ارزیابی یادگیری ساختواژی زبان فارسی با استفاده از یک الگوریتم تخمین بیزی

الهام سادات سبط
کارشناس ارشد زبان‌شناسی رایانشی
دانشگاه تهران
e.sebt@ut.ac.ir

مصطفی صالحی
استادیار
دانشکده علوم و فنون نوین
دانشگاه تهران
mostafa_salehi@ut.ac.ir

مزدک انوشه
استادیار
دانشکده ادبیات و علوم انسانی
دانشگاه تهران
mazdakanushe@ut.ac.ir

چکیده

پردازش ساختواژی با استفاده از الگوریتم‌های داده‌محور، رویکردی است که طیف وسیعی از پژوهش‌ها را به خود اختصاص داده است و می‌تواند در پردازش متون تخصصی، کهن و غیررسمی زبان فارسی مانند شبکه‌های اجتماعی، کاربرد داشته باشد. علی‌رغم گستردگی این حوزه، پژوهش‌ها در زمینه ساختواژه زبان فارسی بیشتر به روش‌های قاعده‌مند اختصاص دارد و کارایی این الگوریتم‌ها تاکنون برروی داده‌های زبان، ارزیابی نشده است. در این پژوهش یادگیری ساختواژی بی‌نظارت و نیمه‌نظارتی با استفاده از تخمین بیزی بررسی شده، کارایی آن برروی داده‌هایی از مجموعه داده‌های استاندارد وبلاگ‌های ایران، پیکره همشهری و پیکره متنی زبان فارسی ارزیابی شد. به دلیل دردسترس نبودن داده استاندارد آموزش، با شناسایی و اصلاح پارامترهای تأثیرگذار مانند نویز و طول واژه‌ها در داده و کنترل نقش نیم‌فاصله در فرآیند آموزش و تقطیع، نتایج در هر دو معیار صحت و فراخوانی بهبود یافته و معیارF به نتایجی که پژوهش‌ها برای زبان انگلیسی ثبت کرده‌اند، نزدیک شده است.

کلیدواژه‌ها: پردازش ساختواژی، یادگیری ماشین، تخمین بیزی، الگوریتم داده محور

۱. مقدمه

پردازش ساختواژی به عنوان پردازش پایه در طیف وسیعی از ابزارهای پردازش زبان طبیعی از جمله سیستم‌های ترجمه ماشینی، پرسش و پاسخ، بازیابی اطلاعات و غلط‌یاب املایی کاربرد دارد. اهمیت بهره‌گیری از این پردازش در کوچک‌کردن حجم واژگان، پوشش‌دادن ساخت‌های گوناگون واژه‌ها و امکان پردازش کلمه‌های خارج از واژگان است.

پردازشی که پردازشگر ساختواژی برروی واژه انجام می‌دهد، می‌تواند بسته به نوع زبان و کاربرد آن در سایر ابزارهای پردازش زبان طبیعی متفاوت باشد و غالبا به‌منظور تحقق یکی از اهداف زیر به کار می‌رود (ویرپیوا[1] و همکاران، ۲۰۱۱):

- تقطیع ساختواژی: هدف تقطیع ظاهر واژه و استخراج تک‌واژها به‌صورت فهرستی از زیررشته‌های سازنده واژه یا صورت ظاهری تک‌واژها است. هرچند در این رویکرد تک‌واژگونگی و فرآیندهای واژ-واجی[2] نادیده گرفته می‌شود، به دلیل کاربرد در حوزه‌های گوناگون پردازش زبان طبیعی از جمله ترجمه

[1] Virpioja
[2] Morphophonemic

ماشینی، بازشناسی گفتار و بازیابی اطلاعات و نیز توسعه الگوریتم‌های بی‌نظارت[۳] و نیمه‌نظارتی[۴] دارای اهمیت است.

- ریشه‌یابی و خوشه‌بندی ظاهری واژه‌ها: شامل ریشه‌یابی و دسته‌بندی واژه‌ها برحسب ریشه یا ستاک است و خروجی الگوریتم برای هر کلمه یک برچسب یا ریشه است.
- تجزیه و تحلیل ساختواژی: خروجی این پردازشگرها فهرستی از برچسب‌های تعیین‌کننده ریشه و وندها است، به طوریکه ساختار واژه را تحلیل و تک‌واژگونه‌ها را شناسایی می‌کند.

در حوزه پردازش ساختواژی دو رویکرد عمده استفاده از روش‌های قاعده‌مند و الگوریتم‌های داده‌محور[۵] است. روش‌های قاعده‌مند نیازمند استخراج تک‌واژها و روابط ساختواژی زبان است. اما در رویکردهای یادگیری ماشین[۶] مبتنی‌بر الگوریتم‌های داده‌محور، الگوریتم با استفاده از واژگانی که برای آموزش در اختیار دارد ساختواژه زبان را می‌آموزد و تک‌واژها را در واژه‌های جدید شناسایی می‌کند. علی‌رغم رشد این حوزه پژوهشی در دهه‌های اخیر و توانایی الگوریتم‌ها در شناسایی و تقطیع ساخت‌های پیچیده زبان، وندهای تصریفی و اشتقاقی، واژه‌بست‌ها، ساخت‌های مرکب و نیز کلمات خارج از واژگان، تاکنون برروی داده‌های زبان فارسی ارزیابی نشده است. با توجه به تنوع واژه‌ها در حوزه‌های مختلف علوم و محدودیت پردازشگرهای قاعده‌مند در پردازش واژه‌های خارج از واژگان، همچنین گسترش پژوهش‌ها در حوزه پردازش زبان طبیعی به‌منظور تحلیل متون محاوره‌ای و غیررسمیِ پیام‌رسان‌ها و شبکه‌های اجتماعی، توجه به الگوریتم‌های داده‌محور در پردازش ساختواژی می‌تواند رویکردی بهینه‌تر و کم‌هزینه‌تر باشد.

