Journal of Sociolinguistics
Payame Noor University

فصلنامهٔ زبان‌شناسی اجتماعی
دورهٔ پنجم (سری جدید)، شمارهٔ ۴، پیاپی ۲۰، پاییز ۱۴۰۱ (صفحات ۵۲-۳۵)

Original Research **مقالهٔ پژوهشی**

تحلیل آماری اخبار جعلی فارسی مربوط به کوید-۱۹

مسعود قیومی*
استادیار زبان‌شناسی، پژوهشگاه علوم انسانی و مطالعات فرهنگی، تهران، ایران

دریافت: **۱۴۰۱/۰۲/۲۱** پذیرش: **۱۴۰۱/۰۳/۱۴**

A Statistical Analysis of Persian Fake News on COVID-19

Masood Ghayoomi*
Assistant Professor of Linguistics, Institute for Humanities and Cultural Studies, Tehran, Iran

Received: **2022/05/11** Accepted: **2022/06/04**

10.30473/il.2023.63989.1537

چکیده

در این پژوهش تلاش می‌شود با استفاده از تحلیل آماری، ویژگی‌های اخبار جعلی فارسی مربوط به کوید-۱۹ بررسی گردد. برای این هدف، ابتدا یک پیکرهٔ زبانی که حاوی اخبار موثّق و جعلی در حوزهٔ کرونا است تهیه می‌شود. سپس الگوهای زبانی این دو دستهٔ داده و همچنین دو تحلیل آماری مقدار اطلاعات و خوانایی اخبار موثّق و جعلی مورد بررسی قرار گرفته و با یکدیگر مقایسه می‌شود. براساس اطلاعات استخراج‌شده و نتایج عملی به‌دست‌آمده از پیکرهٔ خبرهای جعلی، الگوهای زبانی مشترک بین این دو دستهٔ داده وجود دارد. همچنین، مقدار اطلاعات در اخبار موثّق براساس دو معیار آنتروپی و شگفتی بیشتر از اخبار جعلی است. سطح خوانایی خبرهای جعلی با استفاده از تساوی‌های اندازه‌گیری خوانایی متن مورد ارزیابی قرار گرفته‌است و این نتیجه به‌دست آمده است که اخبار جعلی در مقایسه با اخبار موثّق عمدتاً ساده بوده و دشوار نیست. در فرایند برچسب‌گذاری خودکار خبرهای موثّق و جعلی براساس سطح دشواری حجم زیادی از اخبار جعلی ساده تشخیص داده شده‌است و تعداد کمی از اخبار موثّق با سطح زبانی دشوار بود. علاوه‌بر این دستاورد و بررسی آماری ویژگی‌های زبانی براساس میزان اطلاعات و خوانایی اخبار جعلی، جنبهٔ کاربردی این اطلاعات آماری جهت تشخیص خبر جعلی با استفاده از روش‌های یادگیری ماشینی مورد مطالعه قرار گرفت.

کلیدواژه‌ها: زبان رسانه، اخبار جعلی فارسی، کوید-۱۹، نظریه اطّلاعات، آنتروپی، شگفتی، خوانایی.

Abstract

In this research, an attempt is made to investigate the characteristics of Persian fake news related to Covid-19 by using statistical analysis. To this end, first, a language corpus containing reliable and fake news in Persian in the field of Corona is prepared. Then, the language patterns of these two data sets, as well as two statistical analyzes of the amount of information and the readability of reliable and fake news, are examined and compared with each other. According to the exteracted information and the experimental results achieved from the developed corpus on COVID-19 fake news, there are common language patterns in these two datasets. Moreover, the amount of information in reliable news is more than fake news based on two measures of entropy and surprise. Based on the results, the readability level of the fake news is measured based on the readability formulas. According to the results, the text of fake news is simpler than real news. In the process of automatic labeling of reliable and fake news based on the level of difficulty, most news is recognized as simple texts. The results show that fake news is mostly simple and not difficult compared to reliable news. In addition to this achievement, to study linguistic properties of fake news statistically based on the information amount and readability, the applicablity of this statistical information was studied to detect fake news using machine learning methods.

Keywords: Media Language, Persian Fake News, COVID-19, Information Theory, Entropy, Surprisal, Readability.

***Email:** m.ghayoomi@ihcs.ac.ir

مقدمه

در گذشته انتشار اخبار تنها از طریق روزنامه و تلویزیون انجام می‌شد. اما امروزه با گسترش چشمگیر رسانه‌های اجتماعی و وبگاه‌های خبری، حجم بالایی از اطّلاعات، از جمله اخبار، به‌راحتی درمیان کاربران مبادله می‌شود. با افزایش روزافزون تعداد اخبار منتشرشده در این رسانه‌ها، تشخیص درستی و صحّت این اخبار از اهمیّت ویژه‌ای برخوردار است؛ چرا که اخبار منتشرشده در رسانه‌های اجتماعی ممکن است جعلی باشد و به‌سرعت میان افراد دست‌به‌دست شود. به‌عنوان مثال، توییتر[1] یکی از رسانه‌های اجتماعی محبوب برای به‌اشتراک‌گذاشتن اخبار و نظرات درمورد رخدادهای مختلف توسط کاربران و اصحاب رسانه است. طبق آمار[2]، روزانه حدود ۵۰۰ میلیون توییت در توییتر به اشتراک گذاشته می‌شود که این رقم می‌تواند نشان‌دهندهٔ مناسب‌بودن این بستر برای انتشار اخبار جعلی[3] درمیان کاربران باشد. در مطالعهٔ وثوقی[4] و همکاران (۲۰۱۸)، سرعت گسترش اخبار جعلی در توییتر شش برابر بیشتر از سرعت انتشار اخبار موثّق اعلام شده‌است. خبر جعلی، خبری دروغ است که می‌تواند باعث فریب مردم شود. اصطلاح خبر جعلی با اصطلاحات دیگری همچون «شایعه»[5]، «اطّلاعات ناصحیح»[6] یا «تبلیغات»[7] نیز شناخته می‌شود. اخبار جعلی به سه دسته تقسیم می‌شود: جعل جدّی[8]، کلاه‌برداری بزرگ[9] و جعل طنزآمیز[10] (روبین[11] و همکاران، ۲۰۱۵). عوامل دیگری، مانند ظهور یک رسانه یا هواداری یک رسانه از یک گرایش سیاسی، زمینه را برای تحت‌تأثیرقراردادن اخبار موثّق فراهم می‌آورد (وارگو[12] و همکاران، ۲۰۱۸). یک رسانه این اخبار را عمدتاً با هدف گمراه‌کردن و به‌منظور آسیب‌رساندن به یک گروه یا فرد با اهداف مالی یا سیاسی منتشر می‌کند و می‌تواند بر عقاید افراد و تصمیم‌های شخصی آنها تأثیر مستقیم بگذارد. همچنین، اخبار جعلی برروی افکار عمومی کشورها و حتی اقدامات سیاسی دولت‌ها و نهادهای بین‌المللی تأثیر قابل‌توجّه‌ای دارد. نمونهٔ آن، انتخابات سال ۲۰۱۶ ریاست جمهوری آمریکا است که حجم زیادی از اخبار جعلی در ارتباط با کاندیداها در رسانه‌های اجتماعی منتشر شد و تأثیر بسیار زیادی برروی نتیجهٔ این انتخابات داشت (بووت[13] و ماکسه[14]، ۲۰۱۹).

امروزه با گسترش چشم‌گیر رسانه‌های خبری و اجتماعی، حجم بالایی از اخبار و اطّلاعات به‌راحتی در اختیار کاربران قرار می‌گیرد و زمینه را برای وضعیت اینفودمی فراهم می‌آورد. از نظر سازمان بهداشت جهانی، منظور از اینفودمی وجود حجم زیادی از اطلاعات در قالب دیجیتال است که حاوی اطلاعات جعلی، منحرف‌کننده و غیرقابل اعتماد بوده و سبب سردرگمی افراد و بروز رفتارهای خطرناک در آنان می‌گردد و سلامت جامعه را تحت تأثیر قرار می‌دهد.[15] اگرچه این حجم اطّلاعات در ابتدا با هدف آگاهی و اطّلاع‌رسانی ارائه می‌شد، رفته‌رفته به‌واسطهٔ تأثیر زیاد این اطّلاعات بر افکار عمومی و هزینهٔ اندک نشر و گسترش آن، فضایی برای سودجویان فراهم آورد تا با سوارشدن بر موج اخبار و اطّلاعات غلط، به اهداف سیاسی و مادّی خود برسند. علی‌رغم فعّالیت‌های پژوهشی انجام‌شده در این حوزه بر زبان‌های مختلف و به‌طور خاص بر زبان انگلیسی، زبان فارسی در این زمینه مورد توجّه قرار نداشته‌است؛ بنابراین، انجام هرگونه فعّالیت پژوهشی جهت تشخیص اخبار جعلی فارسی چه از نظر زبان‌شناختی و رسانه‌ای و چه از نظر فنّی برای تهیهٔ ابزار تشخیص خودکار خبر جعلی از اهمیّت به‌سزایی برخوردار است. از سوی دیگر، همه‌گیری کوید-۱۹ (کرونا) در دو سال اخیر باعث شده‌است که این موضوع به‌عنوان یکی از موضوعات مهم در اکثر خبرهای روزانه به‌چشم بیاید. از آنجا که این موضوع با مسائل مختلف ازجمله سلامت و بهداشت، اقتصاد، سیاست و حتّی مسائل فرهنگی و اجتماعی و مذهبی عجین شده‌است، انتشار اخبار جعلی در این حوزه می‌تواند ابعاد مختلف زندگی افراد جامعه را تحت‌الشعاع قراردهد. در این پژوهش تلاش می‌شود با استفاده از تحلیل‌های آماری و روش‌های محاسباتی، ویژگی‌های کلّی اخبار جعلی فارسی مربوط به کوید-۱۹ مورد بررسی قرار گیرد و از این اطّلاعات جهت تشخیص خودکار خبر جعلی استفاده گردد.

1. Twitter
2. https://www.internetlivestats.com/twitter-statistics/
3. fake news
4. S. Vosoughi
5. rumor
6. misinformation
7. propaganda
8. serious fabrication
9. large-scale hoaxes
10. humorous fakes
11. V. L. Rubin
12. C. Vargo
13. A. Bovet
14. H. A. Makse
15. https://www.who.int/health-topics/infodemic

پیشینهٔ پژوهش

پیشینهٔ پژوهش را می‌توان به دو دسته تقسیم کرد: الف) مطالعات انجام‌شده برروی تشخیص خبر جعلی به‌صورت خودکار؛ و ب) مطالعات انجام‌شده با استفاده از داده‌های مربوط به کوید-۱۹.

