Biannual scientific-research journal of language and linguistics
Fall & Winter 2024, 19 (37), 173-192
https://lsi-linguistics.ihcs.ac.ir/
DOI: 10.30465/LSI.2024.47711.1729

Original Research
Received: 27/Mar/2023 *** Accepted: 06/Mar/2024
ISSN: 2423-3847

Design and Preparation of Persian Labeled Dataset from COVID-19 News for Fake News Detection

Zahed, Forough [1]
Department of Computer Science, Faculty of Statistics, Mathematics and Computer, Allameh Tabataba'i University, Tehran, Iran

Bahrani, Mohammad[2]
Department of Computer Science, Faculty of Statistics, Mathematics and Computer, Allameh Tabataba'i University, Tehran, Iran

Mansouri, Alireza[3]
ICT Research Institute (ITRC), Tehran, Iran

Abstract

Fake news detection using content features have attracted many researchers in the last few years. These approaches rely mainly on news datasets and analyzing their style and content. Although there are some fake news datasets in English, fake news detection in the Persian language suffers from the lack of suitable datasets. This article introduces a manually labeled Persian fake news dataset, containing about 5000 posts related to COVID-19 and extracted from Telegram messenger. The process of building the dataset is done in two stages: 1) data collection and pre-processing; and 2) labeling manually using a settled rule set and an established framework. In the labeling stage, seven tasks have been used for labeling, including: 1) Factual; 2) Hate, blame, and negative speech; 3) Rising moral, encouragement, and advise; 4) Political news; 5) Death statistics; 6) Cure, medicine, and health care; and 7) Worth to be considered for fact checking. For each labeling task, 3 labels including "Yes", "No", and "Can't decide" are used. The main labeling task, i.e. "Factual" task is assigned to two annotators and in case of disagreement between annotators, the label assigned by third annotator is accepted. The kappa measure for inter-annotators agreement obtained equal to 0.706 that is in substantial range. This dataset is about 10 times larger in comparison to similar Persian datasets and can be used for not only fake news studies but also some other Persian Natural Language Processing (NLP) studies.

Keywords: fake news, COVID-19 pandemic, labeled dataset, social networks.

1. forooghzhd@gmail.com
2. bahrani@atu.ac.ir (Corresponding Author)
3. amansuri@itrc.ac.ir
How to cite: Zahed, F., Bahrani, M., & Mansouri, A. (2024). Design and Preparation of Persian Labeled Dataset from COVID-19 News for Fake News Detection. *Language and Linguistics*, *19*(37), 173 - 192. doi: 10.30465/LSI.2024.47711.1729

مقاله پژوهشی

تاریخ دریافت: ۱۴۰۲/۰۱/۰۷ *** تاریخ پذیرش: ۱۴۰۲/۱۲/۱۶

ISSN: ۲۳۲۲-۳۸۴۷

دوفصلنامه علمی- پژوهشی زبان و زبان‌شناسی
دوره ۱۹، شماره ۳۷، پاییز و زمستان ۱۴۰۲، ۱۷۳- ۱۹۲
lsi-linguistics.ihcs.ac.ir
DOI: 10.30465/LSI.2024.47711.1729

طراحی و تهیهٔ دادگان برچسب‌خورده فارسی از اخبار مرتبط با همه‌گیری کووید-۱۹ به منظور تشخیص اخبار جعلی

زاهد، فروغ گروه رایانه، دانشکده آمار، ریاضی و رایانه، دانشگاه علامه طباطبائی، تهران، ایران
بحرانی، محمد گروه رایانه، دانشکده آمار، ریاضی و رایانه، دانشگاه علامه طباطبائی، تهران، ایران
منصوری، علیرضا پژوهشگاه ارتباطات و فناوری اطلاعات، تهران، ایران

چکیده

در این مقاله یک پیکرهٔ برچسب‌خورده، به منظور به‌کارگیری در تشخیص اخبار جعلی با حدود ۵۰۰۰ پست مربوط به اخبار همه‌گیری کووید-۱۹ از پیام‌رسان تلگرام استخراج شده و برچسب‌زنی می‌شود. فرایند ساخت پیکره در دو مرحله انجام می‌پذیرد. مرحلهٔ اول شامل جمع‌آوری و پیش‌پردازش داده‌ها و مرحلهٔ دوم شامل برچسب‌گذاری آنها می‌باشد. در مرحلهٔ اول، داده‌ها فیلتر می‌شوند و بعد از انجام پردازش‌های لازم برروی آنها، در مرحلهٔ دوم، بر اساس یک شیوه‌نامه، اقدام به برچسب‌گذاری می‌شود. در مرحلهٔ برچسب‌گذاری، از هفت عنوان موردنظر برای وظایف، استفاده می‌گردد و هر پست خبری با توجه به این هفت وظیفه برچسب‌گذاری می‌شود. ایجاد یک چهارچوب مناسب (شیوه‌نامه) برای برچسب‌زنی یکی از اقدامات مهم در این مرحله می‌باشد. شیوه‌نامه در اختیار دو برچسب‌زن خبره که بدین‌منظور آموزش دیده‌اند قرار می‌گیرد و اخبار از لحاظ هفت وظیفهٔ ۱) صحیح یا جعلی بودن ۲) سیاسی بودن ۳) بالا بردن سطح آگاهی عمومی، دادن روحیه یا دادن یک توصیه به خواننده ٤) مطالب مربوط به دارو و درمان یا مراقبت‌های بهداشتی ٥) آمار مرگ و میر ٦) داشتن محتوای حاوی مطالب تنفرآمیز، سرزنش، عیب‌جویی، منفی بافی و ۷) ارزش داشتن برای بررسی واقعیت، مورد بررسی قرار گرفته و بر این اساس، برچسب درست، نادرست یا خنثی می‌گیرند. در صورت عدم توافق بین دو برچسب‌زن، از برچسب‌زن سوم نظرخواهی می‌شود. برچسب‌دهی اخبار طوری انجام می‌گیرد که در نهایت، دسته‌های متوازنی در وظیفهٔ صحیح یا جعلی بودن اخبار به دست آید.

کلیدواژه: اخبار جعلی، همه‌گیری کووید-۱۹، دادگان برچسب‌خورده، شبکه‌های اجتماعی

۱ مقدمه

در سال‌های اخیر سکوهای[1] رسانه‌های اجتماعی مختلف مانند تلگرام بسیار پرطرفدار و محبوب شده‌اند. زیرا باعث تسهیل در کسب اطلاعات می‌شوند و بستری را برای به اشتراک‌گذاری اخبار و موضوعات مختلف فراهم می‌کنند. در دسترس بودن داده‌های غیرمعتبر در سکوهای رسانه‌های اجتماعی توجه گسترده‌ای را در میان محققان به خود جلب کرده است. اخبار جعلی[2] به دلیل تاثیر منفی که دارد، مورد توجه محققان، روزنامه‌نگاران، سیاست‌مداران و عموم مردم قرار گرفته است (کالی‌یار[3] و همکاران، ۲۰۲۱). بعد از آغاز شیوع بیماری کووید-۱۹[4]، اخبار مربوط به این بیماری به طور فزاینده‌ای در کانال‌های خبری مختلف درحال گسترش است که بسیاری از آنها نادرست و گمراه‌کننده هستند. گسترش این اطلاعات غلط منجر شده که مردم به برخی از درمان‌های پزشکی و واکسن‌ها بی‌اعتماد شوند و حتی از کارهای ضروری مانند فاصله‌گذاری اجتماعی و زدن ماسک خودداری کنند. برای مقابله با گسترش اطلاعات غلط، بسیاری از سازمان‌ها در سراسر جهان مانند PolitiFact، FactCheck و Snopes تلاش‌های قابل‌وجهی برای تایید مقالات خبری و انتشار خبر در رسانه‌های اجتماعی انجام داده‌اند. با این حال تلاش‌های این سازمان‌ها برای پوشش و بررسی همهٔ اخبار و پست‌هایی که روزانه به اشتراک گذاشته می‌شود جوابگو نیست. پس نیاز به تشخیص خودکار اخبار جعلی برای کاهش بار بر روی مفسران اخبار جعلی وجود دارد (عامر[5] و علیان[6]، ۲۰۲۱).

