سومین همایش ملی
زبان شناسی رایانشی
۲۷ و ۲۸ آبان ۱۳۹۲
دانشگاه صنعتی شریف

دانشگاه صنعتی شریف
انجمن زبان‌شناسی ایران

پیکره‌ی فارسی تحلیل احساس سِنتی‌پِرس:
توسعه‌ی یک پیکره‌ی تحلیل احساس متنی برای زبان فارسی

پدرام حسینی
دانشجوی کارشناسی ارشد
گروه مهندسی کامپیوتر
دانشکده فنی، دانشگاه گیلان
hosseinip@msc.guilan.ac.ir

حسن ملکی گلندوز
فارغ التحصیل کارشناسی
گروه مهندسی کامپیوتر
دانشکده فنی، دانشگاه گیلان
ha.maleky@gmail.com

علی احمدیان رمکی
فارغ التحصیل کارشناسی ارشد
گروه مهندسی کامپیوتر
دانشکده فنی، دانشگاه گیلان
ahmadianrali@msc.guilan.ac.ir

منصوره انواری رستمکلایی
فارغ التحصیل کارشناسی ارشد
گروه مهندسی کامپیوتر
دانشگاه آزاد اسلامی واحد رشت
mansoureh.anvari.r@gmail.com

سید ابوالقاسم میرروشندل
استادیار و عضو هیئت علمی
گروه مهندسی کامپیوتر
دانشکده فنی، دانشگاه گیلان
mirroshandel@guilan.ac.ir

کلیدواژه‌ها: تحلیل احساس، پیکره احساس، فرایند نشانه‌گذاری

سومین همایش ملی
زبان شناسی رایانشی
۲۸ و ۲۹ آبان ۱۳۹۲
دانشگاه صنعتی شریف

دانشگاه صنعتی شریف
انجمن زبان‌شناسی ایران

پیکره‌ی فارسی تحلیل احساس سِنتی‌پِرس:
توسعه‌ی یک پیکره‌ی تحلیل احساس متنی برای زبان فارسی

چکیده

تحلیل احساس یکی از زمینه‌های مطالعاتی با اهمیت در پردازش زبان‌های طبیعی، متن کاوی و همچنین زبان‌شناسی رایانشی به شمار می‌آید. با توجه به رشد فزاینده‌ی علاقمندی به این زمینه پژوهشی در سالهای اخیر، نیاز به در اختیار داشتن منابع داده‌ی مناسب برای آن نیز به خوبی احساس می‌شود. در این مقاله، مراحل کامل توسعه‌ی یک پیکره‌ی تحلیل احساس با نام سِنتی‌پِرس شرح داده خواهد شد. بر اساس اطلاعات موجود، می‌توان اظهار نمود که این پیکره در نوع خود اولین پیکره‌ی مربوط به تحلیل احساس با چنین ویژگی‌هایی برای زبان فارسی است. این پیکره شامل بیش از ۲۶۰۰۰ جمله بوده و از مشخصات ویژه‌ای بهره می‌برد. به عنوان مثال، نه تنها مثبت یا منفی بودن جملات بلکه شدت بار معنایی آنها نیز با استفاده یک بازه‌ی عددی در این پیکره نشانه‌گذاری شده‌اند. همچنین در پایان نیز آمار و ارقام مربوط به پیکره و نحوه‌ی محاسبه‌ی درصد توافق مابین نشانه‌گذارها ارائه خواهد شد.

کلیدواژه‌ها: تحلیل احساس، پیکره احساس، فرایند نشانه‌گذاری

۱. مقدمه

با توجه به گسترش روز افزون فضای مجازی در قالب سایت‌ها، وبلاگ‌ها، فروم‌ها، انجمن‌ها و شبکه‌های اجتماعی، منبع عظیمی از داده‌ها شامل نظرات و نقدهای کاربران و مشتریان در رابطه با انواع کالاها و خدمات بروی شبکه جهانی اینترنت موجود است. این نظرات به طور فزاینده‌ای توسط گروه‌های مختلفی از افراد مورد استفاده قرار می‌گیرند. به عنوان مثال، سازمان‌های بزرگ در تعیین و تبیین سیاست‌های کلان، نظرات کاربران و مشتریان خود را بسیار مورد توجه قرار می‌دهند. همچنین مردم نیز همواره تمایل دارند نظرات سایر افراد را در رابطه با یک خدمت و یا کالای مشخص، پیش از استفاده از آن بدانند (Liu:2012). به طور خلاصه می‌توان گفت که "آنچه دیگران فکر می‌کنند" همواره در فرایند تصمیم‌گیری بسیار حائز اهمیت بوده است (Pang, Lee:2008). به علاوه، این منبع عظیم و مفید اطلاعات همواره در کارهای علمی و پژوهشی هم مورد نیاز و توجه بوده است. تحلیل احساس[۱] از جمله وظایف مهم در خانواده‌ی بزرگ پردازش زبان‌های طبیعی به شمار می‌آید و اخیراً نیز به طور فزاینده‌ای در حال تبدیل شدن به زمینه‌ی تحقیقاتی مورد علاقه‌ی محققین است (Liu:2012). تحلیل احساس فرایندی است که در آن نظرات ابراز شده در مورد موجودیت‌های مختلف جهت مشخص نمودن بار معنایی نظرات مرتبط با آن موجودیت مورد تحلیل قرار می‌گیرند (Liu:2012). در تحلیل احساس بدون تردید دسترسی به منابع داده متناسب با هدف مد نظر بسیار مهم و حائز اهمیت است.

تاکنون اغلب پژوهش‌های صورت گرفته در رابطه با توسعه‌ی پیکره‌ی تحلیل احساس در پردازش زبان‌های طبیعی برای زبان انگلیسی بوده است، در حالیکه نیاز به کار در رابطه با زبان‌های غیر انگلیسی نیز وجود دارد. به طور ویژه، در حال حاضر پیکره‌ای مربوط به تحلیل احساس به زبان فارسی توسعه داده نشده است.

---

[۱] Sentiment Analysis

همچنین به این نکته نیز باید اشاره داشت که تنها تعداد محدودی از پیکره‌های توسعه داده شده (چه در انگلیسی و چه سایر زبان‌ها) به طور عمومی برای همگان جهت انجام تحقیقات علمی و پژوهشی قابل دسترس هستند.

