تشخیص زبان در شبکه‌های اجتماعی

ندا ناصری[۱]، مصطفی صالحی[۲]، محمود بی‌جن‌خان[۳]، هادی ویسی[۴]، وحید رنجبر[۵]

[۱] دانش‌آموخته کارشناسی ارشد زبان‌شناسی رایانشی، دانشکده علوم و فنون نوین، دانشگاه تهران،
nasseri.n@ut.ac.ir

[۲] استادیار گروه بین‌رشته‌ای فناوری، دانشکده علوم و فنون نوین، دانشگاه تهران،
Mostafa_salehi@ut.ac.ir

[۳] استاد گروه زبان‌شناسی، دانشکده ادبیات و علوم انسانی، دانشگاه تهران،
mbjkhan@ut.ac.ir

[۴] استادیار گروه بین‌رشته‌ای فناوری، دانشکده علوم و فنون نوین، دانشگاه تهران،
h.veisi@ut.ac.ir

[۵] دانشجوی دکتری، دانشکده علوم و فنون نوین، دانشگاه تهران،
vranjbar@ut.ac.ir

چکیده

با فراگیر شدن اینترنت و وب حجم زیادی از متون که به هر زبانی نوشته می‌شوند در دسترس است. مسئله تشخیص زبان یکی از مراحل اصلی برای هر نوع پردازش دیگر بر روی متن است. تاکنون پژوهش‌هایی بر روی تشخیص زبان در متون انگلیسی انجام شده اما کارها روی زبان فارسی در این حوزه محدود است. از طرفی با گسترش رسانه‌های اجتماعی متون محاوره بیشتر مورد استفاده قرار می‌گیرند که روش‌های تشخیص زبان ارائه شده برای زبان رسمی دقت خوبی برای این نوع متون ندارند. در این مقاله روشی جهت تشخیص زبان محاوره ارائه شده و تمرکز بر تشخیص زبان فارسی و زبان‌های با رسم‌الخط مشابه یعنی کردی مرکزی، عربی، پشتو و اردو و فضای مورد مطالعه، شبکه‌های اجتماعی است. فرآیند طراحی شده از یک مرحله سنجش آماری (الگوی زبانی مبتنی بر Nتایی‌ها) و دو مرحله غیرآماری (نویسه‌های ویژه و ایست‌واژه‌ها) تشکیل شده است. برای ارزیابی روش پیشنهادی مجموعاً تعداد ۱۰۰۰ جمله از هر پنج زبان به صورت تصادفی از پیام‌های شبکه‌های اجتماعی استخراج شده است. نتایج به دست آمده از ارزیابی روش پیشنهادی بر روی مجموعه داده جمع‌آوری شده نشان می‌دهد که روش پیشنهادی که تلفیق روش آماری و غیرآماری است بهبود قابل‌توجهی نسبت به روش آماری به تنهایی داشته است. همچنین نتایج به دست آمده نشان‌دهنده عملکرد خوب روش پیشنهادی در مقایسه با ابزارهای قدرت‌مندی مانند گوگل و زیراکس است. این ابزار همچنین توانایی تشخیص زبان پشتو و کردی مرکزی را دارد که برای سایر ابزارهای معروف تشخیص زبان، قابل شناسایی نیست.

کلمات کلیدی

پردازش زبان طبیعی، تشخیص زبان، الگوسازی زبانی، شبکه‌های اجتماعی، زبان فارسی محاوره

۱- مقدمه

امروزه با افزایش حجم اطلاعات در دسترس روی اینترنت –که هر یک به زبانی نگاشته شده‌اند– نیاز فوق‌العاده به ابزاری که توانایی تشخیص زبان[۱] را داشته باشد کاملاً محسوس است. تشخیص زبان، فرآیند نگاشت یک سند به زبانی است که سند به آن نوشته شده است [۱]. به طور خلاصه می‌توان گفت ابزار تشخیص زبان باید بتواند با توجه به الگوی زبانی که به آن شناسانده شده، و شناختی که از قواعد و چگونگی وقوع انواع پاره‌گفتار دارد از بین کاندیدهای مختلف، زبان متن مورد نظر را به درستی تعیین کند.

در متون استخراج شده از شبکه‌های اجتماعی، انواع غلط‌های املایی، نوواژه‌ها و مخفف‌ها به چشم می‌خورد. گاهی کاربران این شبکه‌ها به دلیل محدودیت فضای در دسترس یا زمان و حتی گاهی نیز از سر تفنن و یا خلاقیت، اقدام به تغییر ظاهر واژه‌های مورد استفاده و یا خلق واژه‌های جدید می‌کنند. بدین ترتیب مشاهده می‌شود که آنچه که در شبکه‌های اجتماعی به

عنوان زبان فارسی به کار برده می‌شود، چندان با فارسی معیار همخوانی ندارد و به عبارتی این فضاها دارای ادبیات خاص خود هستند.

تشخیص زبان در شبکه‌های اجتماعی همواره از چالش‌های مطرح بوده و این مسئله زمانی بیشتر رخنمون دارد که متون مورد بررسی، کوتاه و حاوی نویسه‌های خارج از الفبا یا متشکل از چند زبان باشند. علاوه بر این، متون دارای واژه‌های ساختگی و نوواژه‌ها، املا نامناسب متون، متون دارای واژه‌های مخفف از دیگر مشکلات موجود در مسیر تشخیص زبان هستند. قبل از انجام عملیات مختلف روی متن نظیر ترجمه، خلاصه‌سازی، دسته‌بندی[۲]، بهنجارسازی[۳]، تقطیع متن[۴] و غیره ابتدا باید زبان آن مشخص باشد. از طرفی بازیابی اطلاعات[۵] و موتورهای جستجو[۶]، عقیده کاوی[۷] در متون و تحلیل احساسات[۸]، ماشین‌های ترجمه[۹] و سامانه‌های پرسش و پاسخ [۱۰] از دیگر حوزه‌هایی هستند که نیاز به تشخیص زبان را آشکار می‌سازند.