بنابراین مسئلهٔ تشخیص اخبار درست و نادرست مربوط به بیماری کووید-۱۹ با استفاده از رویکردهای یادگیری ماشین[7] با کمک پردازش زبان طبیعی[8] و پیاده‌سازی روشی برای حل مسئلهٔ تشخیص اخبار جعلی در کانال‌های خبری، به خصوص در پیام‌رسان تلگرام[9] می‌تواند به آگاهی افراد جامعه در ابعاد مختلف سیاسی، اجتماعی، اقتصادی و پزشکی و غیره کمک شایانی بکند (ونگ[10]، ۲۰۱۷).

در این پژوهش، ایجاد یک دادگان برچسب‌خورده به زبان فارسی، برای به‌کارگیری در فرایند آموزش و ارزیابی در روش‌های یادگیری ماشین برای تشخیص خودکار اخبار جعلی مدنظر

---

1. Platforms
2. Fake News
3. Kaliyar
4. COVID-19
5. Ameure
6. Aliane
7. Machine Learning (ML)
8. Natural Language Processing (NLP)
9. Telegram
10. Wang

است. بررسی اخبار، فقط به درست و غلط بودن آن ختم نمی‌شود. اینکه یک خبر از لحاظ سیاسی و یا از لحاظ روانی و چه تاثیراتی روی خوانندهٔ خبر داشته است، می‌تواند به علل و اهداف ایجاد اخبار نادرست کمک شایانی بکند. علاوه بر این بسیاری از ابعاد دیگر را نیز می‌توان در بررسی مسئلهٔ پیش‌رو دخیل دانست، پس حدود مسئله فراتر از تشخیص صرفاً درست یا غلط بودن خبر خواهد بود و یک خبر را در ابعاد مختلف تجزیه و تحلیل کرده و با بررسی آنها در کنار بررسی رویکردهای مختلف یادگیری ماشین و انتخاب بهترین رویکرد، اخبار مورد ارزیابی قرار خواهد گرفت.

۲- پیشینهٔ پژوهش

بر اساس تحقیقات اخیر، مجموعه‌ای از روش‌های یادگیری ماشینی برای تشخیص انواع اخبار جعلی در رسانه‌های اجتماعی وجود دارد. تاکنون تشخیص اخبار جعلی بیشتر برای زبان انگلیسی انجام شده است. یکی از این تحقیقات، تشخیص اخبار جعلی را با به‌کارگیری روش‌های بیز ساده و شبکه‌های عصبی انجام می‌دهد (آفی‌ونگسوفون[۱] و چانگزتیتواتانا[۲]، ۲۰۱۸). همچنین در پژوهشی دیگر، محققان تشخیص خودکار اخبار جعلی را با استفاده از مدل‌های TF-IDF و Word2Vec (برای تبدیل متن به نمایش عددی) و الگوریتم‌های مختلف مانند بیز ساده و LSTM[۳] انجام دادند (ویجایاراغاوان[٤] و همکاران، ۲۰۲۰).

کومار[٥] و همکاران (۲۰۱۸) یک بررسی کلی از جنبه‌های مختلف اخبار جعلی انجام دادند. در پژوهش آنها شاخه‌های مختلفِ اخبار جعلی و الگوریتم‌های موجود برای شناسایی اخبار جعلی مورد بررسی قرار گرفته است. در یکی دیگر از تحقیقات، شین[٦] و همکاران (۲۰۱۸) دربارهٔ نظریه‌های بنیادی در رشته‌های مختلف برای تقویت مطالعهٔ میان‌شته‌ای اخبار جعلی تحقیق کرده‌اند که در این مطالعه اخبار جعلی را از چهار دیدگاه مورد بررسی قرار داده‌اند.

پوساداس–دوران[۷] و همکاران (۲۰۱۹) با استفاده از وب‌گاه‌های خبری، یک پیکره از اخبار جعلی را برای زبان اسپانیایی توسعه دادند. پیکرهٔ آنها شامل دو برچسب اخبار صحیح و اخبار جعلی می‌باشد. با استفاده از این پیکره، آنها یک سامانه تشخیص اخبار جعلی مبتنی بر سبک را ارائه دادند.

---

1. Aphiwongsophon
2. Chongstitvatana
3. Long-short term memory
4. Vijayaraghavan
5. Kumar
6. Shin
7. Posadas-Durán

شو[1] و همکاران (۲۰۲۰) رابطه‌ای را بین اخبار جعلی و واقعی در دسترس در سکوهای برخط از توییتر به‌دست آورده که شامل یو.آر.ال[2]های حاصل از بررسی واقعیت‌ها است. در بررسی‌ها دریافتند که یو.آر.ال‌ها شناخته‌شده‌ترین راهبرد برای به اشتراک‌گذاری مقالات خبری در مراحل مختلف هستند.

سینگ[3] و همکاران (۲۰۱۷) ویژگی‌هایی را با استفاده از روش‌های یادگیری ماشین سنتی برای طبقه‌بندی اخبار جعلی به دست آوردند. آنها مشکل اخبار جعلی را با ماشین بردار پشتیبان[4] به عنوان یک طبقه‌بندی‌کننده که دقت ۸۷ درصدی دارد، برطرف کردند.

کرستانی[5] و روسو[6] (۲۰۲۰) مدل جدیدی را ارائه کردند که بر اساس شبکهٔ عصبی پیچشی[7] و ترکیبی از کلمات با ویژگی‌ها است که الگوهای زبانی کاربر را نشان می‌دهد.

الحداد[8] و همکاران (۲۰۲۱) یک مجموعه دادهٔ توییت‌های عربی و انگلیسی تفسیر شده‌ای را به نام "COVID-19" منتشر کردند که از ٤ فوریه تا ۱۰ مارس سال ۲۰۲۰ جمع‌آوری شده و با استفاده از ۱۳ الگوریتم یادگیری ماشینی و هفت روش استخراج ویژگی تفسیر شده‌اند.