در این مقاله روند توسعه‌ی یک پیکره مربوط به تحلیل احساس به نام سِنتی‌پِرس[۲] به طور مفصل شرح داده خواهد شد. این پیکره از بیش از ۲۶.۰۰۰ جمله‌ی فارسی که به طور دستی نشانه‌گذاری شده‌اند تشکیل شده است. یکی از ویژگی‌های این پیکره آن است که شامل جملاتی از فارسی به هر دو صورت رسمی و غیر رسمی (محاوره‌ای) می‌شود. به علاوه، در این پیکره بار معنایی جملات با استفاده از یک بازه، شامل پنج عدد، نشانه‌گذاری شده است. استفاده از چنین بازه‌ای برای تعیین بار معنایی جملات امکان خوبی را فراهم می‌آورد تا در کارهای بعدی بتوان با استفاده از الگوریتم‌های یادگیری، رابطه‌ای معنادار بین این بار معنایی و تعداد کلمات نظر[۳] موجود در جمله یافت نمود. از دیگر ویژگی‌های پیکره‌ی سِنتی‌پِرس، نشانه‌گذاری کلمات کلیدی در هر جمله است. این کلمات کلیدی ممکن است که خود نیز یک کلمه‌ی نظر و همچنین دارای بار معنایی باشند. یکی از موارد کاربرد این کلمات کلیدی می‌تواند کمک به یافتن دلیل منتسب نمودن یک بار معنایی مشخص توسط نشانه‌گذار به جمله باشد. یکی از موارد دیگری که می‌توان در ارتباط با این پیکره به آن اشاره نمود آن است که نشانه‌گذاری جملات در آن در هر سه سطح، سند، جمله و ویژگی صورت پذیرفته است (Liu:2012).

ساختار مقاله پیش رو به این ترتیب است. ابتدا در بخش ۲ به مرور کارهای مربوط به توسعه‌ی پیکره برای تحلیل احساس می‌پردازیم. سپس در بخش ۳ به منبع داده‌ی مورد استفاده از پیکره و همچنین روند توسعه آن اشاره خواهیم نمود. در بخش ۴ به تشریح برخی از مفاهیمی که در فرایند نشانه‌گذاری اسناد با آنها مواجه می‌شویم، می‌پردازیم. برخی از مهم‌ترین اجزا و مؤلفه‌های اسناد موجود در پیکره را در بخش ۵ معرفی خواهیم نمود. در بخش ۶ به تعدادی از مهم‌ترین چالش‌های فرایند نشانه‌گذاری جملات اشاره می‌کنیم. در بخش ۷ در خصوص ابزار نشانه‌گذاری جملات موجود در اسناد و وضعیت در دسترس بودن پیکره به طور مختصر توضیحاتی را عنوان خواهیم کرد. در بخش ۸ آمار جامعی در خصوص پیکره و نحوه‌ی محاسبه‌ی میزان توافق میان نشانه‌گذارها ارائه خواهیم داد و در نهایت در بخش ۹ به یک جمع‌بندی و کارهایی که در آینده می‌توان به آنها مبادرت ورزید، اشاره خواهد شد.

۲. پژوهش‌های مرتبط

در زمینه‌ی تحقیقاتی تحلیل احساس، در اختیار داشتن یک منبع داده‌ی مناسب و قابل اعتماد از اهمیت بالایی برخوردار است. از میان پیکره‌های توسعه داده شده در این زمینه تحقیقاتی، تنها تعدادی از آنها به طور عمومی قابل دسترسی هستند و به علاوه از میان آنها تنها تعداد کمی به زبان‌های غیر انگلیسی اختصاص دارند. در این بخش، ابتدا به مرور پیکره‌های توسعه داده شده برای انگلیسی خواهیم پرداخت.

---

[۲] SentiPers
[۳] Opinion words

سپس پیکره‌های مربوط به سایر زبان‌ها را بررسی خواهیم کرد. در پایان نیز به مواردی از پیکره‌های چند زبانی اشاره خواهیم نمود.

تعدادی از کارهای مربوط به توسعه پیکره پیش از سال ۲۰۰۰ در (Wiebe, et al:2005) آمده‌اند و تاریخ مختصری در خصوص پیکره‌های احساس در آن مقاله شرح داده شده است. سال ۲۰۰۰ سالی بود که از آن پس، عقیده کاوی مبدل به یکی از زمینه‌های پژوهشی مورد علاقه‌ی محققان در زمینه پردازش زبان‌های طبیعی شد (Liu:2012). در برخی از کارهای مربوط به توسعه‌ی پیکره نظیر (Kim, et al:2004)، (Bethard, et al.:2004) و (Yu, Hatzivassiloglou:2003) نشانه‌گذاری‌ها در سطح جمله و کلمه بوده و تنها در برخی از آنها نظیر (Hu, liu:2004) نشان‌ها[۴] و کلمات نظر نیز نشانه‌گذاری شده‌اند. جملات در پیکره‌ی (Hu, liu:2004) شامل نظرات برخط مربوط به پنج وسیله الکترونیکی هستند. این پیکره از ۱۱۳ سند، ۴.۵۵۵ جمله و ۸۱.۸۵۵ نشانه[۵] تشکیل شده است. MPQA یکی دیگر از پیکره‌های محبوب در زمینه عقیده کاوی بوده که تاکنون به طور عمده توسط محققان مورد استفاده قرار گرفته است (Wiebe, et al:2005). این پیکره شامل ۱۰.۶۵۷ جمله در مجموع ۵۳۵ سند است. MPQA عمدتاً شامل مقالات خبری و جملاتی است که به طور دستی نشانه‌گذاری شده‌اند. مجموعه داده نقد فیلم کُرنل[۶] نیز از جمله پیکره‌های مناسب در این زمینه به شمار می‌آید (Pang, Lee:2002). از جمله دیگر پیکره‌های مربوط به تحلیل احساس می‌توان به پیکره JDPA اشاره نمود (Kessler, et al.:2010). این پیکره‌ی غنی شامل جملاتی عمدتاً برگرفته از مطالب و ارسال‌های تعدادی وبلاگ است. در پیکره JDPA علاوه بر انواع نشانه‌گذاری‌ها، آمار کاملی شامل نتایج مربوط به محاسبه‌ی میزان توافق میان نشانه‌گذارها[۷] نیز ارائه شده است. از جمله دیگر کارهای مطلوب انجام شده در زبان انگلیسی می‌توان به پیکره‌ی توسعه داده شده در (Blitzer:2007) اشاره نمود که شامل نظرات مربوط به محصولات برگفته از سایت آمازون است. در این پیکره بار معنایی توسط یک بازه عددی از ۱ تا ۵ مشخص شده است.

برخلاف پیکره‌های اشاره شده برای انگلیسی، تعدادی محدود از پیکره‌ها نیز برای زبان‌های غیر انگلیسی توسعه داده شده‌اند و به طور خاص تاکنون پیکره‌ای که برای زبان فارسی در زمینه‌ی تحلیل احساس توسعه داده شده باشد، وجود ندارد. البته در رابطه با زبان فارسی می‌توان تعدادی پیکره‌ی مناسب مربوط به سایر زمینه‌های تحقیقاتی را نام برد. نظیر پیکره‌هایی برای نشانه‌گذاری نقش اجزای جمله[۸] (Bijankhan, et al.:2011)، تجزیه کردن وابستگی[۹] (Rasooli, et al.:2013) و مشخص کردن موضوع[۱۰] (AleAhmad. et al:2009). اما در رابطه با پیکره برای سایر زبان‌های غیر انگلیسی می‌توان به (Rushdi Saleh, et al. 2011) اشاره نمود که یک پیکره برای زبان عربی به حساب می‌آید. این پیکره شامل ۵۰۰ نقد فیلم برگرفته از

---

[۴] Target
[۵] Token
[۶] Cornell movie review dataset
[۷] Inter-annotator agreement
[۸] POS tagging
[۹] Dependency parsing
[۱۰] Topic detection

