تحلیل احساس پست‌های شبکه‌های اجتماعی در بحران کرونا با استفاده از خوشه‌بندی دو مرحله‌ای

سمیرا عباسی[۱] و فاطمه امیری[۲و*]

[۱]گروه مهندسی پزشکی، دانشگاه صنعتی همدان، همدان، ایران.

[۲]گروه مهندسی کامپیوتر، دانشگاه صنعتی همدان، همدان، ایران

چکیده

در بحران کرونا با طیف وسیعی از افکار، احساسات و نگرش‌ها در شبکه‌های اجتماعی مواجهیم. دستیابی به درک جامعی از نگرش‌های جامعه نیازمند پردازش این داده‌هاست. هدف این پژوهش شناسایی ویژگی پیام‌هایی است که منجر به قطبیت‌های احساسی گوناگون در شبکه‌های اجتماعی می‌شوند. در این پژوهش از پست‌های فارسی توییتر، اینستاگرام، تلگرام و کانال‌های خبری و روش‌های پردازش زبان طبیعی استفاده شده است. در روش پیشنهادی این پژوهش، خوشه‌بندی دو مرحله‌ای مبتنی بر شبکهٔ عصبی خودسازمان‌ده و K-میانگین استفاده شده است. نتایج نشان دادند پست‌های حوزهٔ سلامت و فرهنگ با قطبیت منفی، به احساساتی مانند ترس، تنفر، غم و خشم منجر شده است. پیام‌های مربوط به عملکرد هیجانی و نادرست مردم با احساس غم، ترس و استرس همراه است و امید را در جامعه کاهش داده است.

واژگان کلیدی: کرونا، شبکه‌های اجتماعی، تحلیل احساسات، خوشه‌بندی.

Sentiment Analysis of Social Media Posts in the Corona Crisis using two-stage Clustering

Samira Abasi[1] and Fatemeh Amiri[2,*]

[1]Department of Biomedical Engineering, Hamedan University of Technology, Hamedan, Iran.
[2]Department of Computer Engineering, Hamedan University of Technology, Hamedan, Iran.

Abstract:

During the Corona crisis, we are confronted with a diverse array of thoughts, feelings, attitudes, and behaviors expressed on social media. This data holds valuable insights for individuals and administrators seeking to effectively respond to the crisis. The goal of this study is to identify specific attributes of messages that generate different emotional reactions. The study aims to examine the content shared by Twitter, Instagram, and Telegram users, as well as news pertaining to the COVID-19 pandemic in Iran. The dataset extracted from these social networks focuses on the period between January 21 and April 29, 2020, encompassing content shared within Iran, and in Persian language.

It is important to mention that the dataset and its corresponding labels were published by the Cognitive Sciences and Technologies Council (CSTC) in Iran. Initially, the content of each post underwent pre-processing, which involved the removal of stop words, word normalization, tokenization, and stemming. The emotional labels utilized Plutchik's model and encompassed emotions, such as joy, trust, fear, surprise, sadness, anticipation, anger, disgust, stress, or other emotions. For this study, clustering algorithms were employed to analyze the social media posts.

A two-stage clustering method was employed in this study. The proposed clustering algorithm combined the self-organized neural network and K-means algorithms. Following our proposed algorithm, the data underwent clustering using the self-organized neural network (SOM) as the initial step. The resulting

* Corresponding author

*نویسندهٔ عهده‌دار مکاتبات

• تاریخ ارسال مقاله: ۱۳۹۹/۱۰/۴ • تاریخ پذیرش: ۱۴۰۱/۱۰/۱۸ • تاریخ انتشار: ۱۴۰۲/۵/۲۰ • نوع مطالعه: کاربردی

cluster centers from SOM were then utilized as the starting points for the K-means algorithm. The implementations were developed using Python version 3.7 and MATLAB R2015a. Hazm Tools were utilized for data pre-processing, while the clustering process was performed in MATLAB. The Davies-Bouldin clustering evaluation metric was employed to determine the optimal number of clusters.
The measure was computed for the number of clusters ranging from 2 to 50 within the two-stage clustering method. The findings indicated that the optimal number of clusters was ten. Examination of the outcomes revealed that posts related to health and culture, characterized by a negative sentiment, evoked emotions, such as fear, hatred, sadness, and anger. Messages concerning individuals' emotional well-being and improper functioning induced feelings of sadness, fear, and stress, contributing to a decline in societal hope. The results demonstrated a strong association between anger and disgust, while fear, stress, and sadness exhibited a positive correlation. To mitigate negative emotions and foster trust in authorities, we recommend providing clear, comprehensive information about the coronavirus pandemic.
Keywords: COVID-19, Social media, Sentiment analysis, Clustering

۱- مقدمه

شیوع کوید ۱۹[1] تأثیر عظیمی بر استفاده از شبکه‌های اجتماعی در همهٔ دنیا داشته است. با شیوع این ویروس در کل دنیا، فعّالیّت افراد در شبکه‌های اجتماعی مانند فیس‌بوک، توییتر و ... افزایش یافته است [۱]. بر اساس گزارش‌ها [۲] محبوبیت رسانه‌های اجتماعی در ایالات متحدهٔ آمریکا بالاست و ۶۸ درصد از بزرگسالان آمریکایی اخبار را از رسانه‌های اجتماعی دریافت می‌کنند. این امر به‌ویژه در مورد اطّلاعات بهداشتی و علمی صادق است، و از رسانه‌های اجتماعی برای بررسی آگاهی، نگرش و واکنش عمومی در مورد بیماری‌های خاص استفاده شده است [۳]. با شیوع کرونا نیز، در اواخر سال ۲۰۱۹، مردم در بسیاری از کشورها برای جمع‌آوری اطّلاعات راجع‌به این بیماری از شبکه‌های اجتماعی استفاده کرده‌اند [۴].

داده‌ها و پست‌هایی که به‌صورت عمومی در شبکه‌های اجتماعی در دسترس هستند، می‌توانند برای بررسی افکار، احساسات و نگرش‌ها در رویارویی با بحران کرونا استفاده شوند[۱]. از سوی دیگر، گسترش اطّلاعات نادرست می‌تواند راهبردهای بهداشت عمومی را تضعیف کند [۵] و پیامدهای بالقوّه خطرناکی دارد[۶].

مطالعهٔ پیام‌ها در شبکه‌های اجتماعی می‌تواند به متخصصان، سیاستگذاران و جامعه برای مطالعهٔ موضوعات اساسی و تحلیل آنها به‌صورت مناسب کمک کند. همچنین، می‌تواند در درک وضعیت اجتماعی، مؤثّربودن سیاست‌ها و روش‌های پیاده‌سازی‌شده در جامعه در رویارویی با بحران کرونا، و نیز در طرّاحی گام‌های بعدی بسیار مفید باشد. به‌طور خلاصه، دستیابی به درک دقیق و جامعی از نگرش‌ها و رفتارهای جامعه نیازمند پردازش این داده‌هاست. روش‌های مختلفی برای تحلیل پیام‌های شبکه‌های اجتماعی استفاده شده است [۷-۱۰]، تعدادی از پژوهش‌ها به گروه‌بندی پیام‌ها به انواع مختلف مانند احتیاط و مشاوره، اطلاعیه‌ها و اقدامات، کمک‌های مالی و خدمات، پشتیبانی احساسی و عاطفی، انتقاد و شایعه پرداخته‌اند [۱۱-۱۵]. شناسایی ویژگی‌های کلیدی هر گروه از پیام‌ها به مسئولان این امکان را می‌دهد که بتوانند انواع مختلف اطّلاعات را بر اساس نیاز عمومی جامعه منتشر کنند[۱۶]. به‌عنوان‌مثال، داده‌های گروه انتقاد می‌تواند مسئولان را از نگرانی‌های اصلی جامعه مطلع کند، در نتیجه، پاسخ‌های مناسب آن را ارایه دهند. مطالعهٔ پیام‌هایی از نوع پشتیانی احساسی این امکان را فراهم می‌کند که مسئولان الگوهای حمایت اجتماعی کاربران شبکه‌های اجتماعی را بیاموزند و از چنین منابع داوطلبانه‌ای بهرهٔ بهتری ببرند. مطالعات [۱۷, ۱۸] نشان می‌دهند که مشارکت جامعه به‌عنوان روشی شناخته شده است که نه‌تنها جوامع را در فرایند تصمیم‌گیری دخیل می‌کند، بلکه آنها را به‌سمت مشارکت فعّال هدایت می‌کند و در کنترل سلامت خود توانمند می‌سازد. بنابراین، اهمیت و فوریّت خاصی در مطالعهٔ محتوای رسانه‌های اجتماعی برای کوید ۱۹ وجود دارد.

هدف پژوهش حاضر شناسایی ویژگی پیام‌هایی است که منجر به قطبیت‌های احساسی مختلف در عموم جامعه می‌شود. استفاده از پیام‌ها با قطبیت احساسی مناسب می‌تواند در کنترل همه‌گیری، ایجاد شبکه‌های ایمنی و ارتقای تاب‌آوری در بحران‌های مشابه کرونا مفید و مؤثر باشد.

