Karafan

Quarterly Scientific Journal of Technical and Vocational University

Spring 2023, Vol. 20, No. 1, p. 341-362

Journal Homepage: https://karafan.tvu.ac.ir/?lang=en

TVU

doi:10.48301/KSSA.2023.357227.2251

Introducing a Language Model based on BERT to Analyze Sports Content in the Persian Language

Davood Sotoude[1*], Amin Amiri Tehranizade[2]

[1]Faculty Member, Department of Computer Engineering, Technical and Vocational University (TVU), Tehran, Iran.

[2]Postdoc Researcher, Department of Medical Informatics, Faculty of Medicine, Mashhad University of Medical Sciences, Mashhad, Iran.

ARTICLE INFO

Article Type:
Original Research

Received: 09.04.2022
Revised: 11.13.2022
Accepted: 01.30.2023

Keyword:
Language Models
Natural Language Processing
Sentiment Analysis
Named-entity Recognition
Dataset

***Corresponding Author:**
Davood Sotoude
Email: *dsotoude@tvu.ac.ir*

ABSTRACT

Seljuk Pretrained language models are very important because of their application in issues related to natural language processing. Language models such as BERT have become more popular among researchers. Due to the focus of these language models on English, other languages are limited to some multilingual models. In this article, the PersianSportBERT language model is presented for the purpose of Persian sports analysis in topics related to this linguistic field. This language model is based on the Bert language model and was trained using the collected dataset. Three problems were used to evaluate the new language model: sentiment analysis, named entity recognition and text infilling. In order to train this language model, due to the lack of a suitable dataset, a wide range of sports events and news in the Persian language was prepared from several online sources. Due to the specialization of this model and compared to the language models presented for the Persian language, this model provided better results in all three problems. This model had the best performance with 71.7% and 95.2% in text infilling and named entity recognition, respectively. In sentiment analysis, the sports model presented better results. These findings demonstrate that using a language model related to any specialized field will have better results compared to language models related to the general field of texts.

E-ISSN: 2538-4430
ISSN: 2382-9796

EXTENDED ABSTRACT

Introduction

Language models have become an important element in many natural language processing problems. These models are usually trained on unlabeled data and then fine-tuned using labeled data for use in final problems.

Among Persian speakers, sports news is one of the most important aspects of every Iranian's life. Sports fields such as football, volleyball, and wrestling enjoy a great number of supporters. For this purpose, language model of sports in Persian language was examined in the present research named VarzeshiBERT. This language model is trained based on the ParsBERT language model, which is a general language model in the Persian language, using the data set collected without labels from the collection of Persian language sports fields. This dataset contains 2.7 million documents and 25 million sentences in the field of sports texts.

Methodology

The VarzeshiBERT sports language model is based on ParsBERT model. This model needed to train 110 million parameters. The training of this language model was carried out using the RTX 3060-12GB graphics processor based on the parameters in Table 1. Considering that ParsBERT is a language model based on BERTbase and trained with general content in Persian language, the existing model of this language model on the HuggingFace website was used to teach VarzeshiBERT. The parameters of the optimizer and other items taken from this model were unchanged.

Based on the BERT language model, the training method is a masked language model in which 15% of the words of each sentence are randomly selected. Of these, 80% will be with the [MASK] token, 10% with another random token, and 10% will be considered unchanged. The goal of training the network is to fill the masked tokens with suitable words.

Table 1. Parameters used in network training.

Network parameters	Value	Hyper parameters	Value
Hidden layers	12	Steps	1M
Attention heads	12	Batch size	6
Hidden size	768	Dropout	0.1
Sequence Length	128	learning rate	1e-5

Results and discussion

The VarzeshiBERT was evaluated with three problems in natural language processing compared to existing language models in Persian language. These issues include text infilling, sentiment analysis, named entity recognition. Due to the lack of labeled sports data set in Persian language for the last two problems, a new data set with manual labeling was prepared which was used to evaluate the sports model. In each of these problems, the language model was evaluated with the following models, which due to the

specialization of this model in all evaluations, the sports language model showed better efficiency and results.

- **mBERT:** the multilingual version of the BERT model, which includes 102 different languages, including Persian.
- **XLM-RoBERTa:** also includes 100 languages, including Persian.
- **ParsBERT:** which is the basic model for training sports model with sports data set.

Text infilling

One of the problems that was used for its evaluation is text infilling. The results of this evaluation can be seen in Table 2.

Table 2. Results of the evaluation of text infilling.

Language model	Accuracy	
	Acceptable	True
XLM-RoBERTa	46.45	22.31
ParsBERT	65.17	35.65
mBERT	40.23	24.28
VarzeshiBERT	71.7	40.12

Named entity recognition

The purpose of this problem was to find the identity of each word in the text. In order to train and evaluate sports texts, a new dataset called perSportNer consisting of 4732 sentences was collected. The distribution of classes of the new dataset is shown in Figure 1.

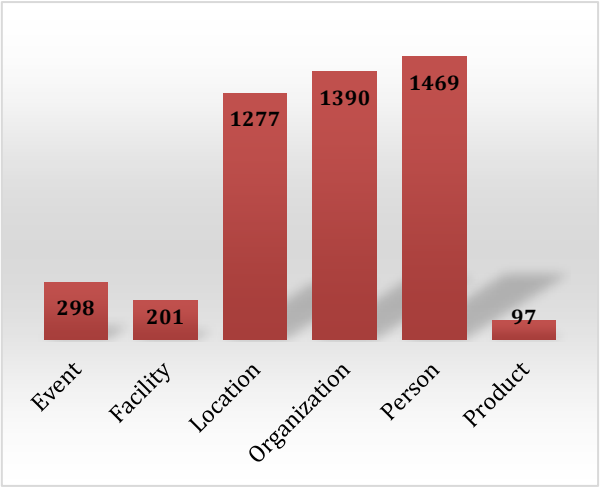


Figure. Frequency of classes of labeled dataset called perSportNer.

Fine-tuning the network and training it on labeled data was carried out using four evaluated models including VarzeshiBERT. Figure 2 shows the NER process.

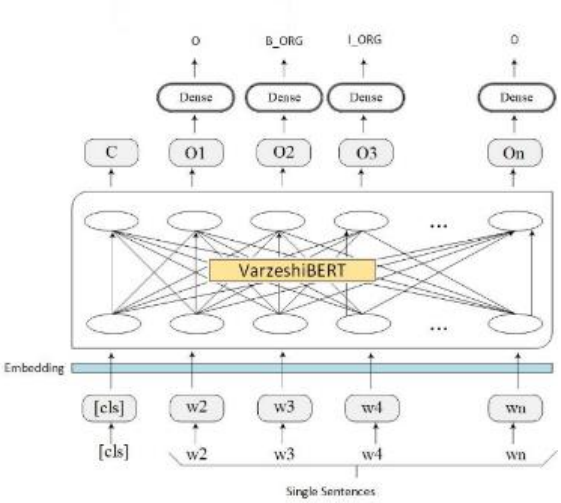


Figure 2. Structure of the presented model for the problem of tagging speech components.

Results of this evaluation can be seen in Table 3.

Table 3. The evaluation of NER.

Language model	F1
	perSportNer
XLM-RoBERTa	80.3
ParsBERT	92
mBERT	83.1
VarzeshiBERT	95.2

Sentiment analysis

User comments on sports websites usually indicate the status of a team or athlete and the level of satisfaction with its performance. One of the uses of VarzeshiBERT is to analyze the sentiments of users on different sites. Due to the lack of an independent data set for the field of sports in Farsi, a data set of comments on the varzesh3 site was collected. The distribution of the collected dataset can be seen in Figure 3.

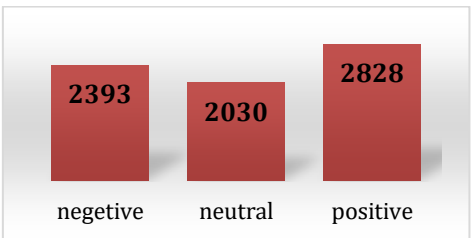


Figure 3. Frequency of labeled classes in the perSportSent dataset.

In order to fine-tune and train the models, a linear prediction layer is placed in the output of [cls] token of the evaluated models. Figure 4 shows the structure used to train the evaluated models.

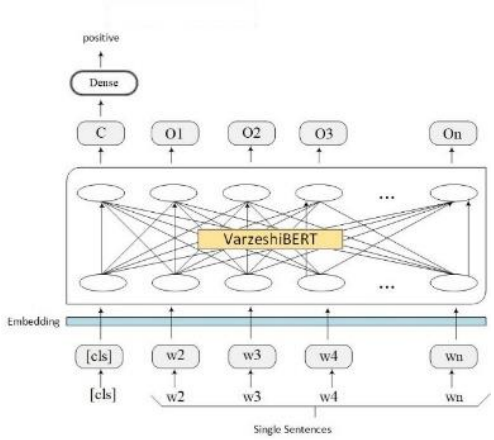


Figure 4. Proposed structure of the model for the sentiment analysis problem.