در این پژوهش، تقطیع ساختواژی داده متنی زبان فارسی با استفاده از مدل تخمین بیزی بررسی شده، چالش‌های داده متنی و دست‌یابی به بهترین نتیجه با توجه به عدم دسترسی به داده آموزش استاندارد و کاستی‌هایی در داده‌های موجود، از جمله حجم و نویز داده و گوناگونی روش‌های نگارشی و کاربرد نویسه غیرزبانی نیم‌فاصله، مورد ارزیابی قرار گرفت. برای ارزیابی دقیق‌تر نتایج به‌دست آمده از الگوریتم‌های بی‌نظارت و نیمه‌نظارتی، داده حاشیه‌نویسی‌شده[۷] و استاندارد ساختواژی زبان فارسی طراحی و تهیه شد.

۲. پیشینه موضوع

در پردازش ساختواژی. دو رویکرد غالب استفاده از روش‌های قاعده‌مند و روش‌های یادگیری ماشین است.

در روش قاعده‌مند، قواعد ساختواژی و واژگان زبان به یک ماشین حالت[۸] یا مبدل حالت محدود[۹] داده می‌شود و سامانه می‌تواند با دریافت یک واژه در ورودی، تجزیه و تحلیل ساختواژی کامل آن را در خروجی نمایش دهد (جورافسکی و مارتین، ۲۰۰۷).

[3] Unsupervised
[4] Semi-supervised
[5] Data-driven algorithm
[6] Machine learning
[7] Annotated data
[8] Finite-state automaton (FSA)
[9] Finite -state transducer (FST)

در حوزه یادگیری ماشین نخستین پژوهش‌ها به رویکردهای بانظارت[10] بر می‌گردد که به حجم بالای داده برچسب‌خورده نیاز دارد و بسیار هزینه‌بر است. بنابراین گرایش‌ها به سمت حوزه پژوهشی بی‌نظارت سوق پیدا کرد که برای زبان‌هایی با منابع زبانی اندک کارایی بالایی دارد. در پژوهش‌های اخیر روش‌های نیمه‌نظارتی به دلیل بایاس شدن برای نوع خاصی از زبان و کاربرد، رشد چشمگیری را نشان می‌دهند. این روش‌ها را می‌توان در چهار دسته کلی طبقه‌بندی کرد:

۱- الگوریتم‌های خانواده مورفسور[11]: خانواده مورفسور، گروهی از پردازشگرهای ساختواژی برپایه روش‌های یادگیری ماشین آماری است که از سال ۲۰۰۲ تاکنون تکامل و بهبود یافته است. یادگیری در این پردازشگرها به عنوان یک روش یادگیری پارامتری با سه مؤلفه توصیف می‌شود: مدل، تابع هزینه، الگوریتم یادگیری و تقطیع (کوریمو[12] و همکاران، ۲۰۱۰a؛ ویرپیوا و اسمیت، ۲۰۱۵).

در بیشتر پردازشگرهای این خانواده از مدل‌های احتمالاتی زایشی[13]استفاده شده است که می‌تواند احتمال توأم[14] $P(A = a, W = w|\theta)$ را برای هر کلمه w و تقطیع ساختواژی آن، که به‌صورت لیستی از برچسب‌های معرف تک‌واژها است $a = (m_1, m_2, \dots, m_n)$، به‌شرط پارامترهای مدل θ محاسبه کند.

$$P(A = a, W = w|\theta) = \prod_{i=1}^{|a|} P(m_i|\theta) \qquad (۱)$$

در این پردازشگرها تابع هزینه وابسته به پارامترها، مدل و مجموعه داده است. در نسخه‌های گوناگون عمدتا برپایه تخمین بیزی[15] یا تخمین بیشینه احتمال پسین[16](MAP) تعریف شده است. در نسخه‌های نخستین از تخمین بیشینه شباهت[17] و تخمین کوتاه‌ترین طول توصیفی[18] نیز استفاده شده است.

۲- دستور تطبیق‌دهنده[19]: دستور تطبیق‌دهنده یک چارچوب مدل‌سازی است که از مدل بیزی غیرپارامتری[20] استفاده و تلاش می‌کند ساخت‌های درختی را برروی رشته نویسه‌های داده ورودی جستجو کرده، قواعد ساختواژی با پیچیدگی‌های گوناگون را بیاموزد (سرتز و گلدواتر[21] ۲۰۱۳).

۳- میدان تصادفی شرطی: روشی برپایه مدل گرافیکی احتمالاتی غیرمستقیم تمایزی[22] است که در آن مدل به‌منظور یادگیری ارتباط بین خروجی الگوریتمی بی‌نظارت مانند مورفسور پایه‌ای با بافت زیررشته‌ای پیرامونش آموزش می‌بینند. در این رویکرد ایده کلی تلاش و تمرکز بر مدل‌کردن مرزهای تک‌واژ به جای شناسایی رشته نویسه‌ها است که در چارچوب مسئله‌ای از دنباله برچسب‌ها مطرح می‌شود (روکولینن[23] و همکاران، ۲۰۱۴).