در این پژوهش، تمرکز بر روی داده‌های استخراج‌شده از شبکه‌های اجتماعی و کانال‌های خبری در ایران است. مجموعه داده‌های استخراج‌شده شامل پست‌ها و اخبار منتشرشدهٔ ایرانیان در شبکه‌های اجتماعی با تمرکز بر کوید ۱۹ است. این محتوا از کانال‌های تلگرامی، صفحه‌های عمومی اینستاگرام و توییتر، و همچنین اخبار داخلی منتشرشده به‌صورت آن‌لاین از تاریخ یکم بهمن ماه سال ۱۳۹۸ تا دهم اردیبهشت ماه سال ۱۳۹۹ استخراج

[1] COVID-19

شده است. در این پژوهش برای تحلیل داده‌ها از الگوریتم خوشه‌بندی دومرحله‌ای که مرحلهٔ نخست آن شبکهٔ عصبی خودسازمان‌ده (SOM)[1] و مرحلهٔ دوم K-میانگین[2] است، استفاده شده است. نتایج آزمایش‌ها نشان می‌دهد بیشتر پیام‌های منتشرشده در بحران کرونا منجر به احساسات منفی مانند خشم، ترس و استرس می‌شوند.

در ادامه، کارهای پیشین در بخش ۲ و روش پیشنهادی شامل زیر بخش‌های جمع‌آوری داده و الگوریتم خوشه‌بندی در بخش ۳ تشریح می‌شود. در بخش ۴ نتایج به دست آمده ارائه و تحلیل می‌شود. در نهایت، بخش آخر به جمع بندی و نتیجه گیری می‌پردازد.

۲- کارهای پیشین

شبکه‌های اجتماعی لنزهایی هستند که از طریق آن افراد اطّلاعات را در موقعیت های مختلف جمع‌آوری و به اشتراک می‌گذارند. از رسانه های اجتماعی برای بررسی آگاهی، نگرش و واکنش مردم در مورد بیماری های خاص استفاده شده است [۱۹]. به‌عنوان‌مثال، داده‌های توییتر برای تحلیل موقعیت‌های بحرانی مانند تحلیل پاندمی‌ها استفاده شده است [۲۰].

چندین پژوهش [۱۷, ۱۸] نشان داده‌اند که مشارکت اجتماعی می‌تواند به‌عنوان روشی برای فرایند تصمیم‌گیری افراد و نیز توانمند کردن آنها در مشارکت فعّال برای افزایش سلامت استفاده شود. پژوهشگران [۲۱] نشان داده‌اند اطّلاعات منتشرشده در موقعیّت‌های بحرانی در شبکه‌های اجتماعی می‌تواند برای سیاستگذاران و افراد جامعه جهت پاسخ‌دهی به همه‌گیری ارزشمند باشد. بنابراین راهبردهای مناسب جهت انتشار اطّلاعات می‌تواند در بحران همه‌گیری کوید ۱۹ مفید و مؤثّر باشد. آنها از روش‌های پردازش زبان طبیعی برای طبقه‌بندی اطّلاعات مربوط به کوید ۱۹ در هفت گروه استفاده کردند. لوپز و همکاران [۲۲] برای درک نحوهٔ واکنش مردم در کشورهای مختلف به سیاست‌های مربوط به کوید ۱۹، یک مجموعه دادهٔ چندزبانه توییتر را تجزیه و تحلیل کردند. در پژوهش [۲۳]، براساس تجزیه و تحلیل نمودار اجتماعی به‌دست‌آمده از فعّالیّت‌ها در توییتر، چارچوبی برای شناسایی رهبران در این شبکهٔ اجتماعی پیشنهاد شده است. رهبران می‌توانند با تأثیر مثبت، به انتشار اطّلاعات معتبر کمک کنند. همچنین در طول یک بحران، کیفیت اطّلاعات در رسانه‌های اجتماعی مهم است. گسترش اطّلاعات نادرست می‌تواند استراتژی‌های بهداشت عمومی را تضعیف کند و عواقب بالقوه خطرناکی را به همراه دارد [۵]. برخی مطالعات در زمینه گسترش اطّلاعات نادرست در شبکه‌های اجتماعی و تأثیر آن بر مردم کار کردند [۲۴, ۲۵]. رایج‌ترین رویکرد برای مبارزه با اطّلاعات نادرست فیلترکردن مبتنی بر محتوا است [۶, ۲۶]. استفاده از روش‌های یادگیری عمیق، این کار را آسان‌تر و مقیاس‌پذیرتر می‌کند [۲۷, ۲۸].

برای تجزیه و تحلیل محتوای رسانه‌های اجتماعی، از روش‌های مختلف یادگیری ماشین استفاده شده است [۲۹-۳۱]. ویجیان و همکاران [۲۹] برای طبقه‌بندی متون و اسناد از رگرسیون خطی استفاده کردند. برخی مطالعات نشان دادند که طبقه‌بندی بیزین ساده [۳۲] روشی اثبات‌شده، ساده و مؤثّر برای طبقه‌بندی متن است.

رگرسیون منطقی[3] یکی از روش‌های محبوب و قدیمی برای طبقه‌بندی است. رگرسیون منطقی چندجمله‌ای می‌تواند احساس کاربران توییتر را با ۷۴٪ دقت پیش‌بینی کند[۳۳]. k-نزدیک‌ترین همسایه[4] (knn) یک طبقه‌بندی متنی غیرشاخصی است که متون یا اسناد را بر اساس اندازه‌گیری شباهت طبقه‌بندی می‌کند [۳۴]. نمس و کیس [۲۱] با استفاده از پردازش زبان طبیعی احساسات کاربران توییتر را تجزیه و تحلیل و با استفاده از شبکهٔ عصبی بازگشتی طبقه‌بندی کرده‌اند.

روش‌های خوشه‌بندی به‌عنوان روش‌های بدون ناظر و مبتنی بر شباهت بین داده‌ها هستند که می‌توانند نتایج مناسبی را از داده‌های آموزشی به دست آورند. در این فرایند داده‌های مشابه با یکدیگر در یک گروه قرار می‌گیرند. فرایند خوشه‌بندی، نمایی کلی از گروه‌ها را در یک مجموعه از مستندات نمایش می‌دهد. با توجّه به هدف پژوهشی ذکرشده، تمرکز ما در این پژوهش بر روی روش‌های خوشه‌بندی است و به این منظور از ترکیب دو الگوریتم خوشه‌بندی K-میانگین و SOM استفاده شده است.

۳- روش پیشنهادی

دیاگرام روش پژوهش پیشنهادی در شکل (۱) نشان داده شده است. بعد از جمع‌آوری پیام‌ها در شبکه‌های اجتماعی انتخابی، لازم است بر روی آنها پیش‌پردازش انجام شود. با توجه به حجم بالای پیام‌ها، یک مجموعه شامل ۶۳۹۹ پست و خبر به‌طور تصادفی انتخاب و به‌صورت دستی برچسب‌گذاری انجام می‌شود. بعد از اعمال الگوریتم خوشه‌بندی بر روی داده، محتوای پست‌های هر خوشه

[1] Self-organizing map
[2] K-means
[3] Logistic regression
[4] k-nearest neighbors

به‌طور مجزّا تحلیل و بررسی می‌شود. پیاده‌سازی در نرم‌افزار پایتون ورژن ۳/۷ و متلب R2015a انجام شده است. برای پیش‌پردازش داده‌ها نیز از ابزار Hazm[1] استفاده، و در ادامه خوشه‌بندی در نرمافزار متلب انجام شده است.

۳-۱- جمع‌آوری داده‌ها و پیش‌پردازش

در این مرحله، پست‌ها و خبرها در رابطه با کوید ۱۹ که در بازه زمانی یکم بهمن ماه ۱۳۹۸ تا دهم اردیبهشت ۱۳۹۹ منتشر شده است، از شبکه‌های اجتماعی جمع‌آوری شده‌اند. در ادامه اطّلاعات شامل متن، تعداد نظرات (کامنت‌ها) و پسندها (لایک‌ها) و همچنین اطّلاعات مالک پست استخراج شدند. همهٔ این پست‌ها و خبرها به زبان فارسی هستند و از کانال های عمومی تلگرام، صفحات عمومی توییتر، اینستاگرام و اخبار داخلی که به‌صورت آنلاین منتشر می‌شوند، استخراج شده‌اند. در بخش‌های بعدی هر پست یا خبر را یک سند می‌نامیم.

پیش از انجام خوشه‌بندی، لازم است محتوای هر سند پیش‌پردازش شود. این پیش‌پردازش شامل حذف وقفه‌ها، نرمال‌سازی کلمات، قطعه‌بندی[2]، و ریشه‌یابی[3] بوده است. نخست، هر سند طبیعی‌سازی شد. به این منظور، در برخی موارد ضروری فاصله‌ها به نیم‌فاصله تبدیل شدند. سپس متن طبیعی‌سازی‌شده، قطعه‌بندی، و توکن‌های کوچک‌تر از دو حرف، کلمات توقف، پیوندها و شناسه‌های ذکرشده در پیام‌ها حذف شدند. سپس کلمات باقی‌مانده توسط تابع مربوط به ریشه‌یابی در ابزار Hazm به ریشه‌هایشان تبدیل شدند و در پایان، کلمات موجود در فهرست، دوباره در کنار هم قرار داده و با عنوان محتوای پیش‌پردازش‌شده ذخیره شدند.