در این مقاله ما یک روش جدید مبتنی بر Nتایی و نویسه‌های ویژه برای تشخیص زبان متون محاوره گرفته شده از شبکه‌های اجتماعی ارائه می‌دهیم. روش پیشنهادی قابلیت تشخیص زبان فارسی و زبان‌های با رسم‌الخط مشابه یعنی کردی مرکزی، عربی، پشتو و اردو می‌باشد. نتایج ارزیابی نشان می‌دهد که روش پیشنهادی برای زبان‌های گفته شده نسبت به ابزارهای موجود کارایی بهتری دارد. در ادامه ابتدا در بخش ۲ روش‌های پیشین را مورد بررسی قرار می‌دهیم و در بخش ۳ روش پیشنهادی را به طور کامل شرح می‌دهیم. در بخش ۴ روش پیشنهادی را ارزیابی می‌کنیم. در نهایت نتیجه گیری و کارهای آتی را در بخش ۵ بیان می‌کنیم.

۲- مروری بر روش‌های پیشین

ایده اولیه در ساخت ابزار تشخیص زبان بدین صورت است که در ابتدا با توجه به ویژگی‌های مخصوص هر زبان الگوی زبانی[۱۱] برای زبان‌های هدف ساخته می‌شود. سپس این الگو یا الگوها به سیستم آموزش داده می‌شوند. سپس برای تشخیص زبان سند جدید با همان روشی که قبلاً مورد استفاده قرار گرفته و با معیارهای مشابه، الگوی زبانی ساخته می‌شود. سپس الگوی زبانی سند جدید با آنچه که سیستم از قبل آموزش دیده مقایسه می‌شود. زبان الگویی که کمترین فاصله و به عبارتی بیشترین شباهت را با الگوی زبانی سند جدید داشته باشد، به عنوان زبان سند جدید تعیین می‌شود [۲]. روش‌های ساخت الگوی زبانی به دو دسته کلی آماری و غیرآماری تقسیم می‌شوند. روش آماری اصولا مبتنی بر Nتایی‌ها[۱۲] و روش غیرآماری مبتنی بر واژه است [۳].

N تایی روشی است که احتمال وقوع یک دنباله از نویسه‌ها را می‌سنجد. به عبارتی مشخص می‌کند احتمال آمدن یک نویسه با توجه به آمدن نویسه دیگری قبل از آن چقدر است [۳]. در متون طولانی معمولاً توالی واژه‌ها و در متون کوتاه توالی حروف بیشتر سنجیده می‌شود. در این دسته از روش‌ها ابتدا لیست توالی‌های نویسه‌ها تهیه شود و سپس با روش‌های دسته‌بندی زبان متن مجهول شناسایی می‌شود [۲]. بهترین تکنیک شناخته شده، سندها را طبق آمار یا امتیازشان بر حسب Nتایی توالی نویسه‌ها بین یک سند و یک زبان دسته‌بندی می‌کند [۱]. در این روش ابتدا برای تولید مدل زبانی، متونی انتخاب می‌شوند که زبان آن‌ها مشخص بوده و توالی نویسه‌های آن‌ها استخراج می‌گردد. این توالی‌ها سپس به ترتیب بسامدشان در متون مرتب می‌شوند . در این‌جا از قانون زیف[۱۳] استفاده می‌شود که طبق آن، بسامد یک رخداد (در این‌جا توالی نویسه‌ها) با رتبه‌ای که به آن اختصاص داده می‌شود نسبت عکس دارد. بدین معنی که پربسامدترین توالی رتبه یک را به خود اختصاص می‌دهند و به همین ترتیب کم بسامدترین توالی رتبه عددی بزرگتری خواهد داشت. کاونار و ترنکل معتقدند سیصد توالی پربسامدتر در هر زبان، نشانه بسیار خوبی برای تشخیص آن زبان است [۴]. بدین ترتیب مدل‌های زبانی آماده می‌شوند که حاوی پربسامدترین توالی‌ها در زبان‌هایی است که ابزار قرار است بتواند آن‌ها را تشخیص دهد. در ادامه، هنگامی که متن جدیدی وارد می‌شود و برای تشخیص زبان آن، پربسامدترین توالی‌های آن استخراج شده و به ترتیبی مشابه آن‌چه در پیش گفته شد مرتب می‌شوند. سپس این لیست جدید با توالی نویسه‌ها و در واقع با الگوهای زبانی از پیش تعیین شده برای سیستم مقایسه می‌شود. مدل زبانی که کمترین فاصله را با لیست توالی‌های متن جدید داشته باشد، تعیین‌کننده زبان متن جدید است [۴].

در مقابل روش‌های آماری، مدل‌های الهام گرفته از زبان‌شناسی نیز برای تشخیص زبان پیشنهاد شده‌اند. برای مثال می‌توان لیست ایست‌واژه‌ها را نام برد که در آن یک سند بر حسب هم‌پوشانی آن با لیست‌های زبان‌های مختلف دسته‌بندی می‌شود [۵]. روش‌های دیگر شامل همبستگی واژه و مقوله دستوری [۶] و مدل‌های دسته گرامری [۷] می‌شود. از روش‌های مبتنی بر واژه، واژه‌های ممنوعه و نویسه‌های ویژه را نیز می‌توان نام برد.

در روش نویسه‌های ویژه تمرکز بر روی نویسه هایی است که بیش از دیگر نویسه‌ها می توانند نماینده یک زبان باشند و در زبان‌های دیگر یا وجود ندارند یا به ندرت یافت می‌شوند [۸]. در روش ایست‌واژه‌ها منبع مقایسه، کلماتی هستند که تعدادشان محدود است اما به فراوانی در متون مختلف اعم از فرهنگی، سیاسی، اجتماعی، علمی و غیره یافت می‌شوند [۹]. این دسته از کلمات شامل حروف ربط، حروف عطف ، ضمایر، حروف اشاره و گاهی نیز افعال پرکاربرد هستند که می‌توانند نماینده خوبی برای زبان باشند.