---

[10] Supervised
[11] Morfessor
[12] Kurimo
[13] Generative probabilistic models
[14] Joint probability
[15] Bayesian estimation
[16] Maximum a posteriori
[17] Maximum likelihood
[18] Minimum description length
[19] Adaptor grammar
[20] Non-parametric Bayesian model
[21] Sirts & Goldwater
[22] Discriminative undirected probabilistic graphical model
[23] Ruokolainen

۴- گوناگونی پسین حروف[۲۴](LSV) و انتروپی پسین حروف[۲۵](LSE): مبنای این روش استفاده از خواص توزیع حروف در داخل واژه است. در این مدل گوناگونی پسین به حروفی اشاره دارد که پس از هر حرف در واژه قرار می‌گیرد و گوناگونی پیشین، حروفی است که پیش از آن قرار می‌گیرد. بنابراین نقطه‌ای که دارای پتانسیل تقطیع بالایی است را می‌توان با توجه به تغییرات قابل ملاحظه‌ای که در شماره پسین‌های حروف، در داخل واژه رخ می‌دهد، شناسایی کرد. LSE نیز روشی برپایه استفاده از انتروپی حروف پسین و پیشین به جای شمارش آنها است. مرزهای تکواژها نوعا دارای LSE بالاتری نسبت به سایر نقاط در واژه هستند (برداگ[۲۶]، ۲۰۰۷).

در حوزه پردازش ساختواژی زبان فارسی اغلب پژوهش‌ها بر روش‌های قاعده‌مند متمرکز بوده و از داده واژگان زایا (اسلامی و همکاران، ۱۳۸۳) استفاده کرده‌اند. این پردازشگرها قادر به پردازش کلمات خارج از واژگان از جمله کلمات تخصصی و ساخت‌های گونه محاوره نیستند (مگردومیان، ۲۰۰۰؛ شمس‌فرد و جعفری، ۲۰۱۰؛ مواجی و همکاران، ۱۳۹۰؛ سرابی و همکاران، ۲۰۱۳). به‌طور محدود مواجی و همکاران (۱۳۹۱) پیاده‌سازی یک الگوریتم یادگیری ساختواژی بی‌نظارت از خانواده مورفسور را برروی داده‌های زبان فارسی گزارش کرده‌اند، اما اشاره‌ای به داده آموزش و آزمون مورد استفاده و نحوه ارزیابی و دقت به‌دست آمده نداشته‌اند.

۳. تقطیع ساختواژی داده‌های زبان فارسی با استفاده از تخمین بیزی

با توجه به کاربرد گسترده تخمین بیزی در حوزه پردازش ساختواژی داده‌محور و دقت‌های بالایی که در پژوهش‌های گوناگون ثبت کرده است (روکولینن و همکاران، ۲۰۱۶؛ کوریمو و همکاران، ۲۰۱۰b)، یادگیری ساختواژی الگوریتم مورفسور پایه‌ای گسترش‌یافته (ویرپیوا و اسمیت، ۲۰۱۵؛ ویرپیوا و همکاران، ۲۰۱۳) بر روی داده‌های زبان فارسی ارزیابی شد.

این الگوریتم مطابق رابطه (۲) از تابع هزینه MAP استفاده و تلاش می‌کند به مدلی بهینه برسد که تابع هزینه را کمینه کند.

$$L(\theta, D_W) = -\log P(\theta) - \log P(D_W|\theta) \qquad (۲)$$

در تخمینMAP فرآیند یادگیری مستلزم یک گمانه یا پیش‌فرض در مورد پارامترهای مدل است که احتمال پیشین[۲۷] $P(\theta)$ خوانده می‌شود. تخمین پارامترهای مدل یا احتمال پسین[۲۸]، در چارچوب آمار بیزی، محصول احتمال پیشین و تخمین شباهت داده به شرط داشتن پارامترهای مدل $P(D_W|\theta)$ است.

[24] Letter successor variety
[25] Letter successor entropy
[26] Bordag
[27] Prior probability
[28] Posterior probability

مدل θ فهرستی از تک‌واژهای زبان است که در داده D_W شناسایی و استخراج شده است. پارامترهایی که در این مدل تعریف و ارزیابی می‌شود، شامل خواص ظاهری هر تک‌واژ و خصوصیات کلی واژگان یا مدل است که به‌صورت زیر تعریف می‌شود:

- تعداد گونه‌های تک‌واژها یا به عبارتی اندازه واژگان μ
- تعداد قطعات مشاهده شده در داده γ
- رشته‌های حروف در تک‌واژها σ_i $(\sigma_1 ، \sigma_2 ، \ldots ، \sigma_\mu)$
- تعداد کل تک‌واژها و رخداد آن‌ها τ_i $(\tau_1 ، \tau_2 ، \ldots ، \tau_\mu)$

برای توزیع احتمال تک‌واژها از پیش‌فرضی استفاده می‌شود که حاصل اشتراک دو متغیر مستقل برمبنای شکل ظاهری[29] تک‌واژ، رابطه (۳) و کاربرد[30] آن در پیکره، رابطه (۴) است. به این ترتیب احتمال بالاتر به تک‌واژهایی با طول کمتر اختصاص می‌یابد.

$$P(\sigma_i) = P(|\sigma_i|) \prod_{j=1}^{|\sigma_i|} P(C = \sigma_{ij}) \qquad (۳)$$

$$P(\tau_1, \ldots, \tau_\mu | \mu, \gamma) = 1 / \binom{\gamma - 1}{\mu - 1} \qquad (۴)$$

در آموزش، استخراج تک‌واژها و تخمین هزینه مدل، از یک الگوریتم حریصانه بازگشتی[31] استفاده می‌شود که در هر جستجو پس از دستیابی به هر نقطه بهینه محلی، پارامترهای مدل را به‌روزرسانی می‌کند. می‌توان با وزن‌دهی نقشی را که تابع پیشین در هزینه نهایی مدل ایفا می‌کند، کنترل کرد.

$$L(\theta, D_W) = -\alpha \log P(\theta) - \log P(D_W | \theta) \qquad (۵)$$

در یادگیری نیمه‌نظارتی علاوه‌بر داده برچسب‌نخورده به حجم اندکی داده برچسب‌خورده نیاز است و لازم است در تخمین هزینه مدل، علاوه‌بر لگاریتم شباهت داده برچسب‌نخورده D_W، لگاریتم شباهت داده برچسب‌خورده D_{W-An} نیز افزوده شود.