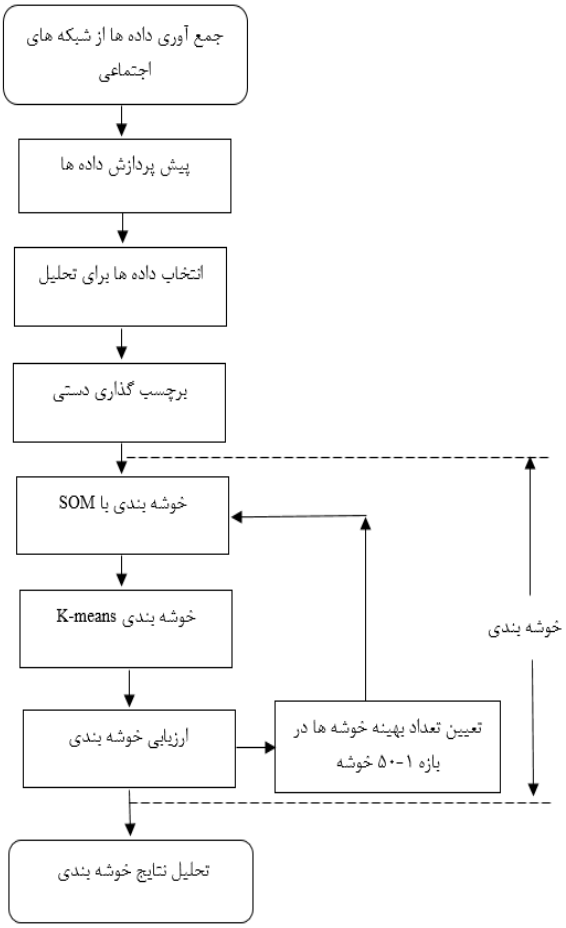
(جدول- ۱): تعداد سندهای استفاده‌شده از هر شبکه اجتماعی.

(Table- 1): Number of posts used from each social media

شبکهٔ اجتماعی	تعداد سندها
توییتر	۱۶۳۰
اینستاگرام	۲۰۷۳
خبر	۷۸۸
تلگرام	۱۸۴۸

به‌منظور تحلیل بیشتر، تعداد ۶۳۳۹ سند به‌طور تصادفی انتخاب شدند که توزیع آنها در شبکه‌های اجتماعی مختلف در جدول (۱) نشان داده شده است. در انتخاب این سندها دقت شده است که نمونه‌های تکراری فقط یک بار انتخاب شوند و همچنین، سندهای تبلیغاتی و خالی که فقط شامل هشتگ بودند، انتخاب نشوند. در ادامه داده‌ها به‌صورت دستی برچسب‌گذاری شده‌اند.[4]

برچسب‌های دسته احساس بر اساس مدل plutchik [۳۵] شامل شادی، غم، ترس، تنفّر، خشم، شگفتی، اعتماد، پیش‌بینی، استرس و سایر هیجانات هستند. برچسب‌ها به‌صورت دوتایی در نظر گرفته شده‌اند. درصورتی‌که سندی دارای برچسب خاصی باشد مقدار یک و در غیر این صورت مقدار صفر به آن اختصاص داده می‌شود.



(شکل- ۱) : روندنمای روش پژوهش

(Figure-1): Flowchart of the research method

۳-۲- خوشه‌بندی

روش‌های تحلیل احساس به دو دسته کلی تقسیم می‌شوند: تجزیه و تحلیل احساسات مبتنی بر فرهنگ لغت[5] و تحلیل احساسات با استفاده از یادگیری ماشین. در تجزیه و تحلیل احساسات مبتنی بر فرهنگ لغت لازم

[1] https://github.com/sobhe/hazm
[2] Tokenize
[3] Stemming

[4] مجموعه داده با حمایت ستاد توسعه علوم شناختی در آزمایشگاه شبکه‌های اجتماعی دانشکدهٔ برق و کامپیوتر دانشگاه تهران جمع‌آوری، و پردازش و برچسب‌زنی دستی این مجموعه در آزمایشگاه پردازش زبان فارسی دانشکدهٔ کامپیوتر دانشگاه شهید بهشتی انجام شده است. دوازده نفر نیز سندها را برچسب زده‌اند. مجموعهٔ داده در نشانی https://covidchallenge.cogc.ir/ قابل دسترسی است.

[5] lexicon-based sentiment analysis

است که واژگان احساسات به‌صورت دستی جمع‌آوری شوند و سپس فرهنگ لغت احساسات ساخته شود. روش‌های تجزیه و تحلیل احساسات مبتنی بر یادگیری ماشین می‌تواند به‌طور خودکار داده‌های متنی در مقیاس بزرگ را بدون مداخلهٔ دستی مدیریت کند. روش‌های تجزیه و تحلیل احساسات مبتنی بر یادگیری ماشین به روش‌های با ناظر و بدون ناظر دسته‌بندی می‌شوند. روش‌های خوشه‌بندی بدون ناظر بسیار متداول هستند [۳۶].

روش‌های مختلفی با رویکردهای متفاوت برای انجام تحلیل خوشه‌ای داده‌ها توسعه داده شده‌اند. هدف اصلی مسأله خوشه‌بندی، افراز یک مجموعه n تایی به k گروه به نحوی که این افراز واریانس درون‌خوشه‌ای کل[1] را کمینه می‌کند و به‌صورت زیر تعریف می‌شود:

فرض کنید $\{x_i, i = 1,2,\dots,n\}$ مجموعه‌ای شامل n داده و x_{ij} نشان‌دهندهٔ ویژگی jام از $i =$ برای x_i است. $W =$ ماتریس $k = 1,2,\dots,K$ و $1,2,\dots,n$ $w_{ik} =$ که به‌طوری ،می‌شود تعریف $[w_{ij}]$ و $w_{ij} \in \begin{cases} 1, if\ x_i\ belong\ to\ kth\ cluster \\ o \quad ,otherwise \end{cases}$ $\{0,1\} and, \sum_{j=1}^{K} w_{ij} = 1$ است.

در این حالت مرکز خوشهٔ kام، $c_k = \frac{\sum_{i=1}^{n} w_{ik} x_{ij}}{\sum_{i=1}^{n} w_{ik}}$ به‌صورت و است $c_{kj} = (c_{k1}, c_{k2}, \dots, c_{kd})$ محاسبه می‌شود. واریانس درون خوشه‌ای خوشه kام به‌صورت $S^{(k)}(W) = \sum_{i=1}^{n} w_{ik} \sum_{j=1}^{d} (x_{ij} - c_{kj})^2$ تعریف می‌شود و واریانس درون‌خوشه‌ای کل نیز عبارت است از:

$$S(W) = \sum_{k=1}^{K} S^{(k)} = \sum_{i=1}^{n} w_{ik} \sum_{j=1}^{d} (x_{ij} - c_{kj})^2 \quad (1)$$

که به آن معیار خطای مربع[2] می‌گویند. هدف در مسأله خوشه‌بندی، یافتن $W^* = [w_{ik}^*]$ که S(W) را کمینه کند یعنی: $S(W^*) = \min_W \{S(W)\}$.

دو روش متداول در خوشه‌بندی، شبکهٔ عصبی خودسازمان‌ده و K-میانگین است. هریک از این روش‌ها مزایا و معایبی دارند. شبکهٔ خودسازمان‌ده یکی از الگوریتم‌های خوشه‌بندی است که به‌طور گسترده استفاده می‌شود، اما زمان‌بر است. از طرفی K-میانگین الگوریتمی ساده است و سرعت بالایی در خوشه‌بندی دارد، اما به نوفه حساس است، و همچنین وابسته به تعیین مراکز اولیهٔ خوشه‌هاست. الگوریتم خوشه‌بندی دومرحله‌ای بر اساس شبکهٔ عصبی خودسازمان‌ده و K-میانگین نقص‌های هریک از دو روش را جبران می‌کند، و در عین حال مزایای هر دو روش را داراست.

بر اساس شکل-۱ الگوریتم خوشه‌بندی دومرحله‌ای به‌صورت زیر پیشنهاد می‌شود:

۱- برای انجام خوشه‌بندی به‌صورت مؤثّر و سریع، نخست، داده‌ها توسط شبکه SOM خوشه‌بندی می‌شوند.

۲- در ادامه الگوریتم K-میانگین اجرا می‌شود. نتایج حاصل از شبکه SOM به‌عنوان مراکز اولیهٔ خوشه‌ها در K-میانگین استفاده می‌شود.

