Journal of Information and Communication Technology

Volume 17, Issue 65, Special Issue on "Artificial Intelligence and data analytics", October 2025, pp 62-72

Developing a Persian Sentiment Analysis Dataset for Social Media Texts

Parisa Mohamadi Kalkhoran[1], Mojgan Farhoodi[2,*]

[1] IT Faculty, ICT Research Institute (ITRC), Tehran, Iran
[2] IT Faculty, ICT Research Institute (ITRC), Tehran, Iran

Received: 03 May 2024, Revised: 15 March 2025, Accepted: 29 March 2025
Paper type: Research

Abstract

This paper presents a Persian dataset for sentiment analysis of texts published on social media. The dataset creation process involved several key stages: First, a comprehensive methodology was developed for data collection and labeling. Next, data was extracted from the Twitter platform using keyword-based methods, hashtags, and high-engagement accounts. During the preprocessing stage, irrelevant data and textual noise were removed and corrected. The data labeling process was conducted manually through a crowdsourcing platform, where each tweet was labeled by three annotators, and the final label was determined based on majority voting. To ensure quality control, a calibrated dataset was prepared, and inter-annotator agreement was evaluated. The final dataset consists of over 5,000 tweets labeled as positive, negative, or neutral. Based on the results obtained from applying various models to this dataset, it can be concluded that this dataset serves as a reliable resource for developing and evaluating sentiment analysis models in the Persian language.

Keywords: Sentiment Analysis, Social Media, Twitter, Persian Dataset, Text Processing, Data Labeling

[*] Corresponding Author's email: farhoodi@itrc.ac.ir

دوفصلنامه
فناوری اطلاعات و ارتباطات ایران

سال هفدهم، شماره ۶۵، مهر ۱۴۰۴، شمارهٔ ویژهٔ «هوش مصنوعی و تحلیل داده»، صفحه ۶۲ الی ۷۲

ایجاد مجموعه دادگان فارسی تحلیل احساس در متون منتشرشده در شبکه‌های اجتماعی

پریسا محمدی کلخوران [۱]، مژگان فرهودی [۲*]

[۱] پژوهشگر پژوهشگاه ارتباطات و فناوری اطلاعات، تهران، ایران

[۲] عضو هیات علمی پژوهشگاه ارتباطات و فناوری اطلاعات، تهران، ایران

تاریخ دریافت: ۱۴۰۳/۰۲/۱۴ تاریخ بازبینی: ۱۴۰۳/۱۲/۲۵ تاریخ پذیرش: ۱۴۰۴/۰۱/۰۹

نوع مقاله: پژوهشی

چکیده

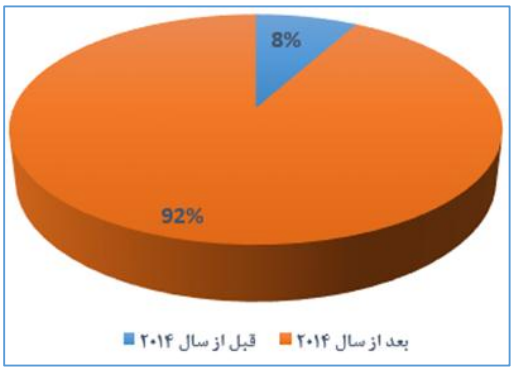
در این پژوهش، یک مجموعه دادگان فارسی برای تحلیل احساسات در متون منتشر شده در شبکه‌های اجتماعی تهیه شده است. فرآیند تهیه دادگان شامل چندین مرحله اساسی بوده است: ابتدا، یک شیوه‌نامه جامع برای جمع‌آوری و برچسب‌گذاری داده‌ها تدوین شد. سپس، داده‌ها از پلتفرم توییتر با استفاده از روش‌های مبتنی بر کلمات کلیدی، هشتگ‌ها و اکانت‌های پرمخاطب استخراج گردید. در مرحله پیش‌پردازش، داده‌های نامرتبط و نویزهای متنی حذف و اصلاح شدند. فرآیند برچسب‌گذاری داده‌ها به‌صورت انسانی و از طریق یک بستر جمع‌سپاری انجام شد، که طی آن هر توییت توسط سه نفر برچسب‌گذاری شد و برچسب نهایی بر اساس رأی اکثریت تعیین گردید. به‌منظور کنترل کیفیت، یک مجموعه داده کالیبره تهیه شد و میزان توافق بین برچسب‌گذاران مورد بررسی قرار گرفت. در نهایت، دادگان نهایی شامل بیش از ۵۰۰۰ توییت با برچسب‌های مثبت، منفی و خنثی آماده گردید. با توجه به نتایج حاصله از اعمال چندین مدل مختلف بر روی این مجموعه می‌توان گفت که این مجموعه داده می‌تواند به‌عنوان یک منبع معتبر برای توسعه و ارزیابی مدل‌های تحلیل احساسات در زبان فارسی مورد استفاده قرار گیرد.

کلیدواژگان: تحلیل احساسات، شبکه‌های اجتماعی، توییتر، مجموعه داده فارسی، پردازش متن، برچسب‌زنی داده‌ها.

* رایانامهٔ نویسنده مسؤول: farhoodi@itrc.ac.ir

۱- مقدمه

در سال‌های اخیر، شبکه‌های اجتماعی به بستری محبوب برای اشتراک‌گذاری افکار و نظرات مردم تبدیل شده‌اند. وبلاگ‌ها، میکروبلاگ‌ها، مجلات برخط، انجمن‌های گفتگو، فرم‌های نظرسنجی و دیگر امکانات ارتباطی مبتنی بر وب به مردم کمک می‌کنند تا احساسات خود را ابراز کرده و دیدگاه‌هایشان را در مورد موضوعات روزمره زندگی، مسائل اجتماعی، سیاسی و فرهنگی در سطح ملی و بین‌المللی با دیگران به اشتراک بگذارند. حجم گسترده و ماهیت متنوع اطلاعات متنی موجود در وب، همراه با تکامل مداوم آن، فرصتی منحصر به فرد برای مطالعه افکار عمومی و چالش‌های خاص موجود در پردازش این اطلاعات به وجود آورده است. درک و تفسیر حجم بالای داده‌های متنی منتشر شده در رسانه‌های اجتماعی، نیازمند ابزار قدرتمندی است که بتوان از بطن این متون، احساسات یا نظرات افراد را شناسایی کرده و آنها را طبقه‌بندی نمود. تجزیه و تحلیل احساسات فرایندی است که در آن، تلاش می‌شود که هیجانات، نظرات یا عواطف انسان از محتوای منتشرشده (مانند متن) به‌صورت خودکار شناسایی شود و کاربردهای زیادی دارد [۱]؛ از جمله کمک به تصمیم‌گیری (به عنوان مثال، برای انتخاب یک رستوران یا خرید یک محصول)، تحلیل میزان رضایت مشتری، انجام نظرسنجی و پیش‌بینی در مسائل تجاری، سیاسی و فرهنگی. روش‌های مختلفی برای تجزیه و تحلیل احساسات وجود دارد که در سال‌های اخیر روش‌های مبتنی بر یادگیری عمیق بیش از همه مورد توجه قرار گرفته‌اند که برای آموزش و آزمایش، وابسته به مجموعه داده‌ای باکیفیت هستند. با وجود آن که دادگان زبان انگلیسی در این حوزه نسبتاً غنی است، در زبان فارسی همچنان نیاز مبرمی به مجموعه داده‌های متنوع و مناسب وجود دارد، چرا که علیرغم وجود اشتراک بسیار در نحوه ابراز احساسات در میان انسان‌ها با نژادها و قومیت‌های مختلف، باید گفت که استفاده از دادگان ترجمه شده یک زبان به زبانی دیگر روش کارآمدی نیست، چرا که با توجه به فرهنگ و شرایط اجتماعی ممکن است یک واژه یا جمله در دو زبان و یا حتی در گویش‌های مختلف یک زبان، معانی متفاوت و گاه متضادی داشته باشد. با توجه به این نیاز، یکی از اهداف آزمایشگاه پارسی‌آزما در پژوهشگاه ارتباطات و فناوری اطلاعات، تهیه مجموعه داده‌ای با محوریت متون به‌دست‌آمده از رسانه‌های اجتماعی است که بتوان از آن برای توسعه و ارزیابی مدل‌های تحلیل احساس در متون فارسی استفاده کرد. در این راستا شبکه اجتماعی توییتر به عنوان منبع جمع‌آوری داده‌ها در نظر گرفته شد، چرا که توییتر با ایجاد ارتباط میان افرادی با پیشینه‌ها، فرهنگ‌ها و سنین مختلف، ارتباطات را متحول کرده است [۲]. شکل ۱ نشان می‌دهد که از سال ۲۰۱۴ به بعد، علاقه محققان به تحلیل احساسات توییتر در جهان در زمینه‌های مختلف مانند فیلم، موسیقی، ورزش، اخبار، سلامت، بازار سهام و غیره افزایش چشمگیری داشته است [۲].