وبلاگ‌ها و سایت‌های عربی است. در این پیکره جملات موجود به دو دسته‌ی مثبت و منفی تقسیم شده‌اند و سپس قابلیت اعتماد پیکره توسط بررسی صحت عملکرد تعدادی از الگوریتم‌های دسته‌بندی نظیر SVM بروی آنها مورد سنجش قرار گرفته است. از جمله دیگر پیکره‌های به زبان غیر انگلیسی می‌توان به ChnSentiCorp اشاره نمود که شامل ۱۰۲۱ سند در سه زمینه‌ی آموزش، فیلم و امور مربوط به مسکن است (Tan, Zhang:2008). به عنوان یک پیکره غیر انگلیسی زبان دیگر می‌توان به MLSA اشاره داشت که یک پیکره‌ی نشانه‌گذاری شده چند لایه (شامل سطوح سند، جمله و عبارت) قابل دسترس به طور عمومی برای زبان آلمانی است (Clematide, et al.:2012). توسعه این پیکره شامل نشانه‌گذاری ۲۷۰ سند به زبان آلمانی بوده است. همچنین از معیار Fleiss در این پیکره جهت سنجش میزان قابلیت آن استفاده شده است. در پایان همچنین می‌توان به تعدادی از پیکره‌های چند زبانی نیز اشاره داشت. از جمله این پیکره‌ها NTCIR بوده که سه زبان ژاپنی، انگلیسی و چینی را شامل می‌شود. فرایند نشانه‌گذاری و همچنین ارزیابی پیکره به طور جداگانه برای هر کدام از زبان‌های موجود در این پیکره به طور مفصل شرح داده شده است (Seki, et al.: 2008). از جمله پیکره‌های دو زبانی دیگر در زمینه‌ی تحلیل احساس شامل زبان‌های انگلیسی و آلمانی می‌توان USAGE را نام برد (Klinger, Climiano:2014).

۳. منبع داده و مراحل توسعه پیکره

بدون تردید در روند توسعه‌ی هر پیکره اولین و شاید یکی از مهم‌ترین گام‌ها انتخاب یک منبع داده‌ی مناسب و جامع است. داده‌ی اولیه‌ی استفاده شده در این پیکره برگرفته از وبسایت دیجی‌کالا[۱۱] بوده که این وب‌سایت یکی از برترین و معروف‌ترین وب‌سایت‌های فروش برخط انواع محصولات الکترونیکی[۱۲] در ایران به شمار می‌آید. در کنار فروش انواع محصولات، روزانه کاربران بسیاری اقدام به بازدید از این وب‌سایت نموده و نظرات خود را در رابطه با محصولات مختلف مطرح می‌نمایند. تمامی این ویژگی‌ها دیجی‌کالا را به یک گزینه‌ی مناسب جهت برگزیدن به عنوان یک منبع داده‌ی مطلوب، جهت تهیه‌ی پیکره‌ی مد نظر تبدیل نموده است. یکی دیگر از ویژگی‌های مطلوب دیجی‌کالا آن است که نقدها و نظرات آن شامل هر دو صورت رسمی و غیر رسمی می‌شوند. به طور مشخص معمولاً نقدی که توسط متخصص وب‌سایت در رابطه با انواع کالاها نوشته می‌شود به صورت رسمی و نظرات عمومی و نقدهای ارائه شده توسط کاربران به صورت غیر رسمی‌تر نگاشته می‌شوند.

پس از انتخاب دیجی‌کالا به عنوان منبع داده‌ی پیکره، این وبسایت توسط یک نرم افزار خزنده[۱۳] به طور کامل مورد پویش قرار گرفت و صفحات HTML مربوط به محصولات مختلف از آن استخراج شد. در نهایت با استفاده از نرم افزاری که توسعه دادیم و شرح آن در بخش ۸ داده خواهد شد و پس از طراحی یک ساختار

---

[۱۱] http://www.digikala.com

[۱۲] با توجه به اینکه انواع متنوعی از محصولات در وبسایت دیجی‌کالا عرضه می‌شوند، در اینجا و در پیکره‌ی مد نظر، تنها از متون مربوط به نظرات و نقدهای کاربران و متخصصین وبسایت دیجی‌کالا در رابطه با لوازم الکترونیکی (شامل گوشی تلفن همراه، دوربین‌های عکاسی و فیلم‌برداری، لپ‌تاپ و کتاب‌خوان و لوازم و تجهیزات رایانه و غیره) استفاده شده است. آمار مربوط به انواع محصولات استفاده شده در پیکره به طور کامل در جدول ۳ ارائه شده است.

[۱۳] Crawler

مشخص برای اسناد موجود در پیکره، اسناد خام XML و اولیه‌ی مربوط به پیکره ایجاد شدند. مهم‌ترین اجزای این اسناد در بخش ۶ شرح داده خواهند شد. در مرحله بعدی، تمامی این اسناد توسط چهار نفر و در طی مدت تقریبی چهار ماه نشانه‌گذاری شدند.

۴. نشانه‌گذاری اسناد پیکره

در فرایند نشانه‌گذاری اسناد برخی از مفاهیم به طور مکرر مورد استفاده قرار گرفته‌اند. در این بخش بر آن هستیم که به طور مختصر و در عین حال جامع به معرفی این مفاهیم بپردازیم.

۴-۱. انواع برچسب‌ها[۱۴]

در مجموع چهار برچسب با اسامی Target(M)[۱۵]، Target(I)[۱۶]، Opinion، و Keyword در پیکره موجود هستند. از میان این برچسب‌ها، نوع Keyword را در بخشی جداگانه بررسی خواهیم نمود.

طبق تعریف می‌توان گفت که به طور کلی، نشان، موجودیت یا ویژگی‌ای است که نظری در رابطه با آن ابراز شده باشد. در اینجا ما دو نوع نشان داریم. نشان اصلی که با Target(M) نشان داده می‌شود و نشان نمونه که آن‌را با Target(I) نشان می‌دهیم. تفاوت مابین این دو نوع را با ذکر یک مثال شرح می‌دهیم. جمله‌ی "*این تلفن همراه زیبا رو هفته پیش خریدم. این گوشی واقعاً خوب هست*" را در نظر بگیرید. در این جمله ترکیب "*تلفن همراه*" و کلمه‌ی "*گوشی*" در واقع هر دو به یک موجودیت مشترک اشاره دارند. در اینگونه موارد، برای پرهیز از تعدد درج نشان‌های مختلف، از مفهوم نشان اصلی استفاده می‌کنیم. بدین معنی که اگر هر بار نشانه‌گذار موجودیت جدیدی را شناسایی کند، در مرتبه‌ی اول یک نشان اصلی از آن موجودیت ایجاد می‌شود و اگر در دفعات بعدی و در دیگر جملات همان نظر یا نقد در سند مربوطه مجدداً همان نشان را شناسایی کرد، به جای درج یک نشان جدید یک نشان نمونه از نشان اصلی که تاکنون ایجاد کرده به وجود آورد. با انجام اینکار رابطه‌ی بین تمامی موجودیت‌ها و نشان‌ها در سراسر جملات یک نظر یا نقد که در واقع مبدأ و ماهیت یکسانی دارند حفظ می‌شود و کلمات نظر منتسب به همه‌ی آنها را یکجا در اختیار خواهیم داشت.