مطالعات انجام‌شده برروی تشخیص خودکار خبر جعلی

احمد[1] و همکاران (۲۰۱۷) با استفاده از تحلیل و بررسی متن و روش مبتنی‌بر چندتایی[2] و نمایش برداری مبتنی بر «بسامد واژه-معکوس بسامد سند»[3]، از الگوریتم‌های متداول یادگیری ماشین، مانند ماشین بردار پشتیبان[4]، ماشین بردار پشتیبان خطی[5]، نزدیک‌ترین کای همسایه[6]، درخت تصمیم[7]، گرادیان کاهشی تصادفی[8] و رگرسیون لجستیک[9]، برای تشخیص اخبار جعلی استفاده کرده‌اند. آنها با ایجاد ترکیب‌های متفاوت از توالی واژه‌ها به‌صورت تکی و دوتایی و غیره، اطّلاعات آماری هر ترکیب را با استفاده از روش مبتنی بر واژه برای مجموعهٔ یادگیری، شامل اخبار جعلی و موثّق، بازنمایی کردند. با انجام آزمایش‌ها، بهترین نتیجه با استفاده از الگوریتم ماشین بردار پشتیبان خطی با دقّت ۹۲ درصد به‌دست آمد.

در پژوهشی که توسط رمضانی[10] و همکاران (۲۰۱۹) انجام شده‌است، سعی شده‌است با استفاده از «شبکه‌های عصبی بازگشتی»[11] و معرفی یک «تابع هزینهٔ»[12] جدید، دنبالهٔ اخبار به‌صورت یک پیوستار زمانی بررسی شود و در هر مقطع زمانی با یک احتمال، برچسب خبر مشخص گردد. این پژوهش برروی دادگان جمع‌آوری‌شده از دو رسانهٔ اجتماعی توییتر و «سینا وایبو»[13] آزمایش شده‌است.

خطار[14] و همکاران (۲۰۱۹) از «خودکدگذار وردشی»[15] استفاده کردند تا اخبار جعلی را تشخیص دهند. برای این کار از مجموعه دادهٔ توییتر استفاده کرده‌اند که شامل یک بخش متن و یک عکس به‌همراه آن است. مدل آنها از سه بخش کدگذارِ[16] متن و تصویر، کدگشای[17] متن و تصویر و تشخیص‌دهندهٔ خبر جعلی تشکیل شده‌است. مجموعهٔ کدگذار و کدگشا در فرایند یادگیری، ویژگی‌های نهان اخبار جعلی را با توجّه به متن و عکس‌ها یاد می‌گیرد؛ سپس با استفاده از این ویژگی‌های نهان توسط بخش تشخیص‌دهنده، جعلی‌بودن آن را تشخیص می‌دهد. دقّت نهایی به‌دست‌آمده برای دادگان توییتر و وایبو به ترتیب ۷۴/۵٪ و ۸۲/۴٪ بوده‌است.

یانگ[18] و همکاران (۲۰۱۹) برای تشخیص اخبار جعلی از روش یادگیری «بدون نظارت»[19] استفاده کردند که در آن با استفاده از یک مدل احتمالاتی گرافیکی، صحّت اخبار و اعتبار کاربران مدل شده‌است. لیو[20] و همکاران (۲۰۱۹) یک مدل دومرحله‌ای براساس مدل برت[21] (دولین[22] و همکاران، ۲۰۱۹) ارائه کردند که خبر جعلی با استفاده از تعبیهٔ اطّلاعات فراداده، مانند نام گوینده، شغل و غیره به‌همراه متن اصلی، تشخیص داده می‌شود. این مدل بر روی دادگان لیار[23] آزمایش شده و دقّت ۲۹/۰۷٪ را بدون استفاده از اطّلاعات فراداده و دقّت ۴۰/۵۸٪ را با استفاده از اطّلاعات فراداده به‌دست آورده‌است.

ژوا[24] و همکاران (۲۰۱۹) نیز با استفاده از پیکرهٔ اخبار حاصل از خبرگزاری‌های سی.ان.ان[25] و «دیلی میل»[26]، مدل برت را آموزش دادند و از دادگان مرحلهٔ اوّل مسابقهٔ اخبار جعلی[27] که با عنوان اف.ان.سی.وان[28] شناخته می‌شود برای «تنظیم دقیق»[29] مدل استفاده کردند. در این پژوهش، دو مدل برای تشخیص خبر جعلی ارائه شده‌است که در یک مدل از مدل آمادهٔ برت به‌همراه تابع آنتروپی متقاطع وزن‌دار استفاده شده و در مدل دیگر، با استفاده از پیکرهٔ اخبار، مدل برت آموزش داده شده‌است. نتایج آزمایش این دو مدل به

1. H. I. Ahmed
2. *n*-gram
3. term frequency - inverse document frequency
4. support vector machine
5. linear support vector machine
6. *k* nearest neighbor
7. decision tree
8. stochastic gradient descent
9. logistic regression
10. M. Ramezani
11. recurrent neural network
12. loss function
13. Sina Weibo
14. D. Khattar
15. variational auto-encoder
16. encoder
17. decoder
18. S. Yang
19. unsupervised learning
20. C. Liu
21. BERT
22. J. Devlin
23. Liar
24. H. Jwa
25. CNN
26. Daily Mail
27. https://fakenewschallenge.org
28. F.N.C-1
29. fine tune

ترتیب کارایی برابر با ۷۳/۴٪ و ۷۴/۶٪ به‌دست آورده‌است.

ژانگ[1] و همکاران (۲۰۲۰) یک سامانه با عنوان «تشخیص‌دهنده جعل»[2] پیاده‌سازی کردند که از دو بخش اصلی تشکیل شده‌است: «یادگیری ویژگی بازنمایی»[3] و «استخراج ویژگی صریح»[4]. در کنار هر خبر، اطّلاعات اجتماعی متنوع مرتبط با آن خبر وجود دارد که مدل‌سازی در آن توسط یک واحد یادگیرندهٔ ویژگی انجام می‌پذیرد. این واحد پردازشی، امکان استفاده از چندین ورودی متنوع را به‌صورت همزمان ایجاد می‌نماید.

گلدانی[5] و همکاران (۲۰۲۰) به‌منظور دسته‌بندی اخبار جعلی متوسّط و یا طولانی، از چهار شبکهٔ موازی و برای اخبار کوتاه، از دو شبکهٔ موازی به‌منظور استخراج ویژگی‌های سطح بالا در متن استفاده کردند. پس از استخراج ویژگی‌های سطح بالا، هر شبکه به یک لایهٔ چگال[6] متصل شده و در نهایت با استفاده از یک لایهٔ میانگین ادغام[7]، احتمال جعلی‌بودن خبر مشخص شده‌است. دقّت این روش برروی دادگان آی.اس.اُ.تی.[8] (احمد و همکاران، ۲۰۱۷) ۹۹/۸٪ و برروی دادگان لیار با دسته‌بندی ۶ برچسب، ۳۹/۵٪ بوده‌است.

کالیار[9] و همکاران (۲۰۲۰) یک شبکهٔ عصبی عمیق پیچشی را با نام اف.ان.دی.نت[10] برای تشخیص اخبار جعلی ارائه کردند. در این مدل از لایه‌های پیچشی استفاده شده‌است که به‌صورت آبشاری قرار گرفته‌است تا ویژگی‌های مناسبی را برای اخبار تولید کند. درنهایت، با استفاده از لایه‌های چگال، احتمال تعلق هر خبر به دستهٔ جعلی یا موثّق مشخص می‌شود. به‌منظور ارزیابی مدل ارائه‌شده، آنها از دادگان وبگاه کگل[11] که مربوط به انتخابات سال ۲۰۱۶ آمریکا است استفاده کرده و دقّت ۹۸/۳۶٪ را به‌دست آورده‌اند.

در حوزهٔ تشخیص خبر جعلی در فارسی، زمانی[12] و همکاران (۲۰۱۷) با تمرکز بر روی شبکه‌های اجتماعی به تحلیل و بررسی شایعات فارسی در توییتر پرداختند. در این پژوهش، با بهره‌گیری از اطّلاعات مرتبط با شبکه ارتباطی کاربران و ویژگی‌های مرتبط با هر توییت به دسته‌بندی شایعات با استفاده از مدل‌های متداول یادگیری ماشین پرداختند. علاوه بر تحلیل ویژگی‌های پراهمیّت برای تشخیص شایعات در شبکه‌های اجتماعی، یک مجموعهٔ داده در حوزهٔ تشخیص شایعات توییتر ارائه دادند که شامل ۷۸۳ توییت جعلی و ۷۸۳ توییت موثّق است. به‌منظور استخراج شایعات از دو وبگاه ایرانی گمانه[13] و شایعات[14] استفاده کردند.

محمودآباد[15] و همکاران (۲۰۱۸) در پژوهش خود مجموعهٔ داده‌ای با بررسی نظرهای ۱۱۹۸۱ کاربر فارسی‌زبان در رسانهٔ اجتماعی توییتر تهیه کردند. این مجموعهٔ داده متشکل از بیش‌از ۵/۳ میلیون توییت فارسی است که عمدتاً در مورد زلزلهٔ کرمانشاه بوده و با استفاده از وبگاه شایعات برچسب‌گذاری شده‌است. به‌دلیل نامتوازن‌بودن اخبار شایعه و موثّق، با استفاده از الگوریتم بیش‌نمونه‌برداری اسموت[16]، آنها داده‌ها را متوازن کرده و سپس هر توییت را با برداری شامل اطّلاعات زمینه‌ای، اطّلاعات ساختاری و اطّلاعات جمعیتی بازنمایی کردند. درنهایت داده‌ها با استفاده از الگوریتم‌های متداول یادگیری ماشین دسته‌بندی شد.

جهانبخش نقده[17] و همکاران (۲۰۲۰) مجموعهٔ دادهٔ دیگری با تمرکز برروی اخبار منتشرشده در شبکهٔ اجتماعی تلگرام منتشر کردند. این مجموعهٔ داده شامل ۸۸۲ شایعه و ۸۸۲ پست موثّق است که از کانال خبرگزاری‌های خبرگزاری فارس، دانشجویان ایران (ایسنا)، تسنیم، مهر و خبرگزاری جمهوری اسلامی (ایرنا) و همچنین سه وب‌سایت گمانه، شایعات و ویکی هواکس[18] خزش شده‌است. مدل پیشنهادی آنها الهام‌گرفته از مدل ارائه‌شده توسط آلپورت[19] و پستمن[20] (۱۹۴۷) است. در این مدل، یک رابطه به‌عنوان قدرت شایعه معرفی شده‌است که ارتباط مستقیمی با اهمیّت خبر و ابهام آن دارد. باتوجّه‌به این مقاله، جهانبخش و همکارانش به پیاده‌سازی مدلی برای محاسبه این ضریب، تحت عنوان «قدرت انتشار شایعه در زبان فارسی» پرداختند. برای محاسبهٔ این ضریب، از ویژگی‌های مانند احساس خبر، اهمیّت خبر و ابهام خبر استفاده شده‌است.