$$L(\theta, D_W, D_{W-An}) = -\log P(\theta) - \log P(D_W | \theta) - \log(D_{W-An} | \theta) \qquad (۶)$$

۳-۱. تهیه داده حاشیه‌نویسی‌شده ساختواژی

از چالش‌های اصلی پردازش ساختواژی زبان فارسی با استفاده از الگوریتم‌های داده‌محور، تهیه داده مناسب است. به‌منظور آموزش نیمه‌نظارتی و آزمون الگوریتم‌ها نخست می‌بایست داده استاندارد حاشیه‌نویسی‌شده برای تقطیع ساختواژی زبان فارسی فراهم شود که تا حد امکان چشم‌انداز قابل قبولی از واژه‌ها و تجزیه و تحلیل ساختواژی هرکدام ارائه دهد و ساخت‌ها و اشکال واژه‌ای[32] زبان اعم از تصریفی و اشتقاقی و ترکیبی را به‌خوبی پوشش دهد. به این منظور برای تهیه داده استاندارد ساختواژی زبان فارسی از داده‌های مجموعه استاندارد وبلاگ‌های ایران (آل‌احمد و همکاران، ۲۰۱۶) و پیکره متنی زبان فارسی (بی‌جن‌خان و همکاران، ۲۰۱۱) استفاده شد. به دلیل تنوع نویسه‌های مورد استفاده در متون الکترونیکی فارسی و نیز گوناگونی قواعد

[29] Form
[30] Usage
[31] Recursive, greedy search
[32] Word forms

و روش‌های املایی در نگارش ساخت‌های مرکب و وندهای اشتقاقی و تصریفی، لازم است پیش از تقطیع واحدها، متون بهنجارسازی[۳۳] شود. با توجه به اینکه، بهنجارساز[۳۴]های موجود مانند هضم (۱۳۹۳) و ویراستیار (۱۳۹۳) هرکدام تعداد محدودی از نویسه‌های یونیکد، ساخت‌های اشتقاقی و ترکیبی و نیز وندهای تصریفی و واژه‌بست‌های احتمالی متصل به آن‌ها را بهنجار می‌کنند، در این پژوهش برای بهنجارسازی متون، بهنجارساز هضم بهبود داده شد و با اضافه کردن موارد زیر کاستی‌های مورد نظر رفع گردید:

- صد و ده نویسه یونیکد غیراستاندارد که در گذشته برای تایپ فارسی استفاده می‌شد و وبلاگ‌نویسان استفاده می‌کنند و معادل استاندارد آن‌ها به بهنجارساز افزوده شد. مثلا حرف «ی» در زبان فارسی دارای یک نویسه استاندارد یونیکد «\u06cc» و نویسه‌های غیراستانداردی مثل «\ufbaf» «\u06d2» «\ufbff» و غیره، است.
- بهنجارسازی پی‌بست‌های «ی» نکره، کسره اضافه، ربطی و کمکی و ضمیری، مانند: «خانه اش، خانه ست، خانه ی، خانه ای» و وندهای تصریفی متصل به اسم و صفت «ها - تر - ترین - جات» و قطعاتی[۳۵] که شامل وندهای تصریفی و پی‌بست‌های متصل به آن‌ها باشد، مانند: «های: کتابخانه های - هایمان:کتابخانه هایمان و...».
- بهنجارسازی فعل‌های چندجزئی که بیانگر زمان و نمودهای گوناگون فعل هستند مانند ماضی استمراری.
- بهنجارسازی بیش از هفتاد وند و ساخت ترکیبی با توجه به پژوهش‌های سراجی (۲۰۱۳) و اضافه کردن برخی وندهای دیگر از جمله «وش، سالگی، گانگی و...» به همراه وندهای تصریفی و پی‌بست‌های احتمالی متصل به هرکدام مانند: « ده سالگیش – تلاش گری».
- اصلاح درج اشتباه نویسه نیم‌فاصله در ابتدا یا انتهای واژه یا بیش از یک نویسه در فاصله تک‌واژها. اصلاح واژه‌های دارای تکرار حروف مانند«سلاااام» که به عنوان تاکید در صفحات وبلاگی و شبکه‌های اجتماعی رایج است.

بعد از بهنجارسازی و حذف علائم نگارشی و نویسه‌ها و سایر حروف غیرفارسی و نیز جملات و پست‌های دارای ساختارهای محاوره‌ای، فهرست واژه‌ها از متون پیکره متنی زبان فارسی و داده‌های وبلاگی ایران به همراه فراوانی هر واژه استخراج شد. همچنین بخشی از ساخت‌های غیرفارسی از زبان‌هایی مانند ترکی، کردی و عربی که دارای نویسه‌های مشترک با فارسی هستند، به‌صورت نیمه‌خودکار حذف شد. با توجه به حجم بالای غلط‌های املایی و تایپی، در واژگان فراوانی کمتر از ۵ که بیشترین فراوانی غلط‌های املایی را در خود داشت حذف شده، در نهایت از فهرست واژه‌های یکتا با حجم ۴۸۲۰۳۰ قطعه، ۵۰۰۰ واژه به‌صورت تصادفی استخراج و به‌صورت نیمه‌خودکار حاشیه‌نویسی شد. جهت حفظ تنوع و تناسب واژه‌های داده ساختواژی با واژه‌های موجود در پیکره، نخست واژه‌ها برحسب فراوانی دسته‌بندی شدند و سپس از هر دسته با توجه به اینکه چند درصد از کل پیکره را شامل می‌شود، تعدادی واژه به‌صورت تصادفی، با درنظر گرفتن تناسب مذکور، استخراج گردید. به طوریکه داده ساخت‌های پرتکرار، کم‌تکرار و با تکرار متوسط را متناسب

[33] Normalize
[34] Normalizer
[35] Tokens

با فراوانیشان در پیکره پوشش دهد. جدول ۱، درصد حجمی هر دسته از فراوانی‌ها نسبت به حجم کل قطعات یکتا در پیکره را نشان می‌دهد.