دیگر نوع برچسب موجود کلمه‌ی نظر است که آن‌را با نماد Opinion در پیکره نشان می‌دهیم. کلمه‌ی نظر در واقع نظری است که یک فرد در رابطه با یک نشان اظهار می‌دارد. به عنوان نمونه در جمله‌ی مثال در پاراگراف پیشین، کلمات "*زیبا*" و "*خوب*" به ترتیب برای نشان‌های نمونه‌ی "*تلفن همراه*" و "*گوشی*" یک کلمه‌ی نظر محسوب می‌شوند. قابل ذکر است که برای هر نشان نمونه بیش از یک کلمه نظر هم می‌تواند وجود داشته باشد.

---

[۱۴] Tag
[۱۵] Main target
[۱۶] Instance target

۴-۱-۱. نوع برچسب کلمه‌ی کلیدی

کلمات کلیدی در واقع به نوعی مشابه کلمات نظر هستند و به آن دسته از کلماتی اطلاق می‌شوند که می‌توانند در یافتن دلیل انتساب یک بار معنایی به یک جمله ما را یاری نمایند. از طرفی این کلمات می‌توانند پایه و منبع خوبی برای تشکیل یک مجموعه دایره‌ی لغات برای استفاده در تحلیل احساس به شمار آیند. ذکر این نکته ضروری است که یک کلمه‌ی کلیدی می‌تواند یک کلمه‌ی نظر نیز باشد. از طرفی کلمات کلیدی لزوماً دارای بار معنایی نیستند. به عنوان نمونه‌ای از کاربرد کلمات کلیدی جمله‌ی "*این گوشی واقعاً خوبه*" را در نظر بگیرید. با مشاهده‌ی کلمه "*خوب*" نشانه‌گذار تشخیص می‌دهد که بار معنایی جمله مثبت است. اما با حضور کلمه‌ی "*واقعاً*" می‌توان به نوعی استنباط نمود که نظر دهنده سعی در تأکید خوب بودن گوشی دارد و از این رو نشانه‌گذار می‌تواند "*خیلی مثبت*" را به عنوان بار معنایی جمله در نظر بگیرد. در واقع در این مثال کلمه‌ی "*واقعاً*" می‌تواند یک کلمه‌ی کلیدی باشد که ارتباطی معنادار با بار معنایی نسبت داده شده به جمله داشته باشد.

۴-۲. نمره‌دهی[۱۷] در پیکره

مقصود از نمره‌دهی در واقع همان انتساب یک بار معنایی به جملات و یا به یک برچسب است. در کل دو نوع نمره‌دهی در فرایند نشانه‌گذاری موجود است:

- *نمره‌دهی به جملات:* این نمره عددی است از مجموعه‌ی اعداد {۲+، ۱+، ۰، ۱-، ۲-} که به یک جمله بر حسب میزان مثبت یا منفی بودن احساس آن نسبت داده می‌شود. طبیعتاً جمله با نمره‌ی ۲+ مثبت‌ترین و جمله با نمره‌ی ۲- منفی‌ترین جمله تلقی می‌شوند. همچنین جمله‌ی با نمره ۰ از نظر فرد نشانه‌گذار حاوی بار معنایی نیست. البته این بازه می‌توانست شامل اعداد بیشتری نیز باشد، که چنین کاری روند نشانه‌گذاری را پیچیده‌تر و کار را حتی برای انسان سخت‌تر می‌نماید. در ادامه چند نمونه از انواع نمره‌های نسبت داده شده به جملات را خواهیم دید.
  - *نمره‌ی ۲-*: این گوشی فاجعه هست و کاملاً نا امیدم کرد.
  - *نمره‌ی ۰:* این گوشی رو ماه پیش از دیجی‌کالا خریداری کردم.
  - *نمره‌ی ۱+:* مصرف انرژی گوشی خوبه. در مجموع ازش راضیم.

- *نمره‌دهی به برچسب‌ها:* از میان انواع برچسب‌ها تنها دو نوع برچسب کلمه‌ی نظر و کلمه‌ی کلیدی می‌توانند حاوی یک بار معنایی باشند. در صورتی که بار معنایی این دو نوع کلمه برچسب مثبت باشد نمره‌ی مثبت و در صورتی که بار معنایی آنها منفی باشد، نمره‌ی منفی به آنها نسبت داده می‌شود.

---

[۱۷] Rating

۵. اجزای اسناد موجود در پیکره

همانطور که پیشتر نیز اشاره شد، اسناد موجود در پیکره در قالب فایلهای XML ذخیره شده‌اند. در این بخش قصد داریم تا مهم‌ترین اجزای این فایلها را با جزییات بیشتری معرفی نماییم. در تمامی این اجزا برخی از ویژگی‌ها کاملاً مشترک هستند. ویژگی ID شناسه‌ی منحصر به فرد مربوط به جملات از هر نوع (نظر متخصص، نظرات عمومی کاربران، و نقدهای کاربران) است. ویژگی Value در سراسر اسناد به معنای نمره‌ای است که به بار معنایی یک کلمه یا جمله نسبت داده می‌شود. همچنین ویژگی Holder دربردارنده‌ی نام شخصی است که نظر داده شده منتسب به اوست. در ادامه به بررسی مهمترین اجزای اسناد خواهیم پرداخت.

ابتدا به معرفی عناصری می‌پردازیم که حاوی جملات هر سند هستند. عنصر Review شامل جملات نقدی است که توسط متخصص وبسایت در مورد یک محصول مشخص داده شده است. جملات درون این نقد نیز به تفکیک درون برچسب‌هایی با نام Sentence ذخیره شده‌اند. ساختار این عنصر به شکل زیر است:

```
<Review ID="" Value="">
   <Sentence ID="" Value=""></Sentence>
</Review>
```

عنصر General_Reviews در یک نگاه سطح بالا دربردارنده‌ی مجموعه نظرات عمومی کاربران است که بدنه‌ی هر نظر به تفکیک جملات جداگانه در عنصری با نام General_Review ذخیره شده است. همانند عنصر Review در اینجا نیز جملات درون برچسب‌هایی با نام Sentence ذخیره شده‌اند. ساختار این عنصر را به ترتیب زیر داریم:

```
<General_Reviews>
   <General_Review ID="" Holder="" Value="">
      <Sentence ID="" Value=""></Sentence>
   </General_Review>
</General_Reviews >
```

عنصر Critical_Reviews نیز ساختاری کاملا مشابه با General_Reviews دارد با این تفاوت که این عنصر حاوی نقدهایی است که کاربران در رابطه با یک محصول مشخص اظهار داشته‌اند. همچنین دو ویژگی Voters و Score به ترتیب نشان دهنده‌ی مجموع تعداد آرای داده شده به نقد مورد نظر و تعداد آرای مثبت داده به آن نقد هستند. ساختار این عنصر به ترتیب زیر است:

```
<Critical_Reviews>
   <Critical_Review ID="" Holder="" Score="" Voters="" Value="">
      <Sentence ID="" Value=""></Sentence>
   </Critical_Review >
</Critical_Reviews>
```

اشاره به این نکته خالی از لطف نیست که یکی از دلایل جداسازی جملات در قالب سه عنصر متفاوت، آن است که این جملات از برخی جهات با یکدیگر تفاوت‌هایی دارند. به عنوان مثال، جملات موجود در عنصر Review در واقع دارای ساختاری رسمی اما دو دسته‌ی دیگر حاوی جملاتی هستند که عموماً رسمی به

شمار نمی‌آیند. همچنین جملات موجود در عنصر General_Reviews نسبت به جملات موجود در Critical_Reviews کوتاه‌تر بوده و بیشتر به زبان عامیانه نزدیک‌ترند.