1. J. Zhang
2. fake detector
3. representation feature learning
4. explicit feature extraction
5. M. H. Goldani
6. dense
7. average pooling
8. I.S.O.T
9. R. K. Kaliyar
10. F.N.D.N.E.T
11. Kaggle
12. S. Zamani
13. https://gomaneh.net/
14. http://shayeaat.ir/
15. S. D. Mahmoodabad
16. SMOTE
17. Z. Jahanbakhsh-Nagadeh
18. https://wikihoax.org/
19. G. W. Allport
20. L. Postman

صمدی[1] و همکاران (۲۰۲۱) مدل پردازشی برای تشخیص خبر جعلی فارسی پیشنهاد دادند که مبتنی‌بر بازنمایی سند و بازنمایی توالی است که یک بازنمایی برای هر واژه سند فراهم آمده‌است. در این مدل از بازنمایی مبتنی‌بر بافت جایگاهی که مبتنی‌بر برت (دولین و همکاران، ۲۰۱۹) است در دو مدل استفاده کرده‌است. در مدل اوّل از شبکهٔ عصبی پرسپترون تک‌لایه و در مدل دوم از شبکهٔ عصبی پیچشی استفاده شده‌است. برای این هدف، یک مجموعهٔ داده از رسانه‌های خبری تهیه شده که شامل ۱۸۶۰ خبر جعلی و ۱۸۶۰ خبر موثّق است. اخبار جعلی از خبرگزاری‌های مختلف جمع‌آوری شده و اخبار موثّق از پنج خبرگزاری معتبر آنلاین فارسی، یعنی ایرنا، ایسنا، فارس‌نیوز، همشهری و مهرنیوز خزش شده‌است.

جهانبخش و همکاران (۱۴۰۰) در پژوهش دیگری به ارائه مدلی برای تشخیص شایعات فارسی مبتنی بر ویژگی‌های با ارزش اطّلاعات محتوایی در متن رسانه‌های اجتماعی پرداخته‌اند. در این پژوهش، از داده‌های جمع‌آوری‌شده توسط جهانبخش و همکاران (۲۰۲۰) و زمانی و همکاران (۲۰۱۷) استفاده شده‌است. علاوه‌بر این داده‌ها، از داده‌های صیفی‌کار[2] و همکاران (۲۰۱۸) که مربوط به زلزلهٔ کرمانشاه است نیز استفاده شده‌است.

مطالعات انجام‌شده با استفاده از داده‌های مربوط به کوید-۱۹

تن[3] (۲۰۲۰) به بررسی واژهٔ «fear» (ترس(در اخبار جعلی مربوط به کوید-۱۹ پرداخته‌است. در این پژوهش، از بخش اوّل پیکرهٔ خبرگزاری رویترز[4] که حاوی ۱/۸ میلیون واژه است استفاده شده‌است. براساس تحلیل انجام‌شده، واژهٔ «ترس» در خبرهای جعلی کاربرد خیلی زیادی داشته‌است؛ درحالی‌که این واژه در اخبار موثّق کاربردی نداشته‌است.

باتلر[5] و سیمون-وندنبرگن[6] (۲۰۲۱) به تحلیل پیکره‌بنیان سه اصطلاح «social distance»، «physical distance» و «physical distancing» در زبان انگلیسی کشور انگلستان پرداختند. هدف آنها در این پژوهش این بوده‌است که تا چه حد این دو واژه به‌طور ثابت کنار یکدیگر باقی می‌ماند.

حسینی[7] و همکاران (۲۰۲۰) در پژوهش خود مجموعه‌ای از نظرات منتشرشده در توییتر را گردآوری کرده و در قالب یک پیکره مورد بررسی قرار داده‌اند. آنها بیش‌از ۵۳۰هزار توییت یکتای بدون بازنشر را تحلیل محتوایی کرده و ۲۵ موضوع را در این مجموعهٔ داده تشخیص داده‌اند.

محمودی دهکی[8] و همکاران (۲۰۲۰) با جستجو در وب، اصطلاحات جدیدی که به‌واسطهٔ همه‌گیری کوید-۱۹ وارد زبان انگلیسی شده‌است را جمع‌آوری کرده و معادل‌های این مفاهیم را در اخبار فارسی جستجو کرده‌اند. خروجی این پژوهش، یک فرهنگ اصطلاح‌شناسی دوزبانهٔ مربوط به کوید-۱۹ است که می‌تواند به‌عنوان واژگان و اصطلاحات در حوزهٔ کوید-۱۹ مورد استفاده قرار گیرد.

مذهب و شهیدی‌تبار (۱۳۹۹) در پژوهش خود به بررسی استعاره‌های مفهومی در اخبار برخط فارسی مرتبط با بیماری کوید-۱۹ پرداختند. برای این هدف، آنان پیکرهٔ محدودی که شامل ۵۰ خبر از اخبار فارسی مربوط به کوید-۱۹ است را جمع‌آوری کرده و از نظر استعاری مورد بررسی قرار داده‌اند.

قیومی (۱۴۰۰) در پژوهش خود مجموعه داده‌ای مربوط به کوید-۱۹ از رسانه‌های اجتماعی اینستاگرام و توییتر با بیش‌از ۶۰هزار نظر یکتا را تهیه کرده و به بررسی موضوعات مطرح‌شده در این داده‌ها و همچنین هشتگ‌ها پرداخته‌است. در این پژوهش، ارتباط بین هشتگ و موضوع نظر از منظر معیار آماری ضریب همبستگی پیرسون مورد بررسی قرار گرفته‌است.

همانگونه که در مطالعات انجام‌شده مشخص است، بررسی خبر جعلی موضوعات متنوعی را از جنبهٔ نظری و کاربردی در بر می‌گیرد که بیانگر ظرفیت پژوهشی این حوزه است. با این حال، بررسی ویژگی‌های نهفته در خبرهای جعلی در مقایسه با خبر موثّق از اهمیت به‌سزایی برخوردار است که در این پژوهش به آن پرداخته خواهد شد.

چارچوب نظری

رویکردهای زبانی معرفی‌شده برای تحلیل اخبار جعلی به چند دسته قابل تقسیم است که در ادامه این روش‌ها توضیح داده می‌شود:

الف) روش «کیسه واژه»[9] برای نمایش متون هر واژه

1. M. Samadi
2. M. Seifikar
3. K. H. Tan
4. https://www.reuters.com
5. C. S. Butler
6. A. M. Simon-Vandenbergen
7. P. Hosseini
8. M. Mahmoudi-Dehaki
9. bag of word

به‌عنوان یک واحد مورد توجّه قرار می‌گیرد. در کنار آن، توالی چندواژه‌ای‌های موجود در متن استخراج می‌شود. براساس بسامد تکرار واژه‌ها و یا چند واژه‌ای‌ها متن تجزیه و تحلیل می‌شود تا نشانه‌هایی مبنی بر وجود باهم‌آیی در ترکیبات مشخص و آشکار شود. در نتیجه استفاده از این بسامدها در دادهٔ آموزش یک دسته‌بند می‌توان به پیش‌بینی اخبار جعلی پرداخت (کونروی[1]، ۲۰۱۵).

ب) اخبار موثّق حاوی واژه‌هایی است که با دقّت به بیان جزئیات می‌پردازد. بنابراین، کاربرد واحدهای زبانی کمّی مشخّص، مانند اعداد و ارقام، و یا ساخت زبانی مقایسه‌ای، مانند صفت تفضیلی، در اخبار موثّق بیشتر ظاهر می شود (لوگی[2]، ۲۰۲۱).

ج) اخبار جعلی که ماهیت فریبنده دارد از زبان ساده برای بیان اطّلاعات استفاده می‌کند و دشواری متن یک خبر جعلی کمتر از متن موثّق است. ازاین‌رو، می‌توان با استفاده از تساوی‌هایی که با اصطلاح «خوانایی»[3] معرفی می‌شود به تعیین سطح زبانی و بررسی سطح دشواری متون متعلق به پیکرهٔ خبر جعلی پرداخت. شایان ذکر است تساوی‌های متنوعی برای محاسبهٔ دشواری متن وجود دارد که در ادامه به نمونه‌های آنها اشاره می‌شود.

علاوه بر موارد فوق، برای تمیز خبر جعلی از خبر موثّق، در پژوهش‌ها ویژگی‌های دیگری، مانند استفاده از واژه‌های اغراق‌آمیز یا ویژگی‌های نحوی (بایسونگر[4] و استورر[5]، ۲۰۰۸؛ وایسر[6]، ۲۰۱۶؛ جین[7]، ۲۰۱۷) مطرح شده‌است. در پژوهش حاضر، دو رویکرد آماری مبتنی‌بر معیارهای آماری خوانایی و همچنین نظریه اطّلاعات برای تحلیل متون خبری و بررسی ویژگی جعلی‌بودن در آنها به‌کار می‌رود. ازاین‌رو، در ادامه، این دو رویکرد معرفی می‌گردد.

معیار خوانایی

تجزیه و تحلیل آماری متن توسط شرمن[8] (۱۹۸۳) در قرن ۱۸ آغاز شد. در اوایل قرن نوزدهم، لایولی[9] و پرسی[10] (۱۹۲۳) اوّلین تلاش را برای معرفی تساوی‌های تعیین سطح دشواری کتاب‌های درسی دورهٔ متوسطه انجام دادند. در پایان دههٔ ۱۹۸۰، بیش از ۲۰۰ تساوی خوانایی معرفی شد (دوبای[11]، ۲۰۰۴: ۲). در میان آنها، تساوی دال–شال[12] (دال[13] و شال[14]، ۱۹۴۸) و تساوی فلش–کینسایید[15] (کینسایید[16]، ۱۹۷۵) متداول‌ترین آنها است. دیانی (۱۳۶۶؛ ۱۳۶۹) و موسوی (۱۳۹۷) تساوی‌های متنوعی ازجمله دو تساوی متداولی که به آنها اشاره شد را برای اندازه‌گیری خوانایی متن‌های کتب درسی فارسی استفاده کرده‌اند. دیانی (۱۳۶۶؛ ۱۳۶۹) تلاش کرده‌است تعدادی از تساوی‌های خوانایی که عمدتاً برای زبان انگلیسی طراحی شده‌است را براساس متن فارسی منطبق کند و به ارزیابی سطح دشواری متون درسی فارسی بپردازد.

در تساوی دال–شال (دال و شال، ۱۹۴۸) برای ارزیابی خوانایی متن انگلیسی و سطح دشواری، مجموعه‌ای از واژه‌های ازپیش‌تعریف‌شدهٔ انگلیسی به‌عنوان «واژه‌های دشوار» استفاده می‌شود. این فهرست مهم‌ترین نقش را ایفا می‌کند به‌طوری‌که این لیست به‌شدت به زبان مقصد وابسته است و برای توسعه نیاز به کار دستی فشرده دارد. در این فهرست، منظور از واژه‌های دشوار مواردی است که تعداد هجاهای آنها سه و بیشتر باشد. ازاین‌رو، با دانستن تعداد هجاهای واژه‌ها می‌توان این فهرست را برای هر زبانی تهیه کرد. تساوی‌های خوانایی دیگری معرفی شده‌است که بر ویژگی‌های زبانی عمومی در یک متن تکیه می‌کند. این تساوی‌ها عمدتاً نیاز به پارامترهای وزن‌دار دارد و ثابت است. دشواری فلش[17] (۱۹۷۹) در تساوی (۱) از نسبت تعداد واژه‌های متن به تعداد جملات و نسبت تعداد هجاها به تعداد واژه‌ها استفاده می‌کند.

$$FRE = ۲۰۶٫۸۳۵ - ۱٫۰۱۵\left(\frac{N_w}{N_s}\right) - ۸۴٫۶\left(\frac{N_t}{N_w}\right) \qquad (۱)$$

1. N. J. Conroy
2. J. Lugea
3. readability
4. M. Beißwenger
5. A. Storrer
6. M. Weisser
7. Z. Jin
8. L. A. Sherman
9. B. A. Lively
10. S. L. Pressey
11. W. H. DuBay
12. Dale-Chall
13. E. Dale
14. J. S. Chall
15. Flesch-Kincaid
16. J. P. Kincaid
17. R. Flesch

که در این تساوی، N_w کل واژه‌ها، N_s مجموع جملات و N_l کل هجای واژه‌ها در یک متن است.