شکل ۱. میزان علاقه پژوهشگران به تحلیل احساسات در شبکه اجتماعی توییتر

در ایران نیز با این که به دلیل محدودیت دسترسی، تنها ۹٫۲۴ درصد از مردم از توییتر استفاده می‌کنند [۳]، اما این شبکه اجتماعی در میان اقشار مختلف کاربران ایرانی، از جمله روزمره‌نویس‌ها، دانشجویان و احزاب سیاسی پایگاه قابل توجه و تاثیرگذاری دارد (هرچند که درصد نفوذ آن کمتر از برخی شبکه‌های اجتماعی دیگر است).

با این که توییتر امکان به اشتراک‌گذاری فیلم و عکس را نیز برای کاربرانش فراهم کرده است، اما از آنجا که هنوز متن رایج‌ترین شکل ارتباط محسوب می‌شود [۴]، در مقاله جاری، فقط به تجزیه و تحلیل احساسات مبتنی بر متن توجه شده است.

در مقاله حاضر درصددیم تا ضمن شرح فرآیند آماده‌سازی این مجموعه دادگان، یک مطالعه مقایسه‌ای از مجموعه داده‌های موجود در زبان فارسی نیز داشته باشیم. این مقاله در ادامه شامل بخش‌های زیر است: بخش ۲ مروری مختصر است بر کارهای مشابه قبلی. بخش ۳، روال تهیه مجموعه دادگان را به تفصیل شرح می‌دهد در بخش ۴ نتایج حاصل از چندین مدل بر روی این مجموعه دادگان آورده است و در نهایت در بخش ۵ به نتیجه‌گیری و پیشنهاد فعالیت‌های آتی پرداخته خواهد شد.

۲- کارهای انجام شده

تجزیه و تحلیل احساسات یا عقیده‌کاوی فرآیندی است برای شناسایی و تشخیص یا طبقه‌بندی احساسات یا نظرات کاربران برای هرگونه خدمتی مانند فیلم، محصول، رویداد و مانند این‌ها که می‌تواند مثبت، منفی یا خنثی باشد [۱]. اولین مجموعه داده منتشر شده در زمینه تحلیل احساس به سال ۲۰۰۲ برمی‌گردد که Polarity Dataset نام دارد. این مجموعه داده براساس نظرات ۱۴۴

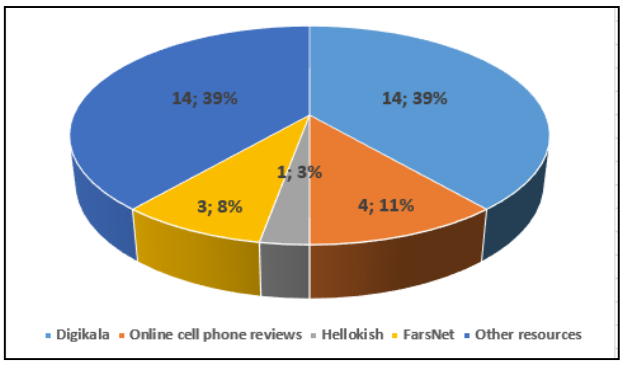
کاربر در مورد فیلم‌های موجود در پایگاه اینترنتی IMDB تهیه گردیده است. داده‌های این مجموعه در سه دسته مثبت، منفی و خنثی طبقه‌بندی شده‌اند [۶]. مجموعه داده IMDB که در سال ۲۰۱۱ ارائه شد، شامل ۵۰۰۰۰ نقد فیلم انگلیسی با برچسب مثبت یا منفی است [۷]. مجموعه داده Stanford Sentiment Treebank (SST) که در سال ۲۰۱۳ منتشر شد، بر اساس نقدهای فیلم Rotten Tomatoes است و شامل ۱۰۶۶۲ جمله می‌باشد که از این میان، نیمی منفی و نیم دیگر مثبت هستند. مجموعه داده Yelp Reviews حاوی نظرات مربوط به بررسی محصولات یا خدمات ارائه شده در سایت آمازون است [۸].

هر چند شروع تحلیل احساس در شبکه‌های اجتماعی به سال‌های ابتدایی قرن ۲۱ برمی‌گردد، اما تنها با ظهور برخی رسانه‌های اجتماعی مانند توییتر، فیسبوک، اینستاگرام و تلگرام بود که این موضوع به طور گسترده‌ای محبوبیت پیدا کرد. یکی از اولین مجموعه داده‌ها برای تحلیل احساس در رسانه‌های اجتماعی Sentiment140 نام دارد. این مجموعه داده که توسط دانشگاه استنفورد در سال ۲۰۰۹ ارائه شد، حاوی ۱,۶ میلیون توییت با برچسب مثبت، منفی و خنثی است که قطبیت توییت‌ها در آن، با ۰ برای منفی، ۲ برای خنثی و ۴ برای مثبت مشخص شده است [۱۴]. از ارزشمندترین مجموعه داده‌ها برای تجزیه و تحلیل احساسات، مجموعه دادگانی است که برای مسابقات SemEval تهیه شده‌اند. این مسابقات وظایف مختلفی را برای شرکت‌کنندگان مطرح می‌کنند که باید با استفاده از یک مجموعه داده مشترک انجام شوند. SemEval-2013 Task 2 [۱۵] و SemEval-2014 Task 9 [۱۶] نمونه‌هایی از مجموعه داده‌های با مقیاس بزرگ برای تجزیه و تحلیل احساسات در توییتر هستند. SemEval-2016 Task 4 [۱۷] شامل پنج زیرمسابقه مربوط به تجزیه و تحلیل احساس در توییت‌های منتشر شده در توییتر می‌باشد، در حالی که SemEval-2018 Task 1 [۱۸] با هدف طبقه‌بندی احساسات برای توییت‌ها به زبان‌های مختلف انگلیسی، عربی و اسپانیایی تهیه گردیده است. مجموعه داده RuSentiment، که در سال ۲۰۱۸ منتشر شد، شامل ۳۱۱۸۵ پست منتشر شده در شبکه اجتماعی VKontakte است که بزرگترین مجموعه داده در نوع خود برای زبان روسی می‌باشد. این مجموعه داده شامل سه دسته اصلی احساسات - مثبت، منفی و خنثی - و همچنین دو دسته اضافی است؛ که یکی از آنها دسته skip می‌باشد که شامل پست‌هایی نویزدار، پست‌های مبهم و یا پست‌هایی است که به زبان غیرروسی مانند اوکراینی منتشر شده‌اند و دسته دیگر، Speech Act نام دارد که در واقع زیرمجموعه‌ای از پست‌های مثبت است که رفتارهای رایج گفتاری مثبت مانند ابراز قدردانی، احوالپرسی و تبریک را پوشش می‌دهند [۱۹]. مجموعه داده چندوجهی MELD که در سال ۲۰۱۸ توسعه یافته، از افزایش و گسترش مجموعه داده متنی EmotionLines [۲۰] ایجاد شده است. EmotionLines شامل ۲۹۲۴۵ گفته از ۲۰۰۰ دیالوگ است که از گفتگوهای مجموعه تلویزیونی Friends و پیام‌رسان فیس‌بوک جمع‌آوری و در بستر جمع‌سپاری MTurk برچسب‌گذاری شده‌اند. MELD شامل همان نمونه گفتگوهای موجود در EmotionLines است، اما حالات صوتی و تصویری را نیز در کنار متن در بر می‌گیرد. MELD شامل بیش از ۱۴۰۰ گفتگو و ۱۳۰۰۰ گفته از گویندگان مختلف است. در هر گفتگو، هر گفتار غیر از برچسب هیجان، دارای حاشیه نویسی احساس (مثبت، منفی و خنثی) نیز هست [۲۱]. مجموعه دادگان GoEmotions، که در سال ۲۰۲۰ ارائه شده، شامل ۵۸۰۰۰ متن از شبکه اجتماعی ردیت از سال ۲۰۰۵ (شروع ردیت) تا ژانویه ۲۰۱۹ است که در ۲۷ طبقه برچسب‌گذاری شده‌اند (۱۲ طبقه مثبت، ۱۱ طبقه منفی، ۴ طبقه مبهم و یک طبقه خنثی) [۲۲]. مجموعه داده Twitter US Airline Sentiment که توسط CrowdFlower ایجاد شده است، حاوی بیش از ۱۴۰۰۰ توییت است که مشکلات شش شرکت هواپیمایی ایالات متحده را تجزیه و تحلیل می‌کند. این توییت‌ها از فوریه ۲۰۱۵ استخراج شده‌اند و دارای برچسب‌های مثبت، منفی یا خنثی هستند [۲۳].

