



Semnan University

Journal of Modeling in Engineering

Journal homepage: <https://modelling.semnan.ac.ir/>

ISSN: 2783-2538



Type of Article

The Role of Social Networks in Predicting the Coronavirus Infection Rate

Fahimeh Hafezi ^a, Maryam Khodabakhsh ^{a,*}

^a Faculty of Computer Engineering, Shahrood University of Technology, Shahrood, Iran

PAPER INFO

Paper history:

Received: 2024-07-08

Revised: 2025-06-09

Accepted: 2025-09-23

Keywords:

Coronavirus,
Social media,
Incidence rate estimation,
Synergy,
Sentiment analysis.

ABSTRACT

Since the coronavirus was recognized as a pandemic infectious disease in 2019, most people were forced to stay at home during the pandemic. Given that social networks are a popular medium among people and during the pandemic, the analysis of user-generated social content can provide new insights and be effective for tracking the occurrence of the pandemic over time. This study aimed to provide a model to predict the incidence rate of COVID-19 in the first wave of the pandemic in Iran, through the analysis of Persian Instagram posts. Using the synergetic technique, three features of semantic similarity, fear feeling, and hope feeling were extracted from Instagram posts. For this purpose, word embedding techniques (Word2Vec, Glove, FastText) were used to calculate semantic similarity, and a BERT-based classifier model was used to identify fear and hope feelings. To improve performance, the SBERT model was also used instead of classical embedding methods. Then, a support vector regression (SVR) model was trained using statistical indices based on these features to predict the daily incidence rate of COVID-19. The results showed that the synergy of semantic similarity and fear sentiment features using SBERT in the SVM model provided the highest performance with a coefficient of determination (R^2) of 0.52, which showed a significant improvement over the baseline methods. These findings indicate that the automatic combination of semantic and sentiment features can be an effective indicator for monitoring epidemics through social networks.

DOI: <https://doi.org/10.22075/jme.2025.34695.2701>

© 2026 Published by Semnan University Press.

This is an open access article under the CC-BY 4.0 license. (<https://creativecommons.org/licenses/by/4.0/>)

* Corresponding author.

E-mail address: m_khodabakhsh@shahroodut.ac.ir

How to cite this article:

F hafezi and M khodabakhsh, "The Role of Social Networks in Predicting the Coronavirus Infection Rate," Journal of Modeling in Engineering, 24 85 (2026): 187-200, doi: 10.22075/jme.2025.34695.2701

نقش شبکه‌های اجتماعی در پیش‌بینی نرخ ابتلا به کروناویروس

فهیمه حافظی^۱، مریم خدابخش^{*}

چکیده	اطلاعات مقاله
از آنجایی که کروناویروس به‌عنوان یک بیماری عفونی همه‌گیر در سال ۲۰۱۹ شناخته شد بیشتر مردم در طول این همه‌گیری مجبور به ماندن در خانه‌های خود شدند. با توجه به این که شبکه‌های اجتماعی به‌عنوان یک رسانه محبوب در بین مردم و در زمان همه‌گیری شناخته شده است، بنابراین تحلیل محتوای اجتماعی تولید شده توسط کاربران می‌تواند بینش‌های جدیدی را ارائه کند و برای ردیابی وقوع همه‌گیری در طول زمان موثر باشند. این پژوهش با هدف ارائه مدلی برای پیش‌بینی نرخ بروز کووید-۱۹ در موج اول همه‌گیری در ایران، از طریق تحلیل پست‌های اینستاگرامی فارسی انجام شد. با استفاده از تکنیک هم‌افزایی، سه ویژگی شباهت معنایی، احساس ترس و احساس امید از پست‌های اینستاگرامی استخراج شد. برای این منظور، تکنیک‌های تعبیه کلمه (Word2Vec, Glove, FastText) برای محاسبه شباهت معنایی و مدل طبقه‌بند مبتنی بر BERT برای شناسایی احساسات ترس و امید به کار گرفته شد. همچنین برای بهبود عملکرد، از مدل SBERT به‌جای روش‌های کلاسیک تعبیه نیز استفاده شد. سپس مدل رگرسیون بردار پشتیبان (SVR) با استفاده از شاخص‌های آماری مبتنی بر این ویژگی‌ها برای پیش‌بینی نرخ بروز روزانه کووید-۱۹ آموزش داده شدند. نتایج نشان داد که هم‌افزایی ویژگی‌های شباهت معنایی و احساس ترس با استفاده از SBERT در مدل SVM بالاترین عملکرد را با ضریب تعیین (R^2) ۰.۵۲ ارائه می‌دهد، که نسبت به روش‌های پایه بهبود قابل توجهی نشان می‌دهد. این یافته‌ها نشان می‌دهند که ترکیب خودکار ویژگی‌های معنایی و احساسی می‌تواند شاخصی مؤثر برای پایش بیماری‌های همه‌گیر از طریق شبکه‌های اجتماعی باشد.	دریافت مقاله: ۱۴۰۳/۰۴/۱۸ بازنگری مقاله: ۱۴۰۴/۰۳/۱۹ پذیرش مقاله: ۱۴۰۴/۰۷/۰۱
	واژگان کلیدی: کروناویروس، شبکه اجتماعی، پیش‌بینی نرخ بروز کرونا، هم‌افزایی، تحلیل احساسات.
DOI: https://doi.org/10.22075/jme.2025.34695.2701	
© 2026 Published by Semnan University Press. This is an open access article under the CC-BY 4.0 license. (https://creativecommons.org/licenses/by/4.0/)	

۱- مقدمه

در سال ۲۰۱۹، یک بیماری تنفسی ناشی از Sars-COV-2 در ووهان، چین [۱] گزارش شد و سازمان بهداشت جهانی (WHO) آن را کروناویروس (کووید-۱۹) نام‌گذاری کرد. کووید-۱۹ یک نگرانی حیاتی برای سلامت عمومی است. ارائه‌دهندگان مراقبت‌های بهداشتی باید در مورد بهداشت عمومی و شیوع کووید-۱۹ که جوامع خود را

تحت تاثیر قرار می‌دهد به‌روز شوند تا به سرعت تصمیم‌های صحیحی بگیرند. این امر آن‌ها را قادر می‌سازد تا فعالیت‌های بهبود یافته را به شیوه‌ای کارآمد ارائه دهند که در شرایط فعلی بسیار مهم است. از زمان شروع همه‌گیری، اکثر دولت‌ها فعالیت‌های سریعی را انجام دادند تا از گسترش کووید-۱۹ جلوگیری کنند. وضعیت قرنطینه در بسیاری از کشورهای جهان اعمال شد، مسافرت‌ها محدود شد، انزوای

* پست الکترونیک نویسنده مسئول: m_khodabakhsh@shahroodut.ac.ir

۱. دانشکده مهندسی کامپیوتر، دانشگاه صنعتی شاهرود، سمنان، ایران

این کلمات کلیدی مرتبط با کرونا را نداشته باشد، سیستم آن را نادیده می‌گیرد. همچنین هیچ تضمینی وجود ندارد که هر محتوا از جمله با کلمه‌های کلیدی تعیین شده به کرونا مرتبط باشد.

به همین دلیل محققان [۹] یک مدل پیش‌بینی برای نرخ بروز کرونا با استفاده از پست‌های مرتبط با کرونا در شبکه‌های اجتماعی پیشنهاد کردند. هدف آن‌ها مشخص کردن این بود که آیا پست‌های منتشر شده در طول شیوع کرونا می‌تواند شاخصی برای تخمین نرخ ابتلا به کرونا باشد یا خیر. از این جهت یک رویکرد معنایی مبتنی بر تکنیک‌های تعبیه کلمه را برای مدل‌سازی کرونا پیشنهاد کردند و سپس یک ویژگی جدید به نام شباهت معنایی را برای اندازه‌گیری شباهت یک پست داده شده به کرونا در فضای معنایی معرفی کردند. علاوه بر شباهت معنایی، دو ویژگی دیگر یعنی احساس ترس و احساس امید را برای شناسایی پست‌های مرتبط با کرونا پیشنهاد کردند. این ویژگی‌ها به عنوان شاخص‌های آماری در یک مدل رگرسیونی برای تخمین موارد جدید مبتلایان استفاده می‌شوند. آن‌ها نشان دادند که در نظر گرفتن ویژگی‌های پیشنهادی منجر به بهبود عملکرد تخمین نرخ بروز به کرونا می‌شود.