۳- روش پیشنهادی

در روش پیشنهادی از الگوی زبانی سه بخشی مشتمل بر Nتایی نویسه‌ها، نویسه‌های ویژه و ایست‌واژه‌ها استفاده می‌شود. در نتیجه الگوی زبانی حاصل از مزایای هر سه روش سود می‌برد و مشکلات ناشی از کاربرد هر روش نیز برطرف می‌شود. بدین ترتیب هر متن مجهول طی سه مرحله سنجیده شده و با الگوهای هر پنج زبان مقایسه می‌شود و نتایج حاصل از مقایسه به صورت مقدار صحیح مثبت بیان می‌شود. الگویی که کوچکترین عدد را داشته باشد نشانگر زبان متن مجهول است.

۳-۱- جمع‌آوری دادگان

با طراحی و ساخت پیکره زبان فارسی محاوره شبکه‌های اجتماعی، دادگان قابل قبولی جمع‌آوری و بهنجارسازی شده و ناخالصی‌های آن نیز زدوده شد. از این پیکره جهت ساخت الگوی زبانی فارسی در این پژوهش استفاده شده است. برای جمع‌آوری دادگان مورد نیاز برای ساخت الگوی زبانی کردی مرکزی با راهنمایی افراد کرد زبان، گروه‌ها و کانال‌هایی در شبکه‌های اجتماعی انتخاب شده و با اطلاع اعضای گروه‌ها، از مطالب آن جهت تهیه دادگان زبان کردی استفاده شد. این دادگان سپس بهنجارسازی شده و برای تهیه الگو آماده شد. مراحل مشابهی جهت تهیه دادگان موثق از زبان اردو، عربی، پشتو نیز طی شد. این مطالب شامل انواع نوشته‌ها از قبیل مکاتبات روزمره، دل‌نوشته‌ها، شعر، خبر، مطالب طنز بود. در شبکه‌های اجتماعی، حجم

زیادی از محتوای دیداری و شنیداری ارائه می‌شود. از این روی تمام پیام‌های ارسال شده در این شبکه‌ها قابلیت استفاده در دادگان مورد نظر را نداشته و تهیه حجم مناسبی از دادگان هر زبان نیازمند گذر زمان و ارسال تعداد کافی پیام نوشتاری توسط کاربران بود. همین پیام‌های نوشتاری نیز باید به دقت بررسی شده تا عاری از اطلاعات مستقل از زبان مانند شکلک، تارنما و غیره باشند.

۳-۲- ساخت الگوی زبانی

در مرحله اول تولید الگوی زبانی، Nتایی نویسه‌های دادگان هر زبان استخراج می‌شود. اما چرا تمرکز بر روی نویسه است و نه واژه؟ با توجه به قوانین واج‌آرایی و قواعد تصریف، محدودیت‌هایی برای همنشینی واج‌ها و در نهایت تشکیل واژه وجود دارد که اجازه ساخت هر گونه واژه دلخواهی را نمی‌دهد. از طرفی گویشوران هر زبان، حتی افرادی که به طور تخصصی به مطالعه زبان نپرداخته اند و افراد بدون تحصیلات، این قوانین را رعایت می‌کنند. این مطلب در هنگام ساخت واژه‌های جدید نیز مصداق دارد و از این روست که اگر قرار باشد یک فرد فارسی زبان واژه جدیدی تولید کند، نتیجه آن چیزی مشابه «پیچگزشخ» نخواهد بود! با این استدلال، اگر مثلاً با محاسبه Nتایی واژه‌هایی که اکنون در زبان فارسی وجود دارند، احتمال رخداد توالی‌های مختلف نویسه‌ها را داشته باشیم، از آنجا که نویسه‌های جدید نیز طبق قوانین قبلی در کنار یکدیگر قرار می گیرند، با تقریب خوبی Nتایی واژه‌های جدید نیز مشابه واژه‌های قبلی موجود در زبان خواهد بود و بدین ترتیب می‌توان با دقت بالایی آن‌ها را به فارسی نسبت داد. از این طریق مشکل وجود نوواژه‌ها و واژه‌های ساختگی در زبان شبکه‌های اجتماعی برطرف شده و نیاز به وجود تک تک آن‌ها در دادگان مورد استفاده در ابزار تشخیص زبان نیست. نکته‌ای که در این میان باید مورد توجه قرار داد، هم اندازه بودن حجم دادگان برای هر زبان است تا از سوگیری به زبان خاصی پرهیز شود.

بعد از استخراج Nتایی نویسه‌های دادگان فارسی، باید آن‌ها را به ترتیب بسامدشان در متون فارسی مرتب کرده و به پرتکرارترین آنها رتبه یک و توالی با تعداد تکرار کمتر رتبه دو و الی آخر نسبت داد. تعداد مشخصی توالی پرتکرار از ابتدای لیست انتخاب شده در الگوی زبانی مورد استفاده قرار می گیرند. این روند برای زبان‌های با رسم‌الخط مشابه فارسی یعنی کردی مرکزی، اردو، پشتو، عربی نیز تکرار شده و صورت اولیه الگوی زبانی مبتنی بر Nتایی برای هر یک از این زبان‌ها ساخته شد.