شاهی[9] و ناندینی[10] (۲۰۲۰) یک مجموعه دادهٔ چند زبانهٔ "Fake Covid" را که از ۱۰۵ کشور جمع‌آوری شده، منتشر کردند که شامل ٤۰ زبان است. این مجموعه داده شامل ۵۱۸۲ مقالهٔ خبری کووید-۱۹ است. مقالات از ۲۹ وبگاه مختلف بین آوریل تا سپتامبر ۲۰۲۰ جمع‌آوری شده و به دو دستهٔ "غلط" و "دیگر" برچسب‌دهی شده‌اند. از جمله پژوهش‌های انجام‌گرفته برروی زبان فارسی نیز پژوهش سقایان و همکاران (۲۰۲۰) می‌باشد. آنها یک پیکره از اخبار فارسی مرتبط با همه‌گیری کووید-۱۹ موجود در توییتر شامل ۵۰۰ سند و ٤ برچسب (طنز، اخبار غلط، اخبار درست و اخبار خنثی) را تهیه کردند و با استفاده از روش‌های یادگیری ماشین به دسته‌بندی آنها پرداختند. آنها همچنین تأثیر ترجمه اخبار را بر دقت تشخیص اخبار جعلی مورد بررسی قرار دادند.

از دیگر پژوهش‌های انجام‌گرفته در زبان فارسی، پژوهش قیومی (۱٤۰۱) می‌باشد. در پژوهش مذکور یک پیکره شامل ٥٦٤ سند و ۲ برچسب (دادهٔ جعلی، دادهٔ مؤثق) از اخبار مرتبط

1. Shu
2. Uniform Resource Locator
3. Singh
4. Support Vector Machines
5. Crestani
6. Rosso
7. Convolutional Neural Network (CNN)
8. Elhadad
9. Shahi
10. Nandini

با همه‌گیری کووید-۱۹ با استفاده از خزش وب جمع‌آوری شده و سپس با استفاده از این پیکره به تحلیل آماری متون اخبار جعلی مرتبط با کووید-۱۹ پرداخته شده است.

۳- چارچوب نظری

رویکردهای مختلفی برای تشخیص اخبار جعلی وجود دارد که با توجه به آن رویکردها، روش تهیهٔ پیکره‌های اخبار جعلی متفاوت است. ژو[۱] و زعفرانی[۲] (۲۰۲۱) در مقالهٔ مروری خود در مورد روش‌های مختلف تشخیص اخبار جعلی، سه رویکرد مختلف را برای تشخیص خودکار اخبار جعلی برمی‌شمارند:

- رویکرد مبتنی بر سبک[۳]: این رویکرد بر تحلیل سبک و محتوای[۴] اخبار تمرکز دارد.
- رویکرد مبتنی بر انتشار[۵]: این رویکرد نحوهٔ انتشار اخبار را در شبکه‌های اجتماعی و کانال‌های خبری مورد مطالعه قرار می‌دهد و بر اساس نحوهٔ انتشار به تشخیص صحیح یا جعلی بودن اخبار می‌پردازد.
- رویکرد مبتنی بر منبع[۶]: در این رویکرد بر اساس منبع یا منابع انتشار خبر، صحیح یا جعلی بودن خبر حدس زده می‌شود.

در پژوهش حاضر، طراحی و تهیهٔ پیکره به‌گونه‌ای است که برای رویکردهای مبتنی بر سبک مناسب باشد. بنابر تعریف ارائه‌شده از «سبک» در پژوهش ژو و زعفرانی (۲۰۲۱)، سبک اخبار جعلی، مجموعه‌ای از خصیصه‌های[۷] قابل اندازه‌گیری است که به خوبی محتوای اخبار جعلی و تفاوت آن با محتوای اخبار صحیح را نمایش می‌دهند. این خصیصه‌های قابل اندازه‌گیری به صورت مجموعه‌ای از ویژگی‌ها[۸] به الگوریتم‌های یادگیری ماشین داده می‌شوند. بنابراین مسئلهٔ تشخیص خودکار اخبار جعلی با استفاده از رویکردهای مبتنی بر سبک، به یک مسئلهٔ طبقه‌بندی متن (با دو یا چند برچسب) تبدیل می‌شود. برای حل این مسئلهٔ طبقه‌بندی لازم است یک مجموعه دادهٔ آموزشی برچسب‌خورده از اخبار جعلی و صحیح در حوزهٔ موردنظر تهیه شود. در این پژوهش حوزهٔ موردنظر، پست‌های فارسی منتشرشده در کانال‌های تلگرامی مربوط به همه‌گیری کووید-۱۹ می‌باشد.

---

1. Zhou
2. Zafarani
3. Style-based
4. Content
5. Propagation-based
6. Source-based
7. Characteristics
8. Features

٤- روش کار

١-٤- مرحلهٔ آماده سازی داده‌ها

موضوعی که قرار است برای آن برچسب‌گذاری انجام شود، خبرهای مربوط به بیماری کووید-۱۹ می‌باشد. مهم‌ترین مسئله در مورد این بیماری در جامعهٔ حال حاضر، موثق بودن یا نبودن اخبار مربوط به این بیماری است چراکه یکی از چالش‌برانگیزترین موضوعات در حال حاضر بوده و موارد سیاسی و اقتصادی و یا حتی دینی و مذهبی را درخور این مسئله کرده و باعث ایجاد فضای پر از اخبار ضدونقیض در کشور شده است. بنابراین باید برچسب‌هایی را انتخاب کنیم که بتواند شفافیت لازم را ایجاد کند. به عنوان مثال:

۱. اگر فرد متخصصی به جهت کسب منافع شخصی و جهت‌گیری‌های سیاسی در مورد واکسیناسیون نظری می‌دهد، باید موثق بودن یا نبودن سخنان این فرد را بررسی کرد. پس شایعه بودن می‌تواند برچسب انتخابی باشد.

۲. اخباری که احساسات و افکار مردم را تحریک می‌کند و باعث ترس و تهدید افراد می‌شوند.

۳. مواردی که برای سرگرم کردن و ایجاد فضای آرام‌تر جامعه، با محتوای طنز منتشر می‌شوند.

٤. اخباری که با سوءاستفاده از اخبار کرونا باعث گمراهی افراد می‌شوند و حس اعتماد افراد را متزلزل می‌کنند.

۵. خبرهایی که نتیجهٔ مطالعات اولیه درمورد موضوع خاصی (به عنوان مثال داروی x در درمان کرونا) است که هنوز به طور کامل مورد تایید سازمان‌های بهداشتی قرار نگرفته‌اند.

و بسیاری از مسائلی که با توجه به موضوع مهم بیماری کووید-۱۹ ممکن است در انتشار خبرها رخ دهد.

بنابراین با توجه به موارد گفته شده می‌توانیم برچسب‌های زیادی حتی بیشتر از موضوعات عادی جامعه در نظر بگیریم اما به دلیل زیاد شدن تعداد برچسب‌ها ناگزیر هستیم که برچسب‌های پیشنهادی اولیه را از فیلترهای مختلف عبور داده و به مهم‌ترین آنها برسیم.

نوع داده‌های موردنظر، پست‌های منتشرشده در کانال‌های تلگرامی می‌باشد. بنابراین ابتدا تعداد زیادی پست تلگرامی مربوط به اخبار کووید-۱۹ جمع‌آوری می‌شود. در این مرحله با انجام پیش‌پردازش‌هایی همچون فیلتر کردن پست‌های کانال‌های تلگرامی و استخراج اخبار مربوط به کووید-۱۹ و حذف اخبار تکراری و ناقص (به صورت دستی)، روند کار پیش گرفته می‌شود، طوری که در نهایت به صورت خالص بتوان ۵۰۰۰ پست تلگرامی مربوط به کووید-۱۹ داشت.