The Macro-F1 criterion was also used to evaluate this problem. Table 4 shows the comparison of all proposed models using the perSportSent dataset.

Table 4. Evaluation results of sentiment analysis for proposed language models.

Language model	MACRO-F1
XLM-RoBERTa	73.39
ParsBERT	83.43
mBERT	75.12
VarzeshiBERT	89.32

Conclusion

The present research is the first implementation of a BERT language model in the specialized field of sports. BERT language models were trained on extensive datasets. The sports language model was prepared using the ParsBERT trained model and its retraining using 2.7 million sports documents. This model was evaluated using three important problems in natural language processing and demonstrated that in the field of sports texts it provided higher efficiency than all related Persian models.

کارافن
فصلنامه علمی دانشگاه فنی و حرفه‌ای
بهار ۱۴۰۲، دوره ۲۰، شماره ۱، ۳۶۲-۳۴۱
آدرس نشریه: https://karafan.tvu.ac.ir/
doi:10.48301/KSSA.2023.357227.2251

مدل زبانی مبتنی بر BERT جهت تحلیل محتوای ورزشی در زبان فارسی

داود ستوده[1*]، امین امیری طهرانی‌زاده[2]

۱- عضو هیات علمی، گروه مهندسی کامپیوتر، دانشگاه فنی و حرفه‌ای، تهران، ایران.

۲- محقق پسادکترا، گروه انفورماتیک پزشکی، دانشکده پزشکی، دانشگاه علوم پزشکی مشهد، مشهد، ایران.

اطلاعات مقاله

نوع مقاله: مقاله پژوهشی

دریافت مقاله: ۱۴۰۱/۰۶/۱۳
بازنگری مقاله: ۱۴۰۱/۰۸/۲۲
پذیرش مقاله: ۱۴۰۱/۱۱/۱۰

کلید واژگان:
مدل زبانی
پردازش زبان‌های طبیعی
تحلیل احساسات
تشخیص نهادهای نامگذاری شده
مجموعه داده

***نویسنده مسئول:** داود ستوده
پست الکترونیکی:
dsotoude@tvu.ac.ir

چکیده

مدل‌های زبانی آموزش دیده، به دلیل کاربرد آنها در مسائل مرتبط با حوزهٔ پردازش زبان‌های طبیعی دارای اهمیت فراوانی هستند. مدل‌های زبانی مانند BERT از محبوبیت بیشتری میان محققان برخوردار شده‌اند. به دلیل توجه این مدل‌های زبانی به زبان انگلیسی، دیگر زبان‌ها به برخی از مدل‌های چند زبانه محدود می‌شوند. در این مقاله، مدل زبانی PersianSportBERT به منظور تحلیل محتوای ورزشی فارسی در مسائل مرتبط با این حوزهٔ زبانی ارائه شده است. این مدل زبانی بر پایهٔ مدل زبانی Bert و با استفاده از مجموعهٔ داده جمع‌آوری شده، آموزش دیده است. سه مسأله برای ارزیابی مدل زبانی جدید استفاده شده است: تحلیل احساسات، تشخیص نهادهای نامگذاری شده و پرکردن جای خالی. برای آموزش این مدل زبانی با توجه به عدم وجود مجموعه داده‌ای مناسب، یک مجموعه داده گسترده از رویدادها و اخبار ورزشی زبان فارسی از چندین مرجع برخط تهیه شده است. با توجه به تخصصی بودن حوزه این مدل و در مقایسه با مدل‌های زبانی ارائه شده برای زبان فارسی، این مدل در هر سه مساله، نتایج بهتری را ارائه داده است. این مدل با ۷۱.۷٪ و ۹۵.۲٪ بهترین عملکرد را به ترتیب در بخش‌های پرکردن جای خالی و تشخیص نهادهای نامگذاری شده داشته است. در تحلیل احساسات نیز مدل ورزشی، نتایج بهتری را به همراه داشته است. این نتایج نشان می‌دهد، به‌کارگیری مدل زبانی مرتبط با هر حوزهٔ تخصصی، نتایج بهتری در مقایسه با مدل‌های زبانی مرتبط اما با حوزهٔ عمومی متون، خواهد داشت.

شاپای الکترونیکی: ۲۵۳۸-۴۴۳۰
شاپای چاپی: ۲۳۸۲-۹۷۹۶

مقدمه

در بین فارسی زبانان، ورزش و اخبار ورزشی یکی از مهم‌ترین بخش‌های زندگی هر ایرانی را تشکیل می‌دهد. حوزه‌های ورزشی از جمله فوتبال، والیبال، کشتی و ... هر کدام دارای طرفداران فراوانی می‌باشند. تحلیل رفتار طرفداران تیم‌های مختلف ورزشی، تشخیص میزان رضایت طرفداران، دریافت اخبار ورزشی به زبان‌های غیر فارسی و ترجمه آن به زبان فارسی و ... همگی از مواردی است که توسط حوزهٔ پردازش زبان طبیعی[1] مورد بررسی قرار می‌گیرند. لذا، ایجاد یک مدل زبانی تخصصی که آموزش آن بر اساس متون ورزشی بوده است، می‌تواند نتایج بهتری در موارد ذکر شده نسبت به مدل‌های زبانی عمومی به همراه داشته باشد.

مدل‌های زبانی در حال حاضر تبدیل به یک عنصر مهم در بسیاری از مسائل پردازش زبان‌های طبیعی شده است. این مدل‌ها معمولا بر روی داده‌های بدون برچسب، آموزش[2] داده می‌شوند و سپس با استفاده از داده‌های برچسب‌گذاری شده جهت استفاده در مسائل نهایی، تنظیم دقیق[3] می‌شوند. مدل‌های متنی مبتنی بر تبدیل [۱] مانند BERT [۲] به دلیل آن‌که تنظیم دقیق آنها برای استفادهٔ نهایی، بسیار ساده می‌باشد، موفقیت‌های چشمگیری را کسب کرده‌اند. این مدل زبانی در ابتدا برای محتوای انگلیسی و پس از آن برای زبان‌های بسیاری ارائه شده است [۳-۵]. از جمله مدل‌های چند زبانی مرجع می‌توان به mBERT [۲] و XLM-R [۶] اشاره کرد. این مدل‌های زبانی با آموزش هم‌زمان بر روی متون زبان‌های مختلف برگرفته از ویکی‌پدیا و دیگر منابع، ساخته می‌شوند. زبان فارسی نیز توسط این مدل‌های زبانی، پشتیبانی می‌شود. کارایی این مدل‌های زبانی در مقالات بسیاری مورد بررسی قرار گرفته است. در زبان فارسی نیز برخی مقالات، مدل‌های زبانی مبتنی بر BERT را ارائه داده است. از جمله ParsBERT [۷] به عنوان یک مدل زبانی عمومی و SinaBERT [۸] به عنوان یک مدل زبانی حوزهٔ تخصصی پزشکی.

به دلیل عدم وجود مجموعه داده‌ای مناسب در زبان فارسی، پردازش سنگین و طولانی بودن آموزش شبکه، یکی از مدل‌های محدود زبانی تخصصی ارائه شده در زبان فارسی، مدل زبانی تخصصی حوزهٔ پزشکی، با نام SinaBERT می‌باشد. برای آموزش شبکهٔ SinaBERT، از متون تخصصی پزشکی استفاده شده است. عدم وجود مدل زبانی در حوزهٔ ورزشی، انگیزهٔ اصلی ایجاد چنین مدلی در زبان فارسی می‌باشد. در این مقاله با استفاده از معماری ارائه شده توسط BERT، یک مدل زبانی تخصصی با محتوای ورزشی ارائه شده است. آموزش شبکهٔ این مدل زبانی بر اساس محتوای استخراج شده از سایت‌های ورزشی و یا بخش ورزشی سایت‌های خبری انجام شده است. با توجه به آن‌که این مدل زبانی، ادامهٔ آموزش بر روی شبکه ParsBERT است، لذا استفاده از مدل زبانی PersianSportBERT در دیگر حوزه‌های زبان فارسی امکان پذیر است، اما مطمئناً بهترین نتیجه زمانی حاصل می‌شود که در یکی از مسائل پردازش زبان طبیعی در حوزهٔ متون ورزشی مورد استفاده قرار گیرد. این مدل زبانی بر اساس مدل زبانی ParsBERT که یک مدل زبانی عمومی در زبان فارسی می‌باشد، با استفاده از مجموعه داده جمع‌آوری شده بدون برچسب، از مجموعه حوزه‌های ورزشی زبان فارسی، آموزش می‌بینند. این مجموعه داده شامل ۲.۷ میلیون سند و ۲۵ میلیون جمله در حوزهٔ متون ورزشی می‌باشد. مراحل ایجاد این مدل زبانی دارای بخش‌های زیر می‌باشد:

- آماده سازی متون ورزشی، جهت آموزش شبکه
- دریافت مدل از پیش آموزش دیده شده (مدل ParsBERT و یا BETR) از منابع معتبر از جمله Huggingface
- تنظیم پارامترهای مورد نیاز جهت آموزش مجدد شبکه با مجموعهٔ متون جدید

[1] Natural language processing
[2] Pre-training
[3] Fine-tuning

- اجرای آموزش شبکه (این فرآیند بسته به تنظیمات شبکه و سخت افزار مورد استفاده، از چند روز تا چند هفته طول خواهد کشید).
- اعتبار سنجی مدل نهایی با استفاده از کد و یا بارگذاری در سایت HuggingFace
- به‌کارگیری مدل نهایی در سه مسئله به عنوان ارزیابی مدل زبانی

ارزیابی این مدل زبانی در سه حوزهٔ مطرح در پردازش زبان‌های طبیعی از جمله تشخیص نهادهای نامگذاری شده[۱]، تحلیل احساسات[۲]، پرکردن جای خالی[۳] انجام شده است. جهت انجام فرآیند ارزیابی نیز، بنابر نیاز از مجموعه داده‌ای ورزشی برچسب گذاری شده در هر کدام از حوزه‌ها استفاده شده است. در نهایت کارایی این مدل با مدل‌های زبانی موجود برای زبان فارسی مورد بررسی قرار گرفته است.

پیشینهٔ تحقیق

با توجه به افزایش محبوبیت مدل‌های زبانی، بسیاری از مسائل حوزهٔ پردازش زبان‌های طبیعی به این مدل‌ها وابستگی پیدا کرده‌اند. برخی از تحقیقات، مدل‌هایی در سطح کاراکتر ارائه داده‌اند [۷]. از جمله یک مدل زبانی سطح کاراکتر[۴] با استفاده از شبکه‌های عصبی بازگشتی[۵] در [۹] ارائه شده است. در تحقیق دیگری از مدل چند وظیفه‌ای سطح کاراکتر برای موضوعات پزشکی جهت رفع مشکلات «خارج از واژگان»[۶] استفاده شده است [۱۰].

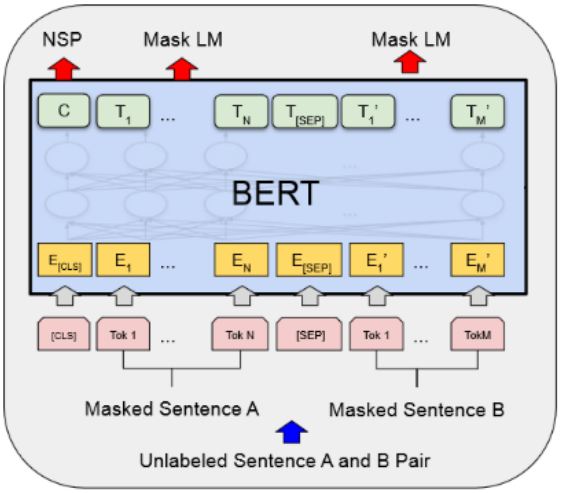
مدل‌سازی زبانی متنی، بر این پایه که هر کلمه در متون مختلف می‌تواند معنی متفاوت داشته باشد، ارائه شد. مدل‌های زبان رمزگذار-رمزگشا[۷]، رمز گذارهای خودکار ترتیبی[۸] و مدل‌های ترتیبی[۹] در این بخش معرفی شدند [۱۱-۱۳]. ELMo و ULMFiT [۱۴] مدل‌های معرفی شده در این حوزه بودند که هر دو مبتنی بر شبکه‌های LSTM[۱۰] [۱۵] هستند. ULMFiT از یک شبکه LSTM چند لایه بهره می‌برد در حالی که ELMo از یک ساختار دوطرفه LSTM جهت پیش بینی کلمات قبلی و بعدی در دنباله‌ای از کلمات استفاده می‌کند. هر دو روش، بهبود مشخصی را در مسائل مطرح در حوزه پردازش زبان‌های طبیعی ارائه دادند [۷].

نمونهٔ دیگر برای مدل‌های ترتیبی، مدل‌های انتقالی هستند [۱] که توجه اصلی آنها به تشخیص ارتباط میان داده‌های ورودی و خروجی است. برخلاف LSTM دراین مدل، هیچ شبکه بازگشتی مورد استفاده قرار نگرفته است. در این مدل از دو موجودیت به نام‌های رمزگذار[۱۱] و رمزگشا[۱۲] استفاده می‌شود. رمزگذار توالی ورودی‌ها را دریافت و آنها را به برداری با ابعاد بالاتر نگاشت می‌کند. این بردار در نهایت به یک توالی خروجی توسط رمزگشا تبدیل می‌شود. چندین مدل زبانی بر اساس مدل‌های انتقالی تاکنون مطرح شده‌اند که می‌توان به GPT [۱۶] و BERT [۲] اشاره کرد.

GPT دارای پشته‌ای از ۱۲ رمزگشای متوالی است. البته ساختار این مدل به صورت یک طرفه مطرح شده است به این معنا که هر کلمه تنها نسبت به کلمات قبلی خود مورد بررسی قرار می‌گیرد. از طرف دیگر BERT کلمات

[1] Named-entity recognition
[2] Sentiment analysis
[3] Text infilling
[4] Character-level model
[5] Recurrent neural network
[6] Out-of-vocabulary
[7] Encoder-decoder language models
[8] Sequence autoencoders
[9] Sequence-to-sequence models
[10] Long short-term memory
[11] Encoder
[12] Decoder

قبل و بعد از هر کلمه را با تعریف یک مدل زبانی ماسک شده پردازش می‌کند. همچنین از پشته‌ای از رمزگذارها به همراه پشته رمزگشا استفاده می‌کند. با این روش نمونه‌ای از شبکهٔ عمیق دو طرفه از پیش آموزش دیده شده، با دقت بالا معرفی شده است. بر این اساس ساختارهای مبتنی بر انتقال مانند شبکه‌های XLNet [۱۷]، ROBERTA [۱۸]، XLM [۱۹]، T۵ [۲۰] و ALBERT [۲۱] تاکنون معرفی شده است که همگی نتایج مثبتی در مسائل به روز پردازش زبان‌های طبیعی ارائه داده‌اند. مدل زبانی BERT در واقع یک مدل مبتنی بر انتقال می‌باشد که بر اساس مدل‌های زبانی ماسک شده و تشخیص جملهٔ بعدی، آموزش اولیه را دریافت می‌کند. مرحلهٔ آموزش این مدل زبانی در شکل ۱ آمده است.



شکل ۱. مدل زبانی BERT [۲].

پس از آموزش اولیهٔ شبکه، تنظیم دقیق BERT برای آموزش شبکه جهت استفاده در مسائل خاص پردازش زبان‌های طبیعی، مورد استفاده قرار می‌گیرد.

در حوزهٔ متون تخصصی، مدل زبانی BERT تاکنون در چندین مورد استفاده شده است. برای مثال SciBERT [۲۲] در حوزهٔ تخصصی موضوعات علمی و هوش مصنوعی، FinBERT [۲۳] در حوزهٔ تخصصی داده‌های مالی، BioBERT [۲۴] در حوزهٔ تخصصی زیست پزشکی، ClinicalBERT [۲۵] در حوزهٔ تخصصی کیلینیکی، LegalBERT [۲۶] در حوزهٔ تخصصی حقوق.

در خصوص زبان مورد استفاده، مدل‌های زیادی تا کنون برای زبان‌های غیر انگلیسی ارائه شده است. ELMo برای زبان پرتغالی، ژاپنی و آلمانی [۱۴] ارائه شده است. همچنین با استفاده از شبکه BERT تا کنونBERTje [۲۷] برای زبان هلندی، AlBERTo [۲۸] برای زبان ایتالیایی، AraBERT [۲۹] برای زبان عربی و مدل‌های دیگری برای زبان‌های فنلاندی، روسی و پرتغالی و برخی زبان‌های دیگر ارائه شده است.