در زبان فارسی، با توجه به منابع [۲۴] و [۳۶]، در طول سال‌های ۲۰۰۷ تا ۲۰۲۱، بیشتر مجموعه داده‌های واژگان احساسی مورد استفاده در مقالات از ترجمه مجموعه داده‌های انگلیسی به دست آمده‌اند. سایر مجموعه داده‌های متنی موجود، عمدتاً بر اساس نظرات مشتریان سایت‌ها درخصوص بررسی فیلم‌ها، محصولات فروشگاه‌های اینترنتی (بیش از همه دیجی‌کالا)، خدمات هتل، سرویس‌های سفارش غذا، فرسته‌های شبکه‌های اجتماعی (توییتر و اینستاگرام) و نظرات مردم در آنها و تعداد محدودی بر مبنای سایت‌های خبری، سایت سازمان بورس و اوراق بهادار، اشعار فارسی و یا مجموعه داده‌های حاصل از ترجمه ماشینی یا ترجمه نظرات مشتریان در سایت‌های خرده فروشی آنلاین تهیه شده‌اند. از جمله این مجموعه داده‌ها می‌توان به HelloKish [25]، MirasOpinion [26]، Iranian Stock Market [27]، Pars- ABSA [28], [29]، Digikala Sentiment [30]، JAMFA Corpus [31]، Snappfood [32], [24]، SentiPers [33]، DeepSentiPers [۳۴] اشاره کرد. شکل ۲، مجموعه داده‌های مورد استفاده در مقالات منتشر شده طی سال‌های ۲۰۱۸ تا ۲۰۲۲ در زمینه تحلیل احساسات در زبان فارسی را نشان می‌دهد [۲۴].

جدول ۱: خلاصه‌ای از ویژگی‌های مجموعه داده‌های متنی فارسی مبتنی بر شبکه‌های اجتماعی

نام مجموعه داده	منبع	سال	موضوع	حجم	روش برچسب‌گذاری
Political tweets [9]	توییتر	۲۰۱۶	توییت‌های سیاسی فارسی درخصوص مذاکرات هسته‌ای	یک میلیون توییت که ۳۰۰۰تای آنها برچسب‌گذاری شده است	هر توییت در مقیاس ۱ تا ۵ (۱،۲: منفی، ۳: خنثی، و ۴،۵: احساسات مثبت) برچسب گذاری شده است.
Foreign-Vaccine and Homegrown Vaccine [10]	توییتر	۲۰۲۲	توییت‌های فارسی در خصوص واکسیناسیون کرونا از ۱ آپریل ۲۰۲۱ تا ۳۰ دسامبر ۲۰۲۱	دو مجموعه داده شامل به ترتیب ۴۰۰۸۳۹ و ۴۰۲۴۳۹ توئیت	برچسب‌گذاری خودکار/ سه دسته: مثبت (۱+)، منفی (۱-) و خنثی (۰)
Large-Scale Colloquial Persian 0.5 (LSCP) [11]	توییتر	۲۰۲۰	توییت‌های فارسی	۱۲۰ میلیون جمله فارسی مستخرج از ۲۷ میلیون توییت	روش جمع‌سپاری نیمه خودکار/ برچسب قطبیت احساسات به صورت عددی بین ۰ و ۱
Insta-Text [12]	اینستاگرام	۲۰۲۰	نظرات کاربران در مورد پست‌های منتشر شده در صفحه اینستاگرام برنامه تلویزیونی «حالا خورشید»	۸۵۱۲ نظر	روش جمع‌سپاری/ سه دسته: مثبت (۱+)، منفی (۱-) و خنثی (۰)
Persian-English Code-mixed Texts [3]	توییتر	۲۰۲۱	توییت‌های ترکیبی فارسی-انگلیسی	۳۶۴۰ توییت	در این مجموعه داده ازسه برچسب‌زن استفاده شده و در نهایت رای اکثریت به عنوان برچسب نهایی لحاظ شده است.
Insta-MultiDSenti [13]	اینستاگرام	۲۰۲۲	نظرات کاربران در مورد پست‌های منتشر شده در صفحه اینستاگرام برنامه تلویزیونی «حالا خورشید»	نظرات کاربران در مورد یک یا دو سلبریتی و استعدادیابی تلویزیونی عصر جدید (مجموعه آموزش شامل ۱۸۱۸۲ نمونه و مجموعه آزمون شامل ۶۷۰۲ نمونه)	داده‌ها دردو دسته برچسب‌گذاری شده‌اند: مثبت و منفی
Our dataset	توییتر	۲۰۲۳	توییت‌های فارسی از ۲۲ مارس ۲۰۲۰ تا ۲۲ مارس ۲۰۲۲	۵۲۵۳ توییت	سه دسته: مثبت (۱+)، منفی (۱-) و خنثی (۰)



شکل ۲: مجموعه داده‌های مورد استفاده در مقالات منتشر شده فارسی در سال‌های ۲۰۱۸ تا ۲۰۲۲ (تعداد و درصد)

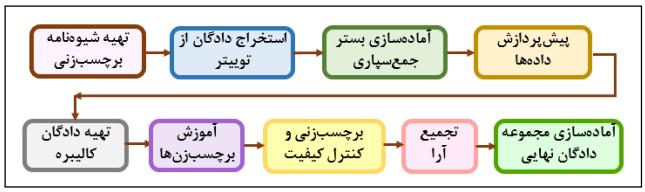
در سال‌های اخیر، تمرکز بر تحلیل احساسات در رسانه‌های اجتماعی در زبان فارسی بیشتر شده است. از جمله مجموعه داده‌های مرتبط با این موضوع که شامل متون توییتر و یا اینستاگرام هستند، می‌توان به مجموعه دادگان Foreign-Vaccine and Homegrown Vaccine [10]، LSCP [11]، Insta-Text [12]، Insta-MultiDSenti [۱۳] و مجموعه Persian-English Code-mixed Texts [۳] اشاره کرد. خلاصه‌ای از ویژگی‌های مجموعه داده‌های فارسی برگرفته از رسانه‌های اجتماعی در جدول ۱ آمده است. همان‌طور که ملاحظه می‌شود این مجموعه داده‌ها از یکی از دو شبکه اجتماعی اینستاگرام و یا توییتر استخراج شده‌اند که این دو شبکه از نظر طیف کاربران و مطالب منتشر شده در آنها فضای متفاوتی با هم دارند. تعدادی از این مجموعه‌های مورد اشاره، محدودیت موضوعی دارند (مثلا فقط به حوزه خاصی مثل نظرات سیاسی درباره مذاکرات هسته‌ای و یا نظرات درباره واکسن‌های کرونا و یا نظرات درباره یک برنامه تلویزیونی و یا بازیگران مشهور محدود شده‌اند) ودر نتیجه کاربردهای خاص‌تری دارند.