لیست کانال‌های شبکه اجتماعی تلگرام که در این پروژه مورد خزش قرار گرفته‌اند، شامل ۹۹ کانال است که مرتبط با بیماری کووید-۱۹ بوده که در اینجا نشانی ۵ کانال به عنوان مثال ذکر می‌گردد:

@OfficialPersianTwitter
@Razcom
@akafiha
@etelaate_omoomi
@SerumPlus

۲-٤- دورهٔ زمانی جمع‌آوری داده

سعی بر این بوده است که دورهٔ زمانی انتخاب بشود که جامعه نیاز مبرمی به کسب اطلاعات درمورد موضوعات خاص این بیماری دارد. در دوران انجام این پژوهش با توجه به تمام اتفاقات مربوط به ویروس کووید-۱۹، مسئله واکسن و واکسیناسیون افراد و جهش‌های ویروس مورد توجه بوده است. اخبار این دوره حاکی از شک و تردید مردم سراسر دنیا در مورد مسئلهٔ واکسن و جهش‌های این ویروس است. پس بهتر است دورهٔ زمانی جمع‌آوری داده با شروع مسئلهٔ واکسیناسیون در کشور باشد. علاوه بر این می‌توان گفت بیش از ۷۰ درصد اخبار شبکه‌های اجتماعی در هنگام شروع واکسیناسیون، در مورد موضوع واکسن و جهش‌ها است. موضوع دیگر در تعیین بازهٔ جمع آوری داده‌ها، داشتن اخبار جدید و با حداقل میزان تکراری بودن است. چراکه در یک بازهٔ زمانی کوتاه‌مدت، میزان تکراری بودن خبرها به علت شیوع آن خبر در پایگاه‌های خبری مختلف، امری طبیعی است. بنابراین با توجه به موارد ذکر شده، بازهٔ زمانی جمع‌آوری داده‌ها از ۲۰۲۱/٦/۲۲ تا ۲۰۲۲/٤/۲۰ انتخاب شد.

۳-٤-پیش‌پردازش داده‌ها

در هنگام خزش و جمع‌آوری پست‌ها از کانال‌های مختلف، برای جلوگیری از وارد شدن پست‌های غیرمرتبط با شیوع بیماری کووید-۱۹ در داده، لازم است فقط پست‌هایی جمع‌آوری شوند که دارای کلمات کلیدی یا هشتگ‌های مرتبط باشند. متأسفانه کلمات کلیدی مانند کووید-۱۹ و Covid-19 به صورت‌های متنوعی نوشته می‌شوند. در زیر کلمات کلیدی که در هنگام جمع‌آوری پست‌ها باید در نظر گرفته شوند، آمده است و سعی شده تا حد امکان صورت‌های مختلف نوشتاری آنها پوشش داده شود:

کلمات کلیدی فارسی (با در نظر گرفتن انواع نویسه‌های "ی" و "ک" فارسی و عربی):

- کرونا

- کوید
- کووید
- کوئید
- کوییید

لازم به ذکر است که نیازی به در نظر گرفتن پسوند ۱۹ یا -۱۹ در کلمات کلیدی نیست و وجود یکی از کلمات کلیدی فوق در متن، معمولا انواع مختلف پسوند ۱۹ را نیز به دنبال دارد. کلمات کلیدی انگلیسی (با در نظر گرفتن حروف بزرگ و کوچک انگلیسی):

- Corona
- Korona
- Covid
- Kovid

در هنگام خزش و جمع‌آوری پست‌ها از کانال‌های تلگرام، لازم است اطلاعات تکمیلی در مورد آنها نیز استخراج شده و در دادگان وارد شود. این اطلاعات شامل تاریخ نشر، شناسه کانال در تلگرام، شناسه پست در آن کانال و تعداد مشاهدات[1] می‌باشد. به منظور استفاده از دادگان در استخراج الگوی انتشار اخبار، لازم است که تعداد مشاهدات هر پست از یک آستانه مشخصی کمتر نباشد. برای کانال‌های مرتبط با کرونا، حد آستانه ۱۰۰ بدین‌منظور مناسب تشخیص داده شد، ولی برای کانال‌های عمومی بهتر است این حد آستانه خیلی بالاتر (مثلاً ۱۰۰۰) در نظر گرفته شود.

٤-٤- برچسب‌گذاری داده‌ها

در این مرحله اخبار و پست‌های تلگرامی برچسب‌گذاری می‌شوند. با بررسی مقالات و پژوهش‌های اخیر، اهمیت موضوع موردنظر و با یک جمع‌بندی مناسب و همچنین با بررسی داده‌های اولیه و منابع خبری، هفت وظیفه برای استفاده در برچسب‌گذاری داده‌ها مورد استفاده قرار گرفت. این وظیفه‌ها عبارتند از:

أ) factual (آیا خبر واقعی است یا جعلی؟): خبری واقعی[2] است که بدون هیچ جهت‌گیری و صرفا از لحاظ علمی و پزشکی کاملا مورد تائید باشد، یعنی صرفاً بر اساس نتایج و یا آمار یا بررسی‌های علمی و غیره به طور واضح مورد تایید سازمان‌های مربوط قرار گرفته باشد. درغیراین‌صورت آن خبر یا پست، غیرواقعی[3] یا جعلی است. به عنوان مثال:

1. view
2. factual
3. non-factual

– تعداد مبتلایان روزانه و بهبود یافتگان مقدار x است: اگر این خبر از طرف سازمان‌های مورد تایید، منتشر شده باشد، برچسب yes و در غیر این صورت برچسب no خواهد خورد. اگر نتوان رد یا تایید را تشخیص داد، برچسب can’t decide می‌خورد.

– خبرهایی که در مورد نتایج واکسن‌ها، میزان جهش بیماری و مانند آن که توسط سازمان‌های بهداشتی منتشر شده است، اگر صحت داشته باشند، برچسب yes، در صورت رد خبر، برچسب no و اگر نتوانیم تشخیص دهیم که درست است یا غلط است، برچسب can’t decide خواهد خورد.

ب) hate, blame, negative Speech (آیا خبر حاوی محتوای تنفرآمیز، سرزنش، عیب‌جویی، منفی بافی یا مطالب ناامیدکننده هست یا خیر؟): اگر از روی خبر خوانده‌شده بتوانیم تشخیص دهیم که آن خبر، احساس سرزنش و تنفر و عیب‌جویی و در کل احساس منفی به مخاطب منعکس می‌کند، برچسب yes و اگر کاملا مشهود باشد که القای چنین حسی را برای مخاطب در خود ندارد، برچسب no و اگر نتوان این احساسات را تشخیص داد، برچسب can’t decide خواهد خورد. به عنوان مثال:

– عوارض واکسن‌ها بعد از دو سال نمایان می‌شود که در بیشتر موارد موجب مرگ افراد می‌گردد: کاملا واضح است که این خبر، ناامیدی برای خواننده القا می‌کند و هر فرد می‌تواند این احساس را دریافت کند؛ بنابراین در این وظیفه برچسب yes می‌خورد.

– فوت ٤١٢ نفر در تاریخ ۲۰۲۱/۱۰/۰۱ : این خبر نیز احساس ناامیدی به خواننده القا می‌کند؛ بنابراین برچسب yes می‌خورد.

– واکسن فایزر نسبت به جهش جدید ایمنی بالایی دارد: این خبر برای هر فردی که آن را می‌خواند، نه تنها احساس بدی القا نمی‌کند بلکه بسیار امیدوارکننده است بنابراین برچسب no می‌خورد.