جدول ۱. درصد حجمی هر دسته از فراوانی‌ها در مجموع دو پیکره

فراوانی قطعات یکتا	۵ تا ۲۰	۲۰ تا ۵۰	۵۰ تا ۱۰۰	۱۰۰ تا ۱۰۰۰	بیش از ۱۰۰۰
درصد حجمی هر دسته در دو پیکره	٪۶/۶	٪۳/۵	٪۳/۶	٪۲۱	٪۶۴/۴

۳-۲. تهیه داده آموزش ساختواژی بی‌نظارت

داده آموزش بی‌نظارت فهرستی از اشکال و ساخت‌های واژه‌ای گوناگون و صحیح زبان است که حجمی بین چندصدهزار تا بیش از یک‌میلیون واژه دارد (روکولینن و همکاران، ۲۰۱۶؛ کوریمو و همکاران، ۲۰۱۰b). با توجه به اینکه چنین فهرستی برای زبان فارسی موجود نیست، داده آموزش از پیکره‌های موجود زبان تهیه شد.

- پیکره متنی زبان فارسی با حجمی حدود ۱۰۰میلیون قطعه
- مجموعه داده استاندارد وبلاگ‌های ایران با حجمی حدود ۴۰۰میلیون قطعه
- پیکره همشهری (آل‌احمد و همکاران، ۲۰۰۹) با حجمی حدود ۱۵۰میلیون قطعه

پس از پیش‌پردازش اولیه شامل جداسازی ساخت‌های محاوره، حذف نویسه‌های غیرفارسی، اعداد و علائم نگارشی، بهنجارسازی کدگذاری متون، اصلاح نویسه نیم‌فاصله و حذف برخی ساخت‌های پرتکرار زبان‌های دیگر مانند کردی و عربی و غیره در داده، که در بخش پیش توضیح داده شد، واژگان هر پیکره استخراج شد. داده‌های به‌دست آمده از پیکره‌ها، نسبت به داده استاندارد آموزشی که در الگوریتم‌های یادگیری ساختواژی استفاده می‌شود، کاستی‌هایی دارد که می‌توان در موارد زیر دسته‌بندی کرد:

- حجم داده صحیح: تعداد اشکال و ساخت‌های واژه‌ای گوناگون و صحیح زبان. این حجم نسبت به حجم داده آموزش مورد استفاده در الگوریتم‌ها کمتر است.
- نویز داده: غلط‌های املایی و ساخت‌های نادرست زبان در داده یا واژگان. در داده‌های مورد بررسی نویز داده بیشتر ناشی از سرهم‌نویسی گروهای اسمی و حرف‌اضافه‌ای، و نیز ساختارهایی از دیگر زبان‌ها است که نویسه‌های مشابه با زبان فارسی دارند مانند ترکی، عربی و غیره.
- تنوع شیوه‌های املایی: گوناگونی روش‌های املایی و شیوه‌های نگارش واژه‌های قرضی و ساخت‌های تصریفی، اشتقاقی و ترکیبی به عنوان مثال: «کتابخانه، کتاب‌خانه؛ خوبترین، خوب‌ترین».
- نویسه نیم‌فاصله: نویسه پرکاربرد غیرزبانی و صرفا تایپی در متون الکترونیکی زبان فارسی که در الگوریتمی مانند مورفسور به عنوان یک نویسه زبانی ارزیابی می‌شود. این نویسه هم در واژه‌های غیربسیط فارسی و هم در واژه‌های بسیط قرضی به کار می‌رود مانند: «جهان‌دیده» و «اس‌ام‌اس». در واژه‌های غیربسیط نمایانگر مرز تک‌واژ است، اما در ساخت‌ها قرضی بسیط، جزئی از تک‌واژ به‌حساب می‌آید.

۴. ارزیابی

کارایی مدل‌های ساخته‌شده با سه معیار صحت[۳۶]، فراخوانی[۳۷] و معیار F[۳۸]، برمبنای BPR[۳۹] یا تشخیص درست یا نادرست مرز تک‌واژ در واژه، ارزیابی شد. این روش، از رایج‌ترین روش‌های ارزیابی الگوریتم‌های پردازش ساختواژی داده‌محور است و به‌صورت زیر محاسبه می‌شود (کوریمو، ۲۰۰۷):

- صحت (P): نسبت تعداد مرزهای تک‌واژی صحیحی که توسط مدل تشخیص داده شده، به کل مرزهای تک‌واژی یافت شده
- فراخوانی (R): نسبت تعداد مرزهای تک‌واژی صحیحی که توسط مدل تشخیص داده شده به کل مرزهای درست تک‌واژی در داده
- معیار F (F1): $2 * \frac{P*R}{P+R}$

در ارزیابی مدل‌های ساخته‌شده از داده ثابت آزمون استفاده شده که مجموعه‌ای ثابت از ۳۰۰۰ واژه است و به‌طور تصادفی از داده حاشیه‌نویسی‌شده استاندارد انتخاب شده است.

۴-۱. ارزیابی یادگیری ساختواژی بی‌نظارت

به‌منظور یافتن پارامترهایی که در بهبود یادگیری الگوریتم با توجه به داده‌های موجود موثر باشد، نقش حجم، نویز، طول واژه‌ها و نویسه نیم‌فاصله در داده، مورد ارزیابی قرار گرفت.