تساوی دیگر که برای ارزیابی خوانایی متن به‌کار می‌رود توسط گانینگ[1] (۱۹۵۲) پیشنهاد شده است که در تساوی (۲) معرفی شده است.

$$GFI = ۰/۴\left[\frac{N_w}{N_s} + ۱۰۰\left(\frac{N_{cw}}{N_w}\right)\right] \qquad (۲)$$

که N_{cm} تعداد واژه‌های پیچیده، N_w تعداد واژه‌ها و N_s تعداد جملات است. ویژگی‌های استفاده‌شده در این تساوی عبارت است از نسبت تعداد واژه‌ها به تعداد جملات و نسبت تعداد واژه‌های پیچیده به تعداد کل واژه‌ها.

تساوی خوانایی دیگر، «شاخص خوانایی خودکار»[2] است که توسط اسمیت[3] و سنر[4] (۱۹۶۷) معرفی شده‌است که در تساوی (۳) نشان داده شده‌است.

$$ARI = ۴/۷۱\left(\frac{N_c}{N_w}\right) + ۰/۵\left(\frac{N_w}{N_s}\right) - ۲۱/۴۳ \qquad (۳)$$

که N_c تعداد حروف،N_w تعداد واژه‌ها و N_s تعداد جملات است. ویژگی‌های استفاده‌شده در این تساوی عبارت است از نسبت تعداد واژه‌ها به تعداد جملات و نسبت تعداد حروف به تعداد واژه‌ها. مزیت این تساوی در مقایسه با تساوی‌های قبلی این است که تعداد حروف در نظر گرفته می‌شود که نسبت به شمارش تعداد هجا ساده‌تر است.

دیانی (۱۳۶۶) تساوی فلش (۱۹۷۹) در تساوی (۱) را تغییر داد و با اقتباس از آن، شاخص خوانایی فلش-دیانی که در تساوی (۴) نمایش داده شده‌است را برای ارزیابی خوانایی یک متن فارسی معرفی نمود.

$$FDRI = ۲۶۲/۸۳۵ - ۱/۱۰۵\left(\frac{N_w}{N_s}\right) - ۰/۸۴۶\left(\frac{N_t}{N_w}\right) \qquad (۴)$$

معیار نظریهٔ اطّلاعات

اخبار جعلی همانا خبری است که عامدانه با هدف انحراف افکار برای بهره‌برداری اهداف سیاسی، اقتصادی و غیره توسط گروهی تولید و نشر می‌یابد. بنابراین، بین تولیدکنندهٔ خبر، دریافت‌کنندهٔ خبر و خود خبر نوعی ارتباط وجود دارد. شانون[5] (۱۹۴۸) نظریهٔ اطّلاعات را معرفی کرده‌است؛ به این صورت که در یک مدل ریاضی‌گونه مقدار اطّلاعات در کانال ارتباطی بین گوینده و شنونده با استفاده از آنتروپی اندازه‌گیری می‌گردد. هرچقدر آنتروپی بالا باشد، میزان اطّلاعات درمورد آن کم بوده و پیش‌بینی‌پذیری را کمتر می‌کند؛ و بالعکس. برای محاسبهٔ آنتروپیِ یک متغیرِ تصادفی، از تساوی (۵) استفاده می‌شود.

$$H(X) = -\sum_{x \in X} P(x) log_۲ P(x) \qquad (۵)$$

که در این تساوی X متغیر تصادفی است، $P(x_i)$ مقدار احتمال متغیر است و این متغیر می‌تواند یک واژه باشد. در محاسبهٔ مقدار اطّلاعات یک جمله ابتدا بایستی مقدار اطّلاعات هر واژه محاسبه شود و سپس با تجمیع اطّلاعات واژه‌های یک متن، به مقدار اطّلاعات یک متن دست یافت.

رویکرد دیگری که برای محاسبه مقدار اطّلاعات یک متن به‌کار می‌رود معیار شگفتی[6] است که توسط تریبوس[7] (۱۹۶۱) معرفی شده و روش محاسبهٔ آن در تساوی (۶) ذکر شده‌است.

$$S(Y) = log_۲ \frac{۱}{P(y)} \qquad (۶)$$

که در این تساوی Y متغیر تصادفی عمومی است، $P(y)$ مقدار احتمال متغیر است. همانند آنتروپی، هرقدر امتیاز شگفتی بیشتر باشد اطّلاعات موجود در آن بیشتر است و شگفتی بیشتری را بر می‌انگیزد و برعکس. از مقایسهٔ آنتروپی و شگفتی چنین می‌توان نتیجه گرفت که این دو معیار در راستای یکدیگر است با این تفاوت که شگفتی از ابتدا برروی کل متن محاسبه می‌شود و آنتروپی از تجمیع اطّلاعات واژه‌های یک متن محاسبه می‌گردد.

داده‌های پژوهش

از آنجا که برای زبان فارسی داده‌ای برای اخبار جعلی مرتبط با کوید-۱۹ موجود نیست و منبع معتبری نیز برای این اخبار وجود ندارد تا بتوان به‌صورت مستقیم آنها را استخراج کرد، این داده بایستی برای اهداف مورد نظر این پژوهش تهیه گردد.

1. R. Gunning
2. automated readability index
3. E. A. Smith
4. R. J. Sener
5. C. E. Shannon
6. surprisal
7. M. Tribus

استخراج اخبار جعلی بدون آگاهی از محتوای این اخبار همراه با چالش است؛ چراکه تصمیم‌گیری درمورد موثّق‌بودن یک خبر نیاز به اطّلاعات و دانش فراتر از متن یک خبر دارد. بنابراین نمی‌توان از فردی که کار نشانه‌گذاری داده را انجام می‌دهد انتظار داشت صرفاً با خواندن متن یک خبر امکان تشخیص جعلی‌بودن آن خبر داشته باشد. درنتیجه، باید روشی برای استخراج اخبار جعلی ارائه دهیم که از یکسو طیف وسیعی از موضوعات را در بر گیرد و از سوی دیگر قضاوت در مورد جعلی‌بودن آن خبر در فرایند نشانه‌گذاری داده توسط برچسب‌زن را میسر نماید. برای همین منظور، در این پژوهش، یک روش ۶ مرحله‌ای که شبیه روش معرفی‌شده توسط صمدی و همکاران (۲۰۲۱) است را برای استخراج اخبار جعلی کوید-۱۹ ارائه می‌دهیم. این مراحل عبارت است از:

۱. باتوجّه‌به نبود یک منبع اختصاصی و مطمئن برای ذخیرهٔ اخبار جعلی منتشرشده در فضای مجازی مربوط به کوید-۱۹، امکان دسترسی مستقیم به این اخبار وجود ندارد. بنابراین ما از یک فرض عموماً درست در اخبار فارسی استفاده کردیم تا بتوانیم اخبار جعلی را به‌صورت عمومی و در گسترهٔ وسیعی از موضوعات استخراج کنیم. عمدتاً اخبار جعلی فارسی پس‌از مدتی توسط نهادهای رسمی و یا وبگاه‌های خبری تکذیب می‌شود. بنابراین برای هر خبر جعلی انتظار می‌رود که یک تکذیبیه وجود داشته باشد. بر همین اساس، از رویکرد معکوس برای رسیدن به اخبار جعلی استفاده نمودیم. در گام اوّل با یک پرس‌وجو کلی مانند «تکذیب خبر» با استفاده از رابط کاربری برنامه گوگل، فهرستی از اخبار تکذیب‌شده را استخراج می‌کنیم.

۲. با داشتن عنوان اخبار تکذیب‌شده که در مرحلهٔ قبل استخراج شده‌است می‌توانیم با حذف کلمات خاص، مانند «تکذیب» و «شایعه»، به عنوان خبری برسیم که از لحاظ متنی به‌عنوان خبر جعلی اوّلیه شباهت دارد.

۳. پس از استخراجِ عنوانِ احتمالیِ خبر جعلی، با استفاده از موتور جست‌وجوی گوگل تلاش می‌کنیم تا اصل خبر جعلی را در وب‌سایت‌ها بیابیم. در این مرحله، برای هر خبر تکذیب‌شده فهرستی از اخبار جعلی محتمل به‌دست خواهد آمد.

٤. پس‌از جمع‌آوری اخبار احتمالی جعلی، با استفاده از کلیدواژه‌های «کرونا»، «کوید» و «کووید» اخبار مربوط به حوزهٔ کوید-۱۹ را از سایر اخبار تفکیک می‌کنیم و تنها برروی داده‌های مرتبط با کوید-۱۹ تمرکز می‌نماییم.

۵. پس از استخراج اخبار جعلی محتمل در حوزهٔ کوید-۱۹، به‌ازای هر خبر تکذیب‌شده، تمامی اخبار به‌صورت دستی توسط عوامل انسانی برچسب‌گذاری می‌شود.

٦. پس از مشخص‌شدن تعداد اخبار جعلی در گام قبلی، به همان تعداد خبر موثّق به‌صورت تصادفی از خبرگزاری‌های معتبر متفاوت استخراج می‌کنیم و پس از تأیید نیروی انسانی به مجموعهٔ دادهٔ گردآوری‌شده اضافه می‌نماییم.

نمونه‌هایی از خبرهای جعلی در رسانه‌های خبری و رسانه‌های اجتماعی و پیام‌رسان تلگرام ارائه می‌گردد:

-کرونا جان الهام شیخی بازیکن جوان فوتسال را گرفت. الهام شیخی بازیکن پیشین تیم ملی فوتسال زنان ایران بر اثر کرونا درگذشت. به گزارش رسانه‌های ایران، الهام شیخی ۲۳ ساله بود و سابقه بازی در تیم ملی فوتسال زنان و تیم سپاهان اصفهان را داشت. این ورزشکار جوان به دلیل ابتلا به ویروس کرونا در شهر قم درگذشت.

-اقدام بزرگ رونالدو برای مبارزه با کرونا؛ هتل‌های کریستیانو به بیمارستان تبدیل می‌شود. فوق ستاره یوونتوس در اقدامی خیرخواهانه برای مبارزه با ویروس کرونا همه هتل‌هایش را در اختیار دولت پرتغال قرار داد.

-درمان قطعی ویروس کرونا!

-کی به کیه، منم واکسن کرونا رو کشف کردم.

-کروناویروس درمان شد. البته باید تأیید بشه ولی دور از انتظار نیست.