ج) rise moral, give advise (آیا خبر حاوی مطالبی در جهت بالابردن سطح آگاهی عمومی، دادن روحیه یا دادن توصیه به افراد هست یا خیر؟): در این موارد بررسی می‌کنیم که آیا موارد ذکرشده به خواننده القا می‌گردد یا خیر. به عنوان مثال آیا توصیه‌ای به فرد برای مصرف فلان مواد غذایی دارد؟ خبر برای فرد امیدبخش است و لبخند رضایت برای فرد به ارمغان می‌آورد؟ یا آیا خبر باعث افزوده شدن اطلاعات درست به خواننده می‌شود؟ که در صورت مثبت بودن هر یک از این موارد برچسب yes و در صورت منفی بودن برچسب no و در صورتی که نتوان تشخیص داد برچسب can’t decide تخصیص می‌دهیم. به عنوان مثال:

– پژوهش‌های اخیر از تاثیر مثبت ویتامین D در بهبود بیماران کرونایی خبر می‌دهد: کاملا واضح است این خبر، یک خبر نوید بخش در رابطه با بیماری کرونا است بنابراین برچسب yesمی‌خورد.

د) political (آیا خبر حاوی مطالب سیاسی هست یا خیر؟): اگر خبر به هر نحوی رنگ و بوی سیاسی داشته باشد برچسب yes و اگر نداشته باشد برچسب no و در صورت عدم تشخیص برچسب can’t decide خواهد خورد. به عنوان مثال:

– شروع بیماری کووید-۱۹ از یک کشور ابرقدرت اقتصادی بود!: همان‌طورکه از این خبر بر می‌آید هر شخصی که این خبر را بخواند ناخودآگاه ذهنش به سمت و سوی عمدی بودن شیوع این بیماری برای سود کشور چین می‌رود پس پای سیاست به میان می‌آید بنابراین برچسب yes می‌خورد.

– ثبت نام افراد بالای ۳۸ سال برای تزریق واکسن از فردا: کاملا مشخص است که خبر هیچ ارتباطی به سیاست ندارد و صرفا اعلام یک خبر مهم برای آگاهی مردم است؛ پس برچسب no می‌خورد.

ه) cure (آیا خبر در مورد مراقبت‌های بهداشتی یا بحث دارو و درمان هست یا خیر؟): اگر خبر به هر نحوی مربوط به دارو و درمان و مراقبت‌های بهداشتی و پزشکی بود، برچسب yes و غیر اینصورت no و اگر قابل تشخیص نبود برچسبcan’t decide می‌خورد. به عنوان مثال:

–۲۰ میلیون واکسن امروز وارد کشور شد: این خبر آمار یک مورد مهم در بخش درمان و پیشگیری را به ما می‌دهد بنابراین بدیهی است که برچسب yes می‌خورد.

– واکسن آسترازنکا باعث ایجاد ناباروری برای خانم‌های جوان می‌شود: این خبر فارغ از اینکه به دلیل ایجاد امتناع افراد برای واکسیناسیون منتشر می‌شود یا اینکه بدون هیچ هدف خاصی آن را منتشر کرده‌اند، مربوط به دارو و موارد بهداشتی است بنابراین برچسب yes می‌خورد.

– فوت ۲۰۰ نفر در فلان روز: این خبر هیچ مطلبی درمورد دارو و درمان و مسائل بهداشتی نمی‌دهد؛ بنابراین برچسب no می‌خورد.

و) mortality (آیا خبر در مورد مرگ و میر یا آمار مرگ و میر و ابتلا به کرونا هست یا خیر؟): اگر خبر مربوط به آمار و ارقام فوتی‌ها بود yes و اگر نبود برچسب no می‌خورد.

ز) worth fact checking (آیا خبر ارزش بررسی درستی یا نادرستی را دارد یا خیر؟): اطلاعاتی که بدون هیچ جهت‌گیری و هدف خاصی منتشر می‌شوند ولی هنوز از لحاظ علمی کاملا مورد تایید یا رد، قرار نگرفته‌اند، ارزش بررسی درستی یا نادرستی را دارند. به بیان دقیق‌تر اخبار و اطلاعاتی که درمورد هر مسئله‌ای حول بیماری کرونا وجود دارد و پس از بررسی به

این نتیجه می‌رسیم که طبق مطالب شبکه‌های خبری معتبر هنوز این خبر کاملا تایید نشده و در دست مطالعه و بررسی است، برچسب yes خواهد خورد و اگر رد یا مورد تایید بودن خبر، مشخص بود، برچسب no می‌خورد چراکه با بررسی منابع خبری کاملا به نتیجه می‌رسیم. به عنوان مثال:

– احتمال بروز مشکل لختگی خون از هر ۱۰۰۰ نفر برای یک نفر برای واکسن سینوفارم وجود دارد: اگر این خبر را در چندین منبع معتبر داخلی وخارجی بررسی کنیم متوجه خواهیم شد که هنوز در رابطه با آن به طور قطعی به یک نتیجهٔ کلی نرسیده‌اند پس می‌توان در این وظیفه برچسب yes را تخصیص داد.

– تا ماه مِی قرص جایگزین واکسن وارد می‌شود: اخباری نظیر این خبرها که مواردی را در آینده بیان کرده است، نمی‌توان کاملا موثق دانست و بهتر است برچسب yes بخورد.

در کل همان‌طورکه مشخص است، برای هر وظیفه سه برچسب yes، no و can’t decide پیشنهاد شده است و به‌طور کلی در تمامی این هفت وظیفه، برچسب‌ها به صورت زیر تخصیص می‌یابند:

- اگر خبر مورد نظر با توجه به بررسی‌ها، شامل مطالب و مفاهیم موردنظر در آن وظیفه بود، برچسب yes می‌خورد.
- اگر خبر مورد نظر با توجه به بررسی‌ها، شامل مطالب و مفاهیم موردنظر در آن وظیفه نبود، برچسب no می‌خورد.
- اگر نتوان درمورد برچسب‌زدن یک خبر به صورتyes یا no تصمیم گرفت، برچسب can’t decide می‌خورد.

در جدول ۱ خلاصه‌ای از وظایف آمده است. لازم به ذکر است که هر خبر در هر وظیفه فقط یک برچسب می‌خورد.