در حوزهٔ زبان فارسی با استفاده از مدل زبانی BERT تا کنون چندین مدل با پشتیبانی از این زبان ارائه شده است. mBERT و XLM-ROBERTA [۶] از جمله اولین مدل‌های زبانی با پشتیبانی از زبان فارسی است که البته حوزهٔ تخصصی متون و اندازهٔ مجموعه داده‌ای آنها (برای زبان فارسی) مشخص نیست. از دیگر مدل‌های زبان فارسی، همچنین می‌توان به PARSBERT [۷] اشاره کرد. این مدل، یک مدل زبانی فارسی در حوزهٔ عمومی می‌باشد که با بیش از ۲.۸

میلیون سند اخبار و داده‌های فروشگاه‌های فروش برخط آموزش دیده است. مدل دیگر ارائه شده،SINABERT [۸] می‌باشد که در حوزهٔ تخصصی متون پزشکی مدل خود را ارائه داده است. این مدل با استفاده از مدل PARSBERT و با بیش از ۲.۸ میلیون سند تخصصی پزشکی آموزش دیده شده است. در دیگر حوزه‌های تخصصی فارسی تا کنون مدل جدیدی ارائه نشده است که دلایل آن می‌تواند عدم وجود مجموعه داده‌ای مناسب و همچنین سنگین بودن حجم پردازش‌های مراحل آموزش BERT باشد.

مدل‌های زبانی مرتبط

در این مقاله از ۳ مدل زبانی مطرح که زبان فارسی را پشتیبانی می‌کنند، جهت ارزیابی نهایی استفاده شده است. این مدل‌ها عبارتند از mBERT [۲]، xlm-roberta-base [۶] و parsBERT [۷]. دو مدل اولیه به‌صورت عمومی بسیاری از زبان‌های رایج دیگر را نیز پشتیبانی می‌کنند اما مدل سوم، به صورت خاص جهت زبان فارسی تهیه شده است. mBERT با پشتیبانی از ۱۰۴ زبان مطرح دنیا که برای بسیاری از کارهای چند زبانه مفید است، یکی از مدل‌های زبانی با بیشترین زبان مورد پشتیبانی را معرفی کرده است. برای آموزش این مدل زبانی، مجموعه داده عظیم از وبسایت ویکی‌پدیا، مورد استفاده قرار گرفته است. مجموعه داده مورد استفاده، به صورت متن خام و حاوی جملات می‌باشد. هیچ گونه نشانه‌گذاری دستی توسط انسان یا پویا توسط ماشین بر روی این مجموعه داده انجام نشده است. این مدل با استفاده از مدل زبانی ماسک شده (MLM) و پیش‌بینی جملهٔ بعدی(NSP) آموزش دیده است. مدل MLM، ۱۵ درصد کلمات هر جمله ورودی به صورت تصادفی، مخفی شده است. در فرآیند آموزش، شبکه باید کلمات مخفی شده را پیش‌بینی نماید. در این نوع از مدل‌های زبانی، آموزش شبکه به صورت دو سویه[1] در جملات می‌باشد، بدین معنی که علاوه بر تأثیر کلمات اولیه جمله بر کلمات بعدی، کلمات آخر جمله نیز بر کلمات ابتدایی جمله تأثیرگذار خواهند بود. در مدل NSP نیز، شبکه دو جمله را به‌صورت هم‌زمان از ورودی دریافت می‌کند. این دو جمله گاهی اوقات در متن واقعی، دو جمله پشت سر هم می‌باشند و گاهی اوقات هیچ ارتباطی به هم ندارند. شبکه می‌بایست طی فرآیند آموزش تشخیص دهد آیا این دو جمله، می‌توانند دو جمله پشت سر هم باشند یا خیر؟ xlm-roberta-base نیز با پشتیبانی از ۱۰۰ زبان و آموزش شبکه خود با حدود ۲.۵ ترابایت، به عنوان یکی از مدل‌های زبانی مطرح، معرفی شده است. این مدل نیز با استفاده از مدل زبانی ماسک شده (MLM) آموزش دیده است. پس از تنظیم دقیق مدل زبانی، می‌توان از آن به عنوان دسته‌بندی جملات، تحلیل احساسات، تشخیص نهادهای نامگذاری شده، و ... استفاده کرد. این مدل‌های زبانی در پایگاه اینترنتی HuggingFace بارگذاری شده و استفاده از آنها و تنظیم دقیق آنها برای وظایف دیگر به سادگی امکان‌پذیر می‌باشد. ParsBERT نیز یک مدل زبانی برگفته از BERT می‌باشد که با توجه به آن‌که برای زبان فارسی تهیه شده است، بسیار سبک‌تر از مدل‌های ذکر شده بوده و در بسیاری از ارزیابی‌های انجام شده نتایج بهتری را در بر داشته است. برای آموزش این مدل زبانی، از تنظیمات مرتبط با BERT (Base) استفاده شده است. تنها تفاوت این مدل با مدل BERT چند زبانه، در محتوای آموزش شبکه بوده است که فقط جملات فارسی برای آموزش شبکه مورد استفاده قرار گرفته است. برای آموزش این شبکه از هر دو روش MLM و NSP استفاده شده است. در ارزیابی انجام شده در [۷] در سه حوزه مطرح در پردازش زبان‌های طبیعی شامل تحلیل احساسات، تشخیص نهادهای نامگذاری شده و دسته‌بندی متون، مدل زبانی ParsBERT بهترین نتایج را در مقایسه با دیگر مدل‌ها به خود اختصاص داده است که دلیل اصلی آن طبق تحلیل ارائه شده، تخصصی بودن زبان مورد استفاده در این مدل زبانی می‌باشد.

بخش‌های بعدی این مقاله به ترتیب زیر خواهد بود:

- ایجاد مجموعه داده‌ای جدید تشکیل شده از متون ورزشی به زبان فارسی (بدون برچسب)

[1] Bidirectional

- آموزش مجدد ParsBERT با استفاده از مجموعه داده‌ای جدید
- ارزیابی مدل زبانی ورزشی با سه موضوع تخصصی پردازش زبان‌های طبیعی

ضمناً مدل زبانی نهایی جهت استفاده در تحقیقات بعدی در این حوزه در وبسایت تخصصی حوزۀ پردازش زبان‌های طبیعی huggingface[1] قرار داده شده است.

روش‌شناسی مدل

در این بخش، روش شناسی مدل مورد استفاده مطرح شده است که شامل سه بخش جمع‌آوری داده، پیش‌پردازش داده‌ها و آماده‌سازی جملات می‌باشد.

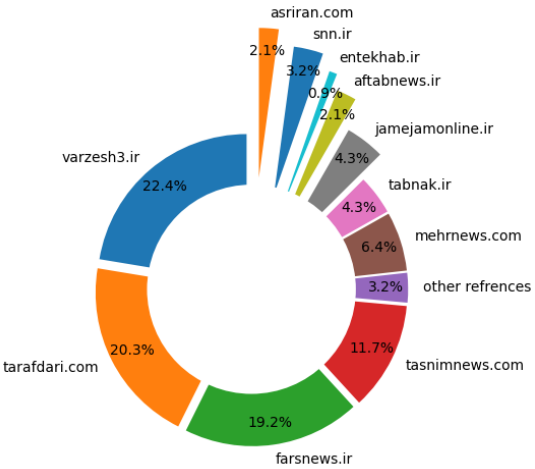
جمع‌آوری داده

تمام مدل‌های زبانی که بر روی زبان فارسی ارائه شده است هیچ کدام بر روی محتوای ورزشی آموزش ندیده است. با وجود آن‌که محتوای ورزشی بسیاری به صورت برخط وجود دارد اما هیچ مجموعه داده منسجمی از این نوع نوشتار وجود ندارد. لذا برای آموزش یک مدل زبانی ورزشی در زبان فارسی، یک مجموعه دادۀ ورزشی از مجموعه مقالات، اخبار و متون ورزشی برخط تهیه شده است. این متون شامل خبرهای حوزۀ ورزشی داخلی و خارجی می‌باشد. البته با توجه به آن‌که بیشترین اخبار ورزشی و خوانندگان این اخبار را علاقه‌مندان به ورزش فوتبال تشکیل می‌دهد، حجم زیادی از مجموعه داده جمع آوری شده نیز در این حوزه قرار دارند. این مجموعه شامل ۲.۷ میلیون سند می‌باشد که از منابع زیر جمع آوری شده است:

- وبسایت‌های خبری ورزشی
- بخش ورزشی خبرگزاری‌های فارسی
- وبسایت‌هایی با تولید محتوای ورزشی
- مجموعه مقالات ورزشی در دسترس که امکان تهیه فایل متنی از آنها وجود داشته است.
- کتاب‌های ورزشی با متون قابل دسترس
- انجمن‌های ورزشی برخط
- برخی گروه‌ها و کانال‌های ورزشی تلگرام

درصد فراوانی اسناد با موضوعات ورزشی با نام persianSportCorpus در تصویر ۲ آمده است.

[1] https://huggingface.co/montazeri/bert-base-persian-sport-bert-uncased



شکل ۲. فراوانی منابع در جمع آوری اسناد با متون ورزشی.