برای جمع‌آوری داده‌های موثّق، از خزش در وب[1] برای جمع‌آوری داده‌ها استفاده می‌شود. هدف از این کار، جمع‌آوری داده خبری در حوزهٔ کوید-۱۹ می‌باشد تا یک پیکره در این دامنه تهیه شود. نتیجهٔ به‌دست‌آمده از فرایندهای شرح داده‌شده برای جمع‌آوری و برچسب‌گذاری اخبار جعلی کوید-۱۹ به فارسی در جدول (۱) گزارش شده‌است. در این دادگان، تا کنون ۲۸۲ خبر جعلی و موثّق درمورد کوید-۱۹ از داده‌های خبری و رسانه‌های اجتماعی

1. Web crawling

به‌دست آمده‌است و پیکره‌ای با حجم ۵۶۴ خبر که شامل خبر جعلی و موثّق به تعداد مساوی است تهیه شده‌است. اگرچه حجم مجموعهٔ دادهٔ تهیه‌شده زیاد نیست، همین مقدار داده نیز با بسیاری از دادگان موجود در حوزهٔ تشخیص خبر جعلی، مانند بازفید[1] (پادهاست[2] و همکاران، ۲۰۱۸) با تعداد ۲۹۹ خبر جعلی برای زبان انگلیسی یا جرمن‌فیک‌ان‌سی[3] (فوگل[4] و ژیانگ[5]، ۲۰۱۹) با تعداد ۴۹۰ داده برچسب‌گذاری‌شده برای زبان آلمانی، قابل رقابت است. این پیکره در دسترس عمومی قرار دارد.[6]

جدول ۱. اطّلاعات آماری پیکرهٔ خبر جعلی کوید-۱۹

تعداد کل داده			دادهٔ جعلی			دادهٔ موثّق		
واژه	واژه یکتا	سند	واژه	واژه یکتا	سند	واژه	واژه یکتا	سند
۱۰۸۸۲۷	۱۰۶۵۴	۵۶۴	۵۴۰۴۴	۶۷۷۵	۲۸۲	۵۴۷۸۳	۷۲۹۷	۲۸۲

تجزیه و تحلیل داده‌ها

بررسی پیکرهٔ خبر جعلی کوید-۱۹

ابتدا تحلیلی از الگوهای زبانی به‌کاررفته در خبرهای موثّق و جعلی ارائه می‌دهیم. برای این هدف، داده‌های خبری موثّق و جعلی در پیکرهٔ خبر جعلی کوید-۱۹ را از یکدیگر جدا کرده و سپس به استخراج فهرست واژه‌ها و توالی ۲ تا ۴ واژه از این دو مجموعهٔ داده و مقایسهٔ الگوی زبانی آنها با یکدیگر می‌پردازیم.

برای شروع، تعداد ۱۰۰ واژهٔ پربسامد را از دو مجموعهٔ داده مربوط به خبرهای جعلی و موثّق استخراج کرده و با هم مقایسه کردیم.[7] از مقایسهٔ این دو فهرست، ۷۳ درصد واژه‌های پربسامد خبرهای موثّق و جعلی مشترک بود و ۲۷ درصد متفاوت بود. در جدول ۲، واژه‌های غیرمشترک خبر موثّق و جعلی فهرست شده‌است. واژه‌های پربسامد غیرمشترک خبر موثّق مربوط به کوید-۱۹ بیشتر برروی مسائل بهداشت و درمان و اطّلاع‌رسانی پیرامون موضوع کوید-۱۹ متمرکز است؛ درحالی‌که واژه‌های پربسامد غیرمشترک خبر جعلی متنوع بوده و انسجام چندانی ندارد.

جدول ۲. مقایسهٔ واژه‌های غیرمشترک در ۱۰۰ واژهٔ پربسامد خبرهای موثّق و جعلی

خبر موثّق	خبر جعلی
استفاده	اثر
اشاره	اسلامی
افراد	آموزش
افزود	او
آمریکا	بانک
باشد	بعد
بیماران	بیمارستان
پزشکی	پرداخت
تعداد	پس
تولید	تهران
جدید	تومان
حال	جان
دارند	حقوق
داشته	خبر
درصد	دانشگاه
درمان	درباره
شرایط	دلیل
طرح	دولت
کاهش	شبکه
کند	شهر
کووید۱۹	قم
مقابله	یکی
میلیون	کار
مورد	گفته
نظر	مجلس
واکسن	نیست
وضعیت	وجود

در بررسی بعدی، توالی دو-واژه‌ای‌ها را از این دو مجموعه استخراج کردیم و به بررسی ۵۰ توالی پربسامد در هر دو مجموعهٔ داده پرداختیم. ۶۸ درصد عبارات دو-واژه‌ای به‌کاررفته در دو دستهٔ خبر مشترک بود. عبارات متفاوت در این دو دستهٔ خبر در جدول ۳ گزارش شده‌است. عبارات غیرمشترک خبر موثّق بیشتر حاوی واژه‌های محتوایی مربوط به کوید-۱۹ است؛ درحالی‌که دو-واژه‌ای‌های خبرهای جعلی متشکل از واژه‌های نقشی است.

1. BuzzFeed
2. M. Potthast
3. GermanFakeNC
4. I. Vogel
5. P. Jiang
6. https://www.ihcs.ac.ir/corpora/fa/page/3376
۷. به‌دلیل محدودیت در صفحات مقاله، از ارائهٔ فهرست‌های بلند صرفه‌نظر کردیم.

جدول ۳. مقایسهٔ واژه‌های غیرمشترک در دو-واژه‌ای‌های خبر موثّق و جعلی

خبر موثّق		خبر جعلی	
واژهٔ اوّل	واژهٔ دوم	واژهٔ اوّل	واژهٔ دوم
این	بیماری	ادامه	داد
این	کشور	از	دست
با	این	آموزش	و
بیش	از	با	کرونا
به	کووید۱۹	بر	اثر
در	حالی	به	ویروس
در	کشور	بود	که
شود	و	در	ایران
شیوع	ویروس	شد	و
علوم	پزشکی	فضای	مجازی
که	از	کرده	است
مبتلا	به	کرونا	به
نفر	از	هزار	تومان
هزار	و	و	از
هلال	احمر	و	این

در تحلیل بعدی، ۵۰ توالی سه-واژه‌ای پربسامد به‌عنوان الگوی زبانی را از خبرهای موثّق و جعلی استخراج کرده و با هم مقایسه کردیم. ۵۲ درصد عبارات سه-واژه‌ای به‌کاررفته در دو دستهٔ خبر مشترک بود. عبارات متفاوت در جدول ۴ گزارش شده‌است. اگرچه خبر جعلی با قاطعیت راجع به رویدادها صحبت نمی‌کند، عباراتی مانند «اظهار کردن»، «خاطرنشان کردن»، «خبر دادن»، «تصریح کردن»، «با اشاره به»، «با توجّه به»، «به نقل از» و مانند آن در اخبار جعلی کوید-۱۹ مشاهده شد که با کاربرد این عبارات تلاش می‌شود به‌صورت ضمنی، خبر جعلی به‌عنوان خبر موثّق جلوه کند و سبب به خطا افتادن نه‌تنها ماشین بلکه انسان در تمیزدهی خبرها شود. در عبارات غیرمشترک خبر موثّق از افعّالی مانند «بیان کردن» و ««تأکیدکردن» استفاده شده‌است که این فعل‌ها در خبر جعلی موجود نبود. همچنین در اخبار موثّق از نظر محتوایی، اهمیّت بیشتری به مسائل بهداشتی و درمانی داده شده‌است؛ درحالی‌که موضوعات متنوعی در عبارات استخراج‌شده از داده‌های متعلق به اخبار جعلی پوشش داده شده‌است.

جدول ۴. مقایسهٔ واژه‌های غیرمشترک در سه-واژه‌ای‌های خبر موثّق و جعلی

خبر موثّق			خبر جعلی		
واژهٔ اوّل	واژهٔ دوم	واژهٔ سوم	واژهٔ اوّل	واژهٔ دوم	واژهٔ سوم
۲۴	ساعت	گذشته	ابتلا	به	ویروس
از	این	تعداد	اثر	ابتلا	به
استفاده	از	ماسک	اظهار	داشت	:
اشاره	به	اینکه	با	ویروس	کرونا
بیان	کرد	:	بر	اثر	ابتلا
برای	مقابله	با	به	بیماری	کرونا
بهداشت	و	درمان	به	پرداخت	اقساط
به	کووید۱۹	در	تصریح	کرد	:
تأکید	کرد	:	خاطرنشان	کرد	:
حالی	است	که	خبر	داد	و
خبر	داد	.	دانشگاه	آزاد	اسلامی
در	این	زمینه	در	فضای	مجازی
در	پاسخ	به	سقف	نقل	و
در	حالی	است	شده	است	.
در	کشور	به	شیوع	کرونا	در
رسیده	است	.	عامل	حوزه	مسکن
سخنگوی	وزارت	بهداشت	عضو	کمیسیون	عمران
قرار	دارند	.	علوم	پزشکی	مشهد
کرونا	در	کشور	کرده	است	.
کووید۱۹	در	کشور	کرونا	در	ایران
مبتلا	به	کووید۱۹	کمیسیون	عمران	مجلس
نفر	از	بیماران	وجود	دارد	و
نفر	رسید	.	ویروس	کرونا	به
وی	با	اشاره	وزیر	آموزش	و

در ادامه، الگوی زبانی را تا چهار واژهٔ متوالی افزایش داده و ۵۰ توالی چهار-واژه‌ای را از خبرهای موثّق و جعلی استخراج کرده و با هم مقایسه کردیم. ۲۰ درصد عبارات چهار-واژه‌ای به‌کاررفته در دو دستهٔ خبر مشترک بود که عبارت است از «این در حالی است»، «با اشاره به اینکه»، «بیماران مبتلا به کرونا»، «به ویروس کرونا در»، «جان خود را از»، «خود را از دست»، «در حالی است که»، «را از دست داده‌اند»، «مقابله با ویروس کرونا» و «وزیر آموزش و پرورش». واژه‌های به‌کاربرده‌شده در این الگوهای زبانی به‌گونه‌ای است که می‌تواند مدل‌های زبانی مبتنی‌بر آمار را با چالش مواجه کند و به کاهش کارایی الگوریتم یادگیری

بیانجامد. الگوهای زبانی غیرمشترک خبر موثّق با خبر جعلی در جدول ۵ گزارش شده‌است. عبارات به‌کاررفته در این دو دستهٔ اخبار بسیار شبیه یکدیگر است که بیانگر سختی تمیزدهی خبر موثّق از خبر جعلی و بالعکس است. شایان ذکر است با افزایش تعداد واژگان در بررسی الگوهای زبانی مشکل «تنک‌بودن داده»[1] بیشتر احساس شده و اشتراک و افتراق دو مجموعهٔ داده از نظر بسامدی تفاوت زیادی ندارد.