جدول ۱- خلاصه‌ای از وظایف هفت‌گانه

عنوان برچسب (وظیفه)	توضیح برچسب	مقدار برچسب*
1- factual	آیا خبر واقعی است یا جعلی؟	Y/ N/ X
2- hate, blame, negative speech	آیا خبر حاوی محتوای تنفرآمیز، سرزنش، عیب‌جویی، منفی‌بافی یا مطالب ناامیدکننده هست یا خیر؟	Y/ N
3- rise moral, give advise	آیا خبر حاوی مطالبی در جهت بالابردن سطح آگاهی عمومی، دادن روحیه یا دادن توصیه به افراد هست یا خیر؟	Y/ N
4- political	آیا خبر حاوی مطالب سیاسی هست یا خیر؟	Y/ N/ X
5- cure	آیا خبر در مورد مراقبت‌های بهداشتی یا بحث دارو و درمان هست یا خیر؟	Y/ N
6- death, mortality	آیا خبر در مورد مرگ و میر یا آمار مرگ و میر و ابتلا به کرونا هست یا خیر؟	Y/ N
7- worth fact checking	آیا خبر ارزش بررسی درستی یا نادرستی را دارد یا خیر؟	Y/ N/ X

*Y=Yes, N=No, X=Can’t decide

۵-٤- منابع مورد مراجعه برای برچسب‌گذاری واقعی یا جعلی بودن خبر

ازآنجاکه از بین هفت وظیفه مذکور، وظیفهٔ مربوط به تعیین واقعی یا جعلی بودن خبر از همه مهم‌تر بوده و هدف تهیهٔ دادگان نیز در این راستا می‌باشد، در این بخش برخی از منابع مورد مراجعه برای برچسب‌گذاری اخبار در این وظیفه را ذکر می‌کنیم:

- وبگاه‌ها، کانال‌ها و صفحات شبکه‌های اجتماعی خبرگزاری‌هایی که در زمینهٔ نشر اخبار درست، شناخته‌شده هستند.
- سایت‌ها و کانال و صفحات شبکه‌های اجتماعی که جعلی بودن یا نبودن اخبار را بررسی می‌کنند.
- صفحات شبکه‌های اجتماعی که در زمینهٔ نشر اخبار پزشکی موثق هستند. در این معیار از صفحات پزشکی به زبان انگلیسی هم می‌توان استفاده کرد.
- صفحات شبکه‌های اجتماعی متخصصان و پزشکان مجرب که بدون هیچ‌گونه جهت‌گیری و کسب منافع شخصی و با ارتباط با مراکز مهم تحقیقاتی داخل و خارج از کشور، جدیدترین و موثق‌ترین اخبار را در مورد ویروس کرونا نشر می‌دهند و در مورد شایعه بودن یا نبود اخبار، توضیحات درست را در صفحه‌های مجازی خود منتشر می‌کنند؛ به عنوان مثال، ستادهای کرونا و صفحهٔ دانشگاه شهید بهشتی و مانند آن.

- وبگاه‌های معتبر که می‌توان از طریق موتورهای جست‌وجو و لغات کلیدی خبر مورد نظر، به آنها دست یافت.

نکتهٔ قابل توجه در مورد صفحات شبکه‌های اجتماعی متخصصان و مراکز پزشکی داخلی و خارجی، این است که با مقایسهٔ یک خبر مشخص در چند صفحه تخصصی مختلف می‌توان به نتیجهٔ قابل اعتمادی رسید و همین‌طور می‌توان گفت نظر اشخاصی که متخصص بیماری‌های عفونی و میکروبیولوژی و ویروس‌شناسی هستند، می‌تواند بسیار حائز اهمیت باشد به‌خصوص در مواردی که تمام این متخصصان نظر مشابهی در مورد خبرهای کووید-۱۹ داشته باشند؛ پس در مورد خبرهای مهم و حساس که برچسب‌گذاری آنها نیاز به تحقیق بیشتری دارد، دقت عمل بیشتری اعمال خواهد شد.

در جدول ۲، عنوان تعدادی از صفحات وب و شبکه‌های اجتماعی برای تحلیل و بررسی اخبار و صحت‌سنجی آنها آمده است. شایان ذکر است که برای این منظور، نمی‌توان برای یک خبر، فقط به یک منبع برای تحلیل و صحت‌سنجی بسنده کرد و باید بسیاری از این موارد را در کنار هم مورد بررسی قرار داد. صفحات متخصصینی که برخی از آنها در جدول ۲ آمده است، عمدتاً افرادی بدون جهت‌گیری‌های سیاسی و یا ابزاری، برای منافع شخصی یا گروهی هستند.

جدول ۲- برخی از صفحات مورد استفاده در برچسب‌گذاری

NAME	NAME
@Sbmuniversity	@Doctor__online
@dattums	@Dr_behnam_hajihossailoo
@Vira university	@Dr.shojaei.icu
@wikihoax	@Persian_epochtimes
@factnameh	@drmasoudmardani
@who	@scienceORG
@Emergency_iran	@khabarfouri
Tasnimnews.com	shayeaat
@_dr.key_	Irna.ir
@Sam.pournezhad	Isna.ir

٦-٤- شیوهٔ برچسب‌گذاری یک خبر

با توجه به اینکه هفت وظیفه باید برچسب‌گذاری شود، هر یک از این وظیفه‌ها جداگانه بررسی می‌شود. با توجه به نوع وظیفه، اگر به بررسی منابع خبری احتیاج بود، به سراغ منابع ذکرشده رفته و خبر بررسی می‌گردد. در بسیاری از موارد، شخص برچسب‌زن صرفاً با خواندن خبر و تجزیه و تحلیل آن بر اساس شم زبانی، دریافت احساسات و منطق خود، می‌تواند تشخیص دهد که خبر در چه دسته‌ای قرار می‌گیرد؛ بنابراین در این موارد انجام برچسب‌زنی به‌طور ساده‌تری

انجام می‌گیرد. با توجه به موارد گفته‌شده به‌جز یک یا دو وظیفه، بقیهٔ موارد بدون مراجعه به منابع خبری و با سرعت بیشتری برچسب‌گذاری خواهد شد.

۵- نتایج

۱-۵-شیوه‌نامهٔ نحوهٔ برچسب‌گذاری دادگان

با شروع مرحلهٔ برچسب‌گذاری طی ٤ مرحله جمع‌آوری و برچسب‌زنی دادگان، نتایج زیر تحت عنوان شیوه‌نامهٔ برچسب‌زنی اخبار حاصل شد که در بهبود عملکرد برچسب‌زنی موثر خواهد بود:

۱. باید توجه داشت که در هر پست، هر وظیفه به صورت جداگانه و بدون توجه به سایر وظیفه‌ها برچسب‌زنی می‌شود. به عنوان مثال:

احضار دکتر روازاده به دلیل اظهاراتش در ممنوعیت واکسن

این پست در وظیفهٔ factual ممکن است برچسب can’t decide بخورد ولی در وظیفهٔ worth fact checking برچسب no می‌خورد و در واقع موضوع مهمی برای ارزش بررسی نیست. به طور کلی در این مثال خاص نمی‌توان گفت چون خبر واقعی یا جعلی بودن خبر مشخص نیست، حتماً باید ارزش بررسی درستی/نادرستی داشته باشد یا برعکس. یا اینکه اگر خبری کلاً نادرست بود، باز هم وظیفه‌های دیگر آن را بررسی می‌کنیم و برچسب‌زنی می‌کنیم. بنابراین در هنگام برچسب زنی، هر وظیفه را به تنهایی مورد ارزیابی قرار می‌دهیم تا برچسب‌های واقع بینانه و مناسب‌تری داشته باشیم.

۲. در بعضی از وظیفه‌ها ما از برچسب can’t decide استفاده نکرده‌ایم (جدول ۱). برای مثال برای وظیفه mortality که مربوط به مرگ‌ومیر است، می‌توان صریح گفت که یا خبر مشخصاً به مرگ‌ومیر ربط دارد یا خیر! و احتیاجی به برچسب can’t decide نیست. در کل در وظیفه‌های "hate, blame, negative speech"، "rise moral, give advise"، "cure" و "mortality" تنها از دو برچسب yes و no استفاده شده است. لازم به ذکر است که هر خبر در هر وظیفه تنها یک برچسب می‌گیرد.

