگامی که ساختار مدل تغییر می‌کند، امتیازهای مشابهت نیز متفاوت هستند. روش ما بهبود قابل توجهی در آزمون‌های چندزبانه و تک‌زبانه ایجاد می‌کند. توجیه منطقی این موضوع این است که روش ما، علاوه بر فاصله بین فاعل و فعل، بر ارتباط میان این دو تمرکز دارد و PMI عامل وزن دار هم‌خدادی 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

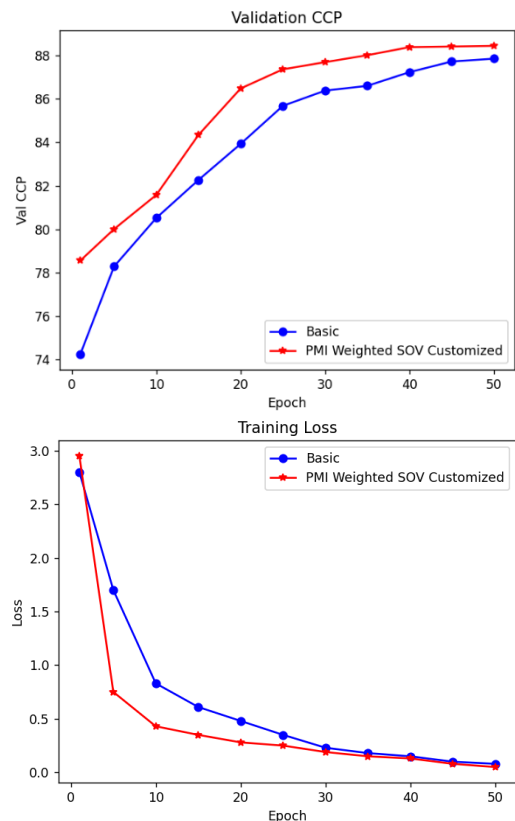
می‌کند. بر اساس این، ما می‌فهمیم که مدل پیشنهادی بازنمایی برداری بهتری را ایجاد کرده و به عبارتی دقت STS بیشتری به دست آورده است.

#### جدول 3-4: مطالعه فرسایشی برای STS

#### ۴-۶- قابلیت تعمیم در مشابهت‌یابی معنایی متون

با توجه به مجموعه داده STS-B فارسی-انگلیسی، شکل 1-4 نمودار هم بستگی اعتبار-آموزش را تحت سه شرط نشان می‌دهد: مدل پایه (نمودار زرد)، سفارشی سازی SOV (نمودار آبی)، سفارشی سازی SOV با وزن PMI (نمودار قرمز) و سفارشی سازی SOV با یادگیری مشترک وزن دار PMI (نمودار سیاه). CCP معیار پیروان را نشان می‌دهد و CCP آموزش نشان دهنده عملکرد مدل در مجموعه اعتباردهی و مجموعه آموزش در طول فرایند آموزش است.

- 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*



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 معرفی کردیم. ابتدا مجموعه داده چندزبانه جامعی با داده‌های موازی برای زبان‌های SOV و SVO ایجاد کردیم تا تنوع زبانی را تضمین کنیم. از رمزگذاری موقعیت نسبی وزن‌دار و غنی‌شده با تزریق اطلاعات هم‌رخدادی از طریق اندازه‌گیری اطلاعات مشترک نقطه‌ای (PMI) برای تقویت بازنمایی متنی استفاده کردیم. علاوه بر این، از یک چارچوب یادگیری مشترک استفاده کردیم که نمونه‌های مشترک بین زبان‌ها را برای بهبود STS بین زبانی به کار می‌برد. با آموزش همزمان بر روی چندین جفت زبان، مدل ما توانایی انتقال دانش را به دست آورد. مدل پیشنهادی ما بر مجموعه داده‌های STS-Benchmarks فارسی-انگلیسی و فارسی-فارسی ارزیابی شد و به ترتیب به ضریب همبستگی پیرسون به مقدار ۰.۸۸،۲۹٪ و ۰.۹۱،۶۵٪ دست یافت.

## مراجع

- [1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional

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

- [18] I. 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.