همان‌طورکه در شکل (۱) مشاهده می‌شود، برای تعیین بهینهٔ خوشه‌ها لازم است مراحل یک و دوی الگوریتم خوشه‌بندی دومرحله‌ای به‌ازای مقادیر مختلف تکرار شود. الگوریتم خوشه‌بندی دومرحله‌ای نه‌تنها ویژگی خودسازمان‌دهی SOM را دارد، بلکه کارایی بالای روش K-میانگین را نیز داراست. بنابراین، زمان طولانی همگرایی SOM و احتمال انتخاب نادرست مراکز اولیهٔ خوشه‌ها در K-میانگین حل خواهد شد. مقایسهٔ نتایج این روش‌ها در مطالعهٔ حاضر نشان دادند که روش خوشه‌بندی دومرحله‌ای نتایج بهتر و سرعت بالاتری دارد، بنابراین، پس از بررسی‌های اولیه مشخص شد روش خوشه‌بندی دومرحله‌ای برای پژوهش حاضر مناسب است. در ادامه این روش‌ها شرح داده شده‌اند.

۳-۲-۱- شبکهٔ خودسازمان‌ده

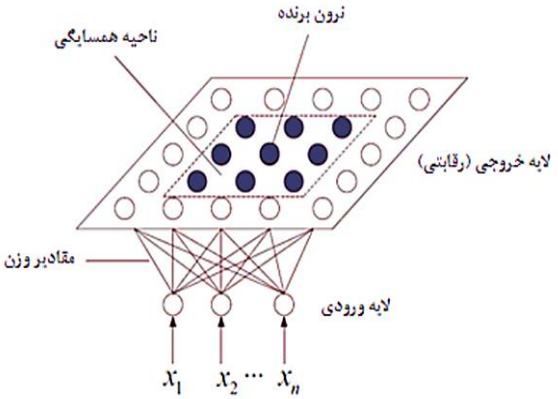
شبکهٔ خودسازمان‌ده نخست، توسط پروفسور کوهنن[3] ارائه شد به‌همین‌دلیل، به‌عنوان شبکهٔ Kohonen نیز شناخته می‌شود. شبکهٔ خودسازمان‌ده نوعی شبکهٔ عصبی با استفاده از روش‌های بدون ناظر آموزش، و یکی از الگوریتم‌های خوشه‌بندی است که به‌طور گسترده برای تحلیل داده‌ها استفاده می‌شود [۳۷].

یکی از ویژگی‌های متمایز SOM این است که می‌تواند فضای ورودی با ابعاد بالا را به فضایی با ابعاد پایین‌تر نگاشت، و در عین حال توپولوژی داده‌های اصلی را حفظ کند.

ساختار شبکهٔ SOM شامل یک لایهٔ ورودی و یک لایهٔ رقابتی است. لایهٔ ورودی، داده‌هایی را که ابعاد بالا دارند، دریافت، آن‌ها را جمع‌آوری و از طریق وزن‌ها به لایهٔ رقابتی ارسال می‌کند. لایهٔ رقابتی از نورون‌های پیوسته به هم تشکیل شده است و برای ارسال اطّلاعات به خروجی استفاده می‌شود. آرایش نورون‌های لایهٔ رقابتی می‌تواند به شکل‌های مختلفی باشد: یک‌بعدی خطی، دوبعدی آرایه‌ای، و شبکهٔ سه‌بعدی. متداول‌ترین ساختار،

[1] total within-cluster variation
[2] Square-error (SE) measure
[3] Kohonen

آرایهٔ دوبعدی است شکل(۲)؛ که شبیه تصویر قشر مغز است [۳۸].



(شکل -۲): ساختار شبکه SOM.
(Figure- 2): Structure of SOM network.

فرایند اجرای شبکهٔ SOM شامل دو مرحله است: آموزش و نگاشت کردن. در طول فاز آموزش بردارهای داده‌ها به‌طور تصادفی به شبکه ارائه می‌شوند. برای یک الگوی ورودی خاص، یک نورون برنده در لایهٔ خروجی وجود خواهد داشت که بزرگ‌ترین پاسخ را تولید خواهد کرد. در آغازِ فاز آموزش مشخص نیست که کدام نورون در لایهٔ خروجی بیشترین پاسخ را تولید خواهد کرد. زمانی‌که الگوی ورودی تغییر می‌کند، نورون برندهٔ فضای ورودی نیز تغییر خواهد کرد. به دلیل اثرات تحریک متقابل جانبی، نورون‌های اطراف نورون برنده پاسخ قوی‌تری خواهند داشت، بنابراین، تمامی نورون‌ها و همسایه‌های نورون برنده در سطوح گوناگونی تعدیل می‌شوند. شبکهٔ SOM وزن‌های لایهٔ خروجی را با تعداد زیادی از داده‌های آموزشی تنظیم می‌کند و در نهایت، هر نورون در لایهٔ خروجی به الگوی ردهٔ خاصی از داده‌ها حساس خواهد بود. وقتی ویژگی‌های دو خوشه به هم نزدیک باشند، موقعیت نورون‌های این دو رده نیز نزدیک خواهد بود [۳۷]. بنابراین، SOM آموزش‌داده‌شده نگاشتی را از فضای داده‌ها به فضای دوبعدی در اختیار می‌گذارد و داده‌های مشابه، به یکدیگر نزدیک خواهند بود.

۳-۲-۲- خوشه‌بندی K-میانگین

خوشه‌بندی k-میانگین یکی از رایج‌ترین روش‌های خوشه‌بندی است که در اواخر سال ۱۹۶۰ میلادی ارائه شد. هدف K-میانگین یافتن گروه‌هایی از داده‌ها با ویژگی‌های مشابه و اختصاص آنها به تعداد K خوشه است. این روش، خوشه‌ها را به‌وسیلهٔ مراکز آنها مشخص می‌کند. هر نقطه از داده‌ها به خوشه‌ای اختصاص داده می‌شود که فاصلهٔ مرکز خوشه تا نقطه (مثلا فاصلهٔ اقلیدسی) کمترین مقدار باشد. به عبارت دیگر، الگوریتم K-میانگین تعداد K مرکز را مشخص می‌کند و سپس هر نقطه را به نزدیک‌ترین خوشه اختصاص می‌دهد. کلمه «میانگین» در نام روش به محاسبهٔ میانگین داده‌ها یعنی یافتن مرکز اشاره دارد. در حین آموزش شبکه، میانگین خوشه‌ها در هر تکرار محاسبه و سپس جابه‌جا می‌شود. الگوریتم برای هر خوشه، میانگین وزن‌دار همهٔ داده‌های خوشه را محاسبه می‌کند و به‌عنوان مرکز جدید در نظر می‌گیرد. تعداد خوشه‌ها باید نخست توسط کاربر تعیین شود. K-میانگین یک الگوریتم تکراری است و شامل مراحل زیر است:

۱. مجموعه‌ای از داده‌ها به‌عنوان ورودی به شبکه داده می‌شود و تعداد خوشه‌ها (K) تعیین می‌شود.
۲. تعداد K مرکز به‌صورت تصادفی انتخاب و فاصلهٔ نقاط از مراکز حساب می‌شود. انتخاب نوع فاصله به داده‌ها و تابع هدف بستگی دارد.
۳. هر نقطهٔ داده به نزدیک‌ترین خوشه که کمترین فاصله را از مرکز آن دارد، اختصاص داده می‌شود.
٤. الگوریتم مراکز خوشه‌ها را با محاسبهٔ میانگین جدید به‌روزرسانی می‌کند.

مراحل سه و چهار تا زمان هم‌گرایی تکرار می‌شود؛ یعنی مرحلهٔ آموزش زمانی تمام می‌شود که مراکز خوشه‌ها دیگر تغییر نکند، یا اینکه می‌توان از قبل، تعداد تکرارها را مشخص کرد. پیاده‌سازی الگوریتم K-میانگین ساده است و از بسیاری از الگوریتم‌های خوشه‌بندی سریع‌تر است، اما نقص‌هایی هم دارد [۳۹]. برای مثال، نتیجه خوشه‌بندی تا حد زیادی وابسته به مراکز اولیه است و نخست به تعیین تعداد خوشه‌ها نیاز دارد. همچنین این الگوریتم برای داده‌هایی با ابعاد بالا مناسب نیست.

۳-۲-۳- خوشه‌بندی دومرحله‌ای

همان‌طورکه گفته شد، از برتری‌های الگوریتم K-میانگین پیاده‌سازی آسان و محاسبهٔ سریع است، اما نتایج خوشه‌بندی تحت تأثیر مراکز اولیهٔ خوشه‌ها هستند و این احتمال وجود دارد که نتایج خوشه‌بندی در کمینه‌های محلی گیر بیفتد [۴۰]. یک الگوریتم دومرحله‌ای بر اساس SOM و K-میانگین، از برتری خوشه‌بندی خودکار SOM استفاده می‌کند. برای تعیین سریع‌تر مراکز خوشه‌ها، داده‌های ورودی، نخست، با SOM خوشه‌بندی می‌شوند، نتایج حاصل از SOM به‌عنوان بردار مراکز اولیه خوشه‌ها برای روش K-میانگین استفاده می‌شود. ترکیب دو روش SOM و K-میانگین می‌تواند نقص‌های دو روش را جبران کند و نتایج خوشه‌بندی را بهبود بخشد [۳۸]. آموزش

الگوریتم دومرحله‌ای به‌صورت زیر است:

۱. ابتدا باید الگوریتم SOM پیاده‌سازی و داده‌ها به‌عنوان ورودی به آن ارائه شود. شبکه با این داده‌ها آموزش داده می‌شود و خروجی این مرحله بردارهای وزن شبکه خواهد بود. در این مرحله زمان آموزش را می‌توان کاهش داد. برای مثال، لازم نیست SOM تا هم‌گرایی کامل ادامه دهد؛ و به‌عنوان‌مثال، دویست سیکل کافیست.