از این جهت هدف اصلی ما در این مقاله، مطابق با کار محققان [۹] پیشنهاد یک مدل پیش‌بینی برای نرخ بروز کرونا با استفاده از پست‌های مرتبط با کرونا در شبکه‌های اجتماعی است. تفاوت کار ما با [۹] استفاده از رویکرد هم‌افزایی ویژگی‌های پیشنهادی است. بنابراین ما مشخص می‌کنیم که آیا ویژگی‌های پیشنهادی در این پژوهش رفتار مکمل و هم‌افزایی بر یکدیگر دارند یا خیر. فرضیه ما بر این اساس است که ویژگی‌های احساس ترس یا احساس امید می‌توانند به عنوان شاخص‌های قوی در هنگام شناسایی پست‌های مرتبط با کرونا در نظر گرفته شوند. به همین دلیل، پیشنهاد هم‌افزایی دو ویژگی احساس ترس و احساس امید با ویژگی شباهت معنایی ممکن است بتواند منجر به بهبود عملکرد شود. برای انجام این کار، سه گونه از هم‌افزایی‌ها را معرفی می‌کنیم که به آن‌ها (۱) شباهت معنایی + احساس ترس، (۲) شباهت معنایی + احساس امید، (۳) شباهت معنایی + احساس ترس + احساس امید اشاره می‌شود. ما ویژگی‌های پیشنهادی را بر روی یک مجموعه داده فارسی از اینستاگرام که در طول موج اول

اجتماعی وجود داشت که در آن مردم در خانه می‌مانند و امکان کار از خانه برای بسیاری از کارمندان وجود داشت [۲]. در طول یک بیماری همه‌گیر با نرخ عفونت بالا، فعالیت‌های کاهش‌دهنده سریع نقش مهمی در کاهش سرعت شیوع و جلوگیری از کانون‌های جدید بیماری ایفا می‌کند [۳].

در این پژوهش، ما به طور خاص علاقه‌مند به ایجاد مدلی برای پیش‌بینی موج کرونا با استفاده از شبکه‌های اجتماعی هستیم. پلتفرم‌های رسانه‌های اجتماعی یکی از پرکاربردترین منابع اطلاع‌رسانی در جهان هستند، دسترسی آسان و ارزان به اینترنت و تعداد زیاد کاربران ثبت‌شده در این پلتفرم‌ها، آن‌ها را به یکی از ساده‌ترین و مؤثرترین راه‌های انتشار داده‌ها تبدیل کرده است. در طول رویدادهای بزرگ مانند مسابقات ورزشی، شیوع بیماری‌ها یا بلایای طبیعی، هدف اصلی جمع‌آوری اطلاعات بیشتر است. رسانه‌ها تلاش می‌کنند تا اطلاعات بیشتری به دست آورند و آن‌ها را گزارش دهند تا مردم بهتر در جریان وقایع قرار بگیرند [۴].

رسانه‌های اجتماعی از طریق فعالیت‌های متعددی از جمله انتقال مداوم بروزرسانی‌های محلی و جهانی، و همچنین صدور هشدارها و دستورالعمل‌هایی برای مقابله با همه‌گیری و پیامدهای آن، نقش مهمی در انتشار داده‌ها در مورد بیماری‌ها از جمله کووید-۱۹ و اثرگذاری آن ایفا می‌کنند [۵]. این دانش ممکن است به پیش‌بینی عفونت‌های کووید-۱۹ کمک کند، که امکان پیش‌بینی هدفمندتر جمعیت‌های در معرض خطر را فراهم می‌کند.

کارهای اخیر در زمینه پیش‌بینی شیوع کرونا سعی در ردیابی حجم محتوای مرتبط با کرونا و مرتبط کردن آن با موارد رسمی مبتلایان یا مرگ و میر دارد. روش اولیه برای بازیابی محتوای مرتبط با کرونا در میان تمام محتوایی که در شبکه‌های اجتماعی به اشتراک گذاشته می‌شود، بر جستجوی کلمه کلیدی متکی است [۳، ۶-۸]. به عنوان مثال، غروی و همکاران [۳] برای تشخیص علائم هشداردهنده Covid-19 پیام‌های تویتر را براساس دو کلمه "سرفه" و "تب" فیلتر کردند. اگرچه این روش عملکرد چشمگیری از خود نشان داده است، اما از مشکل شکاف واژگانی رنج می‌برد، به این معنی است که روش مورد نظر تنها محتوایی را تشخیص می‌دهد که شامل کلمات کلیدی مشخصی باشد. بنابراین، اگر محتوایی درباره کرونا باشد ولی

برای درک بیماری و انتشار داده‌ها در طول شیوع بهره بردند. در مطالعه‌ای [۱۵]، چارچوبی برای پیش‌بینی شیوع دنگی در سطح شهری در برزیل ارائه شده است. توییت‌ها را با کلمه‌کلیدی دنگی برای تشخیص زودهنگام و مدل پایش دنگی در برزیل جمع‌آوری می‌کنند. آن‌ها توییت‌هایی را شناسایی کردند که حاکی از تجربه شخصی با دنگی بود و توییت‌های تقلید، دیدگاه‌ها، داده‌ها و بازاریابی توسط الگوریتم‌های یادگیری ماشین را حذف کردند. آن‌ها دریافتند که این توییت‌ها ارتباط قوی با موارد گزارش شده در سطح ملی دارد. لی و همکاران [۱۶] داده‌های روند را از گوگل ترند، فهرست جستجو بایده و سیناوبیو با جستجوی "coronavirus" و "pneumonia" برای مطالعه پیش‌بینی اولیه موارد مبتلا به کرونا از طریق جستجوهای اینترنتی و داده‌های شبکه‌های اجتماعی بازاریابی کردند. آن‌ها داده‌ها را با استفاده از همبستگی تاخیر، تجزیه و تحلیل کردند و دریافتند که حداکثر همبستگی بین داده‌های روند با کلمه‌کلیدی کرونا و تعداد تشخیص‌ها در ۸ تا ۱۲ روز قبل از افزایش موارد تایید شده کروناویروس در سه پلتفرم وجود دارد.

اولری [۱۷] از رگرسیون برای مدل‌سازی تعداد موارد جدید مبتلایان و مرگ‌ومیر کروناویروس توسط سه متغیر جستجوی گوگل ترند، بازدید از صفحه ویکی‌پدیا و تعداد توییت‌ها برای کروناویروس و COVID-19 استفاده کرد. تحلیلی توسط یوسفینقانی و همکاران [۱۲] با استفاده از فیلتر کردن داده‌های توییت و جستجوهای آنلاین گوگل برای تشخیص امواج در تعداد موارد تایید شده در ایالات متحده و کانادا انجام شد. در مطالعه دیگری [۳]، با تحلیل فراوانی کلمات کلیدی مانند "سرفه" و "تب" در توییت، مشخص شد شیوع واقعی COVID-19 می‌تواند ۵ تا ۱۹ روز زودتر از گزارش‌های رسمی باشد. این پژوهش نشان داد که تجزیه و تحلیل توییت‌های مربوط به علائم رایج ارتباط قوی با موارد گزارش شده توسط CDC دارد. با وجود تاخیر در آزمایش و تشخیص، نویسندگان نتیجه گرفتند که داده‌های توییت می‌تواند به عنوان سیستم هشدار اولیه برای ارزیابی سریع شیوع بیماری‌ها و اطلاع‌رسانی تصمیمات سیاستی عمل کند.

در مطالعه [۶] برای پیش‌بینی میزان بروز کروناویروس در ایران از داده‌های جمع‌آوری شده از وب‌سایت گوگل ترند استفاده شده است. محققان از رگرسیون خطی و حافظه

کرونا در ایران جمع‌آوری می‌شود، ارزیابی می‌کنیم، نتایج خود را از دیدگاه‌های مختلف بررسی می‌کنیم و قسمت‌هایی را که ویژگی‌ها می‌توانند به طور قابل توجهی عملکرد را افزایش دهند، نشان می‌دهیم.

۲- کارهای مرتبط

پژوهش‌ها نشان می‌دهد که رسانه‌های اجتماعی ممکن است ابزارهای ارزشمندی برای نظارت بر بیماری‌ها باشند که در جهت بهبود توانایی دولت‌ها و متخصصان بهداشت عمومی در تشخیص سریع‌تر شیوع بیماری نسبت به روش‌های سنتی و افزایش واکنش به شیوع آن استفاده می‌شوند. به عنوان مثال، استفاده از شبکه‌های اجتماعی در آموزش بهداشت عمومی به دلیل توانایی آن در حذف موانع فیزیکی که به طور سنتی دسترسی به پشتیبانی و منابع مراقبت‌های بهداشتی را مختل می‌کند، افزایش یافته است [۱۰]. در طول رویداد H1N1 در سال ۲۰۰۹، مراکز کنترل و پیشگیری از بیماری‌های ایالات متحده (CDC) به صفحه فیس‌بوک خود مراجعه کردند تا به مردم در مورد این بیماری و اهمیت واکسیناسیون آموزش دهند. ده‌ها هزار وب‌سایت شبکه‌های اجتماعی مرتبط با ارتقای سلامت وجود دارد که در حال حاضر در دسترس عموم هستند [۱۱]. شبکه‌های اجتماعی نشان داده‌اند که می‌توانند دسترسی عمومی به داده‌های نظارت بر سلامت عمومی را با جستجوی داده‌های به موقع و قابل اعتماد در مورد شیوع یا شدت آنفولانزا، ابولا، ویروس دنگی و کرونا ویروس، سندرم تنفسی خاورمیانه بهبود بخشند [۳، ۱۲-۱۷].

