تشخیص پیام‌های درخواست و غیر درخواست در شبکه‌های اجتماعی با رویکردهای ترکیبی

پردیس مرادبیکی[1]، علیرضا بصیری[*3]

دانشجوی دکترای دانشکدهٔ مهندسی برق و کامپیوتر دانشگاه صنعتی اصفهان، ایران[1]

استادیار دانشکدهٔ برق و کامپیوتر، دانشگاه صنعتی اصفهان، اصفهان، ایران[*2]

چکیده

امروزه با رشد روزافزون استفاده از شبکه‌های اجتماعی، حجم داده‌های تولیدشده روبه‌افزایش است؛ از طرفی کسب‌وکارهای زیادی در شبکه‌های اجتماعی مختلف فعالیت دارند؛ به همین دلیل تشخیص نیازمندی‌های کاربران برای بازاریابان در شبکه‌های اجتماعی یکی از نیازمندی‌های توسعهٔ کسب‌وکارهای اینترنتی و تجارت الکترونیکی است؛ ازاین‌رو تشخیص خودکار پیام‌های درخواست و به‌نوعی فیلترینگ آن‌ها در متون فارسی با اهمیت است. پژوهش حاضر با هدف بهبود تشخیص پیام‌های درخواست در مجموعه پیام‌های ارسال‌شده در پیام‌رسان‌ها انجام شده‌است. امروزه شبکه‌های اجتماعی به‌راحتی در دسترس‌اند؛ ازاین‌رو پیام‌ها در شبکه‌های اجتماعی متفاوت با متون ادبی است. پیام‌ها در شبکه‌های اجتماعی دارای داده‌های اضافی و عامیانه‌اند؛ از طرف دیگر واژه‌ها نیز شامل غلط‌های املایی فراوان‌اند؛ بنابراین مقابله با این پیام‌ها یک چالش محسوب می‌شود. در این پژوهش ابتدا پیش‌پردازش و حذف داده‌های اضافی مورد بررسی قرار گرفته‌است. برای مقابله با دیگر چالش‌های مطرح‌شده روش پیشنهادی در حین حفظ ارزش واژه‌ها، با غلط‌های املایی نیز مقابله کرده‌است. پس از استخراج ویژگی‌های مناسب، یک مدل ترکیبی مبتنی بر شبکه‌های عصبی عمیق برای فرایند تشخیص پیام‌های درخواست و طبقه‌بندی طراحی شد. در مرحلهٔ ارزیابی، آزمایش‌های جامعی برای تحلیل عملکرد مدل پیشنهادی پیاده‌سازی شد. مطابق نتایج به‌دست‌آمدهٔ precision، recall و f-score روش پیشنهادی، به‌طورتقریبی برابر نوددرصد است و در مقایسه با روش‌های پیشین ارائه‌شده، به طور میانگین پنج درصد بهبود یافت.

واژگان کلیدی: تجارت الکترونیکی، تشخیص درخواست، شبکه‌های اجتماعی، پیام‌رسان، روش مبتنی بر یادگیری عمیق

Recognizing request and non-request messages in social networks with combined approaches

Pardis moradbeiki[1], alireza basiri[2*]

Phd Student, Department of Electrical and Computer Engineering, Isfahan University of Technology, Isfahan, Iran[1]

Assistant Professor of Electrical and Computer Engineering Department, Isfahan University of Technology, Isfahan, Iran[2]

Abstract

The aim of the request recognition task in social networks is to understand the intent behind the posts, comments, or messages shared by users. Many businesses are actively present on various social networks, making it crucial to identify user needs for marketers in this space to foster the growth of online businesses and e-commerce. Detecting request messages automatically and filtering them is essential. However, social network messages often contain slang and numerous spelling errors, posing challenges for research in this domain. While extensive research has been conducted in English, studies on this task in Persian are limited. Telegram stands out as the most popular social network in Iran, with

[*] نویسندهٔ عهده‌دار مکاتبات

[*] Corresponding author

• تاریخ ارسال مقاله: ۱۴۰۳/۱/۳۰ • تاریخ پذیرش: ۱۴۰۳/۹/۱۴ • تاریخ انتشار: ۱۴۰۴/۳/۲۸ • نوع مطالعه: پژوهشی

[Downloaded from jsdp.rcisp.ac.ir on 2026-09-05]

[DOI: 10.61186/jsdp.22.1.39]

a large Persian-speaking user base. This study utilized a standard labeled Persian dataset from Telegram for training and testing purposes, comprising 85741 messages from the platform, evenly split between request and non-request categories. To tackle the significant challenges posed by sarcastic messages and spelling mistakes on social media platforms, we devised a multi-step hybrid strategy.
The initial step involves preprocessing. Social media data typically consists of unstructured and slang-ridden user messages, necessitating preprocessing to enhance Persian text processing and reduce slang usage. The pre-processing phase is crucial when dealing with social media platforms. Because Telegram is unique compared to other platforms the data cleaning process varies. This study's accomplishment includes developing a unique dataset and filtering out noise from Telegram enhancing improvement in the pre-processing phase. Also, this involves normalizing different word forms, such as "beautiful" and "beauty," to maintain the integrity of word meanings.
The subsequent step focuses on feature extraction. Various approaches to feature extraction come with their own set of advantages and drawbacks. Hence, we employed hybrid feature extraction methods to address this complexity. While Tf-Idf methods assess word importance without considering meaning, FastText retains semantic similarity. By combining the bag of words and FastText methods, our research aims to enhance accuracy. The final step involves classification, where deep learning networks are utilized to evaluate these features.
Experimental findings indicate that our final model achieves precision, recall, and f-score rates of nearly 90%, representing a 5% improvement on average compared to previous methodologies.