در مرحله بعد، نویسه‌های خاص هر زبان به الگوی زبانی آن اضافه شد. برای مثال، این نویسه‌ها در فارسی شامل حروف «گ، چ، پ، ژ» می‌شود. این روش مخصوصاً جهت بازشناسی زبان فارسی از عربی که مشترکات زیادی دارند به کار می‌رود. از طرفی، برای برخی نویسه‌ها بیش از یک صورت نوشتاری وجود دارد. برای مثال در اردو برای نویسه «ٹ» صورت‌های مختلف برای ابتدای واژه، وسط واژه و انتهای واژه وجود دارد و هر یک نویسه‌ای جداگانه با کد متفاوتی محسوب می‌شوند. پس اگر فقط یک صورت نوشتاری از این نویسه در لیست نویسه‌های ویژه اردو وجود داشته باشد، ابزار تشخیص زبان در صورت برخورد با دیگر صورت‌های همین نویسه آن‌ها را به عنوان نشانه زبان اردو تلقی نخواهد کرد و این امر نتیجه نهایی را تحت تاثیر قرار خواهد داد. اما در برخی زبان‌ها برای همه صورت‌های نوشتاری یک نویسه تنها یک کد مشترک وجود دارد. از این جهت مهم است که تمامی صورت‌های نوشتاری همه نویسه‌های ویژه هر زبان بررسی شده و در صورت داشتن کدهای متفاوت، تک تک آن‌ها منظور شوند.

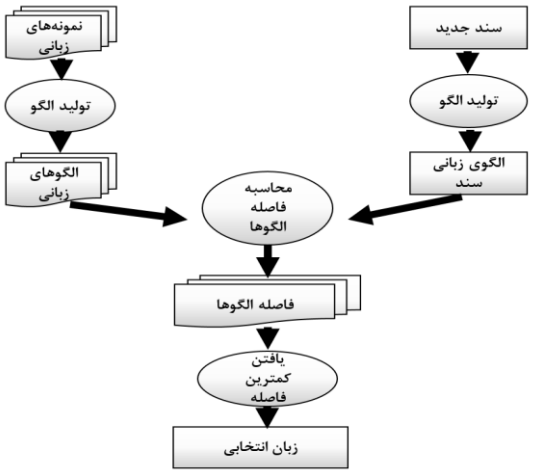
در مرحله سوم لیستی به عنوان ایست‌واژه‌های هر زبان تهیه شد که شامل حروف عطف، حروف ربط، ضمایر، افعال و اسم‌های عام پرکاربرد و مخففها و واژگانی از این دست است. وجود این واژه‌ها در متن می‌تواند نشانگر نوع زبان باشد. نکته حائز اهمیت در این میان، لزوم عدم تشابه در واژگان هر لیست است. به بیان دیگر، یک واژه فقط می‌تواند عضو لیست ایست‌واژه‌های یک زبان باشد. از آن‌جا که زبان‌های مورد نظر در این پژوهش دارای هم‌پوشانی هستند، بعضی واژگان در لیست ابتدایی دو یا سه زبان وجود داشتند. مثلاً «من»، «ما»، «این»، «له» گرچه با تلفظ و معانی متفاوت، در لیست ایست‌واژه‌های فارسی و عربی مشترک هستند. این مطلب می‌تواند باعث بروز خطا در تشخیص زبان شود چون صورت نوشتاری واژه‌ها در این‌جا اهمیت دارد نه معنی و تلفظ آن‌ها. یا مثلاً واژگان «چی» و «به» بین زبان‌های فارسی و کردی مرکزی و پشتو مشترک هستند. واژه «تو» نیز بین فارسی و کردی مرکزی و اردو مشترک است. به عبارتی هر واژه بایستی تنها نماینده یک زبان باشد وگرنه سبب بروز خطا در تصمیم‌گیری خواهد شد. برای رفع این مشکل، تک تک لیست‌ها بررسی شده و واژه‌های مشترک آن‌ها حذف شدند. به علاوه، برای زبان فارسی تعدادی از واژه‌های ساختگی که به مرور عضو ثابت پیام‌های شبکه‌های اجتماعی شده‌اند نیز در لیست ایست‌واژه‌های فارسی منظور شد. بدین ترتیب برای هر زبان طی سه مرحله الگوی زبانی ساخته شد.

۳-۳- تشخیص زبان

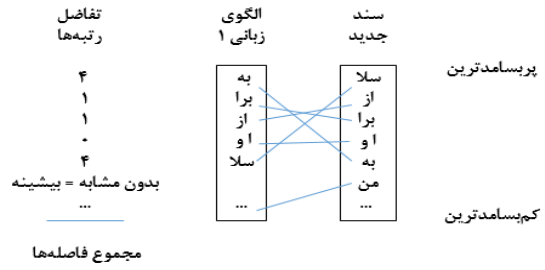
مطابق شکل (۱)، وقتی سند جدیدی برای تشخیص زبان وارد می‌شود، در صورتی که رسم‌الخط متن جدید قابل پذیرش باشد، پربسامدترین توالی‌های آن استخراج شده و به ترتیبی مشابه آن‌چه در پیش گفته شد مرتب می‌شوند. اگر توالی اول در لیست جدید (با زبان نامشخص) با توالی اول در یک الگوی زبانی از پیش تعیین شده (با زبان مشخص) یکسان باشد، قدر مطلق تفاضل فاصله آن‌ها صفر خواهد بود و در غیر این‌صورت، فاصله رتبه این دو توالی مقدار عددی دیگری خواهد داشت. برای مثال اگر اولین توالی نویسه‌ها در الگوی زبانی جدید به صورت ترکیب سه نویسه «سلا» و پنجمین توالی نویسه‌ها در الگوی زبانی پیش‌فرض هم همین ترکیب نویسه‌های «سلا» باشد، تفاضل فاصله این دو برابر با ۴ خواهد بود. سپس برنامه به سراغ توالی دوم رفته و مقایسه مشابهی را انجام داده و مقدار عددی این مقایسه برای همه توالی نویسه‌ها محاسبه خواهد شد. در صورتی که یکی از توالی‌های موجود در متن جدید با هیچ یک از توالی‌های یک الگوی زبانی از پیش تعیین شده مطابقت نداشته باشد چه اتفاقی خواهد افتاد؟ در صورتی که هیچ اقدامی در این مرحله انجام نشود، بدین معنیست که تفاصل دو توالی صفر بوده و به عبارت دیگر، رتبه توالی مورد نظر در متن مجهول و الگوی زبانی پیش‌فرض دقیقاً یکی بوده است و بدین ترتیب خطای زیادی در تشخیص خواهیم داشت. برای رفع این مشکل، به ازای هر توالی بدون مطابقت، مقدار عددی ۱۰۰۰ به عنوان بیشینه تفاضل به آمار قبلی اضافه می‌شود. چنان که از شکل (۲) نیز برمی‌آید، پس از محاسبه تک تک تفاضل‌ها در نهایت لیستی از اعداد صحیح مثبت خواهیم داشت که حاصل‌جمع این اعداد بیان کننده میزان فاصله دو الگوی زبانی مورد مقایسه است. این روند مقایسه میان الگوی زبانی جدید و سایر الگوهای زبانی پیش‌فرض نیز انجام خواهد شد. در نهایت از این مرحله