پیش پردازش داده‌ها

اسناد جمع آوری شده با حذف تگ‌های HTML و پیوندهای موجود به صورت کامل پاکسازی شده است. تبدیل حروف عربی به فارسی، حذف کاراکترهای اضافه و ... از دیگر موارد پاکسازی متن بوده است. مراحل آماده‌سازی اسناد، با استفاده از کتابخانه Hazm[1] و برخی کتابخانه‌های دیگر انجام می‌شود.

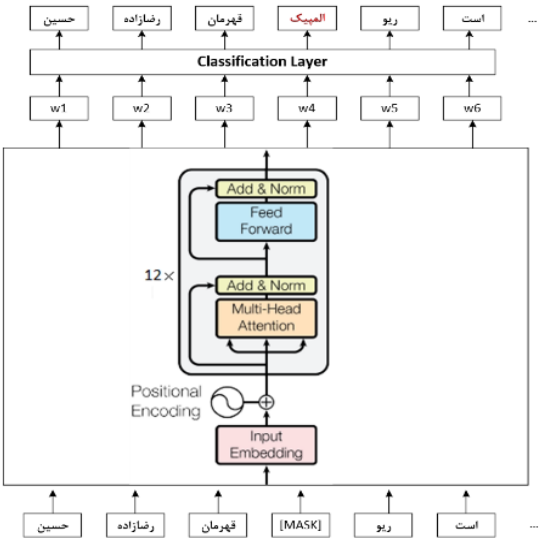
آماده سازی جملات

با توجه به ساختار ورودی BERT، هر سند باید به چندین جمله تبدیل و هر جمله دقیقا در یک خط قرار گیرد. میان اسناد مختلف نیز یک خط خالی قرار داده می‌شود. در نهایت، یک فایل متنی با حجم ۵.۳ گیگابایت شامل ۲۵ میلیون جمله از اسناد جمع‌آوری شده جهت آموزش با مدل BERT آماده شده است.

آموزش PersianSportBERT

مدل زبانی ورزشی PersianSportBERT بر پایهٔ مدل BERTbase می‌باشد. این مدل برای آموزش نیاز به مقداردهی به ۱۱۰ میلیون پارامتر دارد. آموزش این مدل زبانی با استفاده از سیستم با پردازنده گرافیکی RTX ۳۰۶۰- ۱۲GB بر اساس پارامترهای جدول ۱، انجام شده است. با توجه به آن‌که ParsBERT، مدل زبانی مبتنی بر BERTbase و آموزش دیده با محتوای عمومی در زبان فارسی می‌باشد، برای آموزش PersianSportBERT از مدل موجود این مدل زبانی در سایت HuggingFace استفاده شده است. پارامتر های بهینه ساز و سایر موارد بدون تغییر از این مدل برگرفته شده است. مدل مورد استفاده جهت آموزش شبکه که برگرفته از معماری BERT (Base) می‌باشد، در شکل ۳ نشان داده شده است. جملهٔ فارسی ورزشی با حذف یک کلمه و جایگزینی توکن [MASK]، به عنوان ورودی به شبکه داده می‌شود و شبکه به منظور اصلاح اوزان خود در فرآیند پس-انتشار[2]، سعی می‌کند کلمه مخفی شده را پیش‌بینی کند. در شکل ۳ کلمه «المپیک» به عنوان کلمه‌ای با بیشترین احتمال جایگزینی توکن [MASK] نمایش داده شده است.

[1] https://github.com/sobhe/hazm
2 back-propagation



شکل ۳. شبکهٔ مورد استفاده جهت آموزش مدل زبانی متون ورزشی برگرفته از معماری BERT.

همان‌طور که در بخش معرفی BERT عنوان شد، آموزش اولیه BERT با دو مسئله پیش‌بینی جمله بعدی و پر کردن جای خالی انجام می‌پذیرد. اما با توجه به تحقیقات انجام شده در [۳۰] آموزش BERT با پیش‌بینی جملهٔ بعدی در نهایت منجر به کاهش کارایی مدل آموزش دیده برای مسائل دیگر می‌شود. لذا در مدل ورزشی فارسی نیز تنها از روش پر کردن جای خالی استفاده می‌شود. این روش که، مدل زبانی ماسک شده[1] نیز نامیده می‌شود، ۱۵ درصد کلمات هر جمله به صورت تصادفی انتخاب می‌شوند. از این تعداد، ۸۰٪ با توکن [MASK]، ۱۰٪ با توکن تصادفی دیگر و ۱۰٪ نیز بدون تغییر در نظر گرفته می‌شوند. هدف آموزش شبکه، پرکردن (پیش‌بینی) توکن‌های ماسک شده با کلمات مناسب می‌باشد. همان‌طور که در شکل ۳ مشاهده می‌شود این شبکه دارای دو بخش رمزگذار (سمت چپ) و رمزگشا (بخش سمت راست) می‌باشد. بخش رمزگذار، بر اساس الگوی ذکر شده، برخی از کلمات هر جمله را مخفی می‌کند و بخش رمزگشا، با دریافت جمله کامل، سعی می‌کند کلمات مخفی شده را پیش‌بینی نماید. این فرآیند موجب می‌شود ۱۱۰ میلیون پارامتر شبکه آموزش ببینند.

جدول ۱. پارامترهای مورد استفاده در آموزش شبکه.

ساختار شبکه	مقدار	پارامترهای آموزش	مقدار
Hidden layers	۱۲	Steps	۱M
Attention heads	۱۲	Batch size	۶
Hidden size	۷۶۸	Dropout	۰.۱
Sequence Length	۱۲۸	learning rate	۱e-۵

1 masked language modelling

ارزیابی

مدل ورزشی زبان فارسی BERT با سه مسئله در پردازش زبان‌های طبیعی نسبت به مدل‌های زبانی موجود در زبان فارسی، مورد ارزیابی قرار گرفته است. این مسائل عبارت است از: پر کردن جای خالی در جمله، تحلیل احساسات، تشخیص نهادهای نامگذاری شده. به دلیل عدم وجود مجموعه داده برچسب دار ورزشی در زبان فارسی، برای دو مساله آخر، مجموعه داده جدید با برچسب‌گذاری دستی تهیه شده است که از این مجموعه داده‌ها برای ارزیابی مدل ورزشی استفاده شده است. در هر کدام از این مسائل مدل زبانی با مدل‌های زیر مورد ارزیابی قرار گرفته است که با توجه به تخصصی بودن این مدل، در تمامی‌ارزیابی‌ها، مدل زبانی ورزشی، کارایی و نتایج بهتری را نمایش داده است.

- **mBERT**: نسخه چند زبانی مدل BERT که شامل ۱۰۴ زبان مختلف از جمله فارسی می‌باشد.
- **XLM-RoBERTa**: نیز شامل ۱۰۰ زبان از جمله زبان فارسی می‌باشد.
- **ParsBERT**: که مدل پایه برای آموزش مدل ورزشی با مجموعه داده ورزشی می‌باشد.
- تمامی مدل‌های اشاره شده در بالا، با محتوای کاملاً عمومی آموزش دیده‌اند.

پس از آنکه مدل PersianSportBERT آموزش دید، برای ارزیابی آن در دو مساله اشاره شده آخر، این مدل و تمامی مدل‌هایی که در ارزیابی شرکت می‌کنند، نیاز به آموزش مجدد یا تنظیم دقیق شبکه دارند. اما در مساله ارزیابی پر کردن جای خالی، به دلیل آن‌که مدل اولیه بر اساس تشخیص جای خالی جملات آموزش دیده است، نیاز به هیچ تغییری در مدل زبانی وجود ندارد. پارامترهای مورد نیاز برای این بخش آموزش نیز مانند بخش قبل در نظر گرفته شده است.