۲. بردار وزن حاصل از خوشه‌بندی با SOM به‌عنوان مراکز اولیهٔ خوشه‌ها در خوشه‌بندی با روش K-میانگین در نظر گرفته می‌شود و خوشه‌بندی با K-میانگین انجام می‌شود.

الگوریتم خوشه‌بندی دومرحله‌ای بر اساس SOM و K-میانگین هم ویژگی خودسازمان‌دهی SOM را دارد و هم کارایی بالای روش K-میانگین را داراست. بنابراین مشکل زمان طولانی هم‌گرایی SOM و احتمال انتخاب نادرست مراکز اولیهٔ خوشه‌ها در K-میانگین حل خواهد شد [۳۸].

در این پژوهش، برچسب سندها که به‌صورت دوتایی بودند، به‌عنوان ورودی به شبکه SOM اعمال شدند. شبکه SOM شامل لایهٔ ورودی (ده نورون)، لایهٔ رقابتی و لایهٔ خروجی است که تعداد نورون‌ها در لایهٔ رقابتی و لایهٔ خروجی به تعداد خوشه‌هاست. برای آموزش شبکه از حالت دسته‌ای[1] استفاده شده است. وزن‌های حاصل از آموزش شبکه SOM پس از دویست تکرار، به‌عنوان مراکز اولیهٔ خوشه‌ها در روش K-میانگین در نظر گرفته شدند. خوشه‌بندی برای تعداد خوشه‌ها در بازه یک تا پنجاه خوشه انجام شد و سپس تعداد بهینهٔ خوشه‌ها تعیین شد.

۳-۲-۴- تعیین تعداد بهینهٔ خوشه‌ها

در بسیاری از روش‌های خوشه‌بندی تعیین تعداد بهینهٔ خوشه‌ها مسأله‌ای اساسی است؛ بنابراین، لازم است این شاخص قبل از انجام خوشه‌بندی تعیین شود. این شاخص می‌تواند توسط کاربر بر اساس اطّلاعات قبلی یا به روش ویژه تعیین شود. در این مطالعه، برای ارزیابی تعداد بهینهٔ خوشه‌ها از معیار ارزیابی Davies-Bouldin استفاده شد که برای تعداد بهینهٔ خوشه‌ها مقدار این شاخص حداکثری خواهد بود.

۴- نتایج و بحث

۴-۱- ضریب همبستگی در مجموعهٔ دادهٔ آزمایشی

آزمایش‌ها بر روی مجموعهٔ داده شامل ۶۳۹۹ سند موجود در شبکه‌های اجتماعی انجام شده است. این سندها به زبان فارسی در بحران کرونا و در فضای مجازی منتشر شده‌اند. همان‌طورکه در جدول (۲) مشاهده شد، بیشینهٔ سندها (۳۲ درصد) در این مجموعه‌داده از اینستاگرام هستند؛ درحالی‌که ۲۹ درصد سندها از تلگرام، ۲۵ درصد سندها مربوط به توییتر و دوازده درصد سندها مربوط به اخبار منتشرشده در فضای مجازی است. بررسی‌ها نشان می‌دهد در کل داده‌ها، حدود ۷۲ درصد از پیام‌ها در حوزهٔ بهداشت، فرهنگی و اجتماعی، شش درصد در حوزه مذهبی، سه درصد در حوزهٔ علمی و باقی در حوزهٔ سیاسی یا طنز هستند. پیام‌ها با قطبیت غیرمثبت (منفی یا خنثی) حدود پنجاه درصد کل پیام‌ها هستند و ۲۷ درصد کل پیام‌ها در حوزهٔ اخبار، اطّلاعات و آمار با قطبیت منفی هستند. بررسی پیام‌ها نشان می‌دهد حدود چهل درصد پیام‌ها دربارهٔ عملکرد نهادها و مردم در بحران کرونا است که منجر به احساسات مختلف در جامعه شده است.

در جدول (۳) ضریب همبستگی بین هر زوج احساس مشاهده می‌شود. در هر ستون، ضریب همبستگی بین یک احساس با بقیه احساسات ثبت شده است. به‌عنوان‌مثال، در ستون شادی، ضریب همبستگی بین احساس شادی و باقی احساسات مشاهده می‌شود. همان‌طورکه در ستون شادی مشاهده می‌شود، همبستگی بین احساس شادی و اعتماد از بقیه بیشتر است. درحالی‌که در ستون غم، همبستگی بین غم و ترس و استرس از بقیه بیشتر است و ترس با استرس همبستگی بالایی دارد. در ستون تنفر مشاهده می‌شود که همبستگی بالایی بین تنفر و خشم وجود دارد و همبستگی خطی بین تنفر و استرس ناچیز و حدود صفر است.

۴-۲-نتایج خوشه‌بندی

روش پیشنهادی شاخص ورودی k را دریافت و داده‌ها را در k خوشه گروه‌بندی می‌کند. به‌منظور تعیین تعداد خوشه‌های بهینه، از معیار Davies-Bouldin استفاده شده است که برای k بهینه مقدار معیار مذکور بیشینه می‌شود. به این منظور، مقدار شاخصk در بازه ۱ تا ۵۰ تغییر می‌کند، نتایج محاسبه معیار مذکور در شکل (۳) نشان داده شده است. در این شکل محور افقی مقادیر مختلف k (تعداد خوشه) و محور عمودی مقدار معیار Davies-Bouldin را به‌ازای مقادیر مختلف k نمایش می‌دهد. همان‌طورکه مشاهده می‌شود، بر اساس معیار مذکور، تعداد خوشهٔ بهینه عدد ده است. بنابراین، مقدار شاخص k برابر ده در نظر گرفته شد و روش پیشنهادی با شاخص ورودی ده اجرا، و نتایج آن تحلیل شد.

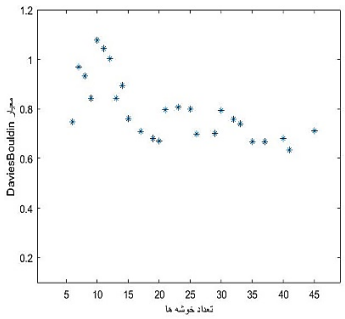
[1] Batch mode

جدول-۳ تعداد سندها و تعداد برچسب‌های احساسات در کل داده‌ها و در هر خوشه را نمایش می‌دهد. به‌عنوان‌مثال، در سطر نخست، کل داده‌ها شامل ۶۳۳۹ سند است؛ که از این تعداد ۴۲۱ سند با برچسب شادی، ۴۳۶ سند با برچسب غم، ۳۳۱ سند با برچسب ترس، ۴۹۸ سند با برچسب تنفر، ۶۳۶ سند با برچسب خشم، ۱۳۸ سند با برچسب شگفتی، ۱۶۰ سند با برچسب اعتماد، ۲۱۰ سند با برچسب پیش‌بینی، ۱۷۶ سند با برچسب استرس و ۱۱۹ سند با برچسب سایر احساسات مشاهده شده است.

همان‌طورکه در جدول (۳) مشاهده می‌شود، ۸۷ درصد سندها در خوشهٔ اول، هفده‌صدم درصد در خوشهٔ دوم، یک و هفت‌دهم درصد در خوشهٔ سوم، چهار و هفت‌دهم درصد در خوشهٔ چهارم، چهارده‌صدم درصد در خوشهٔ پنجم، یازده‌صدم درصد در خوشهٔ ششم، شش‌صدم درصد در خوشهٔ هفتم، چهار و هشت‌دهم درصد در خوشهٔ هشتم، هفتادوهفت‌دهم درصد در خوشهٔ نهم و هجده‌دهم درصد در خوشهٔ دهم قرار گرفته است. در خوشهٔ سوم بیشتر پیام‌ها با احساس غم، ترس و استرس همراه هستند درحالی‌که در خوشه چهارم همه سندها با خشم همراه هستند. در خوشهٔ هشتم همهٔ پیام‌ها حاوی احساس تنفر و خشم هستند. در خوشهٔ نهم نیز، بیشترِ پیام‌ها منجر به احساس غم، تنفر و خشم شده‌اند. همهٔ پیام‌ها در خوشهٔ دهم همراه با احساس تنفر، خشم و استرس هستند.