۴-۱-۱. حجم و نویز داده

جدول ۲، نتایج یادگیری بی‌نظارت الگوریتم را برای هر سه مجموعه داده با بررسی نویز و حجم در هر مجموعه، نشان می‌دهد.

برای تخمین نویز داده از هر مجموعه با فراوانی بیشتر از یک، ۱۰۰۰ واژه تصادفی به‌صورت ۸۰ درصد با فراوانی ۱۵ و کمتر و ۲۰ درصد با فراوانی بیش از ۱۵ انتخاب شد و با شمارش تعداد اشتباهات املایی و تایپی و ساختارهای غیرفارسی برای هر مجموعه، درصد نویز تخمین زده شد. هدف از این تخمین صرفا بررسی نسبی داده‌ها است و معیاری برای سنجش قطعی نویز داده نیست. با توجه به اینکه واژه‌ها و قطعاتی با فراوانی ۱ بیشترین درصد نویز داده را به خود اختصاص داده‌اند، در آموزش الگوریتم از واژه‌ها یا قطعات یکتایی استفاده شد که در واژگان فراوانی بیش از ۱ داشتند.

[36] Precision
[37] Recall
[38] F-score
[39] Boundary precision and recall

جدول ۲. یادگیری ساختواژی بی‌نظارت برروی سه مجموعه داده.

صحت: P، فراخوانی: R، معیار F: F1

	حجم قطعات یکتا	قطعات یکتا با فراوانی بیشتر از ۱	نویز داده	واژه کمتر از ۵ نویسه	P	R	F1
پیکره متنی	۸۱۱۹۱۱	۴۰۴۹۰۷	٪ ۲۸	٪۲۰/۷	۰/۷۵	۰/۵۷	۰/۶۵
داده‌های وبلاگی	۱۸۲۸۹۶۲	۷۲۴۳۸۵	٪ ۳۳	٪۲۳/۳	۰/۷۶	۰/۵۴	۰/۶۳
پیکره همشهری	۵۸۰۸۷۷	۲۵۳۷۷۱	٪۲۲	٪۲۸/۵	۰/۸۳	۰/۵۲	۰/۶۴

هرچند حجم داده همشهری به‌طور قابل ملاحظه‌ای از دو داده دیگر کمتر است در معیارF نتایجی مشابه دو داده دیگر و در معیار صحت بیش از ۷ درصد بالاتر ثبت کرده است. جدول ۲، درصد واژه‌ها با طول ۵ نویسه و کمتر را برای هر سه مجموعه داده، نشان می‌دهد این حجم در داده همشهری بیش از ۵ درصد از داده وبلاگی و نزدیک ۸ درصد از داده پیکره متنی بیشتر است. طول ۵ نویسه با توجه به متوسط طول تک‌واژهای زبان در داده حاشیه‌نویسی‌شده استاندارد انتخاب شده است. الگوریتم در فرآیند مدل‌سازی نخست تمام واژه‌ها را به عنوان تک‌واژ به مدل وارد کرده، سپس تلاش می‌کند با جستجوی محلی برروی امکانات تقطیع واژه‌ها، مدل کم‌هزینه‌تر را انتخاب کند. به دلیل هزینه کم‌تری که به مدل تحمیل می‌شود، تمایل الگوریتم به تقطیع واژه‌های با طول کمتر نسبت به واژه‌های با طول بیشتر، کمتر است و این موضوع درصد تقطیع‌های نادرست در مدل را کاهش می‌دهد.

۴-۱-۲. نویسه نیم‌فاصله

به‌منظور بررسی کارایی نیم‌فاصله، در آزمون بعد، این نویسه از داده حذف و در مرحله پیش‌پردازش با نویسه فاصله جایگزین شد. جدول ۳، حجم هر مجموعه داده بدون نیم‌فاصله و نتایج حاصل از یادگیری الگوریتم را نشان می‌دهد. با توجه به اینکه داده همشهری فاقد نویسه نیم‌فاصله است و وندها و ساخت‌های ترکیبی یا متصل به ستاک نوشته شده، مانند «کتابها» یا با نویسه فاصله از ستاک جدا شده‌اند، مانند «کتاب ها»، در این مرحله داده همشهری و نتایج آن تغییری با مرحله قبل ندارد. در مورد داده‌های دو پیکره دیگر، بهبود در نتایج معیار F مشاهده می‌شود که بیشتر ناشی از بهبود در فراخوانی است.

نکته قابل توجه در نتایج به‌دست آمده در این مرحله این است که حذف نیم‌فاصله هم سبب کاهش بخشی از نویز داده، مربوط به استفاده از نویسه نیم‌فاصله در بین اجزای یک گروه نحوی مانند «آسمان‌وزمین» می‌شود و هم سبب می‌شود برخی از وندهای پرکاربرد اشتقاقی و تصریفی از ستاک‌ها جدا شده، با فراوانی بالا در مدل اولیه قرار گیرد. به این ترتیب احتمال شناسایی و قرارگیری آن‌ها را در مدل نهایی افزایش می‌یابد.

نقش دیگر نویسه نیم‌فاصله در مرحله تقطیع واژه است. در صورتی که بین تک‌واژها نویسه نیم‌فاصله درج شده باشد و الگوریتم تک‌واژ را به‌درستی تشخیص دهد، این نویسه به عنوان بخشی از تک‌واژ قبلی یا بعدی یا به‌صورت یک تک‌واژ مستقل شناسایی می‌شود. این مسئله مغایرتی در خروجی الگوریتم نسبت به داده استاندارد آزمون ایجاد می‌کند. با توجه به نبود این نویسه در داده آموزش لازم است در مرحله تقطیع نیم‌فاصله به عنوان یک نویسه غیرزبانی به الگوریتم معرفی و از درج اشتباه آن در خروجی جلوگیری شود.