پنج عدد صحیح مثبت بدست خواهد آمد که حاصل مقایسه الگوی توالی‌های متن جدید با هر یک از پنج زبان مورد نظر است.

در مرحله دوم، متن جدید با زبان نامشخص، از لحاظ دارا بودن نویسه‌های ویژه بررسی می‌شود. برای مثال اگر متن جدید دارای نویسه‌های «گ، چ، پ، ژ» باشد، می‌دانیم که زبان متن عربی نیست. نویسه‌های ویژه هر زبان در جدول (۱) آمده است. عنوان شد که از مرحله اول، به ازای مقایسه متن مجهول با هر یک از پنج الگوی زبانی، عدد صحیح مثبتی حاصل می‌شود. در مرحله دوم اگر متن مجهول دارای نویسه‌های «گ،چ،پ،ژ» باشد، زبان عربی از روند مقایسه حذف می‌شود. در صورتی که نویسه ویژه زبان پشتو در متن مشاهده شود، به میزان یک درصد از عدد حاصل از مقایسه مرحله اول برای پشتو، کم می‌شود. با هر نشانه جدید و برای هر زبان نیز این روال تکرار می‌شود. در نهایت عدد حاصل از مرحله دوم برای برخی زبان‌ها تغییر کرده که نشان‌دهنده فاصله آن از متن مجهول است.



شکل (۱) نمای روند تشخیص زبان در روش پیشنهادی



شکل (۲) مقایسه توالی‌های سند جدید با الگوهای زبانی

مرحله سوم به بررسی ایستواژه‌ها اختصاص دارد. از آن‌جا که از قبل فهرستی از ایستواژه‌های هر زبان جمع آوری شده، سند جدید از لحاظ دارا بودن چنین واژگانی نیز بررسی می‌شود. اگر واژه‌ای موجود در لیست ایستواژه‌های هر زبان در متن مجهول یافت شود، به ازای هر واژه مجدداً یک درصد از عدد حاصل از مرحله قبل برای آن زبان کم می‌شود که باز هم نشان‌دهنده میزان شباهت آن زبان با متن مجهول است. بدین ترتیب باز هم در مرحله سوم بعضی اعداد حاصل از مرحله اول تغییر کرده و در صورت مشابهت الگو با متن جدید، مقدار عدد کم‌تر می‌شود. در نهایت با داشتن نتیجه حاصل از سه مرحله مقایسه الگوهای زبانی، می‌توان با اطمینان بیشتری در مورد زبان سند جدید تصمیم‌گیری کرد. الگوی زبانی که کمترین فاصله عددی را با الگوی زبانی متن جدید داشته باشد، تعیین کننده زبان متن جدید است. وجود مرحله دوم و سوم از آن جهت است که اگر میان دو یا چند الگوی زبانی که در مرحله اول برای انتخاب به عنوان زبان سند جدید نامزد شده‌اند فاصله بسیار کمی وجود داشته باشد، مراحل بعدی سبب قطعیت در تصمیم‌گیری خواهد شد.

نکته شایان ذکر این است که با این روش، بر مشکلات ناشی از وجود نوواژه‌ها و واژه‌های ساختگی در متن نیز می‌توان فائق آمد. به عبارتی حتی اگر واژه‌ای در عبارت مورد بررسی وجود داشته باشد که قبلاً در پیکره دیده نشده اما توسط یک فارسی زبان ساخته شده باشد، بدلیل ثابت بودن قوانین همنشینی واجها و الگوی زبانی که بر این اساس تعریف شده، قابلیت تشخیص آن به عنوان نمونه‌ای از زبان فارسی وجود دارد. همچنین بخاطر سه مرحله سنجش و عدم اکتفا به الگوی زبانی مبتنی بر Nتایی، حتی با وجود هم‌پوشانی بین زبان‌های مورد نظر و وجود واژگان مشترک و نویسه‌های مشابه، همچنان روند تشخیص زبان با دقت بالایی انجام می‌شود.