دیگر عنصر با اهمیت درون اسناد عنصر Tags بوده و خود حاوی برچسب‌هایی با نام Tag است و ساختار آن را به صورت زیر داریم:

```
<Tags>
   <Tag Type="" ID="" Coordinate="" Relation="" Root="" Synonym="" Value="" />
</Tags>
```

این عنصر دربردارنده برچسب‌هایی است که توسط نشانه‌گذارها در جملات شناسایی شده‌اند. ویژگی Type نشان‌دهنده نوع برچسب بوده و یکی از مقادیر Opinion، Target(I) یا Target(M) را شامل می‌شود. ویژگی Coordinate حاوی مختصات برچسب مورد نظر در جمله است و تنها برای برچسب‌های از نوع Opinion و Target(I) دارای مقدار است. ویژگی Relation برای هر کدام از برچسب‌ها می‌تواند معنای مشخصی داشته باشد. این ویژگی برای برچسب از نوع Target(M) حاوی شناسه‌ی مربوط به نظر و یا نقدی است که این برچسب در آن شناسایی شده است. همچنین مقدار این ویژگی برای برچسب از نوع Target(I) برابر است با شناسه آن Target(M) ای که ریشه‌ی برچسب مورد نظر است. در نهایت برای برچسب از نوع Opinion این ویژگی حاوی شناسه‌ی آن Target(I) ای است که نظر در مورد آن داده شده است. ویژگی‌های Root و Synonym هم به ترتیب حاوی ریشه کلمه‌ی برچسب مورد نظر و مترادف کلمه برچسب مورد نظر هستند. البته دو ویژگی اخیر تنها برای برچسب‌های از نوع Opinion و Target(I) مقدار می‌گیرند.

دیگر عنصر با اهمیت موجود در اسناد عنصر Keywords است. این عنصر حاوی برچسب‌هایی با نام Keyword بوده و ساختار آن به شکل زیر است:

```
<Keywords>
   <Keyword ID="" Coordinate="" Root="" Synonym="" Value="" />
</Keywords>
```

این عنصر در بردارنده‌ی کلمات کلیدی است که در جملات توسط نشانه‌گذارها شناسایی شده‌اند. تمامی ویژگی‌هایی که برای برچسب‌های موجود در این عنصر وجود دارد عیناً دارای معانی مشابه با برچسب‌های موجود در عنصر Tags هستند.

۶. چالش‌های فرایند نشانه‌گذاری

در این قسمت به برخی از چالش‌های موجود و رایج در فرایند نشانه‌گذاری فایلها اشاره خواهد شد. ذکر این نکته ضروری است که بسیاری از این چالش‌ها در حوزه‌ی پردازش زبان‌های طبیعی از جمله مسائل تعریف شده هستند و برخی از آنها ممکن است که در فارسی نسبت به برخی زبان‌های دیگر چالش‌برانگیزتر باشند. از نگاهی دیگر، برخی از این چالش‌ها برآمده از صورت نوشتاری و نگارشی هستند، در حالی که برخی دیگر به محتوی جملات و ارتباط کلمات با یگدیگر مربوط می‌شوند.

وجود اصطلاحات و ضرب‌المثل‌ها می‌تواند یکی از چالش‌های پردازش متون و در نتیجه آن فرایند نشانه‌گذاری در زبان‌های مختلف به شمار آید و زبان فارسی نیز از این قائده مستثنی نیست. به عنوان مثال،

در جملاتی نظیر *"این گوشی گل کاشته"* یا *این "لپ‌تاپ واقعاً غوله"*، اصطلاحاتی نظیر *"گل کاشتن"* و یا *"غول بودن"* صرف نظر از آنچه می‌نمایند به معنی دارا بودن ویژگی‌های بسیار خوب هستند و نوعی تعریف از کالا و یا جنبه‌ی مورد بحث به شمار می‌آیند. در مواجه با چنین مواردی، در پیکره، علاوه بر نشانه‌گذاری اصطلاح مورد نظر، به عنوان نمونه نشانه‌گذاری *"گل کاشته"* و *"غوله"* در دو جمله‌ی مورد نظر، یک ویژگی اضافی نیز برای هر برچسب با نام Synonym که پیشتر در مورد آن توضیح داده شد نیز در نظر گرفته شده است تا نشانه‌گذار بتواند معادل آن اصطلاح را نیز در قالب یک مترادف مطلوب وارد نماید. به عنوان مثال، وارد نمودن معادل *"خیلی خوب"* برای گل کاشته و یا *"خیلی قوی"* یا *"عالی"* برای غوله. وجود چنین امکانی می‌تواند در پردازش خودکار متون در آینده تا اندازه‌ی زیادی یاری‌گر باشد.

از دیگر چالش‌های فرایند نشانه‌گذاری جملات وجود کلمات عامیانه و یا صورت نوشتاری غیر رسمی کلمات است. این عادت نوشتاری به ویژه در متون غیر رسمی رواج بیشتری دارد. به عنوان مثال، به کار بردن واژه‌ی "کوچیک" به جای صورت صحیح *"کوچک"* و یا به کار بردن واژه‌ی *"اصن"* به جای *"اصلاً"*. همانند به کار بردن ویژگی Synonym برای درج مترادف اصطلاحات در اینجا نیز استفاده از ویژگی با نام Root می‌تواند تا اندازه‌ای در فرایندهای بعدی کمک کننده باشد. بدین صورت که در مواقعی که صورت غیر رسمی کلمات را مشاهده نمودیم، در هنگام درج آنها به عنوان یک برچسب، نشانه‌گذار می‌تواند صورت رسمی و صحیح آن کلمه را در ویژگی Root آن برچسب درج نماید تا بدینوسیله تا اندازه‌ای ماشین را در مواجه با کلمات و متون غیر رسمی یاری کند. بنابراین در آینده و در هنگام پردازش متون، به ویژه متون غیر رسمی، هر زمان با کلمه‌ای به ظاهر نا آشنا و عامیانه مواجه شدیم می‌توانیم از طریق جستجو بررسی نماییم که آیا مورد مشابهی از قبل نشانه‌گذاری شده است یا خیر و اگر آری، معادل صحیح آن را از طریق ویژگی Root مورد دسترسی قرار دهیم.