برای بررسی دقیق‌تر احساسات منفی، خوشه‌های سه، چهار و هشت با جزئیات بیشتر بررسی شدند. بیشتر پیام‌ها در خوشهٔ سوم در حوزهٔ فرهنگی هستند. این خوشه حاوی پیام‌هایی در مورد عملکرد نادرست مردم و سهل‌انگاری در مواجهه با این همه‌گیری، آمار مربوط به تعداد مبتلایان و فوت‌شدگان، و نیز شرایط اقتصادی مردم در شرایط کرونا است که منجر به ایجاد احساس غم، ترس و استرس شده است. پیام‌های این خوشه منجر به کاهش امید و ایجاد وحشت در جامعه شده است و قطبیت پیام‌ها در این خوشه منفی است.

پیام‌های خوشهٔ چهارم شامل اخبار، اطّلاعات و آمار در مورد عملکرد نادرست نهادها در حوزهٔ سیاسی و فرهنگی است. این پیام‌ها شامل راهبردهای متناقض نهادها و دولت‌مردان در مواجهه با همه‌گیری کروناست. به‌عنوان‌مثال، از طرفی مردم به ماندن در خانه دعوت می‌شوند؛ از سوی دیگر، برای سفرهای نوروزی سهمیهٔ بنزین در نظر گرفته و دربارهٔ پیش‌فروش بلیط‌های نوروزی اطلاع‌رسانی می‌شود. پیام‌های ناشی از ناهماهنگی بین نهادها و سازمان‌ها و عدم موفقیت مدیران در مدیریت منابع انسانی در مواجهه با بحران کرونا منجر به احساس خشم در جامعه شده است.



(شکل -۳): مقدار معیار Davies-Bouldin به‌ازای تعداد خوشه‌ها در بازه ۱ تا ۵۰ خوشه.

(Figure- 3): Davies-Bouldin criteria value for the number of clusters in the range of 1-50 clusters.

بیشترِ پیام‌های خوشهٔ هشتم در حوزهٔ سیاسی و با قطبیت منفی هستند که اطّلاعات و اخبار درباره عملکرد نادرست نهادها و مردم را منتشر کرده‌اند. عدم دسترسی به تجهیزات درمانی در بیمارستان‌ها و هجوم مردم به داروخانه‌ها در این پیام‌ها مشاهده می‌شود. نتایج نشان می‌دهد عدم ارائهٔ اطّلاعات درست در مورد مبتلایان و عدم شفاف‌سازی در مورد بیماری از سوی دولت‌مردان منجر به احساس تنفر و خشم در جامعه شده است.

نتایج در خوشه‌های مورد بررسی نشان می‌دهد پیام‌ها با قطبیت غیرمثبت که مربوط به عملکرد منفی نهادها و مردم در حوزهٔ بهداشت و درمان و یا فرهنگی و اجتماعی هستند، روی اعتماد منفی افراد مؤثّر هستند. این نوع پیام‌ها منجر به احساسات منفی مانند ترس، غم، خشم و تنفر در افراد شده است. از سوی دیگر، پیام‌های مربوط به عملکرد مثبت نهادها و مردم در حوزه بهداشت و درمان در ایجاد اعتماد نقشی نداشته‌اند و قطبیت غالبِ آنها خنثی بوده است.

تحلیل پیام‌ها نشان می‌دهد مخاطبان با به دست آوردن سطحی از اعتماد، احساس نگرانی کمتری از شرایط خواهند داشت. دین همواره یکی از عوامل مهم اعتمادساز در جوامع محسوب شده است، و تحلیل سندها در پژوهش حاضر نشان می‌دهد که دین با تأکید بر مؤلّفه‌های هویّت ملی توانایی جلب اعتماد را داشته است. داده‌های پژوهش نشان می‌دهد باوجوداین‌که موضوع ویروس کرونا یک بیماری است و انتظار می‌رود پیام‌های علمی بیش از دیگر موضوعات به اشتراک گذاشته می‌شد، اما پیام‌های منتشرشده در این موضوع قابل تأمل نیست و پیام‌ها بیشتر در موضوعات اجتماعی، فرهنگی، بهداشت و سلامت منتشر شده‌اند و پیام‌هایی که در گفتمان اجتماعی و فرهنگی تولید شده‌اند، بیش از دیگر گفتمان‌ها توانایی جلب اعتماد را داشته‌اند.

(جدول ۲): همبستگی بین هر زوج احساس در مجموعه داده.
(Table- 2): The correlation between each emotion pair in data set.

ردیف	احساس	شادی	غم	ترس	تنفر	خشم	شگفتی	اعتماد	پیش‌بینی	سایر هیجانات
۱	شادی	-	-	-	-	-	-	-	-	-
۲	غم	-0.027	-	-	-	-	-	-	-	-
۳	ترس	-0.048	0.2	-	-	-	-	-	-	-
۴	تنفر	-0.063	0.07	0.01	-	-	-	-	-	-
۵	خشم	-0.08	0.13	0.04	0.52	-	-	-	-	-
۶	شگفتی	-0.009	-0.006	0.01	-0.007	-0.013	-	-	-	-
۷	اعتماد	0.1	-0.02	-0.01	-0.03	-0.03	-0.03		-	
۸	پیش‌بینی	0.04	-0.01	0.025	-0.02	-0.02	-0.01	0.01	-	-
۹	سایر هیجانات	-0.03	-0.03	-0.02	-0.04	-0.04	-0.02	-0.01	-0.02	-
۱۰	استرس	-0.03	0.18	0.23	≈ 0	0.02	0.001	-0.02	-0.01	-0.01

(جدول -۳): سندها و تعداد برچسب‌های احساس در کل داده‌ها و در هر خوشه بر اساس روش پیشنهادی
(Table- 3): Posts and number of labels in data set and in each cluster (according to the proposed method)

گروه	تعداد سند	شادی	غم	ترس	تنفر	خشم	شگفتی	اعتماد	پیش‌بینی	استرس	سایر هیجانات
کل داده	6339	421	436	331	498	636	138	160	219	176	119
خوشهٔ اول	5576	417	231	178	168	0	126	155	199	78	117
خوشهٔ دوم	11	0	11	11	11	11	0	0	0	1	0
خوشهٔ سوم	113	3	89	97	1	0	2	0	5	72	1
خوشهٔ چهارم	299	1	36	15	0	299	5	3	11	0	1
خوشهٔ پنجم	9	0	2	1	0	9	0	0	0	9	0
خوشهٔ ششم	7	0	7	7	0	7	0	0	0	0	0
خوشهٔ هفتم	4	0	4	4	0	4	0	0	0	4	0
خوشهٔ هشتم	259	0	0	17	259	254	4	2	4	0	0
خوشهٔ نهم	49	0	49	0	49	40	1	0	0	0	0
خوشهٔ دهم	12	0	7	1	12	12	0	0	0	12	0

۴-۲-۱- مقایسه با سایر روش‌ها

در این بخش نتایج روش پیشنهادی (ترکیب K-میانگین وSOM) با نتایج حاصل از اعمال هر یک از روش های K-میانگین، SOM و C-میانگین فازی[15] [۴۱] به‌طور مجزا مقایسه می‌شود. C-میانگین فازی یک روش‌خوشه‌بندی است که اجازه می‌دهد هر نقطهٔ داده به چند خوشه با درجات مختلف عضویت، تعلق داشته باشد.

در این بخش، بعد از انجام خوشه‌بندی با روش C-میانگین فازی فرض شده است که هر داده به خوشه با بیشترین درجهٔ عضویّت تعلّق دارد.

[15] Fuzzy C-means

نتایج حاصل از خوشه‌بندی به‌ازای شاخص ورودی۱۰ (۱۰ خوشه) در جداول ۴، ۵ و ۶ بیان شده است. جدول-۴ نتایج خوشه‌بندی با استفاده از K-میانگین را بر روی ۶۳۹۹ سند نمایش می‌دهد. هر سطر از جدول، اطّلاعات یک خوشه شامل تعداد سندها و تعداد برچسب احساسات مشاهده‌شده در آن خوشه را نمایش می‌دهد. نتایج نشان می‌دهد برچسب‌ها در همهٔ خوشه‌ها پراکنده هستند. خوشهٔ اول حاوی برچسب‌های غم و شگفتی است. خوشهٔ دوم حاوی برچسب پیش‌بینی است درحالی‌که خوشهٔ سوم حاوی برچسب شادی است. خوشهٔ پنجم حاوی تنفر است و برچسب استرس در خوشهٔ ششم مشاهده می‌شود. در نهایت، خوشهٔ دهم حاوی برچسب ترس است. درحالی‌که بر اساس جدول-۲ انتظار می‌رود خوشه‌هایی ایجاد شود که برچسب ترس و غم و استرس با هم مشاهده شوند. همچنین، انتظار می‌رود خوشه‌هایی ایجاد شود که به‌طور هم‌زمان حاوی برچسب تنفر و خشم با تعدادی قابل‌ملاحظه باشند که در جدول-۴ چنین خوشه‌هایی مشاهده نشده است. در جدول-۵ نتایج اعمال الگوریتم خوشه‌بندی SOM بیان می‌شود. در جدول-۵ مشاهده می‌شود برچسب غم و ترس، هم‌زمان در خوشهٔ چهارم قرار دارند و در خوشهٔ اول برچسب خشم قابل‌توجه است. سایر برچسب‌ها به‌طور پراکنده در خوشه‌ها قرار گرفته‌اند. در جدول-۶ نتایج اعمال الگوریتم C-میانگین فازی بیان شده است. نتایج نشان می‌دهد برچسب‌ها در همهٔ خوشه‌ها پراکنده هستند. در خوشهٔ دوم تنفر و خشم قابل‌توجه است و پیام‌ها در خوشهٔ چهارم با غم و خشم همراه هستند. در خوشه‌های پنجم و هشتم، احساس غم در پیام‌ها مشاهده شده است. در خوشهٔ دهم پیام‌ها حاوی احساس ترس هستند. حال آنکه در روش پیشنهادی (ترکیب K-میانگین و SOM) رابطهٔ همبستگی بین برچسب‌ها در جدول-۲، در خوشه‌بندی قابل‌مشاهده است و می‌توان ویژگی پیام‌ها و سندهای هر خوشه را بررسی و تحلیل کرد؛ که نشان‌دهندهٔ کارآمدی روش پیشنهادی نسبت به هریک از روش‌های K-میانگین، SOM و C-میانگین فازی به‌طور مجزا است.