پرکردن جای خالی

با توجه به آموزشی که مدل در متون ورزشی زبان فارسی دیده است، یکی از مسائلی که جهت ارزیابی آن مورد استفاده قرار گرفته است، پر کردن جای خالی در جملات می‌باشد. در این روش فرض می‌شود، برخی از کلمات یک جمله در دسترس نباشند. هدف اصلی این ارزیابی، پیدا کردن کلماتی است که اگر در جای خالی جمله قرار بگیرند، معنای صحیح جمله حفظ می‌شود. در واقع آموزش BERT با تشخیص جای خالی در جملات انجام شده است، لذا می‌توان از این روش ارزیابی برای تشخیص صحت آموزش یک مدل زبانی نیز استفاده کرد. برای این منظور مجموعه داده جداگانه‌ای از مجموعه داده آموزشی اولیه مورد استفاده قرار گرفته است. این مجموعه داده با استفاده از بیش از ۴۵۰۰ جمله تصادفی تهیه شده از منابع ورزشی فارسی ایجاد شده است. جهت ارزیابی مدل زبانی PersianSportBERT ، در هر جمله، ۱۵ درصد کلمات مخفی شده است. از مدل زبانی PersianSportBERT و مدل‌های مطرح شده در بخش ارزیابی، برای پیش‌بینی کلمات مخفی شده استفاده شده است. در پیش‌بینی کلمات مخفی شده، سه نوع پاسخ از مدل آموزش دیده احتمال دارد: پاسخ کاملاً صحیح، پاسخ قابل قبول و پاسخ نامرتبط. جهت ارزیابی این روش، درصد پاسخ قابل قبول و کاملاً صحیح هر مدل زبانی را به نسبت به کل کلمات مخفی شده مورد بررسی قرار داده‌ایم. در این بررسی مدل ورزشی فارسی، بالاترین کارآیی را در دو نوع ارزیابی 'کاملاً صحیح' و 'قابل قبول' داشته و پس از آن PARSBERT به عنوان یک مدل زبانی عمومی در مرتبهٔ دوم و پس از آن دو مدل دیگر قرار گرفته‌اند. در جدول ۲ نتایج این ارزیابی قابل مشاهده می‌باشد.

جدول ۲. نتیجهٔ ارزیابی مدل ورزشی فارسی در مسأله پر کردن جای خالی کلمات.

مدل زبانی	دقت	
	پیش‌بینی قابل قبول	پیش‌بینی کاملاً صحیح
XLM-RoBERTa	۴۶.۴۵	۲۲/۳۱

مدل زبانی	دقت	
	پیش‌بینی قابل قبول	پیش‌بینی کاملاً صحیح
ParsBERT	۶۵.۱۷	۳۵/۶۵
mBERT	۴۰.۲۳	۲۴/۲۸
PersianSportBERT	۷۱.۷	۴۰/۱۲

در جدول ۳ برخی موارد مورد بررسی، برگرفته از متون ورزشی که در بین این ۴ مدل زبانی مورد مقایسه قرار گرفته است مشاهده می‌شود: (در جملات جدول ۳، مدل زبانی باید کلمه جایگزین [MASK] را پیش‌بینی نماید)

جدول ۳. نمونه جملات ناقص، تکمیل شده با کلمات توسط مدل‌های زبانی.

ردیف	متن جمله	مدل زبانی و تشخیص آن		
		وضعیت پاسخ	کلمه پیش بینی شده	نوع مدل زبانی
۱	حسین [MASK] دو بار مدال طلای مسابقات المپیک را برای ایران به ارمغان آورده‌است	○	طلا	ParsBERT
		○	##ی	mBERT
		○	حسینی	xlm-roBERTa-base
		●●	رضازاده	PersianSportBERT
۲	در مسابقه‌های تکواندو بازی‌های آتن، سهم هر کشور، دو شرکت کننده تعیین شده بود که این سهمیه به هادی ساعی و یوسف [MASK] تعلق گرفت.	●	قادریان	ParsBERT
		○	ایران	mBERT
		○	۲۰۲۰	xlm-roBERTa-base
		●●	کرمی	PersianSportBERT
۳	وحید [MASK] آقای گل فوتسال جهان است	●●	شمسایی	ParsBERT
		○	قهرمانی	mBERT
		○	ی	xlm-roBERTa-base
		●●	شمسایی	PersianSportBERT
۴	دو تیم ذوب آهن و نفت آبادان در ورزشگاه فولاد [MASK] به مصاف هم رفتند.	●	خوزستان	ParsBERT
		●	خوزستان	mBERT
		●●	شهر	xlm-roBERTa-base
		●●	شهر	PersianSportBERT
۵	یوسین بولت دونده [MASK] دو سرعت و سریعترین انسان جهان است.	○	شماره	ParsBERT
		○	با	mBERT
		○	با	xlm-roBERTa-base
		●●	جامایکایی	PersianSportBERT

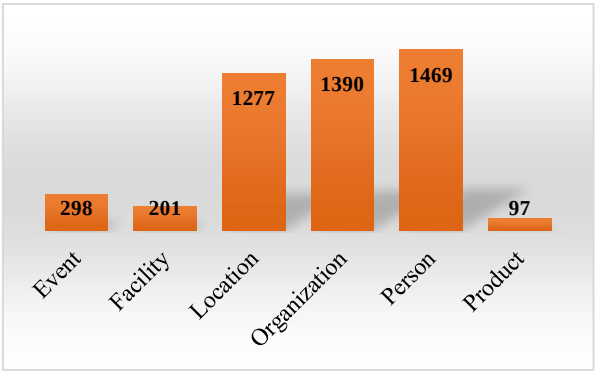
○ : پاسخ نامرتبط

● : پاسخ قابل قبول

●● : پاسخ کاملاً صحیح

تشخیص نهادهای نامگذاری شده (NER)

هدف این مسأله پیدا کردن هویت هر کلمه در متن است. از جمله آن که بتواند تشخیص دهد کلمات متن چه نسبتی با کلاس‌های NER از جمله محل، سازمان، نام اشخاص، .. دارند. به عنوان مثال در جمله «رضازاده متولد اردبیل است» دو نهاد از جمله، مورد شناسایی قرار می‌گیرد، نهاد اول «اردبیل» که از کلاس مکان و «رضازاده» از کلاس اشخاص شناسایی می‌شود. برای آموزش دقیق و ارزیابی مدل زبانی ورزشی فارسی، از یک مجموعه داده جدید در حوزهٔ ورزشی مشابه با کلاس‌های موجود در مجموعه داده Arman استفاده شده است. Arman مجموعه داده‌ای شامل ۷۶۸۲ جمله است که نهادهای آن بر اساس کلاس‌های موجود در شکل ۴ برچسب‌گذاری شده است. نام مجموعه دادهٔ جدید perSportNer و متشکل از ۴۷۳۲ جمله می‌باشد. این مجموعه داده، استخراج شده از متون ورزشی (غیر از مجموعه داده اصلی) بوده و برچسب گذاری آن در ۶ دسته و بر اساس فرمت IOB انجام شده است. در این فرمت، کلماتی که جزو یک موجودیت NER نیستند از نوع "O"، کلماتی که ابتدای یک موجودیت هستند با "B" و باقی کلمات همان موجودیت با "I" تگ‌گذاری می‌شوند. هر دو نوع B ، I با یک آندرلاین از نوع موجویت جدا شده‌اند. توزیع کلاس‌های مجموعه داده‌ای جدید در شکل ۴ آمده است. از هر دو مجموعه داده‌ای ذکر شده، ۸۰٪ مجموعه به صورت تصادفی به عنوان داده‌های آموزشی، ۱۰٪ به عنوان آزمون و ۱۰٪ به عنوان اعتبارسنجی استفاده شده است.

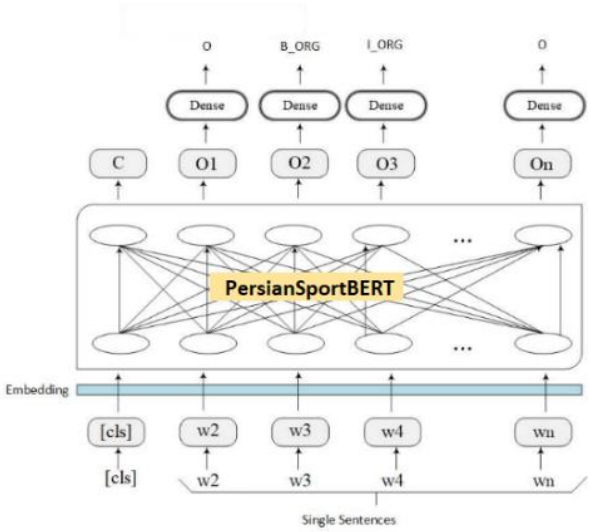


شکل ۴. فراوانی کلاس‌های مجموعه داده‌ای برچسب گذاری شده به‌نام perSportNer.