از دیگر عادات رایج و شاید نادرست در نوشتار غیر رسمی که باعث به وجود آمدن مشکلاتی در فرایند نشانه‌گذاری جملات می‌شود، استفاده از صورت نوشتاری به اصطلاح فینگلیش[۱۸] در متون است. به عنوان مثال، نوشتن کلمه "جی.پی.اس" به جای صورت صحیح و اصلی آن، GPS و یا نوشتن کلمه Perfect به صورت "پِرفِکت" و نظایر آنها. از طرفی مواجه با چنین کلماتی به ویژه در متون غیر رسمی در هر صورت اجتناب ناپذیر بوده و باید چاره‌ای اندیشید تا ماشین در هنگام پردازش این دست کلمات دچار مشکل نشود. در چنین مواردی نیز ویژگی Synonym که پیشتر در مورد آن توضیح داده شد می‌تواند بکار برده شود. بدین ترتیب که در هنگام نشانه‌گذاری کلمات این چنینی معادل صحیح فارسی آنها در جای ویژگی Synonym درج گردد. به عنوان مثال درج کلمه *"کامل"* یا *"بی نقص"* در هنگام درج برچسب پِرفِکت در جمله.

از چالش‌های مربوط به صورت نوشتاری کلمات اگر بگذریم، دسته‌ای از چالش‌های مربوط به محتوی نیز در فرایند نشانه‌گذاری جملات خودنمایی می‌کنند. یکی از این مسائل بحث صورت گرفتن مقایسه بین دو موجودیت در جملات است. به عنوان مثال در جمله‌ای همانند: *گوشی‌های سامسونگ نسبت به نوکیا بهترین* هستند. در این جمله، نظر دهنده به طور مشخص کلمه نظر بهترین را برای موجودیت گوشی سامسونگ بکار برده است. اما مسئله اینجاست که این نظر در مقایسه با نوعی دیگر از گوشی، در اینجا نوکیا، عنوان شده

[۱۸] فینگلیش یا پینگلیش به معنی نوشتن کلمات انگلیسی با استفاده از حروف فارسی است.

است و آیا می‌توان گفت که صفت بهترین برای موجودیت گوشی سامسونگ همواره و در مقایسه با سایر گوش‌ها از شرکت‌های دیگر صدق می‌کند؟ مسئله مقایسه در مبحث تحلیل احساس بسیار حائز اهمیت بوده و در (Liu:2012)، در فصلی جداگانه به طور مفصل به آن پرداخته شده است.

از جمله دیگر چالش‌هایی که می‌توان بدان اشاره نمود، مشکل بودن تشخیص بار معنایی صحیح جمله توسط نشانه‌گذار در برخی موارد است. این مسئه می‌تواند ناشی از بکار برده شدن کلمات نظر متعدد و گاه در تضاد هم باشد. به عنوان مثال در جمله: *کیفیت عکس این گوشی خیلی خوب نیست، اندازه اون هم یه خرده بزرگه، ولی دوسش دارم*، با توجه به اظهار خوب نبودن کیفیت عکس و عدم رضایت از اندازه، و از طرفی ابراز علاقمندی به موجودیت گوشی، شاید کمی مشکل بتوان به طور قطع مشخص نمود که در مجموع نظر دهنده چه نظری در رابطه با موجودیت گوشی دارد.

۷. ابزار نشانه‌گذاری و وضعیت دسترسی به پیکره

جهت انجام فرایند نشانه‌گذاری فایل‌ها یک نرم افزار توسط یکی از اعضای تیم توسعه داده شد. نام این نرم افزار ابزار نشانه‌گذاری احساس گیلان[۱۹] بوده و در وبسایت[۲۰] گروه پردازش زبان‌های طبیعی دانشگاه گیلان قابل دسترسی است. در این نرم‌افزار علاوه بر امکانات مربوط به نشانه‌گذاری اسناد قسمتی نیز برای پیمایش فایل‌های HTML برای دستیابی به برچسب‌های حاوی اطلاعات مورد نظر تعبیه شده است. همچنین اولین قسمت از پیکره تحت عنوان SentiPers V1.0 با دارا بودن بیش از ۸۰۰۰ جمله از طریق آدرس http://nlp.guilan.ac.ir/Dataset.aspx قابل دسترسی برای عموم محققان و علاقمندان است.

۸. آمار مربوط به پیکره

در این بخش قصد داریم که آمار مربوط به پیکره‌ی سِنتی‌پِرس را ارائه دهیم. همچنین در بخشی جداگانه در رابطه با نحوه محاسبه و نتایج مربوط به محاسبه میزان توافق میان نشانه‌گذارها صحبت بعمل خواهد آمد.

جدول ۱ مهم‌ترین آمار مربوط به پیکره شامل تعداد هر کدام از برچسب‌ها را به تفکیک نشان می‌دهد. همچنین در جدول ۲ تعداد هر کدام از برچسب‌های نظر و کلمات کلیدی به تفکیک بار معنایی آنها نشان داده شده است. منظور از خنثی آن دسته از برچسب‌هایی هستند که نه بار معنایی مثبت و نه بار معنایی منفی دارند. همچنین همانطور که در بخش ۵-۲ نیز اشاره شد، نشان‌ها دارای بار معنایی نیستند و به همین دلیل جایی در جدول ۲ نخواهند داشت.

جدول ۱ – آمار کلی مربوط به پیکره سِنتی‌پِرس

نوع	*اسناد*	*جملات*	*کلمه نظر*	*کلمه کلیدی*	*نشان نمونه*	*نشان اصلی*
تعداد	۲۷۰	۲۶.۷۶۷	۲۶.۹۹۶	۳۳.۱۳۶	۳۱.۳۷۵	۱۷.۴۲۲

[۱۹] Guilan Sentiment Annotation Tool (GSAT)

[۲۰] نرم‌افزار از طریق پیوند http://nlp.guilan.ac.ir/Software/GSAT.rar قابل دریافت برای عموم است.

جدول ۲ – تعداد کلمات نظر و کلمات کلیدی به تفکیک بار معنایی

نوع برچسب / بار معنایی	*مثبت*	*خنثی*	*منفی*
کلمه‌ی نظر	۲۱.۴۷۱	۱.۶۶۱	۳.۸۶۴
کلمه‌ی کلیدی	۲۴.۹۱۵	۱.۲۹۳	۶.۹۲۸

همچنین جدول ۳، تعداد هر کدام از انواع محصولاتی که در این پیکره اطلاعات مربوط به آنها نشانه‌گذاری شده است را نشان می‌دهد. همانطور که در جدول نیز مشخص است، از میان محصولاتی که فایلهای مربوط به آنها نشانه‌گذاری شده است، تلفن همراه، دوربین دیجیتال، دوربین فیلم‌برداری و تبلت بیشترین سهم را دارا هستند.