همچنین، بسیاری از این پژوهش‌ها از روش‌های برچسب‌گذاری انسانی بهره برده‌اند و لزوماً به روش‌های خودکار یا نیمه‌خودکار متکی نبوده‌اند. تفاوت کلیدی این پژوهش با مطالعات پیشین، استفاده از یک بستر بومی برای جمع‌سپاری داده‌ها و فرآیند دقیق برچسب‌گذاری انسانی است. در این پژوهش، تمامی مراحل از جمله تدوین شیوه‌نامه، آموزش برچسب‌گذاران، کنترل کیفیت و تهیه

داده‌های کالیبره به‌گونه‌ای طراحی شده‌اند که بالاترین سطح دقت و پایایی در داده‌های برچسب‌خورده تضمین شود. این ویژگی‌ها باعث می‌شود که دادگان ارائه شده، به‌ویژه برای کاربردهای حساس در پردازش زبان طبیعی فارسی، از اعتبار بالاتری برخوردار باشد.

۳- تهیه مجموعه داده

تهیه مجموعه داده معرفی شده در این مقاله طی مراحل مختلفی ایجاد شد که در شکل ۳ نمایش داده شده است. در ادامه به شرح هریک از این مراحل پرداخته می‌شود.



شکل ۳: مراحل مختلف تهیه دادگان

۳-۱ تهیه شیوه‌نامه

به‌منظور اطمینان از کیفیت و یکپارچگی برچسب‌گذاری داده‌ها، پیش از شروع فرآیند جمع‌آوری و برچسب‌زنی، یک شیوه‌نامه جامع تدوین شد. این شیوه‌نامه شامل تعاریف دقیق از دسته‌بندی‌های احساسی، دستورالعمل‌های برچسب‌گذاری، معیارهای انتخاب داده‌ها، و نحوه برخورد با موارد مبهم بود. برخی از مهم‌ترین قواعد این شیوه‌نامه عبارت‌اند از:

- تعریف دقیق برچسب‌ها: توییت‌ها به سه دسته مثبت، منفی و خنثی تقسیم شدند. توییت‌هایی که بیانگر احساسات مثبت واضح (مانند شادی، رضایت یا قدردانی) و یا نزدیک به این موارد بودند، در دسته مثبت قرار گرفتند، در حالی که توییت‌های دارای احساسات منفی (مانند خشم، ناراحتی یا اعتراض) و یا نزدیک به این آنها در دسته منفی جای داده شدند. توییت‌های خبری، اطلاعاتی یا فاقد احساس مشخص در دسته خنثی برچسب‌گذاری شدند.
- نحوه برخورد با متون مبهم: در صورتی که یک توییت حاوی عبارات کنایه‌آمیز، طنز یا استعاره بود، برچسب‌گذاران موظف بودند براساس محتوای کلی و زمینه آن تصمیم‌گیری کنند. اگر همچنان ابهام وجود داشت، این موارد در دسته «داده‌های نامشخص» قرار می‌گرفتند و از مجموعه نهایی حذف می‌شدند.
- سیاست‌های کنترل کیفیت: هر توییت توسط سه نفر بررسی شد و برچسب نهایی بر اساس رأی اکثریت تعیین گردید. علاوه بر این، یک مجموعه داده کالیبره برای ارزیابی دقت برچسب‌گذاران تهیه شد. برچسب‌گذارانی که دقت پایینی داشتند، از ادامه فرآیند حذف شدند.
- محدودیت‌های زبانی و محتوایی: توییت‌هایی که شامل زبان محاوره‌ای بیش‌ازحد، واژگان غیرمعمول یا دارای محتوای نامرتبط بودند، از مجموعه داده حذف شدند. همچنین، توییت‌هایی که حاوی عبارات توهین‌آمیز یا محتوای غیراخلاقی بودند، از داده‌های برچسب‌گذاری‌شده کنار گذاشته شدند.

این شیوه‌نامه، مبنای اصلی برای اطمینان از دقت و صحت فرآیند برچسب‌گذاری بوده و یکی از نقاط قوت این پژوهش در مقایسه با مطالعات پیشین محسوب می‌شود.

در شیوه‌نامه تهیه شده سعی شد تا مثال‌های گوناگونی برای هر سه احساس ذکر شود تا رویکرد واحدی در مواجهه با متون متفاوت اتخاذ شود. در این شیوه‌نامه از فرد خواسته می‌شد که برای تعیین احساس توییت، به احساس کلی گوینده متن توجه کند و نه احساس وی نسبت به یک مقوله یا موجودیت خاص. براساس این دستورالعمل، در صورت مواجهه با جملات پیچیده (شامل احساسات مختلف) یا متون ترکیبی (شامل چند بند یا جمله)، احساس غالب ملاک انتخاب گزینه مناسب قرار می‌گرفت. همچنین از فرد برچسب‌زن خواسته شده بود که در انتخاب احساس، از تفسیر ذهنی و علایق و باورهای شخصی اجتناب کند و فقط بر مبنای تجزیه و تحلیل زبان مورد استفاده در متن تصمیم بگیرد. به عنوان مثال، در جمله «جنگ میلیون‌ها پناهنده ایجاد کرده است.» با توجه به متن، برچسب «منفی» را انتخاب می‌کنیم، هر چند که ممکن است نویسنده در بیان این مطلب، هیچ هیجانی نداشته باشد و قصدش فقط نقل یک خبر باشد.

۳-۲ جمع‌آوری داده‌ها

برای استخراج داده‌ها از توییتر از کتابخانه SNScrape در پایتون استفاده شد. گردآوری توییت‌ها براساس دو رویکرد واژه‌محور و اکانت‌محور صورت پذیرفت. در رویکرد واژه‌محور، ابتدا ۵۰۰ کلمه که دارای بار عاطفی بودند از مجموعه Persian_NRC_EmoLex انتخاب گردید [۳۵]. این مجموعه واژگان که به اختصار به آن NRC Emotion Lexicon یا EmoLex نیز می‌گویند، شامل فهرستی از کلمات انگلیسی و ارتباط آنها با هشت هیجان اصلی (خشم، ترس، انتظار، اعتماد، تعجب، غم، شادی، انزجار) و دو احساس (منفی و مثبت) است که به صورت دستی برچسب‌گذاری شده‌اند. ترجمه این مجموعه واژگان به زبان‌های دیگر از جمله فارسی نیز موجود است. البته برای ترجمه آن از مترجم‌های ماشینی استفاده شده است و به همین دلیل برای زبان فارسی خیلی کیفیت مناسبی ندارد زیرا

بسیاری از واژه‌ها به درستی ترجمه نشده‌اند و یا معادل‌های فارسی آنها بار عاطفی زبان انگلیسی را ندارد. از این‌رو، در پژوهش حاضر، ۵۰۰ کلمه مذکور به صورت دستی و باتوجه به کاربردشان در زبان فارسی انتخاب شدند. از این ۵۰۰ کلمه در دو مرحله برای استخراج توییت‌ها استفاده شد. در مرحله اول، از این کلمات به عنوان یک پرس‌وجو استفاده شد؛ بدین ترتیب ۱۹۰۳۶ توییت استخراج گردید و در مرحله بعدی، از این کلمات به عنوان هشتگ استفاده شد و در هشتگ‌های مندرج در توییت‌ها فرایند جستجو انجام شد که با این روش نیز ۲۰۱۲۱ توییت استخراج گردید. در استخراج توییت‌های این دو مرحله، بازه زمانی یک ساله یعنی کل سال ۱۴۰۱ شمسی (از ۲۰۲۱/۳/۲۱ تا ۲۰۲۲/۳/۲۱) در نظر گرفته شد.