معرفی الگوی نیم‌فاصله، نتایج را در معیارF و صحت به‌طور متوسط نزدیک به ۷ درصد بهبود داده، که در جدول ۳، مشاهده می‌شود.

جدول ۳. ارزیابی یادگیری الگوریتم با داده فاقد نویسه نیم‌فاصله و معرفی الگوی‌نیم‌فاصله

. صحت: P، فراخوانی: R، معیار F: F1

	حذف نیم‌فاصله از داده آموزش				حذف نیم‌فاصله و معرفی الگوی نیم‌فاصله		
	حجم داده	P	R	F1	P	R	F1
پیکره متنی	۲۴۵۸۴۶	۰/۷۷	۰/۶۳	۰/۶۹	۰/۸۶	۰/۶۶	۰/۷۵
داده‌های وبلاگی	۵۴۵۸۹۸	۰/۷۷	۰/۵۷	۰/۶۶	۰/۸۸	۰/۵۸	۰/۷۰
پیکره همشهری	ــــــــــ				۰/۸۳	۰/۶۷	۰/۷۴

۴-۱-۳. وزن‌دهی به تابع احتمال پیشین

جدول ۴، بهبود نتایج را با وزن‌دهی به پارامتر α نشان می‌دهد.

جدول۴. نتایج حاصل از یادگیری الگوریتم با وزن بهینه α و معرفی الگوی نیم‌فاصله

	وزن	صحت	فراخوانی	معیارF
پیکره همشهری	۰/۷۳	۰/۸۷	۰/۷۱	۰/۷۸
داده وبلاگی بدون نیم‌فاصله	۰/۷۹	۰/۸۲	۰/۶۷	۰/۷۴
پیکره متنی بدون نیم‌فاصله	۰/۷۴	۰/۸۶	۰/۷۲	۰/۷۹

به‌منظور تعیین وزن بهینه از داده توسعه‌یافته[40] استفاده شد که شامل ۷۵۰ واژه است که به‌صورت تصادفی و بدون هم‌پوشانی با داده آموزش و آزمون، از داده حاشیه‌نویسی‌شده انتخاب شده است. پس از یادگیری با وزن بهینه، نتایج با استفاده از داده آزمون ارزیابی شد که به‌طور متوسط ۶/۳ درصد بهبود در فراخوانی و ۴ درصد بهبود در معیارF مشاهده می‌شود.

۴-۲. ارزیابی یادگیری ساختواژی نیمه‌نظارتی

در یادگیری نیمه‌نظارتی الگوریتم با ۱۰۰۰ واژه که به‌صورت تصادفی و بدون هم‌پوشانی با داده آزمون، از داده حاشیه‌نویسی‌شده استخراج شد، آموزش دید. جدول ۵، بهبود نتایج یادگیری نیمه‌نظارتی را نسبت به یادگیری بی‌نظارت، نشان می‌دهد. در یادگیری نیمه‌نظارتی نیز معرفی الگو و حذف نیم‌فاصله با بهبود صحت و فراخوانی نتایج را در معیارF به‌طور چشمگیری بهبود داده است. با توجه به اینکه حجم ۱۰۰۰ واژه از داده حاشیه‌نویسی‌شده تقریبا تمام وندهای تصریفی، واژه‌بست‌ها و وندهای اشتقاقی پرکاربرد زبان را پوشش می‌دهد، سبب بهبود مدل ساخته‌شده و افزایش قابل توجه فراخوانی می‌شود. بهبود متوسط ۱۰ درصدی در فراخوانی و ۲/۷ درصدی در معیارF برای یادگیری نیمه‌نظارتی با معرفی الگو و حذف نیم‌فاصله نسبت به یادگیری بی‌نظارت ثبت شده است.

[40] Held out set

جدول ۵. ارزیابی یادگیری نیمه‌نظارتی. صحت: P، فراخوانی: R، معیار F: F1

	یادگیری نیمه‌نظارتی			یادگیری با حذف نیم‌فاصله از داده و معرفی الگوی نیم‌فاصله		
	P	R	F1	P	R	F1
پیکره متنی	۰/۷۱	۰/۷۳	۰/۷۲	۰/۸۵	۰/۸۰	۰/۸۲
داده‌های وبلاگی	۰/۶۵	۰/۷۵	۰/۶۹	۰/۸۲	۰/۸۰	۰/۸۱
پیکره همشهری	۰/۷۳	۰/۷۹	۰/۷۶	۰/۸۵	۰/۸۱	۰/۸۳

۵. جمع‌بندی و نتیجه‌گیری

در این پژوهش به ارزیابی یادگیری ساختواژی زبان فارسی با استفاده از تخمین بیزی پرداخته و داده استاندارد ساختواژی جهت ارزیابی مدل‌ها و آموزش نیمه‌نظارتی برای زبان فارسی طراحی و تهیه شد. با توجه به دردسترس نبودن داده استاندارد آموزش، داده برچسب‌نخورده از سه پیکره بزرگ زبان فارسی انتخاب و یادگیری الگوریتم با توجه به حجم، نویز، طول واژه و نقش نویسه نیم‌فاصله در داده مورد ارزیابی قرار گرفت. حذف نیم‌فاصله از داده و معرفی الگوی نیم‌فاصله به الگوریتم تقطیع، برای جلوگیری از درج اشتباه نیم‌فاصله در تک‌واژها، نتایج را در هر سه معیار صحت، فراخوانی و معیارF به‌طور قابل ملاحظه‌ای بهبود می‌دهد و متوسط نتایج ثبت شده در معیارF برای سه مجموعه داده با اختلاف ۳ درصد در یادگیری بی‌نظارت و ۲ درصد در یادگیری نیمه‌نظارتی به نتایجی که پژوهش‌های دیگر برای زبان انگلیسی ثبت کرده‌اند (روکولینن و همکاران، ۲۰۱۶) نزدیک می‌شود. از دلایل بهبود یادگیری الگوریتم، کوتاه شدن طول واژه‌های داده، معرفی تک‌واژهای بیشتر به الگوریتم، کاهش تنوع الگوهای نگارشی و کاهش نویز در مواردی است که نیم‌فاصله در ساخت گروه‌های نحوی مانند «آسمان‌وزمین» وارد شده است. با توجه به نتایج به‌دست آمده، ارزیابی یادگیری الگوریتم با داده نوشتار غیررسمی مانند متون شبکه‌های اجتماعی، می‌تواند زمینه‌ای برای پژوهش بیشتر باشد.