امین و همکاران [۱۳] با تحلیل داده‌های توییت از سال‌های ۲۰۱۷ تا ۲۰۱۹، پژوهشگران شیوع بیماری‌هایی مانند آنفولانزا و دنگی را با استفاده از روش‌های یادگیری ماشین از جمله SVM، RF، KNN و درخت تصمیم شناسایی کردند. توییت‌ها به دسته‌های مثبت و منفی نسبت به بیماری تقسیم شدند و ویژگی‌های متنی با TF-IDF استخراج شد. نتایج نشان داد که تحلیل احساسات در رسانه‌های اجتماعی می‌تواند ابزاری مؤثر برای شناسایی زودهنگام شیوع بیماری‌ها باشد. در جریان شیوع ابولا در سال‌های ۲۰۱۴ تا ۲۰۱۶ مطالعاتی [۱۴] از توییت‌هایی که ابولا را ذکر می‌کردند برای نظارت بر انتشار داده‌ها، تشخیص زودهنگام بیماری و دانش و نگرش عمومی استفاده کردند. آن‌ها از تجزیه و تحلیل سری‌های زمانی و تجسم جغرافیایی

عمومی مرتبط با کروناویروس از اینستاگرام که در ایران منتشر شده، را داشتند.

در مطالعه [۲۲] نویسندگان به تحلیل قطبیت و طبقه‌بندی پیام‌های کاربران فارسی‌زبان در شبکه‌های اجتماعی (اینستاگرام، توئیتر، تلگرام و کانالهای خبری داخلی) در دوران کرونا پرداخته‌اند. پیام‌ها براساس قطبیت مثبت و منفی طبقه‌بندی شدند و برای این منظور از روش‌های پردازش زبان طبیعی و شبکه‌های عمیق مختلفی شامل CNN، LSTM، LSTM-Fuzzy استفاده شد. بهترین نتیجه از شبکه‌های عمیق کانولوشنی سه‌لایه به دست آمد. همچنین نویسندگان [۹]، یک رویکرد معنایی مبتنی بر تکنیک‌های تعبیه کلمه برای مدل‌سازی کرونا پیشنهاد کردند و سپس یک ویژگی جدید به نام شباهت معنایی را برای اندازه‌گیری شباهت یک پست داده‌شده به کرونا در فضای معنایی معرفی کردند. علاوه بر این، آن‌ها دو ویژگی دیگر یعنی احساس ترس و احساس امید را برای شناسایی پست‌های مرتبط با کرونا پیشنهاد کردند. این ویژگی‌ها به عنوان شاخص‌های آماری در مدل رگرسیونی برای تخمین موارد آلوده جدید استفاده می‌شود. آن‌ها ویژگی‌های خود را در مجموعه داده فارسی پست‌های اینستاگرام، که در موج اول کروناویروس جمع‌آوری شده بود، ارزیابی کردند و نشان دادند که در نظر گرفتن ویژگی‌های پیشنهادی منجر به بهبود عملکرد تخمین نرخ ابتلا به کرونا می‌شود.

۳- تعریف مسئله

پژوهش‌های اخیر [۲۳، ۲۴] نشان می‌دهد که رسانه‌های اجتماعی ممکن است ابزارهای ارزشمندی برای نظارت بر بیماری‌ها باشند که برای بهبود توانایی متخصصان بهداشت عمومی در تشخیص سریع‌تر شیوع بیماری نسبت به روش‌های سنتی و افزایش واکنش به شیوع استفاده می‌شود. مهم‌ترین هدفی که ما دنبال می‌کنیم پیشنهاد یک مدل پیش‌بینی برای نرخ بروز کرونا با استفاده از پست‌های مرتبط با کروناویروس در شبکه‌های اجتماعی است. ما معتقدیم که محتوای تولید شده توسط کاربر و مرتبط با کروناویروس می‌تواند نشانه‌ای برای ردیابی تغییرها و میزان بروز آن در طول دوره اصلی کرونا باشد. محتوای تولید شده توسط کاربر در اینستاگرام پست نامیده می‌شود و به صورت کیسه‌ای از کلمه‌ها نشان داده می‌شود. کاربر از کلمه‌هایی که به طور ضمنی یا صریح با کرونا مرتبط

کوتاه‌مدت بلند مدت (LSTM) برای تخمین تعداد موارد مثبت کروناویروس استفاده کردند و دریافتند که بهترین عوامل غیر از نرخ روز قبل، فراوانی جستجو در مباحث ضدعفونی‌کننده، ضدعفونی‌کننده دست و شستن دست‌ها را شامل می‌شود. هدف محققان در [۱۸] این بود که با استفاده از طیف وسیعی از شاخص‌ها، بررسی کنند که آیا نرخ بروز COVID-19 بین شهرستان‌ها براساس ویژگی‌های اجتماعی-اقتصادی آن‌ها متفاوت است یا خیر. آن‌ها مدل‌های تقویت گرادیان را برای پیش‌بینی نرخ‌های بروز استاندارد شده با سن، با ساختارهای کلان شهرستان‌ها آموزش دادند و از مقادیر (SHAP) برای توصیف ۲۰ ویژگی برجسته از نظر همبستگی منفی/مثبت با متغیر نتیجه استفاده کردند.

اگرچه بیشتر محققان علاقه‌مند به تجزیه و تحلیل محتوای مربوط به همه‌گیری کروناویروس در شبکه اجتماعی توئیتر، بایدو، گوگل ترند و سیناویو هستند، مطالعه‌های کمی هم در شبکه اجتماعی اینستاگرام انجام شده است [۹، ۱۹-۲۱]. محققان در [۱۹] از اینستاگرام به عنوان یک منبع قابل اعتماد برای تجزیه و تحلیل واکنش کاربران شبکه‌های اجتماعی در زمان همه‌گیری COVID-19 استفاده کردند. داده‌های آن‌ها از نظر جنبه‌های مختلف از جمله تجزیه و تحلیل احساس‌ها، تشخیص موضوع، رویکردها و ویژگی‌های جغرافیایی و زمانی مورد بررسی قرار گرفت. در نهایت آن‌ها نشان دادند که واکنش احساس‌های غالب در شبکه‌های اجتماعی خنثی است، در حالی که بیشترین بحث‌های مورد توجه توسط کاربران شبکه‌های اجتماعی مسائل بهداشتی است. همچنین، از تحلیل واژگان با Word2Vec و اندازه‌گیری مناطق تاثیرگذار برای بررسی تغییرات معنایی و تمرکز کاربران بر کشورهای مختلف استفاده شده است. جعفری‌نژاد و همکاران [۲۰] چارچوبی را برای تحلیل رفتار کاربران شبکه‌های اجتماعی در واکنش به اخبار و وضعیت مرتبط با کروناویروس پیشنهاد کردند. آن‌ها از LDA به عنوان ثابت‌ترین مدل موضوع برای آشکار کردن تغییرهای احتمالی در مباحث گفتمان استفاده کردند. پژوهش‌های آن‌ها براساس مجموعه داده‌های فارسی است که از شبکه‌های اجتماعی مختلف و خبرگزاری‌ها در طول موج اول در ایران جمع‌آوری شده است. کار دیگری بر روی داده‌های فارسی توسط نیکنام و همکاران [۲۱] انجام شده است که سعی در توصیف بازنمایی داده‌های بهداشت

علاوه بر شباهت معنایی، احساس ترس و احساس امید نیز به‌عنوان ویژگی‌های موثر در شناسایی پست‌های مرتبط با کرونا معرفی شده‌اند [۹]. پژوهش‌های اخیر [۲۳، ۳۲] نشان داده است که ترس، واکنش احساسی اصلی افراد به کرونا است. همچنین پژوهش‌های [۳۳، ۳۴] نشان داده‌اند که امید هم به‌عنوان یک احساس مثبت در مورد بحران‌هایی مانند بیماری وجود دارد. در مورد کروناویروس نیز، علاوه بر پاسخ‌های منفی، احساس امید نیز در پست‌های کاربران مشاهده شده است [۳۵، ۳۶]. امید مربوط به ساخت واکنش می‌تواند زندگی را به حالت عادی برگرداند. عبارت "ما کرونا را شکست خواهیم داد" نشان‌دهنده احساس امید در پست‌های کاربران ایرانی در مورد کرونا است.