همچنین سعی شد از طریق جستجوی موضوعات داغ در فروردین ۱۴۰۲ نیز تعدادی توییت استخراج گردد که بدین ترتیب، ۹۶۸ توییت استخراج شد. همچنین برای تنوع بیشتر دادگان از جهت محتوا و بار احساسی، ۴۱۶ کلمه دیگر که اغلب آنها بار هیجانی مثبت داشتند نیز انتخاب گردید و از آنها برای جستجو به صورت هشتگ استفاده شد. به ازای جستجوی هر کلمه، حداکثر ۲۰ توییت استخراج شد و این بار محدوده زمانی کل سال ۱۴۰۰ (از ۲۰۲۰/۳/۲۱ تا ۲۰۲۱/۳/۲۱) در نظر گرفته شد. به این روش، حدود ۹۸۳۵ توییت استخراج گردید. در رویکرد دوم که همان رویکرد اکانت محور بود، ۴۰۰ اکانت پربازدید فارسی (در اینجا منظور اکانت‌هایی است که بیش از ۲۰۰۰۰ دنبال‌کننده دارند) به نحوی انتخاب شدند که با توجه به توضیحات صاحب اکانت در قسمت بیوگرافی، اکانت‌ها متنوع باشند؛ بدین ترتیب در بین اکانت‌های منتخب، اکانت‌های خبری و رسمی تا اکانت‌هایی که عموماً متن محاوره‌ای داشته و یا حتی سبک نوشتاری آنها ابداعی (یعنی حاوی کلمات خودساخته و یا املاهای تغییر یافته) است، وجود دارد. همچنین در انتخاب این اکانت‌ها تا حد امکان به جنسیت صاحبان آنها نیز توجه شد و سعی گردید که دادگان از این حیث نیز متعادل باشند. به این روش ۴۹۹۶ توییت، با محدودیت حداکثر ۲۰ توییت برای هر اکانت، استخراج شد.

لازم به ذکر است که در هنگام استخراج کلیه توییت‌ها، غیر از فیلتر بازه زمانی سعی شد موارد زیر نیز لحاظ شوند:

- حذف بازتوییت‌ها، توییت‌های غیرفارسی، توییت‌های غیرمتنی و توییت‌هایی که فقط یک آدرس اینترنتی یا یو.آر.ال بودند.
- حفظ نقل قول‌ها به دلیل این که حاوی اطلاعات متنی بودند.

خلاصه‌ای از فرآیند استخراج داده‌ها در جدول ۲ آورده شده است.

جدول ۲: خلاصه‌ای از فرایند استخراج داده‌ها

رویکرد	روش	تعداد توییت‌های استخراج شده
کلمه-محور	۱- مبتنی بر کلمه کلیدی: - استفاده از ۵۰۰ کلمه کلیدی برگرفته از Persian_NRC_EmoLex در بازه زمانی ۲۱ مارس ۲۰۲۲ لغایت ۲۱ مارس ۲۰۲۳. - استفاده از بیش از ۴۰۰ کلمه (که اغلب آنها دارای احساسی مثبت بودند) با هدف ایجاد تعادل در مجموعه داده‌ها، در کل بازه زمانی ۲۱ مارس ۲۰۲۱ تا ۲۱ مارس ۲۰۲۲. ۲- مبتنی بر هشتگ: - با استفاده از جستجوی همان ۵۰۰ کلمه، این بار به صورت هشتگ، در بازه زمانی ۲۱ مارس ۲۰۲۲ تا ۲۱ مارس ۲۰۲۳. - استفاده از موضوعات پرطرفدار	حدود ۴۹۰۰۰ توییت (حداکثر ۲۰ توییت برای هر کلمه کلیدی)
اکانت-محور	استفاده از ۴۰۰ اکانت پربازدید فارسی	حدود ۵۰۰۰ توییت (حداکثر ۲۰ توییت برای هر اکانت)

۳-۳ آماده‌سازی بستر جمع‌سپاری

یکی از نکات متمایز این پژوهش، استفاده از سامانه جمع‌سپاری بومی آزمایشگاه پارسی‌آزما است. در این بستر، کارهای مختلفی از جمله تعریف فعالیت برچسب‌زنی، تخصیص توییت‌ها به برچسب‌زن‌ها و درج دادگان کالیبره (که جلوتر توضیح داده خواهند شد)، به‌صورت کاملاً شخصی‌سازی شده انجام می‌شود. بومی بودن این بستر، انعطاف و امنیت بالایی را برای گردآوری داده‌های زبانی فارسی فراهم می‌کند و امکان نظارت دقیق بر فرایند برچسب‌زنی را میسر می‌سازد. بنابراین در این مرحله، اجزای مورد نیاز تعیین و تعریف و فرم‌های مربوط به فعالیت مزبور طراحی شدند.

۴-۳ پیش پردازش داده‌ها

این فرآیند از دو مرحله دستی و خودکار تشکیل شده است که در ادامه به شرح هریک پرداخته می‌شود:

الف) در مرحله پیش‌پردازش دستی برای افزایش کیفیت دادگان

کارهای زیر انجام شد:

- حذف توییت‌هایی که متن آنها آنقدر کوتاه یا ناقص بود که تشخیص احساس آنها ممکن نبود.
- حذف توییت‌های صرفاً تبلیغاتی که اطلاعات متنی ارزشمندی نداشتند.
- حذف توییت‌هایی که صرفاً شامل هشتگ بودند.
- حذف توییت‌هایی که به صورت بارز فاقد احساس بودند: این تصمیم به این دلیل اتخاذ شد که در مرحله اولیه جمع‌آوری داده‌ها، بخش قابل توجهی از توییت‌های استخراج‌شده، ماهیت خبری، اطلاع‌رسانی یا تبلیغاتی داشتند و در صورت باقی ماندن در دادگان، می‌توانستند باعث غلبه برچسب خنثی شوند. از آنجا که هدف اصلی این پژوهش، ایجاد مجموعه‌ای متوازن برای تحلیل احساسات در متون فارسی بود، سعی شد تا حد امکان از تسلط داده‌های خنثی بر مجموعه نهایی جلوگیری شود.
- حذف توییت‌هایی که به لحاظ محتوایی مشکل سیاسی، اخلاقی و یا مذهبی داشتند.
- حذف توییت‌هایی که درک احساس آنها منوط به بافت بود و یا مبهم بودند.

ب) در مرحله پیش پردازش خودکار، پیش از بارگذاری داده‌ها در بستر جمع‌سپاری، اقدامات زیر انجام شد:

- حذف شکلک‌ها یا ایموجی‌ها: از آنجا که هدف این پژوهش، تحلیل بر پایه متن خام است و ایموجی‌ها (شکلک‌ها) می‌توانند نقش عاطفی قدرتمندی داشته باشند، حضور آن‌ها ممکن است الگوی زبانی را تحت‌تأثیر قرار دهد و منجر به تفسیر احساسیِ مستقیم (بدون پردازش زبانی متن) و در نهایت سوگیری یا ساده‌سازی بیش از حد شود. افزون بر این، بسیاری از ایموجی‌ها در متون فارسی ممکن است معادل واژگانی روشنی نداشته باشند یا این که در فرهنگ‌های گوناگون احساسات متفاوتی را منتقل کنند. به نظر می‌رسد که با حذف ایموجی‌ها، مدل‌های زبانی صرفاً بر پایهٔ واژگان و عبارات نوشتاری آموزش خواهند دید و به ارزیابی دقیق‌تری خواهند رسید.
- جایگزینی urlها با <URL>
- جایگزینی منشن‌ها با <USERNAME@>
- حذف توییت‌هایی با طول کمتر از ۲۰ کاراکتر
- حذف توییت‌های تکراری
- حذف هشتگ‌های پایانی توییت‌ها برای پرهیز از سوگیری دادگان

۳-۵ تهیه دادگان کالیبره

برای اطمینان از کیفیت و دقت فرآیند برچسب‌گذاری، از دادگان کالیبره استفاده شد. دادگان کالیبره شامل نمونه‌هایی از توییت‌ها بودند که به‌صورت دستی و با دقت بالا توسط دو خبره انتخاب و برچسب‌گذاری شدند. این نمونه‌ها به گونه‌ای انتخاب شدند که احساسات مثبت، منفی یا خنثی را به وضوح و بدون ابهام بیان می‌کردند. هدف اصلی از استفاده از دادگان کالیبره، ارزیابی عملکرد برچسب‌زن‌ها و اطمینان از این بود که برچسب‌ها با دقت و یکدستی لازم اعمال شوند.