جدول (۱)نویسه‌های ویژه هر زبان

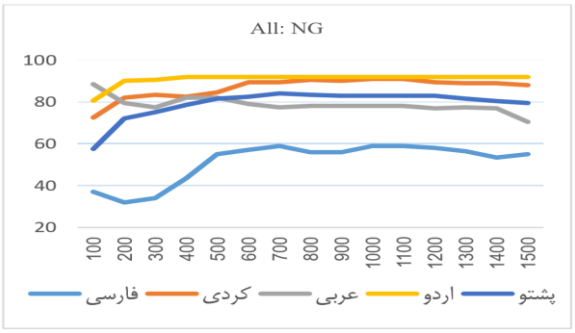
پشتو		اردو	کردی	عربی	تمایز فارسی و عربی
ټ	ښ	ٹ	ێ	أ	ی
څ	ګ	ڈ	ڵ	إ	چ
ځ	ڼ	ڑ	ۆ	ؤ	پ
ډ	ۀ	ں	ڕ		ژ
ړ	ې	ے	ڤ		
ږ	ۍ				

ارزیابی روش پیشنهادی

برای بررسی عملکرد روش پیشنهادی با الگوهای زبانی مختلف، دو مرحله آزمون اجرا گردید. مرحله اول برای همه زبان‌ها (All) فقط با استفاده از الگوی زبانی مبتنی بر Nتایی (NG) و بدون اعمال فیلترهای دوم و سوم بود (الگوی نوع اول) بدین صورت که برای الگوی زبانی، به ازای ۱۰۰ الی ۱۵۰۰ توالی برتر با گام افزایشی ۱۰۰ توالی از ابتدای لیست توالی‌های هر زبان، انتخاب شد و با توالی‌های متن مجهول مورد مقایسه قرار گرفته و نتایج تشخیص زبان حاصل ثبت شد. به عبارتی با ۱۵ بار اجرای آزمون روی هر عبارت و هر بار با الگوی زبانی مبتنی بر Nتایی ولی با طول متفاوت (تعداد توالی‌های جدا شده از هر لیست)، روند تشخیص زبان تکرار شد. نتایج حاصل از آزمون‌های مرحله اول را به تفکیک تعداد توالی‌های مورد استفاده به عنوان الگوی زبانی در شکل (۳) مشاهده می‌نمایید.

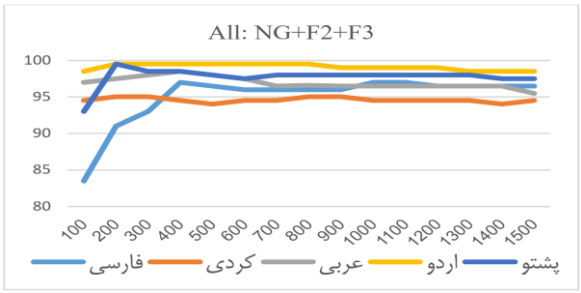
همان‌طور که از شکل (۳) برمی‌آید، بهترین عملکرد کلی در رابطه با زبان اردو مشاهده می‌شود که به طور متوسط برای اکثر توالی‌ها بالاترین نتیجه را بین همه زبان‌ها دارد. زبان اردو با ۴۰۰ توالی پرکاربرد به عنوان الگوی زبانی، بهترین نتیجه را داده و پس از آن نتیجه ثابت می‌ماند. زبان بعدی با بهترین نتایج به طور میانگین در اکثر توالی‌ها، کردی مرکزی (سورانی) است که البته خود با استفاده از ۱۰۰۰ توالی پرکاربرد بهترین نتیجه را بدست داده است. نتایج زبان فارسی در صورت استفاده از ۷۰۰ توالی پرکاربرد بهترین حالت را نشان می‌دهد و به ۵۹ درصد می‌رسد. در رابطه با زبان پشتو نیز چنان که از نمودار مشخص است، نتیجه بهینه با استفاده از ۷۰۰ توالی پرکاربرد به عنوان الگوی زبانی بدست می‌آید. زبان عربی زودتر از بقیه زبان‌ها و با داشتن ۱۰۰

توالی پرکاربرد، بهترین نتیجه را نشان می‌دهد و پس از آن نتایج افت می‌کند. به طور میانگین برای هر پنج زبان، بهترین نتیجه با در نظر گرفتن ۱۰۰۰ توالی پرکاربرد از ابتدای لیست توالی‌های نویسه‌های هر زبان به عنوان الگوی زبانی بدست آمده که به میزان ۸۰٫۶ درصد است. با توجه به نتایج قابل مشاهده در نمودار شکل (۳)، افزایش تعداد توالی‌های هر زبان تا حدی می‌تواند نتیجه را بهبود دهد و پس از آن دقت روش ثابت می‌ماند. این نتایج نشان می‌دهد که تعداد محدودی از توالی‌های پرکاربرد می‌توانند به خوبی نماینده آن زبان باشند.



شکل (۳) نتایج آزمون تشخیص زبان با الگوی نوع اول

در مرحله بعد جهت سنجش میزان تاثیر دو فیلتر دیگر یعنی نویسه‌های ویژه (F2) و ایستواژه‌ها (F3) بر دقت عملکرد ابزار تشخیص زبان یا به عبارتی میزان کارایی الگوی زبانی جدید، آزمون‌های جدیدی انجام شد. در این مرحله علاوه بر استفاده از Nتایی توالی نویسه‌های هر زبان، نویسه‌های ویژه و ایستواژه‌های هر زبان نیز مد نظر قرار گرفته و با احتساب وزن برای این دو فیلتر، عبارات قبلی برای تشخیص زبان به روش پیشنهادی داده شد. همانند مرتبه اول هر جمله ۱۵ بار و هر بار با الگوی زبانی متفاوتی مورد سنجش قرار گرفت. شکل (۴) میانگین تشخیص‌های درست روش پیشنهادی برای هر پنج زبان را برای تعداد توالی‌های متفاوت (۱۰۰ تا ۱۵۰۰ توالی) نشان می‌دهد.