جدول ۵. مقایسهٔ واژه‌های غیرمشترک در چهار-واژه‌ای‌های خبر موثّق و جعلی[1]

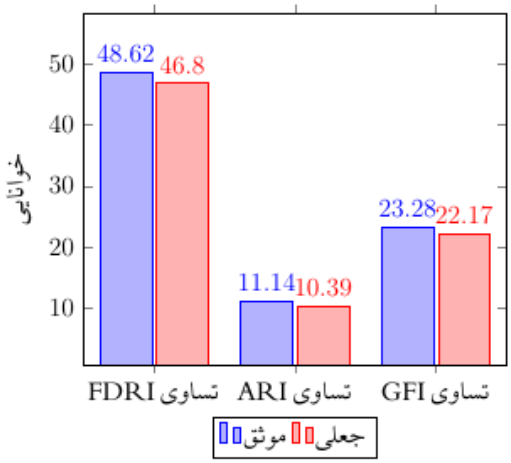
خبر موثّق				خبر جعلی			
واژهٔ اوّل	واژهٔ دوم	واژهٔ سوم	واژهٔ چهارم	واژهٔ اوّل	واژهٔ دوم	واژهٔ سوم	واژهٔ چهارم
ابتلا	به	بیماری	کرونا	از	بیماران	مبتلا	به
ابتلا	به	ویروس	کرونا	آزمایش	تشخیص	کووید۱۹	در
اثر	ابتلا	به	کرونا	اساس	معیارهای	قطعی	تشخیصی
ادامه	داد	:	در	این	بیماری	تحت	مراقبت
از	دست	داد	.	انجام	شده	است	.
اسلامی	ادامه	داد	:	بر	اساس	معیارهای	قطعی
با	توجّه	به	اینکه	بیماران	کووید۱۹	در	کشور
بانک	عامل	حوزه	مسکن	بیماران	مبتلا	به	کووید۱۹
بر	اثر	ابتلا	به	بیماری	تحت	مراقبت	قرار
بر	اساس	تصمیم	بانک	بیمار	جدید	مبتلا	به
به	پرداخت	اقساط	تسهیلات	بهداشت	درمان	و	آموزش
به	دلیل	ابتلا	به	به	کووید۱۹	در	کشور
به	دلیل	شیوع	بیماری	به	کووید۱۹	در	وضعیت
به	ویروس	کرونا	مبتلا	تحت	مراقبت	قرار	دارند
پرداخت	اقساط	تسهیلات	تا	جدید	مبتلا	به	کووید۱۹
تسهیلات	گیرندگان	وام	مسکن	دانشگاه	علوم	پزشکی	ایران
خبر	داد	و	گفت	در	کشور	انجام	شده
خرداد	ماه	سال	جاری	در	کشور	شناسایی	شد
داد	و	گفت	:	درمان	و	آموزش	پزشکی
دانشگاه	علوم	پزشکی	مشهد	در	وضعیت	شدید	این
در	چند	ماه	اخیر	در	وضعیت	قرمز	قرار
در	مجلس	شورای	اسلامی	رئیس	دانشگاه	علوم	پزشکی
دلیل	شیوع	بیماری	کرونا	سازمان	برنامه	و	بودجه
را	از	دست	داد	کرمانشاه	کهگیلویه	و	بویراحمد
را	از	دست	دادند	کشور	انجام	شده	است
ستاد	ملی	مقابله	با	کمیته	بین_المللی	صلیب	سرخ
سقف	برداشت	از	خودپردازها	کووید۱۹	در	کشور	انجام
سقف	نقل	و	انتقال	کووید۱۹	در	کشور	به
شیوع	ویروس	کرونا	در	کووید۱۹	در	کشور	شناسایی
عضو	کمیسیون	عمران	مجلس	مبتلایان	به	ویروس	کرونا
کرونا	جان	خود	را	مبتلا	به	کرونا	در
گیرندگان	وام	مسکن	ملزم	مبتلا	به	کووید۱۹	در

1. data sparsity

خبر موثّق				خبر جعلی			
واژهٔ اوّل	واژهٔ دوم	واژهٔ سوم	واژهٔ چهارم	واژهٔ اوّل	واژهٔ دوم	واژهٔ سوم	واژهٔ چهارم
گفت	:	بر	اساس	مجموع	بیماران	کووید۱۹	در
مسکن	تسهیلات	گیرندگان	وام	میلیون	بشکه	در	روز
مسکن	ملزم	به	پرداخت	نفر	از	بیماران	بهبود
ملزم	به	پرداخت	اقساط	نفر	از	بیماران	مبتلا
ملی	مقابله	با	کرونا	وی	ادامه	داد	:
منطقه	۷	دانشگاه	آزاد	وی	با	اشاره	به
وام	مسکن	ملزم	به	وزارت	آموزش	و	پرورش
ویروس	کرونا	در	ایران	وضعیت	قرمز	قرار	دارند

بررسی متغیرهای خوانایی و اطّلاعات

با استفاده از اطّلاعات پردازش‌شدهٔ حاصل از پردازش‌های توصیف‌شده در قیومی[1] (۲۰۲۲) به بررسی خوانایی خبرهای موثّق و جعلی مربوط به کوید-۱۹ می‌پردازیم. موضوع خوانایی را از دو جنبه بررسی کردیم. جنبهٔ اوّل محاسبهٔ خوانایی متن‌های موثّق و جعلی با استفاده از تساوی‌های معرفی‌شده است. نتایج حاصل از تساوی‌های (۲)، (۳) و (۴) در شکل (۱) قابل مشاهده است.



شکل ۱. مقایسهٔ اندازه‌گیری خوانایی داده‌های خبری موثّق و جعلی کوید-۱۹

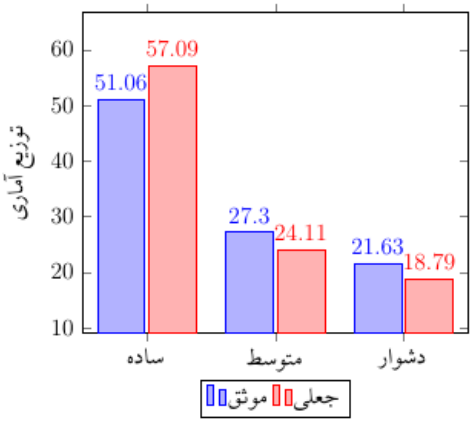
همان‌طور که در این شکل مشاهده می‌شود، به‌طورکلی، براساس معیارهای ارزیابی مختلف، خوانایی متن خبری موثّق نسبت به متن خبری جعلی کمی بیشتر است؛ بنابراین متن خبر موثّق نسبت به متن خبر جعلی کمی دشوارتر است. این نتیجه در فارسی با نظر لوگی (۲۰۲۱) همسو است. شایان ذکر است براساس جدول دشواری-سادگی معرفی‌شده توسط دیانی (۱۳۶۶) برای فارسی، امتیاز به‌دست‌آمده از بررسی خبر موثّق و جعلی طبق تساوی (۴) که به‌ترتیب ۴۸/۶۲ و ۴۶//۸۰ است، سطح «دشوار» برای این متون تشخیص داده می‌شود. همچنین، درجهٔ خوانایی این متون از نظر سطح تحصیلات رسمی باتوجّه‌به قرارگرفتن این امتیاز بین درجهٔ نوشتار ۳۰ تا ۵۰، به گروه «سال‌های اوّل دانشگاه» تعلق دارد.

برای اثبات دشواربودن خبرهای موثّق کوید-۱۹ نسبت به خبرهای جعلی مربوط به این دامنه، تحلیل دیگری را با استفاده از پردازش‌های پایه زبان طبیعی انجام داده‌ایم. در این تحلیل، سطح زبانی خبرهای موثّق و جعلی را در سه سطح ساده، متوسط و دشوار با هم مقایسه کرده‌ایم. برای رسیدن به این هدف از یک دسته‌بند و شیوهٔ استخراج ویژگی در مدل پردازشی معرفی‌شده توسط قیومی (۲۰۲۲) استفاده کرده‌ایم. وی در پژوهش خود به موضوع پیش‌بینی‌پذیری سطح دشواری متون کتاب‌های آموزش فارسی به غیرفارسی‌زبانان پرداخته‌است. ویژگی‌هایی که وی در تهیه مدل پردازشی معرفی کرده‌است عبارت است از اطّلاعات آماری از داده خام بدون هرگونه پردازش، اطّلاعات آماری مستخرج از داده‌ای که در سطح بن‌واژه برچسب‌گذاری شده‌است، اطّلاعات آماری مستخرج از اطّلاعات نحوی در سطح واژه و عبارت که از مقولات دستوری تخصیص‌داده‌شده به واژه و نمودارهای درختی وابستگی و سازه‌ای به‌دست آمده‌است و اطّلاعات معنایی در سطح واژه که از برچسب‌گذاری موجودیت‌های نامدار استخراج شده‌است. ازمیان دسته‌بندهایی که وی آموزش داده‌است، دسته‌بند جنگل تصادفی بالاترین کارایی را به‌دست آورده‌است. بنابراین، در این پژوهش نیز از این دسته‌بند که کارایی مناسبی دارد بهره می‌بریم. پس‌از استخراج ویژگی‌ها

1. M. Ghayoomi

از داده‌ها، با استفاده از الگوریتم یادگیری جنگل تصادفی در مدل قیومی (۲۰۲۲)، کار برچسب‌گذاری هر یک از متن‌های خبری موثّق یا جعلی براساس سه سطح زبانی ساده، متوسط و دشوار انجام شده و سپس براساس تفکیک دوگانهٔ خبری (موثّق یا جعلی)، فراوانی سه سطح زبانی را استخراج کردیم. در شکل (۲) توزیع آماری سطوح زبانی متن‌های خبری نشان داده شده‌است.

همان‌طور که مشخص است، تعداد متن‌های ساده خبر جعلی نسبت به خبر موثّق بیشتر است و تعداد متن‌های متوسط و دشوار خبر جعلی نسبت به خبر موثّق کمتر است. این تفاوت بیانگر این است که دشواری متن‌های موثّق بیشتر است و خبرهای جعلی از نظر زبانی ساده‌تر است. نتیجهٔ حاصل از این پردازش با نتیجهٔ به‌دست‌آمده از اندازه‌گیری خوانایی متن هم‌راستا است.



شکل ۲. مقایسهٔ برچسب‌دهی دسته‌بند جنگل تصادفی خبرهای موثّق و جعلی کوید-۱۹

در بخش ۳-۲ درمورد نظریهٔ اطّلاعات و شگفتی توضیح داده شد. به‌نظر می‌رسد متن خبر موثّق آنتروپی بالاتری را داشته باشد؛ بنابراین، میزان اطّلاعات بیشتر را به خواننده منتقل کرده و با افزایش آنتروپی، شگفتی وی را بر می‌انگیزد. این امر به این دلیل است که اخبار موثّق باتوجّه‌به اطمینانی که از نظر صحّت مطلب در آن وجود دارد، به‌صراحت به بیان توضیحات کامل و جزئیات می‌پردازد. نتایج به‌دست‌آمده برای دو دادهٔ خبر موثّق و جعلی با استفاده از تساوی‌های (۵) و (۶) در جدول ۶ گزارش شده‌است. براساس نتایج به‌دست‌آمده خبر موثّق آنتروپی و شگفتی بالایی را نسبت به خبر جعلی به‌دست آورده‌است؛ به‌عبارتی دیگر، میزان اطّلاعات در خبر موثّق بیشتر از خبر جعلی است.