دادگان کالیبره از میان توییت‌های جمع‌آوری‌شده انتخاب شدند. این توییت‌ها به‌صورت دستی و با توجه به معیارهای زیر انتخاب شدند:

- وضوح احساس: توییت‌هایی که احساسات مثبت، منفی یا خنثی را به‌طور واضح و بدون ابهام بیان می‌کردند.
- تنوع موضوعی: توییت‌هایی که موضوعات مختلفی را پوشش می‌دادند تا اطمینان حاصل شود که برچسب‌زن‌ها با انواع مختلفی از متون مواجه می‌شوند.
- عدم ابهام: توییت‌هایی که حاوی طعنه، کنایه یا عبارات مبهم نبودند انتخب شدند تا از خطا در برچسب‌گذاری جلوگیری شود.

دادگان کالیبره که حدود ۱۰ درصد کل دادگان بودند، به‌صورت تصادفی در میان داده‌های اصلی قرار داده شدند. هر توییت توسط سه برچسب‌زن بررسی شد، و در طول فرآیند برچسب‌گذاری، عملکرد برچسب‌زن‌ها به‌صورت مداوم از طریق این نمونه‌های کالیبره ارزیابی می‌شد. اگر دقت برچسب‌زنی یک فرد در نمونه‌های کالیبره کمتر از ۷۰٪ بود، سیستم به‌صورت خودکار آن فرد را از فرآیند برچسب‌گذاری حذف می‌کرد و کارهای قبلی او به فرد دیگری واگذار می‌شد. این مکانیزم کنترل کیفیت، اطمینان حاصل کرد که برچسب‌زن‌ها با دقت و یکدستی لازم کار خود را انجام می‌دهند.

استفاده از دادگان کالیبره چندین مزیت مهم داشت:

- ارزیابی عملکرد برچسب‌زن‌ها: دادگان کالیبره به‌عنوان معیاری برای ارزیابی دقت و قابلیت اطمینان برچسب‌زن‌ها عمل کردند. این امر به شناسایی برچسب‌زن‌هایی که عملکرد ضعیفی داشتند کمک کرد و اطمینان داد که تنها برچسب‌زن‌های دقیق و قابل اعتماد در فرآیند برچسب‌گذاری مشارکت داشته‌اند.
- شناسایی و اصلاح خطاها: با استفاده از دادگان کالیبره، خطاهای احتمالی در مراحل اولیه فرآیند برچسب‌گذاری شناسایی و اصلاح شدند. این موضوع به کاهش خطاهای سیستماتیک و افزایش کیفیت مجموعه دادگان نهایی کمک کرد.
- افزایش اعتبار مجموعه دادگان: استفاده از دادگان کالیبره و مکانیزم کنترل کیفیت مبتنی بر آن، اعتبار مجموعه دادگان نهایی را افزایش داد. این مجموعه دادگان اکنون می‌تواند به‌عنوان منبعی معتبر برای توسعه و ارزیابی مدل‌های تحلیل احساسات در زبان

فارسی مورد استفاده قرار گیرد.

در نهایت، دادگان کالیبره نه تنها به بهبود کیفیت برچسب‌گذاری کمک کردند، بلکه اطمینان دادند که مجموعه دادگان نهایی از دقت و قابلیت اطمینان بالایی برخوردارند. این رویکرد باعث شد که مجموعه دادگان ایجادشده به‌عنوان یک منبع معتبر برای تحقیقات آینده در حوزه تحلیل احساسات در زبان فارسی مورد استفاده قرار گیرد.

۳-۶ آموزش برچسب‌زن‌ها

در فرایند برچسب‌زنی از ده نفر دعوت به همکاری شد که البته میزان مشارکت آنها متفاوت بود تعدادی از این افراد دانشجوی کارشناسی ارشد رشته زبان‌شناسی بودند و بقیه هم آشنا با حوزه زبان و ادبیات بودند. برای دستیابی به توافق بیشتر در نتایج برچسب‌زنی، ابتدا از این افراد خواسته شد که شیوه‌نامه تهیه شده را مطالعه کرده و سپس برای آشنایی با جوانب کار جلسات آموزشی متعددی به صورت حضوری یا مجازی برگزار شد. در راستای این جلسات، شیوه‌نامه مربوطه تکمیل و یا در مواردی اصلاح گردید.

۳-۷ برچسب‌گذاری دادگان

برای برچسب‌گذاری دادگان از بستر جمع‌سپاری توسعه داده شده در آزمایشگاه پارسی‌آزما استفاده گردید. سپس دادگان کالیبره به طور تصادفی در لابلای سایر داده‌ها قرار گرفت.

روند کار برچسب‌زنی در سامانه جمع‌سپاری پارسی‌آزما به این صورت بود که ابتدا متن توییت به همراه سه گزینه برای برچسب‌زن نمایش داده می‌شد و از وی خواسته می‌شد که از طریق تجزیه و تحلیل زبان مورد استفاده در متن، بهترین گزینه مربوط به احساس نویسنده توییت را انتخاب کند. به افراد تاکید شد که توییت‌ها را براساس پیشینه و علایق خود (مثلاً فرهنگی، مذهبی) برچسب‌گذاری نکنند و فقط احساس یا هیجان نویسنده را در نظر بگیرند. ضمناً گزینه مجزایی هم در صفحه قرار داده شد تا اگر موردی نیاز به حذف داشت و یا بصورت سهوی در مرحله پیش‌پردازش حذف نشده بود، برچسب‌زن آن گزینه را انتخاب کند.

۳-۸ تجمیع آرا

معیار نهایی برچسب احساس هر یک از توییت‌های موجود در مجموعه دادگان، رای اکثریت برچسب‌زن‌ها است. هر توییت توسط سه نفر برچسب‌گذاری گردید و اگر حداقل دو نفر از آنها به یک احساس خاص رای می‌دادند، آن احساس به عنوان برچسب توییت انتخاب می‌شد. برای افزایش سرعت برچسب‌گذاری و همچنین کاهش هزینه‌ها، در صورت کسب دو رای یکسان، توییت از دسترسی دیگران خارج شد و نیازی به رای سوم نبود.

۳-۹ آماده‌سازی مجموعه دادگان نهایی

پس از مرحله تجمیع و کنار گذاشتن داده‌هایی که حداقل یکی از برچسب‌زن‌ها به حذف آنها رای داده بودند و حذف دادگان تکراری، داده‌هایی که حائز رای اکثریت نشده بودند، به برچسب‌زن چهارمی داده شدند. برچسب بسیاری از این دادگان، پس از دریافت آرای فرد چهارم، تعیین شد و تعداد کمی از مواردی که توافقی در موردشان حاصل نشده بود، کنار گذاشته شدند. برای ارزیابی میزان توافق بین برچسب‌زن‌ها از معیار فلیس-کاپا استفاده شد و مقدار ۰,۷۲ برای آن به دست آمد. این معیار با در نظر گرفتن توافق شانسی، امتیاز توافق میان برچسب‌زن‌ها را محاسبه می‌کند.

به این ترتیب، مجموعه دادگان نهایی تحلیل احساس شامل ۵۲۵۳ توییت برچسب خورده است. این مجموعه در قالب CSV ذخیره شده و شامل ستون‌های شناسه توییت (id)، متن پردازش شده توییت (tweet)، نشانی وب (URL) هر توییت و سه ستون Negative، Positive و Neutral است که برچسب هر توییت، با مقدار ۱ در ستون مربوطه مشخص شده است.

توزیع برچسب‌ها در مجموعه داده نهایی به صورت جدول ۳ می‌باشد. همان‌طور که در جدول مشاهده می‌شود، تعداد دادگان خنثی در مقایسه با دو طبقه دیگر بسیار کمتر است و دلیل آن این است که بسیاری از دادگان فاقد احساس در مرحله پیش‌پردازش حذف شده بودند. شکل ۴ ابر کلمات مربوط به این مجموعه دادگان را نشان می‌دهد.