۵- نتیجه‌گیری

مطالعهٔ پیام‌ها در شبکه‌های اجتماعی برای درک نگرش عمومی و رفتارها در طول بحران به‌عنوان راهی برای حمایت اجتماع و ارتقای پیام‌های مربوط به سلامت بسیار مؤثّر بوده است. بررسی اطّلاعات در رویارویی با بحران‌هایی مانند کرونا، حیاتی و کلیدی است زیرا شناختی از نظرات، رفتار و حتّی حالات روانی را فراهم می‌کند. مطالعات پژوهشی نشان داده‌اند رسانه‌های اجتماعی می‌توانند به‌عنوان یک منبع داده برای آگاهی از بیماری همه‌گیر ویروس کرونا و همچنین درک نگرش‌ها و رفتارهای عمومی در طول بحران به شمار آیند. از این رو، تمرکز ما در مطالعهٔ حاضر تحلیل داده‌های استخراج‌شده از شبکه‌های اجتماعی، به‌منظور بررسی احساسات جامعه ایران در برابر بحران کرونا بوده است. به این منظور، اطّلاعات فارسی به‌اشتراک‌گذاشته‌شده در سه شبکهٔ اجتماعی اینستاگرام، توییتر، تلگرام و کانال‌های خبری در ایران جمع آوری و برچسب‌گذاری شده است. در ادامه، با استفاده از الگوریتم خوشه‌بندی دومرحله‌ای مبتنی بر شبکهٔ عصبی خودسازمان‌ده و K-میانگین سندها گروه‌بندی شدند و در نهایت، نتایج تحلیل شدند. نتایج نشان داد روش پیشنهادی بر روش‌های متداول خوشه‌بندی مانند K-میانگین، SOM و C-میانگین فازی برتری دارد.

نتایج آزمایش‌ها نشان دادند بیشتر پست‌های جمع‌آوری‌شده در بحران کرونا قطبیت منفی دارد و پیام‌های مربوط به کرونا منجر به احساس‌هایی با قطبیت منفی مانند ترس، خشم، تنفر شده‌اند و در مواردی امید در جامعه کاهش یافته است. پست‌های مربوط به آمار و اطّلاعات بیماری نیز منجر به احساس ترس و استرس در جامعه شده‌اند. در بررسی ضریب همبستگی بین هریک از احساس‌ها با خروجی خوشه‌بندی مشخص شد احساسات خشم و تنفر بیشترین تأثیر را داشته‌اند.

این پژوهش محدودیت‌هایی نیز داشت. به‌عنوان نمونه، تعداد پست‌های مجموعه داده‌ای که در مرحله اول توسط ستاد توسعهٔ علوم‌شناختی منتشر شده است، محدود است. بنابراین، به‌عنوان کارهای پیشِ رو پیشنهاد می‌شود با استفاده از روش‌های رده‌بندی تعداد بیشتری از پست‌های جمع‌آوری‌شده برچسب‌گذاری شوند. در نتیجه می‌توان ویژگی‌های هر خوشه را با دقت بالاتر بررسی و شناسایی کرد. همچنین می‌توان از روش های یادگیری عمیق برای رده‌بندی و برچسب‌گذاری سندها استفاده کرد.

(جدول -۴): نتایج خوشه‌بندی با استفاده از K-میانگین. سندها و تعداد برچسب‌های احساس در کل داده‌ها و در هر خوشه
(Table- 4): the k-means clustering -Posts and number of labels in data set and in each cluster

گروه	تعداد سند	شادی	غم	ترس	تنفر	خشم	شگفتی	اعتماد	پیش‌بینی	استرس	سایر هیجانات
کل داده	6339	421	436	331	498	636	138	160	219	176	119
خوشهٔ اول	4486	0	207	0	0	0	95	0	0	0	116
خوشهٔ دوم	198	0	0	62	0	11	2	8	198	1	0
خوشهٔ سوم	390	390	14	3	6	1	7	33	0	1	0
خوشهٔ چهارم	584	27	93	23	313	584	10	116	0	12	1
خوشهٔ پنجم	167	0	9	4	167	0	4	1	0	0	0
خوشهٔ ششم	160	2	53	14	2	14	4	0	1	160	1
خوشهٔ هفتم	10	1	10	2	0	0	0	0	10	2	0
خوشهٔ هشتم	116	0	4	4	0	2	7	2	0	0	1
خوشهٔ نهم	10	0	0	1	10	3	0	0	10	0	0
خوشهٔ دهم	218	1	46	218	0	21	9	0	0	0	0

(جدول -۵):نتایج خوشه‌بندی با استفاده از SOM. سندها و تعداد برچسب‌های احساس در کل داده‌ها و در هر خوشه
(Table- 5): the SOM clustering -Posts and number of labels in data set and in each cluster

گروه	تعداد سند	شادی	غم	ترس	تنفر	خشم	شگفتی	اعتماد	پیش‌بینی	استرس	سایر هیجانات
کل داده	6339	421	436	331	498	636	138	160	219	176	119
خوشهٔ اول	317	0	58	24	0	317	5	2	3	13	0
خوشهٔ دوم	304	0	36	15	0	0	5	2	11	6	1
خوشهٔ سوم	174	0	9	5	174	0	4	1	7	0	0
خوشهٔ چهارم	405	1	252	201	0	7	0	8	2	0	0
خوشهٔ پنجم	154	2	54	61	1	7	2	0	4	0	1
خوشهٔ ششم	115	0	7	9	0	0	115	7	2	2	0
خوشهٔ هفتم	4191	0	0	0	317	304	0	104	100	154	116
خوشهٔ هشتم	261	0	5	13	0	0	0	0	62	0	1
خوشهٔ نهم	61	61	2	0	1	1	1	36	28	0	0
خوشهٔ دهم	357	357	13	3	5	0	6	0	0	1	0

(جدول -۶): نتایج خوشه‌بندی با استفاده از Fuzzy C-means. سندها و تعداد برچسب‌های احساس در کل داده‌ها و در هر خوشه
(Table- 6): the Fuzzy C-means clustering -Posts and number of labels in data set and in each cluster

گروه	تعداد سند	شادی	غم	ترس	تنفر	خشم	شگفتی	اعتماد	پیش‌بینی	استرس	سایر هیجانات
کل داده	6339	421	436	331	498	636	138	160	219	176	119
خوشهٔ اول	0	0	0	0	0	0	0	0	0	0	0
خوشهٔ دوم	317	0	58	24	317	317	5	2	3	13	0
خوشهٔ سوم	415	415	15	0	6	1	7	36	28	1	0
خوشهٔ چهارم	289	0	36	0	0	90	5	2	10	6	1
خوشهٔ پنجم	0	0	54	0	0	0	0	0	10	0	0
خوشهٔ ششم	153	0	0	0	6	30	2	5	128	1	0
خوشهٔ هفتم	4075	0	0	0	0	0	0	0	15	0	0
خوشهٔ هشتم	243	1	243	0	9	21	7	4	7	18	0
خوشهٔ نهم	540	0	0	0	154	100	101	107	0	76	117
خوشهٔ دهم	307	5	30	307	6	77	11	4	18	61	1

۶- سپاس‌گزاری

نگارندگان بر خود لازم می‌دانند از ستاد توسعهٔ علوم و فناوری‌های شناختی برای به اشتراک گذاشتن داده‌های لازم جهت انجام این پژوهش تشکر نمایند.

۷- مراجع 7-Refrence

[۱] A. Abd-Alrazaq, D. Alhuwail, M. Househ, M. Hamdi, and Z. Shah, "Top concerns of tweeters during the COVID-19 pandemic: infoveillance study," J Med Internet Res, vol. 22, no. 4, p. e19016, 2020.

[2] M. Smith, D. A. Broniatowski, M. J. Paul, and M. Dredze, "Towards real-time measurement of public epidemic awareness: Monitoring influenza awareness through twitter," in AAAI spring symposium on observational studies through social media and other human-generated content, California, USA, 2016.