جدول ۳- تعداد فایلهای مربوط به هر نوع محصول در پیکره

نوع محصول	*تعداد*
تلفن همراه	۷۳
دوربین دیجیتال	۶۴
دوربین فیلم‌برداری	۳۰
تبلت	۳۰
نوت‌بوک	۱۵
پخش‌کننده موزیک	۱۲
چاپگر	۱۲
تجهیزات مربوط به رایانه	۱۱
تلویزیون	۱۱
کنسول بازی	۶
پویشگر[۲۱]	۵

جدول ۴ نیز تعداد کلمات، نشانه‌ها و همچنین میانگین طول جملات پیکره بر حسب تعداد کلمات را نشان می‌دهد.

جدول ۴ – آمار مربوط به تعداد کلمات در پیکره

عنوان	*تعداد*
کلمه	۵۱۵.۳۸۷
نشانه	۱۷.۶۳۵
میانگین طول جملات (بر حسب کلمه)	۱۹.۲۵

[۲۱] Scanner

۸-۱. توافق مابین نشانه گذارها (ت.م.ن)

به دلیل شرکت داشتن بیش از یک نفر در فرایند نشانه‌گذاری فایلها، محاسبه میزان توافق بین آنها امری مهم به شمار می‌آید. برای محاسبه این مقدار معیارهای متفاوتی وجود دارد که از جمله معروف‌ترین آنها می‌توان به معیارهایی نظیر Cohen's Kappa، Fleiss's K، Cronbach's Alpha، و Krippendorff's Alpha اشاره نمود (Hayes, Krippendroff:2007). جهت محاسبه‌ی ت.م.ن برای انواع برچسب‌های موجود (کلمات نظر، کلمات کلیدی و نشان‌ها)، ما ابتدا این مقدار را برای تمامی زوج فایل‌های نشانه‌گذاری شده‌ی ممکن بین چهار نشانه گذار بدست آوردیم. با فرض در نظر گرفتن اسامی A و B برای دو نشانه گذار، ت.م.ن از طریق رابطه‌ی زیر محاسبه شده است:

$$agr(A||B) = \frac{\left|A \text{ منطبق بر } B\right|}{\left|A \text{ منطبق } B \text{ بر}\right| + \left|A \text{ غیر منطبق } B \text{ بر}\right|}$$

در اینجا منظور از انطباق، یعنی آنکه هر دو نشانه‌گذار، یک کلمه‌ی مشخص با مختصات آغازین یکسان را در یک جمله واحد به عنوان یک برچسب با نوع یکسان نشانه‌گذاری نموده‌اند. در گام بعدی، جهت محاسبه کردن ت.م.ن نهایی از تمامی مقادیر بدست آمده در مرحله قبل برای زوج فایلهای ممکن میانگین ریاضی گرفته شد. جدول ۵ نتایج محاسبه ت.م.ن را برای انواع برچسب‌های موجود نشان می‌دهد. ذکر این نکته حائز اهمیت است که سطح قابلیت اطمینان مقدار بدست آمده برای ت.م.ن در انواع مختلف پیکره‌ها ممکن است متفاوت باشد. بنابراین شاید بهتر آن باشد که در رابطه با میزان قابل اعتماد بودن مقدار ت.م.ن بر حسب وظیفه‌ی موجود در فرایند نشانه‌گذاری قضاوت شود. همچنین بدین نکته نیز باید اشاره داشت که استفاده از معیار Cohen's Kappa جهت محاسبه‌ی ت.م.ن در اینجا چندان مناسب نیست، به دلیل آنکه تعداد برچسب‌هایی که توسط هر فرد در یک جمله واحد شناسایی و نشانه‌گذاری می‌شود می‌تواند متفاوت باشد و در اینجا ما نمی‌توانیم تعداد ثابتی از دسته‌ها را داشته باشیم.

جدول ۵ – مقادیر محاسبه شده مربوط به ت.م.ن

نوع برچسب	*ت.م.ن(٪)*
کلمه‌ی کلیدی	۴۷.۶۱
کلمه‌ی نظر	۴۶.۰۲
نشان نمونه	۴۲.۱۲

تاکنون به این نکته اشاره نموده‌ایم که به هر جمله درون پیکره از سوی نشانه‌گذارها یک بار معنایی نسبت داده شده است. جهت محاسبه‌ی ت.م.ن برای بار معنایی جملات، سه دسته شامل مثبت، خنثی و منفی در نظر می‌گیریم. معادله‌ای که جهت محاسبه‌ی ت.م.ن برای بار معنایی جملات در یک سند استفاده شده در ادامه آمده است:

$$pagr(A||B) = \frac{\text{تعداد انطباقات } A \text{ و } B}{\text{تعداد کل جملات در سند}}$$

در اینجا نیز، پس از بدست آوردن ت.م.ن برای تمامی زوج فایل‌های ممکن، از تمامی این مقادیر برای بدست آوردن نتیجه‌ی نهایی میانگین گرفته شد. نتیجه‌ی نهایی در جدول ۶ نشان داده شده است. با مقایسه‌ی مقادیر بدست آمده در جداول ۵ و ۶ متوجه یک میزان اختلاف در سطح مقادیر بدست آمده برای ت.م.ن می‌شویم. یکی از دلایل آنکه میزان توافق در نشانه‌گذاری برچسب‌ها پایین‌تر از میزان توافق در مشخص کردن بار معنایی جملات است، می‌تواند این باشد که در رابطه با تعیین بار معنایی، تعداد ثابت و مشخصی از اعداد (در اینجا یک بازه شامل پنج عدد) موجود است و انتخاب نشانه‌گذار یکی از این اعداد خواهد بود. در حین محاسبه ت.م.ن برای بار معنایی نیز سه دسته مثبت، منفی و خنثی در نظر گرفته شده است. اما در رابطه با ت.م.ن برای برچسب‌ها، تعداد برچسب‌هایی که در هر جمله توسط نشانه‌گذار شناسایی می‌شود بنا به تشخیص فرد نشانه‌گذار می‌تواند متنوع و متفاوت باشد.

جدول ۶- نتایج محاسبه ت.م.ن برای بار معنایی جملات

بخش	*ت.م.ن بار معنایی (٪)*
نظرات متخصص	۷۹.۴۲
نظرات عمومی کاربران	۷۸.۰۹
نقد کابران	۷۶.۵۳

۹. خلاصه و کارهای آتی

در این مقاله فرایند کامل توسعه‌ی یک پیکره‌ی احساس شامل جملات رسمی و غیر رسمی زبان فارسی به تفصیل شرح داده شد. منابع داده و نحوه‌ی به وجود آوردن پیکره به همراه ساختار فایل‌های موجود در آن نیز مورد بررسی قرار گرفت. همچنین در ادامه به برخی از چالش‌های مربوط به فرایند نشانه‌گذاری فایل‌ها از جمله مواردی که در زبان فارسی اندکی چالش برانگیزترند نیز اشاره شد. در پایان نیز آمارهای مربوط به پیکره و همچنین نحوه‌ی محاسبه‌ی میزان توافق مابین نشانه‌گذارها ارائه داده شد.