جدول ۳: توزیع دادگان در مجموعه دادگان ایجاد شده

برچسب	مثبت	منفی	خنثی
تعداد توییت‌ها در دسته	۱۹۴۸	۳۰۲۱	۲۸۴
درصد توییت‌ها در دسته	۳۶,۲	۵۵,۲	۸,۶
تعداد کل کلمات در دسته	۵۵۰۸۳	۹۶۵۱۲	۷۸۷۶

شکل ۴: ابر کلمات دادگان در مجموعه نهایی. تعداد کل کلمات منحصربه‌فرد: ۲۰۳۶۰

۴- ارزیابی مجموعه دادگان

تهیه مجموعه داده بدون ارزیابی صحت و کارایی آن فاقد اصول علمی لازم برای یک پژوهش است. مجموعه دادگان تهیه‌شده در این پژوهش در رویداد پارسی‌آزما[1] در قالب یک shared-task در اختیار تیم‌های مختلف قرار گرفت تا در یک فضای رقابتی مدل‌های بهینه ارائه کنند [37]. نتایج حاصل از مدل‌های سه تیم برتر روی دادگان تهیه شده در این مقاله در جدول ۴ آورده شده است.

جدول ۴: نتایج حاصل از اعمال مدل‌های مختلف بر روی دادگان

مدل	Word embeding	Precision	Recall	F-Score
Meta Classifier	BERT(PersPoliX[2] CrdiffNLP[3])	۰,۶۷	۰,۷۲	۰,۶۷
XLMRoberta +CNN	XLMRoberta	۰,۶	۰,۶۳	۰,۶
ALBERT	ALBERT	۰,۴۹	۰,۴۸	۰,۴۷

همانطور که در جدول ۴ هم مشخص است، بهترین نتیجه مربوط به مدل متاکلاسیفایر است که در آن از مدل BERT به گونه‌ای استفاده شده است که ابتدا با استفاده از مدل مبتنی بر BERT، text embeddings انجام گردیده و سپس پیش‌بینی‌های نهایی با استفاده از یک شبکه عصبی چندلایه یا شبکه عصبی کانولوشنال[4] صورت گرفته شده است. علاوه بر این، از چندین طبقه‌بند در قالب مدل‌های ترکیبی استفاده شده تا از نقاط قوت مدل‌های فردی بهره‌برداری شود [38]. همچنین این تیم از مدل زبانی PersPoliX که اولین مدل زبانی فارسی با تمرکز بر حوزه‌های اجتماعی-سیاسی است که برای داده‌های مکالمه‌ای و توییت‌ها طراحی شده استفاده نموده است. این مدل بر روی حدود نیم میلیون توییت فارسی آموزش داده شد که منجر به بهبود نتایج شده است.

۵- نتیجه‌گیری و کارهای آتی

عواطف انسانی پدیده‌های روانشناختی پیچیده‌ای هستند که تشخیص دقیق آنها از روی متن تقریباً غیرممکن است. آنچه در کارهای پژوهشی مربوط به تجزیه و تحلیل احساسات انجام می‌شود، در واقع تشخیص تجربه‌ای شکل گرفته از این عواطف است؛ تجربه‌ای که فرد آگاهانه آن را به زبان متن درآورده است.

هنگام شناسایی احساسات از طریق رسانه‌های اجتماعی، باید بر بسیاری از مشکلات غلبه کرد. سبک نگارش غیررسمی کاربران، اشتباهات گرامری و نگارشی و همچنین زبان عامیانه، طعنه و کنایه، استفاده از زبان غیررسمی و کوتاهی طول پیام‌ها از جمله مواردی هستند که در پژوهش‌ها به آنها اشاره شده است. همچنین درک عواطف انسانی به ویژه بر مبنای متن، امری ذهنی و مبهم است. در نتیجه استنباط و تفسیر صحیح حالات عاطفی نویسنده موضوعی چالش برانگیز است. همه این موارد باعث می‌شوند که تشخیص احساسات برای سامانه‌های خودکار امری دشوار باشد. موضوع مهم دیگر آن است که همه سامانه‌های خودکار نیازمند مجموعه داده مناسب هستند، اما به دلیل طیف گسترده موضوعات مورد بحث در رسانه‌های اجتماعی، ایجاد دستی مجموعه داده کاملی از داده‌های برچسب‌گذاری شده که شامل همه شرایط عاطفی قابل تصور باشد، دشوار است، مخصوصاً مجموعه داده‌ای که برای همه کاربردها بهینه باشد. با توجه به این چالش‌ها در پژوهش حاضر، شبکه اجتماعی توییتر که به نحوی همه پیچیدگی‌های ذکر شده را در بر می‌گیرد، برای تهیه دادگان انتخاب شد و با استفاده از بستر جمع‌سپاری که در آزمایشگاه پارسی‌آزما بومی‌سازی شده بود، مجموعه داده مناسبی (بدون محدودیت گستره موضوعی) حاوی بیش از ۵۰۰۰ توییت برچسب‌خورده، آماده شد.

اگرچه نتایج ارزیابی نشان‌دهنده کیفیت مناسب مجموعه دادگان است، اما محدودیت‌هایی نیز وجود دارد، مثلاً، حجم دادگان می‌تواند افزایش یابد تا مدل‌ها با داده‌های بیشتری آموزش ببینند و عملکرد بهتری داشته باشند. همچنین، در کارهای آینده می‌توان شدت احساسات (مثلاً بسیار مثبت، کمی مثبت، خنثی، کمی منفی، بسیار منفی) را نیز در نظر گرفت تا تحلیل احساسات دقیق‌تری انجام شود.

مراجع

[1] S. Kusal, S. Patil, J. Choudrie, K. Kotecha, D. Vora, and I. Pappas, "A Review on Text-Based Emotion Detection -- Techniques, Applications, Datasets, and Future Directions." arXiv, Apr. 26, 2022. Accessed: Oct. 15, 2023. [Online]. Available: http://arxiv.org/abs/2205.03235

[2] A. Kumar and A. Jaiswal, "Systematic literature review of sentiment analysis on Twitter using soft computing techniques," Concurrency and Computation, vol. 32, no. 1, p. e5107, Jan. 2020, doi: 10.1002/cpe.5107.

[3] N. Sabri, A. Edalat, and B. Bahrak, "Sentiment analysis of persian-english code-mixed texts," in 2021 26th International Computer Conference, Computer Society of Iran (CSICC), IEEE, 2021, pp. 1–4. Accessed: Oct. 15, 2023. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9420605/

[4] K. Sailunaz and R. Alhajj, "Emotion and sentiment analysis from Twitter text," Journal of Computational Science, vol. 36, p. 101003, 2019.

[5] P. Mehta and S. Pandya, "A review on sentiment analysis methodologies, practices and applications," International Journal of Scientific and Technology Research, vol. 9, no. 2, pp. 601–609, 2020.

[6] B. Pang and L. Lee, "Sentiment Polarity Dataset Version 2.0," Part of the Natural Language Tool Kit, for the Python computer language, 2002.

[1] Parsiazma.ir

[2] https://huggingface.co/StateOfTheArtAUT/perspolix-persian-political-tweet-xlm-roberta-large

[3] https://huggingface.co/cardiffnlp/twitter-xlm-roberta-base-sentiment

[4] CNN

arXiv, Jun. 02, 2020. Accessed: Oct. 15, 2023. [Online]. Available: http://arxiv.org/abs/2005.00547

[23] F. Rustam, I. Ashraf, A. Mehmood, S. Ullah, and G. S. Choi, "Tweets classification on the base of sentiments for US airline companies," Entropy, vol. 21, no. 11, p. 1078, 2019.

[24] R. Asgarnezhad and S. A. Monadjemi, "Persian sentiment analysis: feature engineering, datasets, and challenges," Journal of applied intelligent systems & information sciences, vol. 2, no. 2, pp. 1–21, 2021.