۳. بسیاری از پست‌ها، مانند آمار فوتی‌های روزانه، به طور مکرر وجود دارند که برچسب‌زنی را با توجه به ماهیتی که دارند راحت‌تر می‌کنند و هر بار برچسب یکسانی خواهند خورد و تغییر نخواهد کرد. به عنوان مثال:

- اخبار مربوط به تعداد واکسیناسیون در وظیفهٔ rise moral, give advise: برچسب yes
- اخبار مربوط به امار فوتی‌ها در وظیفهٔ hate, blame, negative speech: برچسب yes
- اخباری که درمورد واکسیناسیون باشد در وظیفهٔ cure: برچسب yes

– اخبار مربوط به رعایت پروتکل‌های بهداشتی و موارد مربوط به قرنطینه در وظیفه cure نیز برچسب yes می‌خورند.

– اخباری که درمورد رنگ‌بندی شهرها است، اگر حتی یک شهر غیر از آبی باشد، در وظیفهٔ hate, blame, negative speech (به دلیل ناامید کننده بودن خبر) برچسب yes می‌خورد.

٤. بعضی از پست‌ها ممکن است چند خبر را در خود جای دهند ولی فقط یکی از خبرها مربوط به کروناست، ما فقط با توجه به همان یک خبر برچسب‌گذاری می‌کنیم.

۵. بعضی از اخبار با وجود اینکه ممکن است به‌طورکلی به صورت سوالی باشند، ولی قابل برچسب‌گذاری هستند، به عنوان مثال "با وجود اینکه نقشه از رنگ قرمز خارج شده، ولی دلیل پایین نیامدن آمار فوتی‌ها چیست؟" این خبر را می‌توان برچسب‌گذاری کرد چرا که علی‌رغم سوالی بودن، اطلاعات خوبی به ما می‌دهد.

٦. منبع برخی از اخبار درست که منبع موثقی نیز است، در انتهای خبر آمده است. بااین‌حال سعی شده است که از منبع خبری دیگری برای اثبات درستی خبر استفاده گردد.

۷. خبرها و پست‌هایی که در زبان‌ها و گویش‌های دیگری به غیر از فارسی است، حذف می‌شوند.

۲-۵-نتایج اتمام برچسب‌گذاری دادگان

حدود ٤۰۰۰۰ داده در ٤ مرحله مورد بررسی قرار گرفت که بعد از تحلیل و بررسی داده‌ها حدود ۵۰۰۰ خبر قابل برچسب‌گذاری آماده شد. رویهٔ برچسب‌زنی به‌وسیلهٔ دو برچسب‌زن خبره انجام شد. هر دو برچسب‌زن دارای تحصیلات کارشناسی ارشد در زمینه‌های مرتبط با علوم رایانه و زبان‌شناسی رایانشی بودند. هر دو برچسب‌زن کل دادگان را در وظیفهٔ factual برچسب‌زنی کردند. برای برچسب‌زنی شش وظیفهٔ دیگر فقط از یک برچسب‌زن استفاده شد. براساس معیار کاپا، توافق برچسب‌زن‌ها برای وظیفهٔ factual معادل ۰.۷۰٦ محاسبه شد که این میزان توافق در معیار کاپا در محدودهٔ "عملکرد خوب" قرار می‌گیرد (کارلتا[1]، ۱۹۹٦). دو برچسب‌زن برای حدود ۸۰۰ مورد از بین ۵۰۰۰ پست مورد برچسب‌زنی با یکدیگر اختلاف نظر داشتند که در این موارد، نظر برچسب‌زن خبرهٔ سوم ملاک تصمیم نهایی قرار گرفت.

آمار برچسب‌های پیکرهٔ نهایی در هفت وظیفهٔ مختلف، در جدول ۳ آمده است. همان‌طورکه مشاهده می‌شود، در بعضی از وظایف، هیچ برچسبی با عنوان خنثی (can't decide) نداریم. این مورد در وظایفی پیش می‌آید که در آنها معمولاً ابهامی وجود ندارد. به عنوان مثال اینکه خبر دارای محتوایی در مورد مراقبت‌های بهداشتی است یا خیر (cure)، هیچ وقت دارای ابهام نیست

1. Carletta

و همیشه مشخص است که برچسب خبر در این وظیفه بله (yes) یا خیر (no) است. این حالت را ابتدا برای چهار وظیفه از ۷ وظیفه پیش‌بینی کردیم (جدول ۱) ولی در عمل برای وظیفهٔ worth fact checking نیز برچسب خنثی نداشتیم (جدول ۳). با توجه به آمار، میزان خبرها با برچسب yes در وظیفه factual، بیشتر است. این موضوع نشان‌دهندهٔ این است که اخبار درست در ادامهٔ همه‌گیری کرونا، رفته‌رفته بیشتر شده و از جو متشنج فضا در اثر انتشار اخبارجعلی و ناآگاهی افراد جامعه، کاسته می‌شود. به دلیل بیشتر بودن اخبار با برچسب yes در وظیفه factual، سعی کردیم که تا حد امکان از بین ٤٠٠٠٠ دادهٔ اولیه اخبار را طوری انتخاب کنیم که میزان برچسب‌های yes و no در این وظیفه تا حد امکان متعادل[1] باشد. این امر برای بحث آموزش مدل‌های یادگیری ماشین از اهمیت زیادی برخوردار است.

در جدول ٤، خلاصه‌ای از مشخصات پیکرهٔ تهیه‌شده در این پژوهش (شامل حجم پیکره از لحاظ تعداد سندها، تعداد کل کلمات و تعداد برچسب‌ها) آمده است. همچنین در این جدول، مشخصات پیکرهٔ حاضر با پیکره‌های تهیه‌شدهٔ مشابه در زبان فارسی در پژوهش‌های سقایان و همکاران (۲۰۲۰) و قیومی (۱٤۰۱) مقایسه گشته است. هر دو پیکرهٔ مذکور، اخبار مرتبط با همه‌گیری کووید-۱۹ را بر اساس صحیح یا جعلی بودن برچسب‌دهی کرده‌اند. همان‌طورکه مشاهده می‌شود، حجم پیکرهٔ تهیه‌شده در این پژوهش از لحاظ تعداد سندهای برچسب‌دهی‌شده در حدود ده برابر پیکره‌های مشابه قبلی می‌باشد. همچنین پیکرهٔ حاضر از لحاظ تعداد کل کلمات در حدود ۳.۵ برابر پیکرهٔ قیومی (۱٤۰۱) حجم دارد که در مجموع این پیکره را برای آموزش مدل‌های یادگیری ماشین به منظور تشخیص خودکار اخبار جعلی مناسب‌تر می‌سازد.