همان‌طور که گفته شد پژوهش [۹] از ویژگی‌های شباهت معنایی، احساس ترس و احساس امید برای شناسایی پست‌های مرتبط با کرونا استفاده کردند. هدف اصلی ما در این پژوهش نیز، پیشنهاد یک مدل برای پیش‌بینی نرخ بروز کرونا با استفاده از پست‌های مرتبط با کرونا در شبکه‌های اجتماعی با کمک مدل SVM و مدل‌های کلاسیک تعبیه‌کلمه مانند Word2Vec، GloVe و FastText است. منتهی تفاوت کار ما استفاده از رویکرد هم‌افزایی ویژگی‌های شباهت معنایی همراه با احساس ترس و احساس امید همراه با مدل SVM است که در بخش شباهت معنایی علاوه بر مدل‌های کلاسیک تعبیه جملات، از SBERT نیز استفاده می‌شود. که در بخش بعدی به‌طور کامل آن‌ها را شرح خواهیم داد. بنابراین ما مشخص می‌کنیم که چگونه هر یک از این ویژگی‌ها را می‌توان در یک فرآیند پیش‌بینی ادغام کرد تا برآورد موثرتری به دست آورد. ما ویژگی‌های پیشنهادی خود را بر روی یک مجموعه داده فارسی استاندارد طلایی که به‌طور اصولی از اینستاگرام در طول موج اول در ایران جمع‌آوری می‌شود، ارزیابی می‌کنیم. همچنین یافته‌های خود را از دیدگاه‌های مختلف جهت این که آیا ویژگی‌ها باعث بهبود عملکرد می‌شوند یا خیر، تجزیه و تحلیل می‌کنیم.

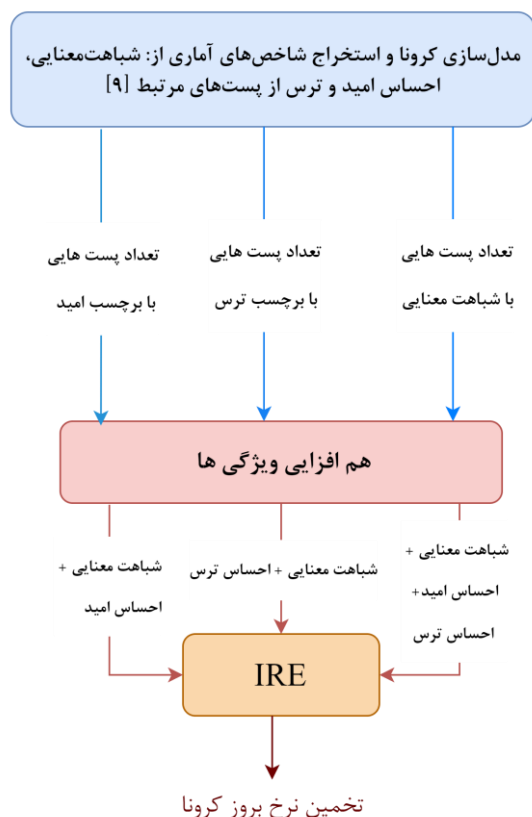
۴- تخمین نرخ بروز کرونا

نتایج پژوهش‌ها در مطالعه [۹] نشان داده است که ویژگی‌های احساس ترس یا احساس امید می‌توانند به عنوان شاخص‌های قوی در هنگام شناسایی پست‌های مرتبط با کرونا در نظر گرفته شوند. به همین دلیل، در این

هستند استفاده می‌کند تا محتوایی در مورد کرونا منتشر کند. لذا، مدل‌سازی مفهوم کرونا به‌صورت کیسه‌ای از کلمه‌های متمایز مرتبط با کرونا امری رایج است [۳، ۲۴]. به همین منظور، کوین و همکاران [۲۴] مجموعه‌ای از کلمه‌های متمایز را براساس علائم کووید-۱۹ از جمله "سرفه خشک"، "تب"، "ناراحتی قفسه سینه"، "کرونا" و "ذات‌الریه" ساختند. همچنین، محققان در [۳] از کلمه‌های کلیدی "تب" و "سرفه" که از شایع‌ترین علائم COVID-19 بود، برای ساخت این مجموعه استفاده کردند. در این پژوهش مفهوم کرونا، به صورت کیسه‌ای از کلمه‌های متمایز مرتبط با کرونا مدل می‌شود و برای بدست آوردن این مجموعه متمایز از کلمه‌ها، از روش معرفی شده در [۹] استفاده شده است.

همان‌طور که در [۲۵، ۲۶] بحث شد، نمایش محتوای پست و کرونا به عنوان کیسه‌ای از کلمه‌ها یک رویکرد متداول است اما مشکل عدم تطابق واژگان همچنان وجود دارد، یعنی با این روش نمی‌توان پست‌هایی را بازیابی کرد که هیچ کلمه‌ای درباره کرونا ندارند در حالی که محتوای آن‌ها با کرونا مرتبط است. بنابراین در این پژوهش از ویژگی شباهت معنایی [۹] برای شناسایی پست‌های مرتبط با کرونا استفاده می‌شود. بدین منظور، از تکنیک‌های معروف تعبیه کلمه‌ها مانند Word2vec [۲۷]، FastText [۲۸] و GloVe [۲۹] استفاده می‌شود تا بتوان پستی را که از نظر اصطلاحی با کلمه‌های کلیدی تعیین شده همپوشانی ندارد، همچنان بازیابی کرد. همچنین، برای غلبه بر محدودیت‌های مدل‌های کلاسیک تعبیه کلمه مانند Word2Vec، GloVe و FastText، از مدل Sentence-BERT (SBERT) [۳۰] استفاده شد. SBERT، که بر پایه معماری ترنسفورمر BERT توسعه یافته است، به‌طور خاص برای تولید تعبیه‌های معنایی در سطح جمله بهینه‌سازی شده و قادر است شباهت معنایی بین پست‌ها و مفهوم کووید-۱۹ را با دقت بیشتری محاسبه کند. همان‌طور که دونگ و همکاران [۳۱] با استفاده از BERT و عمدتاً تجزیه و تحلیل توییت‌ها، رویدادهای نامطلوب دارویی را در داده‌های رسانه‌های اجتماعی یافتند. آن‌ها بازنمایی‌های معنایی توییت‌ها را تولید کردند و شباهت آن‌ها را با کلمات کلیدی در مورد رویدادهای نامطلوب دارویی یافتند. مشخص شد که در توییت‌ها، پیشرفت‌هایی در یافتن محتوای مرتبط با سلامت حاصل شده است.

پیشنهادی برای تخمین میزان بروز کرونا می‌باشد. در این بخش ابتدا مجموعه داده‌ای که برای انجام آزمایش‌ها از آن استفاده شده توصیف می‌شود و سپس به تحلیل نتایج پرداخته می‌شود.



شکل ۱- مروری بر رویکرد پیشنهادی

۵-۱- مجموعه داده

مانند مطالعه قبلی [۹] ما در آزمایش‌های خود، از چهار مجموعه داده از پست‌های اینستاگرام [۳۸، ۳۹] استفاده کردیم. اولین مجموعه در دسترس عموم شامل ۴,۹۱۹,۸۳۹ پست فارسی زبان است که از ۲۱ ژانویه ۲۰۲۰ تا ۲۱ آوریل ۲۰۲۰ جمع‌آوری شده است. مجموعه دوم، مجموعه داده فارسی برچسب‌گذاری شده‌ای است که شامل ۲۰۶۷ پست مرتبط با کروناویروس است که با برچسب "ترس" یا "بدون ترس" برچسب‌گذاری شده‌اند. پست‌های این مجموعه داده توسط کارشناسان بررسی شد تا مشخص شود که آیا مفاهیم مرتبط با کرونا در پست‌ها با ترس ذکر شده یا مورد بحث قرار گرفته است یا خیر. مانند این مجموعه داده، ما از مجموعه داده‌های برچسب‌گذاری شده دیگری از پست‌های فارسی که شامل نمونه‌های «امید» و همچنین نمونه‌های منفی بود، بهره‌مند شدیم. این دو مجموعه برای آموزش

مطالعه فرضیه ما این است که هم‌افزایی دو ویژگی احساس ترس و احساس امید با ویژگی شباهت معنایی می‌تواند منجر به بهبود عملکرد تخمین شود. شکل (۱) رویکرد پیشنهادی ما را نشان می‌دهد که برای انجام این کار، سه گونه هم‌افزایی (۱) شباهت معنایی + احساس ترس، (۲) شباهت معنایی + احساس امید، (۳) شباهت معنایی + احساس ترس + احساس امید را معرفی می‌کنیم. رویکرد ما برای تخمین نرخ بروز کرونا مطابق شکل (۱)، ساخت یک مدل رگرسیونی است که موارد مبتلایان جدید را براساس تعداد پست‌هایی که واقعاً به کرونا مرتبط هستند، پیش‌بینی می‌کند. بدین منظور ابتدا، با در نظر گرفتن یک بازه زمانی T (که در پژوهش ما یک روز است)، سه ویژگی شباهت معنایی، احساس ترس و احساس امید از هر پست منتشر شده در بازه T استخراج می‌شود.