جدول ۶. مقایسهٔ میزان اطّلاعات و شگفتی در خبر جعلی و موثّق

نوع خبر	میزان آنتروپی	میزان شگفتی
جعلی	۲۶۸۴/۳۶	۵۱۵۴/۵۷
موثّق	۲۸۲۲/۰۸	۶۸۷۳/۴۰

بررسی کاربرد متغیرهای خوانایی و اطّلاعات در مدل پردازشی

در گام بعدی پژوهش، تلاش کرده‌ایم از نتایج به‌دست‌آمده خوانایی و میزان اطّلاعات و شگفتی در خبر جعلی و موثّق استفاده کرده و در ساخت یک مدل پردازشی به آن جنبهٔ کاربردی دهیم. برای این هدف باید داده‌های خبر جعلی و موثّق را به دو دستهٔ آموزش و ارزیابی تقسیم نماییم. ازآنجاکه حجم پیکرهٔ تهیه‌شده زیاد نیست، از شیوهٔ ارزیابی متقاطع ۱۰تایی[1] استفاده کردیم به این صورت که در هر دور آزمایش، ۱۰ درصد داده به‌عنوان دادهٔ ارزیابی و بقیه به‌عنوان دادهٔ آموزش استفاده می‌شود. باتوجّه‌به کم‌بودن حجم پیکرهٔ خبر جعلی فارسی مربوط به کوید-۱۹، استفاده از روش‌های یادگیری عمیق نمی‌تواند به نتایج مطلوب منجر شود. بر همین اساس، برای ساخت مدل پردازشی از سه الگوریتم یادگیری ماشینی جنگل تصادفی، رگرسیون لاجستیک و ماشین بردار پشتیبان استفاده کردیم. ازآنجاکه نیاز است دادهٔ ورودی به این دسته‌بندها به‌صورت بردار باشد، از بازنمایی معنایی ایکس‌ال‌ام-روبرتا[2] (کونیو[3] و همکاران، ۲۰۲۰) استفاده کردیم و ۵ مدل برای هریک‌از الگوریتم‌ها ساختیم که در ادامه توضیح داده می‌شود:

مدل پایه: در این مدل، از بردارهای ایکس‌ال‌ام-روبرتا (کونیو و همکاران، ۲۰۲۰) به‌عنوان مدل پایه استفاده می‌شود.

مدل ۱: در این مدل، میزان آنتروپی براساس تساوی معرفی‌شده در بخش ۳-۱ محاسبه شده و به بردار مدل پایه اضافه می‌گردد.

مدل ۲: در این مدل، میزان شگفتی براساس تساوی معرفی‌شده در بخش ۳-۱ محاسبه شده و به بردار مدل پایه اضافه می‌گردد.

مدل ۳: در این مدل، خوانایی هر متن براساس تساوی‌های معرفی‌شده در بخش ۳-۲ محاسبه شده و به بردار مدل پایه اضافه می‌گردد.

1. 10-fold cross validation
2. XML-RoBERTa
3. A. Conneau

مدل ۴: این مدل از ترکیب مدل‌های ۲ و ۳ حاصل شده و خوانایی و میزان شگفتی به بردار مدل پایه اضافه می‌گردد.

ابتدا یک مدل پایه برای مقایسه عملکرد مدل‌ها تهیه کردیم و با استفاده از آن در آزمایش‌های بعدی، ۵ مدل ساختیم و در هر مدل تلاش کردیم مدل پایه را بهبود دهیم. در جدول ۷ نتایج کارایی مدل‌ها برای سه الگوریتم دسته‌بند مورد نظر براساس دو معیار صحّت و امتیاز F گزارش شده‌است. براساس نتایج گزارش‌شده، مدل پایه با استفاده از ماشین بردار پشتیبان در دو معیار صحّت و امتیاز F پایین‌ترین کارایی را به‌دست آورده‌است. از میان روش‌های یادگیری ارزیابی‌شده، بالاترین کارایی مدل پایه توسط الگوریتم یادگیری جنگل تصادفی به‌دست آمده‌است. بنابراین، مدل‌های پیشنهادی دیگر را براساس این الگوریتم یادگیری می‌سنجیم.

استفاده از اطّلاعات آنتروپی به‌همراه مدل پایه (مدل ۱) در تمامی مدل‌ها اثر منفی داشته و سبب کاهش کارایی شده‌است؛ درحالی‌که اضافه‌شدن اطّلاعات شگفتی به مدل پایه (مدل ۲) سبب افزایش کارایی شده و به‌تنهایی اثربخش بوده‌است. استفاده از اطّلاعات خوانایی به‌همراه مدل پایه (مدل ۳) بالاترین میزان کارایی مدل‌ها را به‌دست آورده‌است. این افزایش کارایی در تمامی الگوریتم‌های یادگیری مشهود است. الگوریتم یادگیری جنگل تصادفی ۲/۵ درصد، الگوریتم یادگیری رگرسیون لاجستیک ۴/۴ درصد و الگوریتم یادگیری ماشین بردار پشتیبان ۴/۷ درصد افزایش در معیار F را در مقایسه با مدل پایه به‌دست آورده‌است. در این مدل، بالاترین کارایی به الگوریتم یادگیری جنگل تصادفی تعلق داشت. اگرچه اطّلاعاتی مانند شگفتی و خوانایی با مدل پایه به‌طور جداگانه سبب بهبود مدل در نتایج شد، اضافه‌شدن این اطّلاعات خوانایی و شگفتی به مدل پایه (مدل ۴)، در مقایسه با مدل پایه تأثیرچندانی نداشته و در مقایسه با مدل ۳ نه‌تنها به افزایش کارایی مدل منجر نشده‌است، بلکه به کاهش کارایی مدل انجامیده‌است. این کاهش بیانگر این نکته است که استفاده از ویژگی‌های بیشتر در ساخت مدل الزاماً به نتیجه بهتر نمی‌انجامد. ازاین‌رو، اهمیّت انتخاب ویژگی به‌صورت بهینه در یک مدل پردازشی دوچندان می‌گردد.

پیکرهٔ خبر جعلی را از جنبهٔ دیگری بررسی کردیم و آن بررسی عملکرد الگوریتم‌های یادگیری ماشینی داده‌هایی است که نتیجهٔ مناسبی را به‌دست نیاورده‌است. در این بررسی، داده‌هایی مد نظر است که هیچ‌کدام از الگوریتم‌ها در هیچ‌کدام از مدل‌ها نتوانسته‌است برچسب صحیح را به هر داده تخصیص دهد. با بررسی خروجی الگوریتم‌ها مشخص شد که این حالت در ۵۳ خبر موثّق و ۶۹ خبر جعلی اتفاق افتاده‌است به این معنی که ۵۳ خبر موثّق در پیکره توسط هیچ‌یک از الگوریتم‌ها برچسب صحیح (موثّق) دریافت نکرده و همه به‌عنوان خبر جعلی تشخیص داده شده‌است؛ همچنین ۶۹ خبر جعلی در پیکره توسط هیچ‌یک از الگوریتم‌ها برچسب صحیح (جعلی) دریافت نکرده و همه به‌عنوان خبر موثّق تشخیص داده شده‌است. این شرایط به‌ترتیب ۱۸/۷۹ درصد از خبرهای موثّق و ۲۴/۴۷ درصد از خبرهای جعلی را شامل می‌شود.

جدول ۷. عملکرد الگوریتم‌های یادگیری ماشینی با مدل‌های مختلف

الگوریتم	معیار ارزیابی	مدل پایه	مدل ۱	مدل ۲	مدل ۳	مدل ۴
ماشین بردار پشتیبان	صحّت	۵۰/۱	۵۰/۰	۵۱/۶	**۵۴/۸**	۵۳/۴
	امتیاز F	۴۹/۵	۴۹/۳	۵۰/۴	**۵۴/۳**	۵۲/۶
رگرسیون لاجستیک	صحّت	۵۲/۳	۵۴/۸	۵۴/۱	**۵۶/۵**	۵۶/۰
	امتیاز F	۵۱/۹	۵۴/۵	۵۳/۳	**۵۶/۳**	۵۵/۵
جنگل تصادفی	صحّت	۵۴/۳	۵۳/۰	۵۵/۹	**۵۷/۱**	۵۴/۶
	امتیاز F	۵۴/۱	۵۲/۷	۵۵/۶	**۵۶/۶**	۵۴/۲

یکی‌از علل عدم توانایی ماشین در تشخیص خبر موثّق یا جعلی به‌دلیل اشتراک نسبتاً زیاد بین الگوهای زبانی به‌کاررفته در خبرهای موثّق و جعلی است که مثال‌های این الگوها در بخش ۵-۱ آورده شده و توضیح داده شده‌است. الگوی زبانی مشترک میان این دو دستهٔ اخبار را استخراج کردیم که تعداد ۱۳۱ الگوی زبانی بود. ازمیان این الگوهای زبانی، عبارت‌هایی مانند «ابتلا به کرونا در»، «اثر ابتلا به کرونا»، «بر اثر ابتلا به»، «به کرونا در بیمارستان»، «کرونا در بیمارستان بستری»، «جان خود را از»، «خود را از دست»، «را از دست دادند»، «آمپول تقلبی کرونا در»، «انتظامی غرب استان تهران»، «تقلبی کرونا در شهریار»، «مقابله با ویروس کرونا»، «فروشندگان آمپول تقلبی کرونا»، «ملی مقابله با کرونا» در هر دو دستهٔ خبر موثّق و جعلی مشترک بود. با نگاه محتوایی به این عبارات می‌توان دریافت که اطّلاع‌رسانی درمورد مرگ‌ومیر ناشی از کوید-۱۹، بستری در بیمارستان و اتفاقات و حواشی مرتبط با کوید-۱۹ ازجمله

موضوع‌های مشترک در خبرهای موثّق و جعلی بوده‌است که سبب اثرگذاری منفی بر عملکرد الگوریتم‌های برچسب‌دهی در مدل‌های مختلف شده‌است.

بحث و نتیجه‌گیری

در این مقاله به تحلیل اخبار جعلی در مقایسه با اخبار موثّق مربوط به کوید-۱۹ با استفاده از اطّلاعات آماری نظریهٔ اطّلاعات و خوانایی پرداخته شد. برای این پژوهش، به یک پیکرهٔ زبانی نیاز بود که جعلی یا موثّق بودن خبرها در آن برچسب‌گذاری شده باشد. بر این اساس یک دستورالعمل تهیه شد تا کار استخراج و برچسب‌گذاری داده‌ها انجام پذیرد. ابتدا داده‌ها را از نظر الگوی زبانی بررسی کردیم و سپس درچارچوب نظریه اطّلاعات و خوانایی مورد مطالعه قرار دادیم. براساس نتایج به‌دست‌آمده، اخبار موثّق اطّلاعات بیشتری نسبت به خبرهای جعلی دارد و همچنین شگفتی در اخبار موثّق بیشتر از اخبار جعلی است.