جدول ۳- آمار برچسب‌های پیکره در هفت وظیفه مختلف

وظیفه / برچسب	factual	hate, blame, negative speech	rise moral, give advise	political	cure	mortality	worth fact checking
yes	۲۸۱۳	۲۱۱۸	۱٤۰۲	۳۱۳	۲۰۷۷	۲۱۱	۳٦٤
no	۱۹۹٦	۲۸۸۲	۳٥۹۸	٤٤۳۰	۲۹۲۳	٤۷۸۹	٤٦۳٦
can't decide	۱۹۱	۰	۰	۲٥۷	۰	۰	۰

1. balance

جدول ٤- مشخصات پیکرهٔ تهیه‌شده و مقایسهٔ آن با پیکره‌های مشابه

عنوان پیکره	تعداد کل سندها	تعداد کل کلمات	تعداد برچسب‌ها
پیکرهٔ پژوهش حاضر	٤٩٩١	٣٧٠٤٦٢	۳ در ۷ وظیفه
پیکرهٔ پژوهش سقایان و همکاران (۲۰۲۰)	۵۰۰	-	٤
پیکرهٔ پژوهش قیومی (١٤٠١)	٥٦٤	١٠٨٨٢٧	۲

٦- جمع‌بندی

در این مقاله پیکره‌ای برچسب‌خورده با حدود ۵۰۰۰ پست تلگرامی در مورد همه‌گیری کووید-۱۹ تهیه شد. پیکرهٔ مذکور مربوط به پست‌های تلگرام از کانال‌های مختلف خبری و کانال‌های مربوط به کووید-۱۹ بوده که در بازهٔ زمانی ۲۰۲۱/٦/۲۲ تا ۲۰۲۲/٤/۲۰ جمع‌آوری شده‌اند. در این بازهٔ زمانی علاوه‌بر شیوع بیماری، بحث واکسیناسیون هم در جامعه رایج بوده است. بعد از فیلتر کردن پست‌ها بر اساس کلمات کلیدی مربوط به همه‌گیری و همچنین حذف پست‌های تکراری، اقدام به برچسب‌زنی در هفت وظیفه مربوط به بیماری کرونا گردید. مهمترین وظیفه برچسب‌زنی که هدف عمده پژوهش حاضر بود، جعلی یا صحیح بودن خبر می‌باشد، که برچسب‌ها در این وظیفه شامل Yes، No و Can’t decide بودند. انتخاب پست‌ها طوری انجام شد که در نهایت برچسب‌ها در این وظیفه، متعادل باشند. برچسب‌زنی توسط دو برچسب‌زن خبره انجام شد که در صورت عدم توافق بین آنها، برچسب‌زنی به نفر سوم داده شد. این پیکره از لحاظ تعداد مستندات، حجمی در حدود ده برابر نسبت به پیکره‌های مشابه در زبان فارسی دارد. در مراحل بعدی کار، در نظر است که این دادگان در زمینهٔ تشخیص ماشینی و خودکار اخبار جعلی مرتبط با کرونا به کار رود.

۷- تشکر و قدردانی

این پژوهش با حمایت پژوهشگاه ارتباطات و فناوری اطلاعات (مرکز تحقیقات مخابرات) تحت قرارداد پژوهشی شماره ۵۰۰/٦٩٤١/پ انجام شده است. نویسندگان مقاله از سرکار خانم مهندس سیده فاطمه ابراهیمی که در برچسب‌زنی داده‌ها به نویسندگان کمک کردند، تشکر و قدردانی می‌نمایند.

منابع

قیومی، مسعود (۱٤۰۱). «تحلیل آماری اخبار جعلی فارسی مربوط به کوید-۱۹»، *فصلنامه علمی - پژوهشی زبان‌شناسی اجتماعی*، دوره ۵، شماره ٤، صص ۳۵-۵۲.

Ameur, Mohamed Seghir Hadj, and Hassina Aliane (2021). "Arabic Covid-19 Multi-Label Fake News and Hate Speech Detection Dataset", *Procedia Computer Science*, vol. 189: 232-241.

Aphiwongsophon, S., and P. Chongstitvatana (2018). "Detecting Fake News with Machine Learning Method", In *15th International Conference on Electrical Engineering/Electronics,Computer, Telecommunications and Information Technology* (*Ecti-Con*), Chiang Rai, Thailand.

Carletta, Jean (1996). "Assessing Agreement on Classification Tasks: The Kappa Statistic", *Computational Linguistics,* 22(2): 249–254.

Crestani F., and P. Rosso (2020). "The Role of Personality and Linguistic Patterns in Discriminating Between Fake News Spreaders and Fact Checkers", In *25th International Conference on Applications of Natural Language to Information Systems*, Saarbrucken, Germany.

Elhadad, Mohamed K., Kin Fun Li, and Fayez Gebali (2021). "Covid-19-Fakes: A Twitter (Arabic/English) Dataset for Detecting Misleading Information on Covid-19", In *Leonard Barolli, Kin Fun Li, and Hiroyoshi Miwa, Editors, Advances in Intelligent Networking and Collaborative Systems*, pp. 256–268, Springer International Publishing.

Kaliyar, Rohit Kumar, Anurag Goswami, and Pratik Narang (2021). "Fake News Detection In Social Media with a Bert-Based Deep Learning Approach", *Multimedia Tools & Applications,* vol. 80, 11765–11788.

Kumar, S., and N. Shah (2018). "False Information on Web and Social Media: A Survey", *arXiv:1804.08559*.

Posadas-Durán, J. P., H. Gómez-Adorno, G. Sidorov, and J. J. M. Escobar (2019). "Detection of Fake News in a New Corpus for the Spanish Language", *Journal of Intelligent & Fuzzy Systems*, 36(5): 4869-4876.

Saghayan, M. H., S. F. Ebrahimi, and M. Bahrani (2021). "Exploring the Impact of Machine Translation on Fake News Detection: A Case Study on Persian Tweets about COVID-19", *Proceedings of 29th Iranian Conference on Electrical Engineering (ICEE)*, pp. 540-544, IEEE.

Singh, Vivek K., Rupanjal Dasgupta, Darshan Sonagra, Karthik Raman, and Isha Ghosh (2017). "Automated Fake News Detection Using Linguistic Analysis and Machine Learning", In *International Conference on Social Computing, Behavioral-Cultural Modeling, & Prediction and Behavior Representation in Modeling and Simulation*, pp. 1-3.

Shahi, Gautam Kishore, and Durgesh Nandini (2020). "FakeCovid - A Multilingual Cross-Domain Fact Check News Dataset for Covid-19", *arXiv:2006.11343*.

Shin J., L. Jian, K. Driscol, and F. Bar (2018). "The Diffusion of Misinformation on Social Media: Temporal Pattern, Message, and Source", *Computers in Human Behavior*, vol. 8:278–287.

Shu K., D. Mahudeswaran, S. Wang, D. Lee, and H. Liu (2020). "Fakenewsnet: A Data Repository with News Content, Social Context, and Spatio Temporal Information for Studying Fake News on Social Media", *Big Data* 8(3):171–188.

Vijayaraghavan, S., Y. Wang, Z. Guo, J. Voong, W. Xu, A. Nasseri, J. Cai, L. Li, K. Vuong, and E. Wadhwa (2020). "Fake news Detection with Different Models*", arXiv:2003.04978*.

Wang, William Yang (2017). “Liar, Liar Pants on Fire: A New Benchmark Dataset for Fake News Detection”, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, Vancouver, Canada.

Zhou, Xinyi and Reza Zafarani, (2021), “A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities”, *ACM Computing Surveys*, 53(5): 1-40.

استناد به این مقاله: زاهد، فروغ، بحرانی، محمد و منصوری، علیرضا. (۱۴۰۲). طراحی و تهیهٔ دادگان برچسب‌خورده فارسی از اخبار مرتبط با همه‌گیری کووید-۱۹ به منظور تشخیص اخبار جعلی. *زبان و زبان‌شناسی*، ۱۸ (۳۵)، ۱۷۳-۱۹۲.

doi: 10.30465/lsi.2024.47711.1729