برای محاسبه شباهت معنایی، علاوه بر تکنیک‌های Word2Vec، GloVe و FastText، از مدل Sentence-BERT (SBERT) نیز استفاده شد. SBERT با تولید تعبیه‌های معنایی در سطح جمله، امکان شناسایی دقیق تر پست‌های مرتبط با کووید-۱۹ را فراهم می‌کند. سپس یک مدل پیش‌بینی‌کننده (IRE (Incidence Rate Estimator)، براساس الگوریتم‌های یادگیری ماشین ساخته می‌شود تا نرخ بروز کرونا را در بازه زمانی T تخمین زند. برای دستیابی به این هدف از سه شاخص آماری که مرتبط با این سه ویژگی هستند به نام تعداد پست‌های مرتبط با کرونا از نظر معنایی، تعداد پست‌های دارای احساس ترس، و تعداد پست‌هایی با احساس امید استفاده می‌شود و پیش‌بینی میزان شیوع کرونا براساس ترکیب این سه شاخص آماری در بازه زمانی T صورت می‌گیرد. راه حل ما برای ایجاد IRE آموزش رگرسیون با هر یک از این هم‌افزایی‌ها به عنوان ورودی و تعداد موارد آلوده به عنوان خروجی است. به منظور ایجاد IRE، از رگرسیون بردار پشتیبان (SVR) [۳۷] به عنوان یک تکنیک یادگیری نظارت شده کلاسیک که ثابت شده است یک ابزار موثر در تخمین تابع ارزش واقعی است، استفاده می‌شود. یکی از مزایای SVR قدرت پردازش داده‌های با ابعاد بالا است که به طور موثر از طریق توابع Kernel به دست می‌آید.

۵- آزمایش‌ها و تحلیل نتایج

هدف اصلی ما از انجام آزمایش‌های متعدد، ارزیابی رویکرد

به‌عنوان متغیر وابسته در آموزش مدل رگرسیون برای تخمین میزان بروز کرونا استفاده کردند. از نظر معیارهای ارزیابی، ما عملکرد رویکرد پیشنهادی خود را با کارهای پایه [۳، ۹] با استفاده از ضریب تعیین مقایسه می‌کنیم [۴۱]. ضریب تعیین معمولاً به‌عنوان R-squared (یا R^2) معرفی می‌شود که درصد واریانس را برای تعداد موارد جدید مبتلایان کرونا به‌عنوان یک متغیر وابسته، با در نظر گرفتن تعداد پست‌های مرتبط با کرونا، تعداد پست‌های با احساس ترس یا تعداد پست‌های دارای احساس امید، به‌عنوان متغیرهای مستقل در مدل‌های رگرسیون، پیش‌بینی می‌کند. مقادیر R-squared از ۰ تا ۱ متغیر است و R-squared برابر با ۱ به این معنی است که تعداد موارد جدید مبتلایان به طور کامل با ویژگی‌های پیشنهادی پیش‌بینی می‌شوند. در محاسبه R-squared، ما از یک رویکرد اعتبارسنجی متقابل ۳ برابری برای ارزیابی رویکرد خود و همچنین کارهای پایه استفاده می‌کنیم.

۵-۳- نتایج و یافته‌ها

به منظور تخمین نرخ بروز کرونا، موارد جدید روزانه مبتلایان را از ۱۳۹۸/۱۲/۱ تا ۱۳۹۹/۳/۳۱ در ایران استخراج کردیم. این آمار را از وبسایت^۲ IRANOPENDATA که آمار رسمی کرونا را در ایران منتشر کرده است، به دست آوردیم. به عبارت دیگر هدف اصلی این پژوهش پیش‌بینی موارد جدید مبتلا به کرونا با استفاده از پست‌های اینستاگرامی کاربران ایرانی است. هدف از این بخش پاسخ به این سوال است که آیا ویژگی‌های احساس ترس یا امید می‌توانند در کنار شباهت معنایی به عنوان شاخص‌های قوی برای شناسایی پست‌های مرتبط با کرونا در نظر گرفته شوند. برای پاسخ به این سوال پژوهش، آزمایش‌های متعددی جهت بررسی اثر هم‌افزایی این ویژگی‌ها بر روی عملکرد IRE انجام شده است که در تمام این آزمایش‌ها از ترکیب شاخص‌های آماری معرفی شده به عنوان ورودی استفاده شده است. لازم به ذکر است که ارزیابی ما از مدل IRE به صورت اعتبارسنجی متقاطع ۳ برابری است. نتایج آزمایش‌ها در جدول ۱ نشان داده شده است.

رده‌بند مبتنی بر BERT [۴۰] برای استخراج احساس ترس و امید از پست‌ها ایجاد شده است. به عبارت دیگر، راه‌حل ما، آموزش یک رده‌بند دودویی مبتنی بر BERT است که محتوای پست‌ها را به عنوان ورودی می‌گیرد و وجود احساس ترس یا امید را در محتوای پست‌ها شناسایی می‌کند. در نهایت، به‌منظور ارزیابی رویکرد پیشنهادی، پست‌های حاوی کلمه‌کلیدی "کرونا" را از ابتدای ۲۰ فوریه ۲۰۲۰ تا ۲۰ ژوئن ۲۰۲۰ جمع‌آوری می‌کنیم. به عبارت دیگر، آخرین مجموعه شامل پست‌های فارسی که کاربران در موج اول کروناویروس در ایران منتشر کردند. علاوه بر این، ما پست‌های این مجموعه را مانند اولین مجموعه پیش‌پردازش کردیم.

۵-۲- کارهای پایه و معیارهای ارزیابی

اخیراً پژوهش‌های متعددی به دنبال نظارت بر محتوای تولید شده از شبکه‌های اجتماعی به وسیله کاربران و پیوند آن‌ها با موارد رسمی مبتلا به کرونا بودند [۳، ۶، ۹]. به عنوان مثال، نویسندگان در کار اخیر [۳] معتقد بودند که محتوای توییتر به عنوان یکی از منابع غیررسمی، می‌تواند برای شناسایی و ردیابی شیوع کرونا مورد تجزیه و تحلیل قرار گیرد. آن‌ها داده‌های توییتر را درست قبل و در طول همه‌گیری کرونا در سراسر ایالات متحده در سطح ایالتی، برای رایج‌ترین علائم کرونا از جمله "سرفه" و "تب" بررسی کردند. آن‌ها برای یافتن ارتباط بین تعداد توییت‌های حاوی "سرفه" و "تب" و موارد گزارش شده رسمی، یک تحلیل رگرسیون انجام دادند.

همچنین در مطالعه [۹]، نویسندگان رویکردی برای تخمین نرخ ابتلا به کرونا در شبکه اجتماعی اینستاگرام در ایران، با استفاده از بازایی پست‌های مرتبط با کرونا براساس سه ویژگی شباهت معنایی، احساس ترس و احساس امید ارائه کردند. آن‌ها از شباهت معنایی به‌عنوان اولین ویژگی، برای اندازه‌گیری شباهت کسینوسی بین یک پست داده شده به کرونا در فضای معنایی استفاده کردند. دو ویژگی دیگر نشان می‌دهد که آیا احساس ترس یا احساس امید در یک پست مشخص بیان می‌شود یا خیر. براساس این ویژگی‌ها، تعداد پست‌های روزانه مرتبط با کرونا را تعیین کردند و از آن

^۲ <https://media.iranopendata.org/Sources/IOD350.pdf>

گرفته می‌شوند. نتایج عملکرد IRE برای هم‌افزایی احساس ترس همراه با شباهت معنایی در سطر اول نمایش داده می‌شود. همان‌طور که نشان داده شده است عملکرد IRE در هم‌افزایی شباهت معنایی با احساس ترس برای تکنیک‌های Word2vec و Fasttext نتایج خوبی را نسبت به تکنیک Glove از خود نشان داده است.