[3] P. Hitlin and K. Olmstead, "The science people see on social media. Pew Research Center," ed, pp. 142-148, 2018.

[4] A. R. Ahmad and H. R. Murad, "The impact of social media on panic during the COVID-19 pandemic in Iraqi Kurdistan: online questionnaire study," J Med Internet Res, vol. 22, no. 5, p. e19556, 2020.

[5] J. Zarocostas, "How to fight an infodemic," Lancet, vol. 395, no. 10225, p. 676, 2020.

[6] B. Ghanem, P. Rosso, and F. Rangel, "An emotional analysis of false information in social media and news articles," ACM T Internet Techn, vol. 20, no. 2, pp. 1-18, 2020.

[7] K. P. Murphy, "Naive bayes classifiers," University of British Columbia, vol. 18, p. 60, 2006.

[8] M. Salathé and S. Khandelwal, "Assessing vaccination sentiments with online social media: implications for infectious disease dynamics and control," PLoS Comput Biol, vol. 7, no. 10, p. e1002199, 2011.

[9] Y. Lu, X. Hu, F. Wang, S. Kumar, H. Liu, and R. Maciejewski, "Visualizing social media sentiment in disaster scenarios," 24th International Conference on World Wide Web, 2015, USA, pp. 1211-1215, 2015.

[10] K. Chakraborty, S. Bhattacharyya, and R. Bag, "A Survey of Sentiment Analysis from Social Media Data," IEEE Trans Comput Soc Syst, vol. 7, no. 2, pp. 450-464, 2020.

[11] K. Rudra, S. Ghosh, N. Ganguly, P. Goyal, and S. Ghosh, "Extracting situational information from microblogs during disaster events: a classification-summarization approach," 24th ACM International on Conference on Information and Knowledge Management, 2015, Melbourne, Australia, pp. 583-592, 2015.

[12] S. E. Vieweg, "Situational awareness in mass emergency: A behavioral and linguistic analysis of microblogged communications," University of Colorado at Boulder, 2012.

[13] Q. Zhang, F.-Y. Wang, D. Zeng, and T. Wang, "Understanding crowd-powered search groups: a social network perspective," PloS one, vol. 7, no. 6, p. e39749, 2012.

[14] B. Takahashi, E. C. Tandoc Jr, and C. Carmichael, "Communicating on Twitter during a disaster: An analysis of tweets during Typhoon Haiyan in the Philippines," Comput Hum Behav, vol. 50, pp. 392-398, 2015.

[15] N. Bhuvana and I. A. Aram, "Facebook and Whatsapp as disaster management tools during the Chennai (India) floods of 2015," Int J Disast Risk Re, vol. 39, p. 101135, 2019.

[16] L. Li et al., "Characterizing the propagation of situational information in social media during covid-19 epidemic: A case study on weibo," IEEE Trans Comput Soc Syst, vol. 7, no. 2, pp. 556-562, 2020.

[17] A. Mosam, S. Goldstein, A. Erzse, A. Tugendhaft, and K. Hofman, "Building trust during COVID 19: Value-driven and ethical priority-setting," S Afr Med J, vol. 110, no. 6, pp. 0-0, 2020.

[18] A. Oksanen, M. Kaakinen, R. Latikka, I. Savolainen, N. Savela, and A. Koivula, "Regulation and Trust: 3-Month Follow-up Study on COVID-19 Mortality in 25 European Countries," JMIR Public Health and Surveillance, vol. 6, no. 2, p. e19218, 2020.

[19] X. Ji, S. A. Chun, and J. Geller, "Monitoring public health concerns using twitter sentiment classifications," IEEE International Conference on Healthcare Informatics, 2013, Washington, DC, USA, pp. 335-344: IEEE, 2013.

[20] X. Ye, S. Li, X. Yang, and C. Qin, "Use of social media for the detection and analysis of infectious diseases in China," ISPRS Int J Geo-Inf, vol. 5, no. 9, p. 156, 2016.

[21] L. Nemes and A. Kiss, "Social media sentiment analysis based on COVID-19," J Inf Telecommun, pp. 1-15, 2020.

[22] C. E. Lopez, M. Vasu, and C. Gallemore, "Understanding the perception of COVID-19 policies by mining a multilanguage Twitter dataset," arXiv preprint arXiv:2003.10359, 2020.

[23] D. Pastor-Escuredo and C. Tarazona, "Characterizing information leaders in Twitter during COVID-19 crisis," arXiv preprint arXiv:2005.07266, 2020.

[24] R. Kouzy et al., "Coronavirus goes viral: quantifying the COVID-19 misinformation epidemic on Twitter," Cureus, vol. 12 ,no. 3, 2020.

[25] L. Singh et al., "A first look at COVID-19 information and misinformation sharing on Twitter," arXiv preprint arXiv:2003.13907, 2020.

[26] F. Pierri and S. Ceri, "False news on social media: a data-driven survey," ACM Sigmod Record, vol. 48, no. 2, pp. 18-27, 2019.

[27] N. Ruchansky, S. Seo, and Y. Liu, "Csi: A

hybrid deep model for fake news detection," ACM on Conference on Information and Knowledge Management, 2017, Singapore, pp. 797-806, 2017.

[28] K. Popat, S. Mukherjee, A. Yates, and G. Weikum, "Declare: Debunking fake news and false claims using evidence-aware deep learning," arXiv preprint arXiv:1809.06416, 2018.

[29] V. K. Vijayan, K. Bindu, and L. Parameswaran, "A comprehensive study of text classification algorithms," International Conference on Advances in Computing, Communications and Informatics (ICACCI), 2017, Udupi, Karnataka, India, pp. 1109-1113: IEEE, 2017.

[30] B. Liu, E. Blasch, Y. Chen, D. Shen, and G. Chen, "Scalable sentiment classification for big data analysis using naive bayes classifier," IEEE international conference on big data, 2013, Santa Clara, CA, USA, pp. 99-104: IEEE, 2013.

[31] E. Boiy, P. Hens, K. Deschacht, and M.-F. Moens, "Automatic Sentiment Analysis in On-line Text," in ELPUB, 2007, pp. 349-360, 2007.

[32] K. M. Leung, "Naive bayesian classifier," Polytechnic University Department of Computer Science/Finance and Risk Engineering, vol. 2007, pp. 123-156, 2007.

[33] W. Ramadhan, S. A. Novianty, and S. C. Setianingsih, "Sentiment analysis using multinomial logistic regression," International Conference on Control, Electronics, Renewable Energy and Communications (ICCREC), 2017, Piscataway, New Jersey, pp. 46-49: IEEE, 2017.

[34] K. Kowsari, K. Jafari Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, and D. Brown, "Text classification algorithms: A survey," Information, vol. 10, no. 4, p. 150, 2019.

[35] R. Plutchik, "A general psychoevolutionary theory of emotion," in Theories of emotion: Elsevier, 1980, pp. 3-33.

[36] S. Wu, Y. Liu, J. Wang, and Q. Li, "Sentiment analysis method based on Kmeans and online transfer learning," Comput. Mater. Continua, vol. 60, no. 3, pp. 1207-1222, 2019.

[37] Y.-C. Liu, M. Liu, and X.-L. Wang, Application of self-organizing maps in text clustering: a review. chapter, 2012.

[38] V. Khachidze, T. Wang, S. Siddiqui, V. Liu, S. Cappuccio, and A. Lim, "Contemporary research on E-business technology and strategy," in Conference proceedings iCETS, 2012, p. 43: Springer.

[39] Y. P. Raykov, A. Boukouvalas ,F. Baig, and M. A. Little, "What to do when k-means clustering fails: a simple yet principled alternative algorithm," PloS one, vol. 11, no. 9, p. e0162259, 2016.

[40] Y. Kou, H. Cui, and L. Xu, "The Application of SOM and K-Means Algorithms in Public Security Performance Analysis and Forecasting," International Conference on E-business Technology and Strategy, 2012, Tianjin, China, pp. 73-84: Springer, 2012.

[41] J. C. Bezdek, R. Ehrlich, and W. Full, "FCM: The fuzzy c-means clustering algorithm," Computers & geosciences, vol. 10, no. 2-3, pp. 191-203, 1984.

سمیرا عباسی دارای مدرک دکترا در رشته مهندسی پزشکی از دانشگاه صنعتی سهند تبریز است. ایشان اکنون استادیار دانشگاه صنعتی همدان است. یادگیری ماشین، پردازش سیگنال و الگوسازی سامانه‌های بیولوژیکی از زمینه‌های پژوهشی مورد علاقه وی است.

نشانی رایانامهٔ ایشان، عبارت است از:

samira.abbasi@gmail.com

فاطمه امیری دارای مدرک دکترا در رشتهٔ مهندسی نر مافزار از دانشگاه تهران است. ایشان اکنون استادیار دانشگاه صنعتی همدان است. امنیت شبکه، حریم خصوصی، داده‌کاوی و کلان‌داده از جمله علایق پژوهشی وی است.

نشانی رایانامهٔ ایشان، عبارت است از:

fateme.amiri@gmail.com