تنظیم دقیق شبکه و آموزش آن بر روی داده‌های برچسب‌گذاری شده، با استفاده از ۴ مدل مورد ارزیابی از جمله PersianSportBERT انجام می‌شود. برای این منظور یک لایه پیش‌بینی خطی [1] در خروجی مدلBERT قرار می‌گیرد تا برچسب‌های IOB را برای کلمات هر جمله، کلاس‌بندی نماید. شکل ۵ فرآیند NER را برگرفته از [۲] نمایش می‌دهد. در این شکل، جمله به همراه توکن cls (این توکن در ابتدای هر جمله قرار گرفته و پس از مرحلهٔ آموزش، به عنوان ورودی شبکه، طبق شکل ۵ به شبکه داده می‌شود)، به عنوان ورودی در نظر گرفته می‌شود، سپس بردار معادل کلمات ورودی به مدل داده می‌شود (در شبکه BERT متناظر با هر کلمه از جمله ورودی، یک بردار که توصیف کنندهٔ آن کلمه می‌باشد و در زمان آموزش شبکه، به‌دست آمده است، مورد استفاده شبکه قرار می‌گیرد. این موضوع در شکل ۵ با w داخل مستطیل نمایش داده شده است). پس از تنظیم دقیق شبکه، کلاس هر کلمه(به جز cls) در ner که متناظر با هر کلمه ورودی می‌باشد، به عنوان خروجی شبکه در نظر گرفته می‌شود. به عنوان مثال در شکل

[1] Linear prediction layer

۵، کلاس متناظر با کلمه w3 در جمله ورودی، B_ORG می‌باشد، به این معنا که این کلمه، ابتدای عبارتی است که یک موجودیت از نوع سازمان(Organization) را مشخص می‌کند.



شکل ۵. ساختار مدل ارائه شده برای مساله تشخیص نهادهای نامگذاری شده.

برای ارزیابی این مسئله، بخش آزمون مجموعه داده perSportNer بر روی ۴ مدل آموزش دیده اجرا و نتیجه آن بر اساس شاخص F1 به عنوان معیار ارزشیابی در جدول ۴ قابل مشاهده می‌باشد.

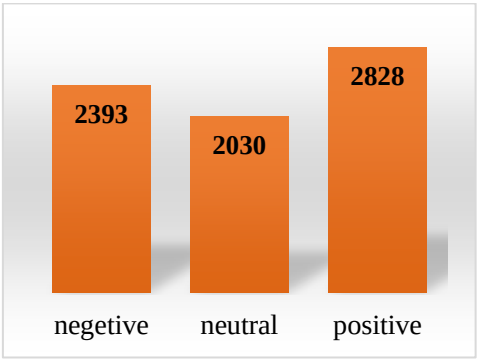
جدول ۴. نتایج ارزیابی تشخیص نهادهای نامگذاری شده برای مدل‌های زبانی مطرح شده.

مدل	F1
	perSportNer
XLM-RoBERTa	۸۰/۳
ParsBERT	۹۲
mBERT	۸۳/۱
PersianSportBERT	**۹۵/۲**

همان‌طور که مشاهده می‌شود مدل زبانی جدید به دلیل آموزش بر روی متون ورزشی فارسی، نتایج بهتری نسبت به دیگر مدل‌های زبانی به خصوص ParsBERT کسب کرده است. مدل‌های چند زبانی mBERT و XLM-RoBERTa به دلیل آن‌که محتوای آموزش آن بر روی زبان فارسی مشخص نیست، پایین‌ترین امتیاز را در این ارزیابی کسب کرده‌اند. ParsBERT به عنوان مدل عمومی زبان فارسی که بخشی از آموزش آن شامل محتوای ورزشی نیز بوده است، در رتبهٔ دوم قرار گرفته است. در متون ورزشی، کلمات کلیدی فراوانی در فرمت NER وجود دارد که توسط متون عمومی ممکن است به هیچ عنوان دیده نشود. لذا تفاوت معنادار مدل ورزشی و مدل ParsBERT به این دلیل می‌باشد.

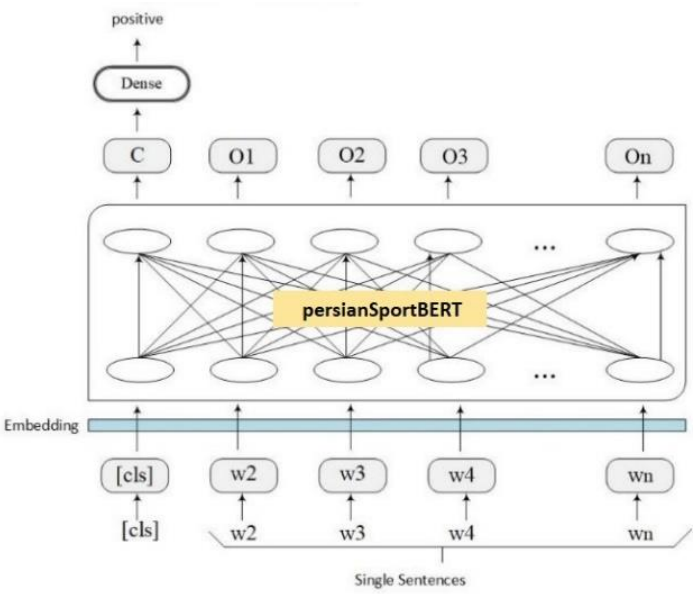
تحلیل احساسات

نظرات کاربران در سایت‌های ورزشی معمولا نشان‌دهندهٔ وضعیت مرتبط با یک تیم یا ورزشکار و میزان رضایت از عملکرد آن می‌باشد. یکی از کاربردهای PersianSportBERT، تحلیل احساسات کاربران در سایت‌های مختلف می‌باشد. هدف این مسئله، طبقه بندی متون، از جمله نظرات، بر اساس محتوای آنها می‌باشد. به دلیل عدم وجود مجموعه داده‌ای مستقل برای حوزهٔ ورزشی در زبان فارسی، مجموعه داده‌ای از نظرات سایت varzesh3 گردآوری شده است. این مجموعه داده به‌نام perSportSent شامل ۷۲۵۰ نظر در خصوص اخبار مندرج در سایت ورزش ۳ می‌باشد، به صورت دستی برچسب گذاری و برای آموزش و ارزیابی مدل فارسی ورزشی مورد استفاده قرار گرفته است. این مجموعه داده به صورت کلاس بندی سه حالته (مثبت، منفی و بی طرف) برچسب گذاری شده است. فراوانی کلاس‌های مجموعه داده گردآوری شده در شکل ۶ قابل مشاهده می‌باشد.



شکل ۶. فراوانی کلاس‌های برچسب‌گذاری شده در مجموعه داده perSportSent.

جهت تنظیم دقیق و آموزش مدل‌های مورد استفاده در ارزیابی با استفاده از مجموعه داده مورد نظر، یک لایهٔ پیش‌بینی خطی در خروجی توکن cls (این توکن در ابتدای هر جمله قرار گرفته و به عنوان ورودی شبکهٔ طبق شکل ۷ به شبکه داده می‌شود)، از مدل‌های مورد ارزیابی، قرار می‌گیرد. شکل ۷ ساختار مورد استفاده جهت آموزش مدل‌های مورد ارزیابی را نشان می‌دهد.



شکل ۷. ساختار مدل ارائه شده برای مسأله تحلیل احساسات.

برای ارزیابی این روش نیز از معیار Macro-F1 استفاده شده است. جدول ۵ مقایسهٔ تمام مدل‌های مطرح شده با استفاده از مجموعه داده perSportSent را نمایش می‌دهد.

جدول ۵. نتایج ارزیابی تحلیل احساسات برای مدل‌های زبانی مطرح شده.

مدل	MACRO-F1
XLM-RoBERTa	۷۳.۳۹
ParsBERT	۸۳.۴۳
mBERT	۷۵.۱۲
PersianSportBERT	۸۹.۳۲

در این مسئله نیز، مدل PersianSportBERT دارای بهترین نتایج در میان مدل‌ها بوده است. با توجه به آن‌که ارزیابی نهایی بر روی مجموعه داده ورزشی perSportSent انجام می‌شود، تفاوت میان دو مدل ParsBERT و PersianSportBERT با حدود ۶ درصد، مشخص است.

نتیجه‌گیری

این مقاله اولین پیاده‌سازی یک مدل زبانیBERT در حوزهٔ تخصصی ورزش می‌باشد. مدل‌های زبانی BERT بر روی مجموعه داده‌های گسترده آموزش دیده‌اند. مدل زبانی ورزشی با استفاده از مدل آموزش دیده ParsBERT و آموزش مجدد آن با استفاده از ۲.۷ میلیون سند ورزشی آماده شده است. این مدل با استفاده از ۳ مسئله مهم در پردازش زبان‌های طبیعی مورد ارزیابی قرار گرفت و مشخص شد در حوزهٔ متون ورزشی، این مدل از تمامی مدل‌های فارسی

مرتبط، کارایی بالاتری را فراهم کرده است. در بخش تحلیل احساسات، PersianSportBERT با مقدار ۸۹.۳۲ (Macro-F1) مدل‌های دیگر نتایج بهتری را رقم زده است. بهترین نتیجه دراین ارزیابی‌ها به بخش پرکردن جای خالی متون ورزشی باز می‌گردد که مدل زبانی جدید با تفاوت بیش از ۶ درصدی نسبت به برترین مدل بعدی در رتبهٔ اول قرار گرفته است. در بخش تشخیص نهادهای نامگذاری شده نیز، این مدل زبانی بهترین نتیجه را به همراه داشته است. البته قابل ذکر است به دلیل استفاده از مدل آموزش دیده شده ParsBERT، مدل جدید به صورت عمومی نیز در تمامی متون قابل استفاده است اما با توجه به تخصصی شدن مرحلهٔ نهایی آموزش این مدل زبانی، همان‌طور که انتظار می‌رود، بهترین نتیجه در متون ورزشی اتخاذ شده است.