منابع

اسلامی، محرم.، شریفی آتشگاه، مسعود.، علیزاده، صدیقه.، زندی، طاهره. (۱۳۸۳). «واژگان زایای زبان فارسی». *اولین کارگاه آموزش زبان فارسی و رایانه*، دانشگاه تهران.

مواجی، وحید.، اسلامی، محرم.، وزیرنژاد، بهرام. (۱۳۹۰). «پارس‌مورف: تحلیلگر ساختواژی زبان فارسی». *دوفصلنامه پردازش علائم و داده‌ها*(۱۰).

مواجی، وحید.، اسلامی، محرم.، وزیرنژاد، بهرام. (۱۳۹۱). «پارس‌مورف: تحلیلگر ساختواژی زبان فارسی». پایان‌نامه کارشناسی ارشد،*گروه زبان‌شناسی رایانشی، دانشگاه صنعتی شریف*

ویراستیار. (۱۳۹۳). [نرم‌افزار]. *دبیرخانه شورای عالی اطلاع‌رسانی*. برگرفته از: http://virastyar.ir.

هضم (۱۳۹۳). [نرم‌افزار]. برگرفته از: http://www.sobhe.ir/hazm.

AleAhmad, A., Amiri, H., Darrud, E., Rahgozar, M., & Oroumchian, F. (2009).“*Hamshahri: A standard Persian text collection”. Knowledge-Based Systems*, 22 (5), 382-387.

AleAhmad, A., Zahedi, M., Rahgozar, M., & Moshiri, B. (2016). "irBlogs: A standard collection for studying Persian bloggers". *Computers in Human Behavior*, 57, 195-207.

Bijankhan, M,. Sheykhzadegan, J., Bahrani, M., & Ghayoomi, M. (2011). "Lessons from building Persian written corpus: Peykare". *Language Resources and Evaluation*, 45(2), 143-164.

Bordag, S. (2007, September). "Unsupervised and knowledge-free morpheme segmentation and analysis. In *Workshop of the Cross-Language Evaluation Forum for European Languages"*,.881-891. Berlin,: Springer.

Jurafsky, D., & Martin ,H.(2007). *Speech and language processing An introduction to natural language processing, computational linguistics, and speech recognition.* New Jersey: Prentice-Hall.

Kurimo, M., Creutz, M., & Varjokallio, M. (2007, September). Morpho challenge evaluation using a linguistic gold standard. "*Workshop of the Cross-Language Evaluation Forum for European Languages"* 864-872. Berlin, Heidelberg: Springer.

Kurimo, M., Virpioja, S., & Turunen, V. T. (2010a). "Proceedings of the Morpho challenge 2010 workshop". *Morpho Challenge Workshop; 2010; Espoo*. Aalto University School of Science and Technology.

Kurimo, M., Virpioja, S., Turunen, V., & Lagus, K. (2010b). "Morpho Challenge competition 2005-2010: evaluations and results". *Proceedings of the 11th Meeting of the ACL Special Interest Group on Computational Morphology and Phonology*,87-95. Association for Computational Linguistics.

Megerdoomian, K. (2000). "Persian computational morphology: A unification-based approach". *Computing Research Laboratory, New Mexico State University*.

Ruokolainen, T., Kohonen, O., Sirts, K., Grönroos, S. A., Kurimo, M., & Virpioja, S. (2016). "A comparative study minimally supervised morphological segmentation". *Computational Linguistics,* 42(1), 91-120

Ruokolainen, T., Kohonenb, O., Virpiojab, S., & Kurimoa, M. (2014). "Painless semi-supervised morphological segmentation using conditional random fields". *Proceedings of th 14th Conference of European Chapter of the Association for Computational Linguistics*, 84-89.

Sarabi, Z., Mahyar, H., & Farhoodi, M. (2013, October). "ParsiPardaz: Persian language processing toolkit". *Computer and Knowledge Engineering (ICCKE)*, 73-79.IEEE.

Seraji, M. (2013). "Preper: A pre-processor for Persian, doctoral dissertation", *Department of Linguistics and Philology, Uppsala University*.

Shamsfard, M., Jafari, H. S., & Ilbeygi, M. (2010, May). "STeP-1: A set of fundamental tools for Persian text processing". *LREC*.

Sirts, K., & Goldwater, S. (2013). "Minimally-supervised morphological segmentation using adaptor" grammars. *Transactions of the Association for Computational Linguistics, 1*, 255-266.

Virpioja, S.,& Smit, P.(2015). *Morfessor documentation*. Retrieved from the Web May, 2017. https://morfessor.readthedocs.io/en/latest.

Virpioja, S., Smit, P., Grönroos, S., & Kurimo, M. (2013). *Morfessor 2.0: Python implementation and extensions for Morfessor Baseline*. Retrieved from the Web May, 2017. http://www.cis.hut.fi/projects/morpho.

Virpioja, S., Turunen, V. T., Spiegler, S., Kohonen, O., & Kurimo, M. (2011). "Empirical comparison of evaluation methods for unsupervised learning of morphology". *TAL, 52*(2), 45-90.