در تحلیل آماری خوانایی نیاز بود که تحلیل‌های زبان‌شناسی به پیکرهٔ تهیه‌شده اضافه شود تا قابلیت برچسب‌گذاری سطح خوانایی داده‌ها به‌صورت الگوریتمی فراهم آید. با استخراج بسامد سطوح خوانایی در دو دستهٔ خبر جعلی و موثّق مشخص شد که اخبار جعلی عمدتاً از نظر خوانایی ساده بوده و اخبار موثّق نسبت به اخبار جعلی دشوارتر است. به نظر می‌رسد سادگی در متن اخبار جعلی موجب افزایش باورپذیری این دسته از خبرها در سطح جامعه می‌شود و می‌تواند در شرایط اینفودمی اثرات مخرب خود را بر جای بگذارد.

برای این‌که بتوانیم اثربخشی این دو دستهٔ معیار آماری را در تشخیص خودکار خبر جعلی بسنجیم، از سه الگوریتم متداول یادگیری ماشین استفاده کردیم. برای یافتن مدل پایهٔ مناسب، ابتدا از روش یادگیری با بازنمایی معنایی ایکس‌ال‌ام-روبرتا استفاده کردیم. سپس مدل‌های مختلف را از ترکیب دو دسته معیارهای آماری با مدل پایه به‌دست آوردیم. بررسی عملکرد مدل‌ها حاکی از اثربخشی شگفتی و خوانایی در تشخیص خبر جعلی بود. انتخاب ویژگی در ساخت مدل از اهمیّت زیادی برخوردار است که به مهندسی ویژگی معروف است. برای به‌دست‌آوردن بهترین نتیجه باید ویژگی‌های مختلف مورد توجّه قرار گیرد و اثربخشی آنها به‌صورت عملی ارزیابی شود. نتایج گزارش‌شده جهت تشخیص خبر جعلی بر روی داده‌های جمع‌آوری‌شده مربوط به کوید-۱۹ در فارسی بیانگر این نکته است که سطح دشواری متن در تشخیص خبر جعلی، چه از منظر نظری با رویکرد زبان‌شناسی و چه کاربردی با رویکرد زبان‌شناسی رایانشی، از اهمیّت به‌سزایی برخوردار است.

تقدیر و تشکر

این پژوهش در چارچوب قرارداد شمارهٔ ۹۹۰۰۹۲۰۴ و با حمایت مالی صندوق حمایت از پژوهشگران و نوآوران کشور انجام پذیرفته‌است.

منابع

جهانبخش نقده زلیخا، فیضی درخشی محمدرضا، شریفی آرش. (۱۴۰۰) ارائه مدلی برای تشخیص شایعات فارسی مبتنی بر تحلیل ویژگی‌های محتوایی در متن شبکه‌های اجتماعی .پردازش علائم و داده‌ها. ۱۸(۱): ۲۹-۵۰.

دیانی، محدحسین. (۱۳۶۶) سه تساوی برای تشخیص سطح خوانایی نوشته‌های ویژه نوسوادان، *روانشناسی و علوم تربیتی*، ۳۹(۱): ۵۹-۸۰.

دیانی، محمدحسین. (۱۳۶۹) معیاری برای تعیین سطح خوانایی نوشته‌های فارسی، *مجله علوم اجتماعی و انسانی*، ۵(۲): ۳۵-۴۸.

قیومی، مسعود. (۱۴۰۰) تحلیل محتوایی موضوع‌ها و هشتگ‌های کرونا در رسانه‌های اجتماعی، *علم زبان*، ویژه‌نامه کرونا، ۸(۱۴): ۸۷-۱۱۵.

Ahmed, H., Traore, I., & Saad, S. (2017). Detection of online fake news using n-gram analysis and machine learning techniques. In *Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments: First International Conference, ISDDC 2017, Vancouver, BC, Canada, October 26-28, 2017, Proceedings 1* (pp. 127-138). Springer International Publishing.

Allport, G. W., & Postman, L. (1947). The psychology of rumor.

Beißwenger, M., & Storrer, A. (2008). 21. corpora of computer-mediated communication. *Corpus Linguistics. An International Handbook. Series:*

Handbücher zur Sprach-und Kommunikationswissenschaft/Handbooks of Linguistics and Communication Science. Mouton de Gruyter, Berlin.

Bovet, A., & Makse, H. A. (2019). Influence of fake news in Twitter during the 2016 US presidential election. *Nature communications*, *10*(1), 7.

Butler, C. S., & Simon-Vandenbergen, A. M. (2021). Social and physical distance/distancing: A corpus-based analysis of recent changes in usage. *Corpus Pragmatics*, *5*(4), 427-462.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., ... & Stoyanov, V. (2019). Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Conroy, N. K., Rubin, V. L., & Chen, Y. (2015). Automatic deception detection: Methods for finding fake news. *Proceedings of the association for information science and technology*, *52*(1), 1-4.

Dale, E., & Chall, J. S. (1948). A formula for predicting readability: Instructions. *Educational research bulletin*, 37-54.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dubay, W. H. (2004). The principles of readability. Impact Information. *Costa Mesa, CA*.

Flesch, R. (1979). *How to write plain English: A book for lawyers and consumer*. Harper & Row.

Ghayoomi, M. (2022). Application of computational linguistics to predicting language proficiency level of Persian learners' textbooks. *Journal of Language Horizons*, *6*(1), 29.

Goldani, M. H., Momtazi, S., & Safabakhsh, R. (2021). Detecting fake news with capsule neural networks. *Applied Soft Computing*, *101*, 106991.

Gunning, R. (1952). *The technique of clear writing*, New York: McGraw-Hill.

Hosseini, P., Hosseini, P., & Broniatowski, D. (2020). *Content analysis of Persian/Farsi Tweets during COVID-19 pandemic in Iran using NLP*. Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020, Association for Computational Linguistics.

Jahanbakhsh-Nagadeh, Z., Feizi-Derakhshi, M. R., Ramezani, M., Akan, T., Asgari-Chenaghlu, M., Nikzad-Khasmakhi, N., ... & Balafar, M. A. (2023). A model to measure the spread power of rumors. *Journal of Ambient Intelligence and Humanized Computing*, *14*(10), 13787-13811.

Jin, Z., Cao, J., Zhang, Y., Zhou, J., & Tian, Q. (2016). Novel visual and statistical image features for microblogs news verification. *IEEE transactions on multimedia*, *19*(3), 598-608.

Jwa, H., Oh, D., Park, K., Kang, J. M., & Lim, H. (2019). exbake: Automatic fake news detection model based on bidirectional encoder representations from transformers (bert). *Applied Sciences*, *9*(19), 4062.

Kaliyar, R. K., Goswami, A., Narang, P., & Sinha, S. (2020). FNDNet–a deep convolutional neural network for fake news detection. *Cognitive Systems Research*, *61*, 32-44.

Khattar, D., Goud, J. S., Gupta, M., & Varma, V. (2019). MVAE: Multimodal variational autoencoder for fake news detection. *Proceedinsg of the World Wide Web Conference*, 2915–2921.

Kincaid, J. P., Fishburne Jr, R. P., Rogers, R. L., & Chissom, B. S. (1975). Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel.

Lively, B. A., & Pressey, S. L. (1923). A method for measuring the" vocabulary burden" of textbooks: Educational administration and supervision,". *A method for measuring the" vocabulary burden" of textbooks: Educational Administration and Supervision*.

Liu, C., Wu, X., Yu, M., Li, G., Jiang, J., Huang, W., & Lu, X. (2019). A two-stage model based on BERT for short fake

news detection. In *Knowledge Science, Engineering and Management: 12th International Conference, KSEM 2019, Athens, Greece, August 28–30, 2019, Proceedings, Part II 12* (pp. 172-183). Springer International Publishing.

Lugea, J. (2021). Linguistic approaches to fake news detection. *Data science for fake news: Surveys and perspectives*, 287-302.

Mahmoodabad, S. D., Farzi, S., & Bakhtiarvand, D. B. (2018, December). Persian rumor detection on twitter. In *2018 9th international symposium on telecommunications (IST)* (pp. 597-602). IEEE.

Mahmoudi-Dehaki, M., Chalak, A., & Heidari Tabrizi, H. (2020). The COVID-19 lingo: societies' responses in form of developing a comprehensive covidipedia of English vs. Persian neologisms (coroneologisms). *Journal of English Language Pedagogy and Practice*, *13*(27), 26-52.

Potthast, M., Kiesel, J., Reinartz, K., Bevendorff, J., & Stein, B. (2017). A stylometric inquiry into hyperpartisan and fake news. *arXiv preprint arXiv:1702.05638*.

Ramezani, M., Rafiei, M., Omranpour, S., & Rabiee, H. R. (2019, August). News labeling as early as possible: Real or Fake?. In *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (pp. 536-537).

Rubin, V. L., Chen, Y., & Conroy, N. K. (2015). Deception detection for news: three types of fakes. *Proceedings of the Association for Information Science and Technology*, *52*(1), 1-4.

Samadi, M., Mousavian, M., & Momtazi, S. (2021). Persian fake news detection: Neural representation and classification at word and text levels. *Transactions on Asian and Low-Resource Language Information Processing*, *21*(1), 1-11.

Seifikar, M., Farzi, S., & Mahmoodabad, S. D. (2018, December). Kermanshah earthquake event tracking through Persian tweets. In *2018 9th International Symposium on Telecommunications (IST)* (pp. 424-428). IEEE.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal*, *27*(3), 379-423.

Sherman, L. A. (1893). *Analytics of literature: A manual for the objective study of English prose and poetry*. Ginn.

Smith, E. A., & Senter, R. J. (1967). Air Force Aerospace Medical Research Laboratory (US): Automated Readability Index. *AMRL-TR. Aerospace Medical Research Laboratories*.

Tan, K. H. (2020). Fear'In Covid-19 Fake News: A Corpus-Based Approach 3l: *The Southeast Asian Journal of English Language Studies*, 26 (2): 1 23).

Tribus, M. (1961). Thermostatics and thermodynamics: an introduction to energy, information and states of matter, with engineering applications. *(No Title)*.

Vargo, C. J., Guo, L., & Amazeen, M. A. (2018). The agenda-setting power of fake news: A big data analysis of the online media landscape from 2014 to 2016. *New media & society*, *20*(5), 2028-2049.

Vogel, I., & Jiang, P. (2019, August). Fake news detection with the new German dataset "GermanFakeNC". In *International Conference on Theory and Practice of Digital Libraries* (pp. 288-295). Cham: Springer International Publishing.

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *science*, *359*(6380), 1146-1151.

Weisser, M. (2016). *Practical corpus linguistics: An introduction to corpus-based language analysis* (Vol. 43). John Wiley & Sons.

Yang, S., Shu, K., Wang, S., Gu, R., Wu, F., & Liu, H. (2019, July). Unsupervised fake news detection on social media: A generative approach. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 33, No. 01, pp. 5644-5651).

Zamani, S., Asadpour, M., & Moazzami, D. (2017, May). Rumor detection for persian tweets. In *2017 Iranian conference on electrical engineering (ICEE)* (pp. 1532-1536). IEEE.

Zhang, J., Dong, B., & Philip, S. Y. (2020, April). Fakedetector: Effective fake news detection with deep diffusive neural network. In *2020 IEEE 36th international conference on data engineering (ICDE)* (pp. 1826-1829). IEEE.