• اثر هم‌افزایی شباهت معنایی + احساس امید

در این پیشنهاد از هم‌افزایی، از تعداد پست‌های مرتبط با کرونا و احساس امید به عنوان دو متغیر مستقل برای پیش‌بینی موارد جدید مبتلایان توسط SVR استفاده می‌شود. ما عملکرد IRE را برای این مورد بررسی می‌کنیم که از ویژگی شباهت معنایی همراه با احساس امید استفاده می‌شود. به عبارت دیگر پست‌های روزانه که از نظر معنایی با کرونا مرتبط هستند و احساس امید کاربران از کرونا را منعکس می‌کنند به عنوان شاخص آماری در آزمایش‌ها در نظر گرفته می‌شوند. مانند ویژگی قبلی، همان‌طور که در سطر دوم

گزارش شده است، بالاترین عملکرد برای هم‌افزایی ویژگی شباهت معنایی با احساس امید در تکنیک‌های تعبیه کلمه Word2vec و FastText نسبت به Glove مشاهده می‌شود. با توجه به این‌که هم‌افزایی شباهت معنایی با احساس ترس از هم‌افزایی شباهت معنایی با احساس امید بهتر است، می‌توان نتیجه گرفت که ویژگی احساس ترس در تخمین موارد جدید مبتلایان موثرتر از احساس امید است.

• اثر هم‌افزایی شباهت معنایی + احساس ترس + احساس امید

با توجه به نتایج دو هم‌افزایی قبلی، هدف از این بخش بررسی این مورد است که، آیا هم‌افزایی سه ویژگی هم منجر به افزایش عملکرد برای IRE می‌شود یا خیر. برای این منظور از تعداد پست‌های روزانه با شباهت معنایی، احساس ترس و احساس امید استفاده می‌شود و عملکرد IRE را براساس این هم‌افزایی در ردیف آخر جدول ۱ گزارش می‌شود. براساس نتایج، می‌توان دریافت که بهترین عملکرد برای تکنیک‌های Word2vec و FastText است. به طور کلی، از بین سه مدل هم‌افزایی، شباهت معنایی + احساس ترس از نظر R-squared در مقایسه با سایر تغییرها برای Word2vec و FastText بهترین نتایج را به دست می‌آورد.

جدول ۱- تاثیر هم‌افزایی ویژگی‌های پیشنهادی بر عملکرد IRE

	تکنیک‌های تعبیه کلمه	R-squared
شباهت معنایی + احساس ترس	Word2vec Glove Fasttext	۰.۴۷ ۰.۳۰ ۰.۴۴
شباهت معنایی + احساس امید	Word2vec Glove Fasttext	۰.۴۲ ۰.۲۶ ۰.۳۶
شباهت معنایی + احساس ترس + احساس امید	Word2vec Glove Fasttext	۰.۴۲ ۰.۲۴ ۰.۳۷

• اثر هم‌افزایی شباهت معنایی + احساس ترس

محققان در مطالعه [۹] دریافتند که استفاده از کلمه‌های کلیدی برای بازیابی پست‌های مرتبط با کرونا منجر به تخمین ضعیف موارد جدید مبتلایان کرونا می‌شود. به این دلیل که، ممکن است هیچ محتوایی مرتبط با کرونا در پست‌های حاوی کلمه "کرونا" وجود نداشته باشد. بنابراین آن‌ها یک رویکرد معنایی مبتنی بر تکنیک‌های تعبیه کلمه برای مدل‌سازی کرونا پیشنهاد کردند و سپس یک ویژگی جدید به نام شباهت معنایی را برای اندازه‌گیری شباهت یک پست داده‌شده به کرونا در فضای معنایی معرفی کردند. هدف آن‌ها از این ویژگی این بود که پست‌های حاوی کلمه "کرونا" که واقعاً به کرونا مرتبط هستند در نظر گرفته شوند و موارد نامربوط حذف می‌شوند که منجر به بهبود عملکرد کار آن‌ها شد. همچنین آن‌ها دریافتند که کاربران اینستاگرام احساس ترس را در دوره شیوع و همه‌گیری بیماری ابراز کرده‌اند. این یافته، نظر محققان را تأیید می‌کند که معتقد بودند خطر درک شده از یک بیماری همه‌گیر باعث می‌شود افراد احساس کنترل شرایط زندگی خود را از دست بدهند که ممکن است باعث ایجاد احساس ترس شود [۴۲].

بنابراین در این مورد از هم‌افزایی، از تعداد پست‌های مرتبط با کرونا و احساس ترس به عنوان دو متغیر مستقل برای پیش‌بینی موارد جدید مبتلایان توسط SVR استفاده می‌شود. به عبارت دیگر پست‌های روزانه که از نظر معنایی با کرونا مرتبط هستند و ترس کاربران از کرونا را منعکس می‌کنند به عنوان شاخص آماری در آزمایش‌ها در نظر

می‌شود) بهترین عملکرد را نسبت به دو هم‌افزایی دیگر در مدل SVM دارد، همچنین در نتایج هم‌افزایی ویژگی‌ها با شباهت‌معنایی که با SBERT بدست آمد، هم‌افزایی احساس ترس با SBERT نیز نسبت به بقیه، عملکرد بهتری از خود نشان داد. تنها نتایج مربوط به این هم‌افزایی‌ها در جدول ۳ نشان داده شده است. در مطالعه [۳۱]، از مدل BERT برای استخراج تعبیه‌های معنایی توییت‌ها و کلمات کلیدی مرتبط با عوارض جانبی داروها و سپس محاسبه شباهت آن‌ها استفاده شده بود. ما نیز از SBERT بهره بردیم که به‌طور خاص برای تولید جاسازی‌های جمله‌ای و محاسبه شباهت کسینوسی بین پست‌ها و کلمات کرونا بهینه‌سازی شده است که عملکرد بهتری در شناسایی پست‌های مرتبط با کووید-۱۹ فراهم کرد.

در [۳]، محققان داده‌های توییت‌ها را درست قبل و در طول همه‌گیری کرونا برای رایج‌ترین علائم کرونا از جمله "سرفه" و "تب" بررسی کردند. ما نیز جهت انجام مقایسه، در آزمایش‌های خود به جای این دو کلمه، تنها از تعداد پست‌های روزانه حاوی کلمه "کرونا" به عنوان ویژگی جهت استخراج پست‌های مرتبط با کرونا استفاده کرده‌ایم. همان‌طور که در جدول ۳ نشان داده شده است بازایی پست‌ها تنها با استفاده از کلمه "کرونا" منجر به تخمین ضعیف موارد جدید مبتلایان می‌شود. زیرا ممکن است هیچ محتوایی مرتبط با کرونا در پست‌های حاوی کلمه "کرونا" مانند پست‌های تبلیغاتی که به‌طور مداوم بر فروش محصول‌های ضدعفونی‌کننده متمرکز هستند وجود نداشته باشد. با توجه به عملکرد ضعیف [۳] در تخمین، محققان در [۹] برای بازایی پست‌های مرتبط با کرونا در کار خود از ویژگی شباهت معنایی استفاده کردند. از منظر شباهت معنایی، پست‌های حاوی کلمه "کرونا" که واقعاً به کرونا مرتبط است در نظر گرفته می‌شوند و موارد نامربوط حذف می‌شوند که منجر به بهبود قابل توجهی نسبت به کار پایه [۳] می‌شود. همچنین آن‌ها دریافتند که علاوه بر شباهت معنایی، ویژگی‌های احساس ترس یا احساس امید نیز می‌توانند به عنوان شاخص‌های قوی در هنگام شناسایی پست‌های مرتبط با کرونا در نظر گرفته شوند. از آنجایی که تخمین نرخ بروز کرونا براساس شباهت معنایی بهتر از دو ویژگی احساس ترس و امید بوده است، در جدول ۳ عملکرد IRE در چهار حالت مختلف گزارش شده است. در حالت اول در سطر اول جدول، که فقط از پست‌هایی که

جدول ۲- تاثیر هم‌افزایی ویژگی‌های پیشنهادی همراه با SBERT و w2v با مدل SVM بر عملکرد IRE.

	هم‌افزایی	R-squared (SVM)
۱	شباهت معنایی (SBERT) + احساس ترس	۰.۵۲
	شباهت معنایی (w2v) + احساس ترس	۰.۴۷
۲	شباهت معنایی (SBERT) + احساس امید	۰.۴۳
	شباهت معنایی (w2v) + احساس امید	۰.۴۲
۳	شباهت معنایی (SBERT) + احساس ترس + احساس امید	۰.۵۰
	شباهت معنایی (w2v) + احساس ترس + احساس امید	۰.۴۲

برای بررسی اثر استفاده از مدل‌های پیشرفته‌تر، شباهت معنایی با استفاده از Sentence-BERT (SBERT) محاسبه شد. مطابق نتایج نشان داد که هم‌افزایی شباهت معنایی (SBERT) با احساس ترس دارای R-squared برابر با ۰.۵۲ در مدل SVM است که عملکرد بهتری نسبت به هم‌افزایی با احساس امید (با R-squared برابر با ۰.۴۳) دارد. همچنین هم‌افزایی هر دو ویژگی یعنی احساس ترس و امید همراه با SBERT دارای میزان R-squared برابر با ۰.۵۰ می‌باشد. با این حال، عملکرد SBERT در مقایسه با Word2Vec، Glove و FastText برتری دارد. استفاده از SBERT به طور محسوسی باعث بهبود دقت پیش‌بینی شد. این بهبود عملکرد نشان می‌دهد که SBERT، به‌عنوان یک مدل مبتنی بر ترنسفورمر، توانایی بیشتری در درک روابط معنایی عمیق و زمینه‌مند در متون فارسی دارد.