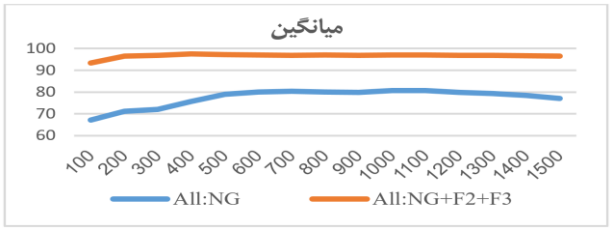


شکل (٤) نتایج آزمون تشخیص زبان با الگوی نوع دوم

با نگاه اول می‌توان متوجه بهبود قابل توجه نتایج آزمون تشخیص زبان شد. این بهبود نتایج مخصوصاً برای زبان فارسی بیش از بقیه زبان‌ها محسوس است. چنان‌که از شکل (۴) برمی‌آید، حداقل نتایج برای فارسی در ۱۰۰ توالی اول با ۸۳٫۵ درصد پاسخ درست، حتی از بهترین نتایج فارسی در مرحله اول آزمون‌ها (شکل ۳) نیز بهتر است. نتایج برای زبان فارسی روند سعودی را طی کرده و در این مرحله بهترین عملکرد ابزار برای زبان فارسی در الگوی ۴۰۰ توالی برتر بدست می‌آید که معادل ۹۷ درصد پاسخ درست است. این بدان معناست که با محاسبات بسیار کمتر و با الگوی زبانی کوچکتری می‌توان به نتایج عالی برای زبان فارسی رسید. همچنین برای زبان‌های دیگر نیز دقت روش با در نظر گرفتن دو فیلتر نویسه‌های ویژه و ایستواژه‌ها افزایش یافته است.

میزان تشخیص درست به طور میانگین برای هر زبان بدین قرار است: فارسی ۹۵ درصد، کردی ۹۴٫۵ درصد، عربی ۹۶٫۹۴ درصد، پشتو ۹۷٫۷۳ و برای اردو نیز ۹۹٫۱ درصد و میانگین هر پنج زبان برای هر ۱۵ الگو ۹۶٫۶۶ درصد است. به طور میانگین برای هر پنج زبان در ۴۰۰ توالی پرکاربرد به بهترین نتیجه می‌رسیم که معادل ۹۷٫۶ درصد است.

شکل (۵) میانگین دقت بدست آمده برای تمام زبان‌ها در هر دو حالت استفاده از Nتایی (NG) تنها و استفاده از دو فیلتر نویسه‌های ویژه (F2) و ایستواژه‌ها (F3) را نشان می‌دهد. همان‌طور مشاهده می‌شود، استفاده از دو فیلتر دیگر، تشخیص درست زبان را بسیار افزایش می‌دهد و الگوی زبانی نوع اول در هیچ یک از توالی‌ها قابل مقایسه با الگوی زبانی نوع دوم نیست. این امر نشان‌دهنده دقت و کارایی بسیار بیشتر الگوی زبانی تکمیل شده است. به عبارتی، با دقیق‌تر کردن الگوی زبانی و اضافه کردن دو فیلتر دیگر، شناخت دقیق‌تری برای ابزار تشخیص زبان حاصل می‌شود و مخصوصاً هنگامی که هم‌پوشانی واژه‌ها سبب نزدیک شدن الگوی توالی نویسه‌ها می‌شود، دو فیلتر دوم و سوم سبب تمایز الگوهای زبانی و در نتیجه تشخیص بهتر می‌شوند. این امر مخصوصاً هنگام مقایسه فارسی و عربی که واژه‌های مشترک بسیاری دارند سودمند است.



شکل (۵) مقایسه نتایج آزمون تشخیص زبان با الگوی نوع اول و دوم

با دقت در شکل (۵) می‌توان دریافت که با استفاده از الگوی زبانی اولیه که تنها مبتنی بر Nتایی است، به طور میانگین برای هر پنج زبان با داشتن ۱۰۰۰ توالی پرکاربرد به بهترین عملکرد ابزار تشخیص زبان دست می‌یابیم. این میزان در ۱۰۰۰ توالی برتر برابر با ۸۰٫۶ درصد است. این میزان برای یک ابزار تشخیص زبان عملکرد ضعیفی را نشان می‌دهد. وقتی الگویی که از زبان‌های مختلف به ابزار تشخیص زبان شناسانده شده، الگوی کاملی نباشد طبعاً ابزار تشخیص زبان در مواجهه با انواع متونی که رسم‌الخط مشابه و هم‌پوشانی واژگانی دارند دچار خطا در تشخیص خواهد شد. بنابراین برای رفع این مشکل بایستی برای شناساندن الگوی دقیق‌تر به ابزار تشخیص زبان تمهیدی اندیشید.

با نگاهی به نتایج حاصل از مرحله دوم آزمون یعنی آزمون با استفاده از الگوی پیشنهادی این پژوهش می‌توان به وضوح مشاهده کرد که میانگین تعداد تشخیص‌های درست برای هر پنج زبان به طرز قابل توجهی افزایش داشته است. به طور کلی بیشینه تعداد تشخیص های درست با داشتن ۴۰۰ توالی پرکاربرد هر زبان حاصل شده و میزان آن ۹۷٫۶ درصد است که افزایش ۱۷ درصدی در میانگین هر پنج زبان به نسبت میانگین الگوی نوع اول را نشان می‌دهد. به علاوه، تعداد توالی‌های مورد نیاز برای رسیدن به بهترین عملکرد نیز کاهش چشمگیری داشته و از ۱۰۰۰ توالی به طور متوسط، به

۴۰۰ توالی کاهش یافته که خود باعث کاهش پیچیدگی محاسباتی و زمانی در هنگام تشخیص زبان می‌شود.

اما ابزار تشخیص زبان در صورت مواجهه با عبارت یا متنی با بیش از یک زبان، چه خواهد کرد؟ در پاسخ باید گفت اگر متن وارد شده، دربردارنده واژگان یکی از زبان‌های مورد بررسی و رسم‌الخط یا زبانی متفرقه باشد، ابزار تشخیص زبان فقط زبانی را که الگوی آن را می‌شناسد تشخیص می‌دهد. برای مثال اگر زبان فارسی به همراه عبارت یا واژه‌ای از انگلیسی یا پینگلیش وارد شود، نتیجه تشخیص زبان فارسی خواهد بود.