یکی از مشکلات متون ورزشی، حجم عظیم اخبار و رویدادهای ورزشی در خصوص ورزش فوتبال می‌باشد. این امر باعث می‌شود دیگر ورزش‌ها به نسبت در مدل زبانی آموزش کمتری ببینند. البته این نسبت به دلیل وجود طرفداران هر ورزش قابل توجیه می‌باشد چرا که بیشترین پیگیری اخبار ورزشی در زمینهٔ فوتبال صورت می‌پذیرد. در عین حال ایجاد مجموعه داده‌ای با ترکیب متعادل‌تری در تمامی ورزش‌ها و همچنین به صورت تخصصی، مدل‌های زبانی برای تحلیل حوزه‌های دیگر از جمله حقوق، علوم دینی، امور مالی از کارهای پیش رو می‌باشد.

References

[1] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017, December 4-9). *Attention is all you need.* 31st Conference on Neural Information Processing System, Long Beach, California, USA. https://doi.org/10.48550/arxiv.1706.03762

[2] Devlin, J., Chang, M-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *Computation and Language*, 1-16. https://doi.org/10.48550/arXiv.1810.04805

[3] Agerri, R., Vicente, I. S., Campos, J. A., Barrena, A., Saralegi, X., Soroa, A., & Agirre, E. (2020, May 11-16). *Give your text representation models some love: the case for basque.* Proceedings of the 12th Conference on Language Resources and Evaluation, Marseille, France. https://doi.org/10.48550/arXiv.2004.00033

[4] Martin, L., Muller, B., Suárez, P. J. O., Dupont, Y., Romary, L., de La Clergerie, É. V., Seddah, D., & Sagot, B. (2019, July 5-10). *CamemBERT: a tasty French language model.* Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Seattle, Washington. http://dx.doi.org/10.18653/v1/2020.acl-main.645

[5] Virtanen, A., Kanerva, J., Ilo, R., Luoma, J., Luotolahti, J., Salakoski, T., Ginter, F., & Pyysalo, S. (2019). Multilingual is not enough: BERT for Finnish. *Computation and Language*, 1-14. https://doi.org/10.48550/arXiv.1912.07076

[6] Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2019, July 5-10). *Unsupervised cross-lingual representation learning at scale.* The 58th Annual Meeting of the Association for Computational Linguistics, Seattle, Washington. https://doi.org/10.48550/arXiv.1911.02116

[7] Farahani, M., Gharachorloo, M., Farahani, M., & Manthouri, M. (2021). ParsBERT: Transformer-based Model for Persian Language Understanding. *Neural Processing Letters*, *53*(6), 3831-3847. https://doi.org/10.1007/s11063-021-10528-4

[8] Taghizadeh, N., Doostmohammadi, E., Seifossadat, E., Rabiee, H. R., & Tahaei, M. S. (2021). SINA-BERT: a pre-trained language model for analysis of medical texts in Persian. *Computation and Language*, 1-9. https://doi.org/10.48550/arXiv.2104.076 13

[9] Huang, G., & Hu, H. (2019). c-RNN: A Fine-Grained Language Model for Image Captioning. *Neural Processing Letters*, *49*(2), 683-691. https://doi.org/10.1007/s11063-018-9836-2

[10] Niu, J., Yang, Y., Zhang, S., Sun, Z., & Zhang, W. (2019). Multi-task Character-Level Attentional Networks for Medical Concept Normalization. *Neural Processing Letters*, *49*(3), 1239-1256. https://doi.org/10.1007/s11063-018-9873-x

[11] Dai, A. M., & Le, Q. V. (2015, December 7-12). *Semi-supervised sequence learning.* Annual Conference on Neural Information Processing Systems 2015, Montreal, Quebec, Canada. https://proceedings.neurips.cc/paper_files/paper/2015/hash/7137debd45ae4d0ab9aa953017286b20-Abstract.html

[12] Ramachandran, P., Liu, P. J., & Le, Q. V. (2017, September 7-11). *Unsupervised pretraining for sequence to sequence learning.* Conference on Empirical Methods in Natural Language Processing 2017, Denmark. https://doi.org/10.48550/arXiv.1611.02683

[13] Sutskever, I., Vinyals, O., & Le, Q. V. (2014, December 8-13). *Sequence to sequence learning with neural networks* 28th Annual Conference on Neural Information Processing Systems 2014, Montreal, Canada. https://proceedings.neurips.cc/paper_files/paper/2014/hash/a14ac55a4f27472c5d894ec1c3c743d2-Abstract.html

[14] Howard, J., & Ruder, S. (2018, July 15-20). *Universal language model fine-tuning for text classification.* Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Melbourne, Australia. https://doi.org/10.18653/v1/P18-1031

[15] Graves, A. (2012). Long Short-Term Memory. In A. Graves (Ed.), *Supervised Sequence Labelling with Recurrent Neural Networks*. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-24797-2_4

[16] Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training. *University of British Columbia*, *12*, 1-12. https://scholar.google.com/citations?view_op=view_citation&hl=en&user=dOad5HoAAAAJ&citation_for_view=dOad5HoAAAAJ:W7OEmFMy1HYC

[17] Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., & Le, Q. V. (2019, December 8-14). *Xlnet: Generalized autoregressive pretraining for language understanding.* Advances in neural information processing systems, Vancouver, British Columbia, Canada. https://proceedings.neurips.cc/paper/2019/hash/dc6a7e655d7e5840e66733e9ee67cc69-Abstract.html

[18] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *Computation and Language*, 1-13. https://doi.org/10.48550/arXiv.1907.11692

[19] Lample, G., & Conneau, A. (2019, December 13-14). *Cross-lingual language model pretraining.* The 33rd Annual Conference on Neural Information Processing Systems, Vancouver, Canada. https://doi.org/10.48550/arXiv.1901.07291

[20] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, *21*(1), 5485-5551. https://arxiv.org/abs/1910.10683

[21] Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2020, April 26-30). *Albert: A lite bert for self-supervised learning of language representations.* 8th International Conference on Learning Representations, Addis Ababa, Ethiopia. https://doi.org/10.48550/arXiv.1909.11942

[22] Beltagy, I., Lo, K., & Cohan, A. (2019). SciBERT: A pretrained language model for scientific text. *Computation and Language*, 1-6. https://doi.org/10.48550/arXiv.1903.10676

[23] Araci, D. (2019). *Finbert: Financial sentiment analysis with pre-trained language models* [Master, Amsterdam]. Netherlands. https://arxiv.org/abs/1908.10063

[24] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2019). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, *36*(4), 1234-1240. https://doi.org/10.1093/bioinformatics/btz682

[25] Huang, K., Altosaar, J., & Ranganath, R. (2020, April 2-4). *Clinicalbert: Modeling clinical notes and predicting hospital readmission.* Conference on Health, Inference, and Learning 2020, Toronto, Ontario, Canada. https://arxiv.org/abs/1904.05342

[26] Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., & Androutsopoulos, I. (2020, November 16-20). *LEGAL-BERT: The muppets straight out of law school.* The 2020 Conference on Empirical Methods in Natural Language Processing, Punta Cana, Dominican Republic. https://arxiv.org/abs/2010.02559

[27] De Vries, W., Van Cranenburgh, A., Bisazza, A., Caselli, T., Van Noord, G., & Nissim, M. (2019). Bertje: A dutch bert model. *Computation and Language*, 1-6. https://doi.org/10.48550/arXiv.1912.09582

[28] Polignano, M., Basile, P., De Gemmis, M., Semeraro, G., & Basile, V. (2019, November 13-19). *Alberto: Italian BERT language understanding model for NLP challenging tasks based on tweets.* 6th Italian Conference on Computational Linguistics,, Bari, Italy. https://iris.unito.it/handle/2318/1759767

[29] Antoun, W., Baly, F., & Hajj, H. (2020, May 11-16). *Arabert: Transformer-based model for arabic language understanding.* Proceedings of the Twelfth International Conference on Language Resources and Evaluation, Marseille, France. https://doi.org/10.48550/arXiv.2003.00104

[30] Joshi, M., Chen, D., Liu, Y., Weld, D. S., Zettlemoyer, L., & Levy, O. (2020). SpanBERT: Improving Pre-training by Representing and Predicting Spans. *Transactions of the Association for Computational Linguistics*, *8*, 64-77. https://doi.org/10.1162/tacl_a_00300