۵-۴- مقایسه کار پیشنهادی با کار پایه

هدف این مطالعه، بررسی تاثیر هم‌افزایی ویژگی‌های شباهت معنایی، احساس ترس و امید برای تخمین موارد جدید مبتلا به کرونا است. انتظار می‌رود که عملکرد، زمانی بهبود یابد که از هم‌افزایی این ویژگی‌ها در تخمین استفاده شود. برای ارزیابی اثر هم‌افزایی، عملکرد IRE با کارهای پایه [۳، ۹] مقایسه شده‌اند که نتایج در جدول ۳ نشان داده شده است. از آنجایی که هم‌افزایی شباهت معنایی با احساس ترس (زمانی که از روش تعبیه Word2vec استفاده

تعبیه معنایی در استخراج روابط عمیق تر و اثر احساسات کاربران در جریان بیماری‌ها است. داده‌های کوتاه و غیررسمی اینستاگرام نیز احتمال بیش‌برازش را افزایش می‌دهد. استفاده از SBERT برای تحلیل داده‌های فارسی اینستاگرامی نیز جنبه نوآورانه پژوهش می‌باشد، که باعث برتری نتایج نسبت به روش‌های معنایی کلاسیک شد. همچنین این پژوهش بر روی داده‌های موج اول کرونا تمرکز دارد که برای تعمیم به سایر بیماری‌ها، دوره‌های زمانی یا پلتفرم‌های دیگر نیاز به بازتنظیم و ارزیابی مجدد خواهد داشت.

۷- نتیجه‌گیری

در این پژوهش، رویکردی برای تخمین نرخ بروز کرونا در پلتفرم‌های میکرو بلاگینگ، به‌ویژه اینستاگرام در ایران، ارائه کرده‌ایم. ما براساس داده‌های جمع‌آوری شده اینستاگرام در پیک اول کرونا در ایران، هم‌افزایی سه ویژگی شباهت معنایی، احساس ترس، و احساس امید را برای بازیابی پست‌های مرتبط با کرونا پیشنهاد کرده‌ایم. براساس این ویژگی‌ها، تعداد پست‌های روزانه مرتبط با کرونا را تعیین کردیم و از آن به‌عنوان متغیر وابسته در آموزش مدل رگرسیون برای تخمین میزان بروز کرونا استفاده کردیم. ما نشان داده‌ایم که رگرسیون آموخته‌شده با استفاده از این متغیر به‌ویژه با مدل SVM از کارهای پایه موجود برحسب R-squared بهتر عمل می‌کند. هم‌افزایی شباهت معنایی همراه با احساس ترس بهترین عملکرد کلی را نشان می‌دهد. آزمایش‌های ما نشان داد که هر ویژگی جنبه خاصی از پست‌ها را به تصویر می‌کشد که می‌تواند در شناسایی پست‌های مرتبط با کرونا موثر باشد بنابراین ادغام هر یک از آن‌ها می‌تواند عملکرد تخمین را بهبود بخشد.

تقدیر و تشکر

نویسندگان از تمامی افرادی که به‌طور مستقیم یا غیرمستقیم در انجام این پژوهش همکاری داشته‌اند، قدردانی می‌کنند.

تعارض منافع

نویسندگان اعلام می‌کنند که هیچ‌گونه تعارض منافعی در ارتباط با انجام این پژوهش و انتشار مقاله وجود ندارد.

تاییدیه اخلاقی

نویسندگان متعهد میشوند که مطالب این مقاله را در هیچ

حاوی کلمات کلیدی به عنوان ورودی مدل استفاده شده است که عملکرد آن از همه حالت‌ها ضعیف‌تر است. در حالت دوم، تنها از شاخص آماری که براساس شباهت معنایی شکل گرفته است به عنوان ورودی مدل استفاده شده است و عملکرد آن به‌عنوان خط پایه مقایسه شده است. در حالت سوم و چهارم رویکرد پیشنهادی ما است، در حالی است که ما برای هم‌افزایی ویژگی شباهت معنایی + احساس ترس علاوه بر شباهت معنایی مبتنی بر Word2Vec از احساس ترس نیز استفاده می‌کنیم، که نتایج مربوط به آن در سطر سوم (شباهت معنایی + احساس ترس) نشان داده شده است. حالت چهارم، از شباهت معنایی به کمک SBERT که با احساس ترس ترکیب شده است و در سطر چهارم جدول ۳ نشان داده شده است. نتایج نشان می‌دهد که مدل SVM، در ترکیب ویژگی‌ها به‌ویژه با استفاده از SBERT، منجر به بهبود قابل توجه در پیش‌بینی نرخ بروز کووید-۱۹ شده است.

جدول ۳- مقایسه بین مدل پیشنهادی و کارهای پایه برحسب

R-squared		R ² (SVM)
کارپایه [۳]		۰.۲۷
کارپایه [۹]	شباهت معنایی	۰.۳۲
رویکرد پیشنهادی ما	شباهت معنایی (w2v) + احساس ترس	۰.۴۷
	هم‌افزایی شباهت معنایی (SBERT) + احساس ترس	۰.۵۲

۶- بحث

مطابق با نتایج آزمایشات، ترکیب شباهت معنایی و احساس ترس و امید در پست‌های اینستاگرامی می‌تواند نرخ بروز کووید-۱۹ را با دقت قابل قبولی تخمین بزند. همان‌طور که مدل SBERT با ویژگی ترس بهترین عملکرد را داشت (با میزان R² برابر با ۰.۵۲ در SVM)، در حالی که Word2Vec عملکرد ضعیف‌تری را از خود نشان داد (با R² برابر ۰.۴۷). این تفاوت به دلیل ماهیت غیررسمی و نویزدار داده‌های اینستاگرامی است، که پس از پیش‌پردازش (حذف ایموجی‌ها و نویز) همچنان چالش برانگیز بودند. هم‌افزایی ویژگی‌های معنایی و احساسی، نسبت به روش‌های کلاسیک [۹] (مانند کار پایه در سطر دوم جدول ۳)، بهبود قابل توجهی نشان داد. این بهبود ناشی از توانایی مدل‌های

مجله دیگری به چاپ نرسانده اند.

مشارکت‌های نویسندگان

فهمیه حافظی: مسئولیت اصلی انجام پژوهش، طراحی و توسعه روش پیشنهادی، پیاده‌سازی و اجرای آزمایش‌ها، تحلیل نتایج و نگارش نسخه اولیه مقاله را بر عهده داشته است.

مریم خدابخش: مسئولیت راهنمایی و نظارت علمی بر

پژوهش، طراحی و توسعه روش پیشنهادی، ارائه پیشنهادها علمی و تخصصی، بازبینی انتقادی و اصلاح محتوای مقاله را بر عهده داشته است.

هر دو نویسنده نسخه نهایی مقاله را مطالعه و تأیید کرده‌اند.

منابع مالی

این پژوهش از حمایت مالی اختصاصی هیچ نهاد یا سازمانی برخوردار نبوده است.