[25] S. Alimardani and A. Aghaie, "Opinion mining in Persian language using supervised algorithms," 2015, Accessed: Apr. 22, 2024. [Online]. Available: https://www.sid.ir/paper/332700/en

[26] S. A. A. Asli, B. Sabeti, Z. Majdabadi, P. Golazizian, R. Fahmi, and O. Momenzadeh, "Optimizing annotation effort using active learning strategies: A sentiment analysis case study in persian," in Proceedings of the Twelfth Language Resources and Evaluation Conference, 2020, pp. 2855–2861.

[27] A. Hatefi Ghahfarrokhi and M. Shamsfard, "Tehran stock exchange prediction using sentiment analysis of online textual opinions," Intell Sys Acc Fin Mgmt, vol. 27, no. 1, pp. 22–37, Jan. 2020, doi: 10.1002/isaf.1465.

[28] T. S. Ataei, K. Darvishi, S. Javdan, B. Minaei-Bidgoli, and S. Eetemadi, "Pars-absa: an aspect-based sentiment analysis dataset for Persian," arXiv preprint arXiv:1908.01815, 2019.

[29] K. Darvishi, S. Javdan, B. Minaei-Bidgoli, and S. Eetemadi, "Pars-ABSA: a Manually Annotated Aspect-based Sentiment Analysis Benchmark on Farsi Product Reviews," in Proceedings of the Thirteenth Language Resources and Evaluation Conference, 2022, pp. 7056–7060.

[30] M. E. Basiri and A. Kabiri, "Words Are Important: Improving Sentiment Analysis in the Persian Language by Lexicon Refining," ACM Trans. Asian Low-Resour. Lang. Inf. Process., vol. 17, no. 4, pp. 1–18, Dec. 2018, doi: 10.1145/3195633.

[31] A. Khodaei, A. Bastanfard, H. Saboohi, and H. Aligholizadeh, "Deep Emotion Detection Sentiment Analysis of Persian Literary Text," 2022, Accessed: Oct. 15, 2023. [Online]. Available: https://www.researchsquare.com/article/rs-1796157/latest

[32] M. Shirghasemi, M. H. Bokaei, and M. Bijankhan, "The impact of active learning algorithm on a cross-lingual model in a Persian sentiment task," in 2021 7th International Conference on Web Research (ICWR), IEEE, 2021, pp. 292–295. Accessed: Apr. 22, 2024.[Online]. https://ieeexplore.ieee.org/abstract/document/9443156

[33] P. Hosseini, A. A. Ramaki, H. Maleki, M. Anvari, and S. A. Mirroshandel, "SentiPers: A Sentiment Analysis Corpus for Persian." arXiv, Jan. 01, 2021. Accessed: Oct. 15, 2023. [Online]. Available: http://arxiv.org/abs/1801.07737

[34] J. P. R. Sharami, P. A. Sarabestani, and S. A. Mirroshandel, "Deepsentipers: Novel deep learning models trained over proposed augmented persian sentiment corpus," arXiv preprint arXiv:2004.05328, 2020.

[35] S. M. Mohammad and F. Bravo-Marquez, "WASSA-2017 Shared Task on Emotion Intensity." arXiv, Aug. 11, 2017. Accessed: Oct. 15, 2023. [Online]. Available: http://arxiv.org/abs/1708.03700

[36] A. Nazarizadeh, T. Banirostam, and M. Sayyadpour, "Sentiment Analysis of Persian Language: Review of Algorithms, Approaches and Datasets," arXiv preprint arXiv:2212.06041, 2022

[37] Farhoodi, M., Mahmoudi, M., & Bokaei, M. H. (2024). ParsiAzma Challenges on Persian Text Analysis in Social Media. International Journal of Information & Communication Technology Research (2251-6107), 16.(۳)

[38] Sobhi, M., & Mazochi, A. (2024). A Comparative Study of BERT-X for Sentiment Analysis and Stance Detection in Persian Social Media. International Journal of Information & Communication Technology Research (2251-6107), 16.(۳)

[7] K. Topal and G. Ozsoyoglu, "Movie review analysis: Emotion analysis of IMDb movie reviews," in 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), IEEE, 2016, pp. 1170–1176. Accessed: Oct. 15, 2023. [Online].: https://ieeexplore.ieee.org/abstract/document/7752387/

[8] "Yelp." Accessed: Oct. 30, 2023. [Online]. Available: https://www.yelp.com/dataset/challenge

[9] E. Vaziripour, C. Giraud-Carrier, and D. Zappala, "Analyzing the political sentiment of tweets in Farsi," in Proceedings of the International AAAI Conference on Web and Social Media, 2016, pp. 699–702. Accessed: Oct. 25, 2023. [Online]. Available: https://ojs.aaai.org/index.php/ICWSM/article/view/14791

[10] Z. B. Nezhad and M. A. Deihimi, "Twitter sentiment analysis from Iran about COVID 19 vaccine," Diabetes & Metabolic Syndrome: Clinical Research & Reviews, vol. 16, no. 1, p. 102367, 2022.

[11] H. Abdi Khojasteh, E. Ansari, and M. Bohlouli, "Large-Scale Colloquial Persian 0.5," https://iasbs.ac.ir/~ansari/lscp/, Feb. 2020, Accessed: Oct. 25, 2023. [Online]. Available: https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-3195

[12] M. Heidari and P. Shamsinejad, "Producing an instagram dataset for persian language sentiment analysis using crowdsourcing method," in 2020 6th International Conference on Web Research (ICWR), IEEE, 2020, pp. 284–287.

[13] M. Panahandeh Nigjeh and S. Ghanbari, "Leveraging ParsBERT for cross-domain polarity sentiment classification of Persian social media comments," Multimedia Tools and Applications, pp. 1–18, 2023.

[14] A. Go, R. Bhayani, and L. Huang, "Twitter sentiment classification using distant supervision," CS224N project report, Stanford, vol. 1, no. 12, p. 2009, 2009.

[15] H. Poursepanj, J. Weissbock, and D. Inkpen, "uOttawa: system description for SemEval 2013 task 2 sentiment analysis in twitter," in Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013), 2013, pp. 380–383. Accessed: Oct. 15, 2023. [Online]. Available: https://aclanthology.org/S13-2062.pdf

[16] B. Velichkov et al., "SU-FMI: System description for SemEval-2014 task 9 on sentiment analysis in Twitter," in Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), 2014, pp. 590–595. Accessed: Oct. 15, 2023. [Online]. Available: https://aclanthology.org/S14-2103.pdf

[17] P. Nakov, A. Ritter, S. Rosenthal, F. Sebastiani, and V. Stoyanov, "SemEval-2016 Task 4: Sentiment Analysis in Twitter." arXiv, Dec. 03, 2019. doi: 10.48550/arXiv.1912.01973.

[18] S. Mohammad, F. Bravo-Marquez, M. Salameh, and S. Kiritchenko, "Semeval-2018 task 1: Affect in tweets," in Proceedings of the 12th international workshop on semantic evaluation, 2018, pp. 1–17. Accessed: Oct. 15, 2023. [Online]. Available: https://aclanthology.org/S18-1001/

[19] A. Rogers, A. Romanov, A. Rumshisky, S. Volkova, M. Gronas, and A. Gribov, "RuSentiment: An enriched sentiment analysis dataset for social media in Russian," in Proceedings of the 27th international conference on computational linguistics, 2018, pp. 755–763. Accessed: Oct. 15, 2023. [Online]. Available: https://aclanthology.org/C18-1064/

[20] S.-Y. Chen, C.-C. Hsu, C.-C. Kuo, Ting-Hao, Huang, and L.-W. Ku, "EmotionLines: An Emotion Corpus of Multi-Party Conversations." arXiv, May 30, 2018. Accessed: Oct. 15, 2023. [Online]. Available: http://arxiv.org/abs/1802.08379

[21] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, "MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations." arXiv, Jun. 04, 2019. Accessed: Oct. 15, 2023. [Online]. Available: http://arxiv.org/abs/1810.02508

[22] D. Demszky, D. Movshovitz-Attias, J. Ko, A. Cowen, G. Nemade, and S. Ravi, "GoEmotions: A Dataset of Fine-Grained Emotions."