با توجه به ویژگی‌های سِنتی‌پِرس این پیکره می‌تواند منبع داده مناسبی جهت انجام امور پژوهشی و علمی برای زبان فارسی و زمینه‌ی کاری تحلیل احساس به شمار آید. یکی از ویژگی‌های شاخص این پیکره آن است که در آن نه تنها مثبت و یا منفی بودن جملات، بلکه شدت بار معنایی جملات نیز نشانه‌گذاری شده است. چنین ویژگی به همراه وجود کلمات کلیدی نشانه‌گذاری شده در جملات، می‌تواند زمینه‌ی مساعدی را برای بدست آوردن یک ارتباط معنادار مابین بار معنایی و برچسب‌ها به وجود آورد. به علاوه می‌توان پژوهش‌هایی را نیز جهت بدست آوردن یک ارتباط معنادار بین نظرات کاربران و امتیازی که آنان به یک محصول مشخص داده‌اند نیز انجام داد. تمامی موارد مذکور می‌تواند سِنتی‌پِرس را به عنوان یک منبع داده مناسب در زمینه عقیده کاوی مطرح نماید. ذکر این نکته نیز در اینجا بسیار ضروری است که درصدهای

مربوط ت.م.ن در رابطه با برچسب‌های مختلف (کلمه‌ی کلیدی، کلمه‌ی نظر، و نشان نمونه) در این پیکره ممکن است در نگاه اول اندکی تأمل برانگیز باشد. اما با توجه به ماهیت فرایند نشانه‌گذاری در پژوهش حاضر باید این را در نظر داشت که ت.م.ن بهتر است تا طبق نوع وظیفه‌ی نشانه‌گذاری مورد تحلیل قرار گیرد و بر این اساس درصد مذکور می‌تواند کاملاً طبیعی و قابل قبول باشد. همچنین صحت نشانه‌گذاری نسخه‌ی اول این پیکره که به طور عمومی در دسترس قرار گرفته، بسیار مطلوب و مطمئن است و می‌تواند با اطمینان بالایی در پژوهش‌های مد نظر استفاده شود. هرچند، قطعاً یکی از کارهای پیش‌رو در آینده‌ی نزدیک، ارتقاء بخشیدن هر چه بیشتر کیفیت نشانه‌گذاری در نسخه‌های بعدی و کامل این پیکره خواهد بود.

به عنوان فعالیت‌های آتی، در نظر داریم تا با استفاده از کلمات نظر و کلمات کلیدی نشانه‌گذاری شده در این پیکره اقدام به توسعه یک دایره‌ی واژگان در زمینه تحلیل احساس نماییم. همچنین قصد داریم تا انواع الگوریتم‌های موجود در زمینه یادگیری ماشین را برای بدست آوردن میزان صحت عملکرد آنها بروی این پیکره آزمایش کنیم. همچنین به عنوان کارهای بعدی می‌توان دامنه اسناد موجود در این پیکره را به زمینه‌هایی همچون ادبی، ورزشی و سیاسی نیز بسط و گسترش داد.

منابع

Liu, B.: Sentiment Analysis and Opinion Mining: Synthesis Lectures on Human Language Technologies, Morgan & Claypool Publishers, (2012).

Pang, B., Lee, L.: Opinion Mining and Sentiment Analysis, Information Retrieval, vol.2, no. 1-2, pp. 1–135, (2008).

Yu, H., and Hatzivassiloglou, V. 2003. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-03).

Kim, Soo-Min, and Eduard Hovy. ”Determining the sentiment of opinions.” Proceedings of the 20th international conference on Computational Linguistics. Association for Computational Linguistics, 2004 .

Bethard, Steven, et al. ”Automatic extraction of opinion propositions and their holders.” 2004 AAAI Spring Symposium on Exploring Attitude and Affect in Text. 2004.

Hu, Minqing, and Bing Liu. ”Mining and summarizing customer reviews.” Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2004.

Wiebe, J., Wilson, T., Cardie, C.: Annotating expressions of opinions and emotions in language, Language Resources and Evaluation, vol. 39, no. 2, pp. 165–210, (2005).

Kessler, J. S., Eckert, M., Clark, L., Nicolov, N: The ICWSM 2010 JDPA Sentiment Corpus for the Automotive, In Proceedings of the 4th International AAAI Conference on Weblogs and Social Media Data Workshop Challenge (ICWSM-DWC 2010), (2010).

Bijankhan, M., Sheykhzadegan, J., Bahrani, M., Ghayoomi, M.: Lessons from Building a Persian Written Corpus: Peykare: Language Resources and Evaluation, vol. 45, No. 2, pp. 143–164, (2011).

Blitzer, John, Mark Dredze, and Fernando Pereira. ”Biographies, bollywood, boomboxes and blenders: Domain adaptation for sentiment classification.” ACL. Vol. 7. 2007.

Klinger, Roman, and Philipp Cimiano. ”The USAGE review corpus for fine-grained, multi lingual opinion analysis.” Proceedings of the Language Resources and Evaluation Conference. 2014.

Rushdi Saleh, Mohammed, et al. ”OCA: Opinion corpus for Arabic.” Journal of the American Society for Information Science and Technology 62.10 (2011): 2045-2054.

Tan, Songbo, and Jin Zhang. ”An empirical study of sentiment analysis for Chinese documents.” Expert Systems with Applications 34.4 (2008): 2622-2629.

Clematide, Simon, et al. ”MLSA-A Multi-layered Reference Corpus for German Sentiment Analysis.” LREC. 2012.

Seki, Yohei, et al. ”Overview of multilingual opinion analysis task at NTCIR-7.” Proc. of the Seventh NTCIR Workshop. 2008.

Pang, B., Lee, L.: Seeing stars: exploiting class relationships for sentiment categorization with respect to rating scales: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics(ACL ’05), pp. 115–124, Michigan, USA(2005).

Yu, H., Hatzivassiloglou, V.: Towards Answering Opinion Questions: Separating Facts from Opinions and Identifying the Polarity of Opinion Sentences, Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing (EMNLP ’03), pp. 129–136, PA, USA, (2003).

Pang, B., Lee, L.:Thumbs up?: sentiment classification using machine learning techniques, Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing (EMNLP ’02), vol. 10, pp. 79–86, PA, USA, (2002).

Rasooli, M. S., Kouhestani, M., Moloodi, A. S.: Development of a Persian Syntactic Dependency Treebank, In Proceedings of the 2013 North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT), pp. 306–314, Atlanta, USA(2013).

AleAhmad, A., Amiri, H., Darrudi, E., Rahgozar, M., Oroumchian, F.: Hamshahri: A Standard Persian Text Collection, Knowledge-Based System, vol. 22, no. 5, pp. 382–387, (2009).

Ghayoomi, M., Momtazi, S., Bijankhan, M.: A study of corpus development for Persian, International Journal on Asian Language Processing, vol. 20, no. 1, pp. 17–34, (2010).

Windfuhr, G. L.: The Iranian Languages, Routledge Language Family Series (Taylor and Francis), pp. 418–419, (2009).

Mahootian, S.: Persian Grammer from the Style Views, Translated by Salmani, M., Second Edition, 1383. (In Persian).

Hayes, Andrew F., and Klaus Krippendorff. ”Answering the call for a standard reliability measure for coding data.” Communication methods and measures 1.1 (2007): 77-89.