هنگامی که متن وارد شده ترکیبی از دو زبان از پنج زبان مورد تمرکز در این پایان‌نامه باشد، زبانی به عنوان نتیجه اعلام خواهد شد که دارای بیشترین تشابه واژگانی و توالی‌های Nتایی با متن ورودی به طور کلی باشد. مثلا اگر متن اردو و پشتو همزمان وارد شوند، هر کدام که دارای تعداد توالی‌های Nتایی بیشتر (به نسبت کل طول متن) و بیشترین تعداد واژگان یا نویسه‌های خاص در متن باشد، به عنوان زبان کلی متن انتخاب خواهد شد. مبنای این گونه عملکرد این است که در نهایت زبانی انتخاب شود که دارای بیشترین تشابه و کمترین فاصله با الگوی زبانی متن جدید باشد. پس اگر درصد بیشتری از متن ورودی پشتو و درصد کمتری اردو باشد، زبان غالب در این متن پشتو بوده و همان انتخاب خواهد شد.

برای مقایسه نتایج ابزار تشخیص زبان تهیه شده در این پژوهش با سایر ابزارهای تشخیص زبان، آزمون‌های مشابهی بر روی دو ابزار تشخیص زبان یعنی گوگل[۱۴] و زیراکس[۱۵] اجرا شد که نتایج آنها در جدول (۲) قابل مشاهده است. عبارت‌های ارائه شده به این ابزارها همان عبارت‌های مورد استفاده در آزمون روش پیشنهادی بوده و زبان‌های مورد آزمون نیز شامل فارسی، کردی مرکزی، عربی، اردو و پشتو بودند. الگوی زبانی مربوط به پشتو و کردی مرکزی برای زیراکس و زبان کردی مرکزی برای گوگل شناخته شده نیست.

جدول (۲) نتایج آزمون ابزار تشخیص زبان گوگل

زبان	گوگل	زیراکس	روش پیشنهادی
فارسی	۱۰۰	۸۹	۹۵
کردی	---	---	۹۴,۵
عربی	۹۸	۹۴	۹۶,۹۴
اردو	۱۰۰	۱۰۰	۹۹,۱
پشتو	۹۶	---	۹۷,۷۳

٤- نتیجه گیری

در این مقاله یک روش جدید برای تشخیص زبان فارسی و زبان‌های با رسم‌الخط مشابه یعنی کردی مرکزی، عربی، پشتو و اردو ارائه شد. روش پیشنهادی علاوه بر استفاده از Nتایی برای ساخت مدل زبانی از دو مرحله سنجش دیگر یعنی در نظر گرفتن نویسه‌های ویژه هر زبان و همچنین ایست‌واژه‌ها نیز بهره می‌برد که در نظر گرفتن این دو مرحله می‌تواند نتایج را بسیار بهبود بخشد. این افزایش و بهبود نتایج با استفاده از الگوی زبانی نوع دوم برای هر پنج زبان مشاهده می‌شود. با مقایسه دو مرحله آزمون، بیشترین تاثیر تکمیل الگوی زبانی را بر تشخیص زبان فارسی و سپس پشتو می‌توان دید؛ اگرچه در سایر زبان‌های مورد بررسی نیز بهبود قابل توجهی به چشم می‌خورد. همچنین در این مقاله یک پیکره از متون محاوره شبکه‌های اجتماعی برای زبان‌های فارسی، کردی مرکزی، عربی، پشتو و اردو جمع‌آوری شده که می‌تواند در پژوهش‌های آینده مورد استفاده قرار گیرد.

مراجع

[1] Baldwin, T., and Lui, M. "Accurate Language Identification of Twitter Messages". *The 5th Workshop on Language Analysis for Social Media (LASM)*, pp. 17- 25. Gothenburg, Sweden, 2014.

[2] Baldwin, T., and Lui, M. "Language Identification: The Long and the Short of the Matter". *Proceedings of Human Language Technologies: The 11th annual conference of the North American chapter of the association for computational linguistics*, pp. 229-237, 2010.

[3] Panich, L. *"Comparison of Language Identification Techniques"*. Heirich Heine Universitaet, Institute Fuer Informatik Datenbanken und Informationssysteme, Duesseldorf, Germany, 2015.

[4] Cavnar, W., and Trenkle, J. "N-gram-based Text Categorization". *The 3rd Annual Symposium on Document Analysis and Information Retrieval*, pp. 161-175. Las Vegas, Nevada, USA, 1994.

[5] Johnson, S. "*Solving the Problem of Language Recognition*". Technical Report, University of Leeds, School of Computer Studies, 1993.

[6] Grefenstette, G. "Comparing Two Language Identification Schemes". *The 3rd International Conference on Statistical Analysis of Textual Data*, pp. 263-268. Rome, Italy, 1995.

[7] Lins , R., and Gonçalves, P. "Automatic language identification of written texts". *Proceedings of the 2004 ACM Symposium on Applied Computing*, pp. 1128–1133. Nicosia, Cyprus, 2004.

[8] Kranig, S. *"Evaluation of Language Identification Methods"*. University of Tübingen International Studies in Computational Linguistics, Tübingen, 2015.

[9] Řehůřek, R., and Kolkus, M. "Language Identification on the Web: Extending the Dictionary Method". *Computational Linguistics and Intelligent Text Processing* CICLing 2009, 357–368, 2009.

پانویس‌ها

[1] language identification
[2] classification
[3] normalization
[4] tokenization
[5] information retrieval
[6] search engines
[7] opinion mining
[8] sentiment analysis
[9] translation machines
[10] Q-A systems
[11] language model
13 N-gram
14 Zipf's law
15 https://translate.google.com
16 https://open.xerox.com/Services/LanguageIdentifier