مراجع

- [1] Guo, Yan-Rong, Qing-Dong Cao, Zhong-Si Hong, et al. "The Origin, Transmission and Clinical Therapies on Coronavirus Disease 2019 (COVID-19) Outbreak—an Update on the Status." *Military Medical Research* 7 (2020): 1–10.
- [2] Wang, Cuiyan, Riyu Pan, Xiaoyang Wan, et al. "Immediate Psychological Responses and Associated Factors During the Initial Stage of the 2019 Coronavirus Disease (COVID-19) Epidemic among the General Population in China." *International Journal of Environmental Research and Public Health* 17, no. 5 (2020): 1729.
- [3] Gharavi, Erfaneh, Neda Nazemi, and Faraz Dadgostari. "Early Outbreak Detection for Proactive Crisis Management Using Twitter Data: Covid-19 a Case Study in the Us." *arXiv preprint arXiv:2005.00475* (2020).
- [4] González-Padilla, Daniel A., and Leonardo Tortolero-Blanco. "Social Media Influence in the COVID-19 Pandemic." *International Brazilian Journal of Urology* 46, no. suppl. 1 (2020): 120–124.
- [5] Wang, Yan, Shuang Yang, and Xinying An. "Opinion Leaders and Crisis Communication During the COVID-19 Pandemic: A Study of Theme Evolution and Emotional Impact on Twitter." *Digital Health* 10 (2024): 20552076241234619.
- [6] Ayyoubzadeh, Seyed Mohammad, Seyed Mehdi Ayyoubzadeh, Hoda Zahedi, et al. "Predicting COVID-19 Incidence through Analysis of Google Trends Data in Iran: Data Mining and Deep Learning Pilot Study." *JMIR Public Health and Surveillance* 6, no. 2 (2020): e18828.
- [7] Huang, Xiao, Siqin Wang, Mengxi Zhang, et al. "Social Media Mining under the COVID-19 Context: Progress, Challenges, and Opportunities." *International Journal of Applied Earth Observation and Geoinformation* 113 (2022): 102967.
- [8] Mahdikhani, Maryam. "Predicting the Popularity of Tweets by Analyzing Public Opinion and Emotions in Different Stages of COVID-19 Pandemic." *International Journal of Information Management Data Insights* 2, no. 1 (2022): 100053.
- [9] Hafezi, Fahimeh, and Maryam Khodabakhsh. "Coronavirus Incidence Rate Estimation from Social Media Data in Iran." *Journal of AI and Data Mining* 11, no. 2 (2023): 315–329.
- [10] Kanchan, Sushim, and Abhay Gaidhane. "Social Media Role and Its Impact on Public Health: A Narrative Review." *Cureus* 15, no. 1 (2023).
- [11] Farsi, Deema. "Social Media and Health Care, Part I: Literature Review of Social Media Use by Health Care Providers." *Journal of Medical Internet Research* 23, no. 4 (2021): e23205.
- [12] Yousefinaghani, Samira, Rozita Dara, Samira Mubareka, et al. "Prediction of COVID-19 Waves Using Social Media and Google Search: A Case Study of the US and Canada." *Frontiers in Public Health* 9 (2021): 656635.
- [13] Amin, Samina, Muhammad Irfan Uddin, Duaa H. AlSaeed, et al. "Early Detection of Seasonal Outbreaks from Twitter Data Using Machine Learning Approaches." *Complexity* 2021, no. 1 (2021): 5520366.
- [14] Guidry, Jeanine P. D., Shana L. Meganck, Paul B. Perrin, et al. "#Ebola: Tweeting and Pinning an Epidemic." *Atlantic Journal of Communication* 29, no. 2 (2021): 79–92.
- [15] Koplewitz, Gal, Fred Lu, Leonardo Clemente, et al. "Predicting Dengue Incidence Leveraging Internet-Based Data Sources: A Case Study in 20 Cities in Brazil." *PLoS Neglected Tropical Diseases* 16, no. 1 (2022): e0010071.

- [16] Li, Cuilian, Li Jia Chen, Xueyu Chen, et al. "Retrospective Analysis of the Possibility of Predicting the COVID-19 Outbreak from Internet Searches and Social Media Data, China, 2020." *Eurosurveillance* 25, no. 10 (2020): 2000199.
- [17] O'Leary, Daniel E., and Veda C. Storey. "A Google–Wikipedia–Twitter Model as a Leading Indicator of the Numbers of Coronavirus Deaths." *Intelligent Systems in Accounting, Finance and Management* 27, no. 3 (2020): 151–58.
- [18] Català, Martí, David Pino, Miquel Marchena, et al. "Robust Estimation of Diagnostic Rate and Real Incidence of COVID-19 for European Policymakers." *PLOS ONE* 16, no. 1 (2021): e0243701.
- [19] Abdukhamidov, Eldor, Firuz Juraev, Mohammed Abuhamad, et al. "Sentiment Analysis of Users' Reactions on Social Media During the Pandemic." *Electronics* 11, no. 10 (2022): 1648.
- [20] Jafarinejad, Fateme, Marziea Rahimi, and Hoda Mashayekhi. "Tracking and Analysis of Discourse Dynamics and Polarity During the Early Corona Pandemic in Iran." *Journal of Biomedical Informatics* 121 (2021): 103862.
- [21] Niknam, Fatemeh, Mahnaz Samadbeik, Farhad Fatehi, et al. "COVID-19 on Instagram: A Content Analysis of Selected Accounts." *Health Policy and Technology* 10, no. 1 (2021): 165–73.
- [22] Tandorost, Hanie, Samira Abbasi, and Fatemeh Amiri. "Classification of Social Media Posts During the Corona Crisis Based on Their Polarity." *Journal of Modeling in Engineering* 21, no. 74 (2023): 81–93.
- [23] Su, Yue, Peijing Wu, Sijia Li, et al. "Public Emotion Responses During COVID-19 in China on Social Media: An Observational Study." *Human Behavior and Emerging Technologies* 3, no. 1 (2021): 127–36.
- [24] Qin, Lei, Qiang Sun, Yidan Wang, et al. "Prediction of Number of Cases of 2019 Novel Coronavirus (COVID-19) Using Social Media Search Index." *International Journal of Environmental Research and Public Health* 17, no. 7 (2020): 2365.
- [25] Baydogan, Cem. "Deep-Cov19-Hate: A Textual-Based Novel Approach for Automatic Detection of Hate Speech in Online Social Networks throughout COVID-19 with Shallow and Deep Learning Models." *Tehnički Vjesnik* 29, no. 1 (2022): 149–56.
- [26] Es-Sabery, Fatima, Khadija Es-Sabery, Junaid Qadir, et al. "A MapReduce Opinion Mining for COVID-19-Related Tweets Classification Using Enhanced ID3 Decision Tree Classifier." *IEEE Access* 9 (2021): 58706–39.
- [27] Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. "Efficient Estimation of Word Representations in Vector Space." *arXiv preprint arXiv:1301.3781* (2013).
- [28] Bojanowski, Piotr, Edouard Grave, Armand Joulin, and Tomas Mikolov. "Enriching Word Vectors with Subword Information." *Transactions of the Association for Computational Linguistics* 5 (2017): 135–146.
- [29] Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. "Glove: Global vectors for word representation." In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532-1543. 2014.
- [30] Reimers, Nils, and Iryna Gurevych. "Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks." *arXiv preprint arXiv:1908.10084* (2019).
- [31] Dong, Fan, Wenjing Guo, Jie Liu, et al. "BERT-Based Language Model for Accurate Drug Adverse Event Extraction from Social Media: Implementation, Evaluation, and Contributions to Pharmacovigilance Practices." *Frontiers in Public Health* 12 (2024): 1392180.
- [32] Murray, Curtis, Lewis Mitchell, Jonathan Tuke, et al. "Symptom Extraction from the Narratives of Personal Experiences with COVID-19 on Reddit." *arXiv preprint arXiv:2005.10454* (2020).
- [33] Wang, Jinping, and Lewen Wei. "Fear and Hope, Bitter and Sweet: Emotion Sharing of Cancer Community on Twitter." *Social Media + Society* 6, no. 1 (2020): 2056305119897319.
- [34] Widmann, Tobias. "Fear, Hope, and COVID-19: Emotional Elite Rhetoric and Its Impact on the Public during the First Wave of the COVID-19 Pandemic." *Political Psychology* 43, no. 5 (2022): 827–50.
- [35] Myrick, Jessica Gall, and Jessica Fitts Willoughby. "A Mixed Methods Inquiry into the Role of Tom Hanks' COVID-19 Social Media Disclosure in Shaping Willingness to Engage in Prevention Behaviors." *Health Communication* 37, no. 7 (2022): 824–32.
- [36] Mousing, Camilla Askov, and Dorthe Sørensen. "Living with the Risk of Being Infected: COPD Patients' Experiences during the Coronavirus Pandemic." *Journal of Clinical Nursing* 30, nos. 11–12 (2021): 1719–29.
- [37] Zhang, Fan, and Lauren J. O'Donnell. "Support Vector Regression." In *Machine Learning*, 123–40. Elsevier, 2020.

- [38] Natural Language Processing Laboratory, Computer Faculty, Shahid Beheshti University. "Data Preprocessing." 2020. <http://nlp.sbu.ac.ir>. (in Persian)
- [39] Social Networks Laboratory, Faculty of Electrical Engineering, University of Tehran. "Data Preparation." 2020. <http://nlp.sbu.ac.ir>. (in Persian)
- [40] Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "Bert: Pre-training of deep bidirectional transformers for language understanding." In Proceedings of the 2019 Conference of the North American Chapter of the Association For Computational Linguistics: Human Language Technologies, Volume 1 (long and short papers), pp. 4171-4186. 2019.
- [41] Andreas, Andreou, Constandinos X. Mavromoustakis, George Mastorakis, et al. "Modified Machine Learning Technique for Curve Fitting on Regression Models for COVID-19 Projections." Paper presented at the 2020 IEEE 25th International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD), 2020.
- [42] Nan, Zhao, Wang Yilin, Li Sijia, et al. "Psychological and Behavioral Impact of Wuhan Lockdown and Suggestions." Bulletin of Chinese Academy of Sciences (Chinese Version) 35, no. 3 (2020): 264–72.