Keywords: e-commerce, request detection, social networks, messaging, deep-learning based method.

۱-مقدمه

امروزه با پیشرفت فناوری، شبکه‌های اجتماعی بسیار محبوب شده‌اند و دسترسی به آن‌ها به راحتی امکان‌پذیر است؛ به‌طوری‌که تا سال ۲۰۲۱ چهار میلیارد و دویست میلیون نفر در شبکه‌های اجتماعی حضور دارند[۱]. این کاربران از شبکه‌های اجتماعی برای ارسال پیام در گروه‌های مختلف استفاده می‌کنند؛ همچنین کسب‌وکارهای زیادی در شبکه‌های اجتماعی مختلف به‌شکل گروه حضور دارند و موضوع فعالیت این گروه‌ها بسیار گسترده‌است. کاربران با توجه به علاقهٔ خود در این گروه‌ها عضو می‌شوند. این محبوبیت در بین مردم، بازاریابان را به‌سمت شبکه‌های اجتماعی جذب کرده‌است؛ از سوی دیگر، داده‌های تولیدشده در این گروه‌ها بسیار زیاد است؛ درنتیجه تجزیه‌وتحلیل تمام داده‌ها دشوار خواهد بود. این مسئله نیاز به داده‌کاوی پیام‌های شبکه‌های اجتماعی را تقویت می‌کند. تجارت الکترونیکی به خریدوفروش محصولات و خدمات در فضای اینترنت گفته می‌شود؛ درواقع هرگونه فعالیت تجاری را در فضای اینترنت که به فروش و انتقال پول ختم شود، تجارت الکترونیکی می‌گویند؛ ازاین‌رو در شبکه‌های اجتماعی تشخیص سریع و صحیح درخواست‌های افراد از سایر پیام‌ها برای کسب‌وکارها و تجارت الکترونیکی امری حیاتی محسوب می‌شود. تشخیص پیام‌های ناخواسته شامل شناسایی و پالایه‌کردن پیام‌های هرز یا نامطلوب است که عمدتاً در سامانه‌های رایانامه یا پیام‌رسان استفاده می‌شود؛ اما در حوزهٔ بازاریابی، این فناوری می‌تواند به تفکیک پیام‌های درخواستی از سایر پیام‌ها کمک کند و در ارتقای فرایند بازاریابی و پیشنهادها مؤثر باشد. پیام‌های شبکه‌های اجتماعی عامیانه و دارای غلط‌های املایی فراوان‌اند که همین مسئله انجام پژوهش در این زمینه را دشوار کرده‌است. در این پژوهش هدف نهایی ما دسته‌بندی پیام‌ها به دو دسته درخواست و عدم درخواست با دقت و صحت مناسب است. برای مدیریت این چالش، کارهای اصلی این پژوهش به شرح زیر خلاصه می‌شوند:

داده‌های استخراج‌شده از شبکه‌های اجتماعی حاوی پیام‌های ارسال‌شده از سمت کاربران است. ماهیت شبکه‌های اجتماعی و در دسترس‌بودن عمومی این شبکه‌ها باعث شده‌است تا بیشتر پیام‌ها عامیانه و بدون ساختار باشند؛ بنابراین نیاز است تا پیش‌پردازش با دقت روی پیام‌ها انجام شود. یکی از تمرکزهای این پژوهش بهبود پیش‌پردازش متون فارسی است تا اثر عامیانه‌بودن پیام‌ها کاهش یابد؛ برای مثال یکی از مراحل پیش‌پردازش به بررسی اینکه واژهٔ «زیبا» به اشکال مختلفی مانند «زیباست»، «زیبایی» نمایش داده شده‌است، می‌پردازد؛ ازاین‌رو برای طبقه‌بندی متون و حفظ ارزش واژه‌ها، باید ریشهٔ آن جایگزین شود.

رویکردهای مختلفی برای کارهای طبقه‌بندی متون مانند روش‌های مبتنی بر یادگیری ماشین که دارای دقت پایین هستند؛ معرفی شده‌است که هر کدام مزایا و معایبی دارند. از طرف دیگر، روش مبتنی بر Tf-Idf با اینکه ارزش واژه‌ها در متن را نشان می‌دهند و ارزشیابی می‌کنند، اما معنا و محتوای واژه‌ها را درنظر نمی‌گیرند؛ همچنین واژه‌ها با غلط‌های املایی زیادی در شبکه‌های اجتماعی وجود دارند که در هنگام استفاده از روش مبتنی بر Tf-Idf برای هر کدام از آن‌ها واژه‌های جداگانه با بردارهای متفاوت در نظر گرفته می‌شوند؛ بنابراین در اینجا باید از رویکردهایی استفاده شود که شباهت معنایی واژه‌ها را نیز حفظ کند. از

[Downloaded from jsdp.rcisp.ac.ir on 2026-09-05]
[DOI: 10.61186/jsdp.22.1.39]

جملهٔ این روش‌ها FastText است؛ همچنین برای پیش‌بینی تشخیص درخواست و عدم درخواست هر پیام، رویکردهای ترکیبی پیشنهاد شده‌است. در این پژوهش برای حل مشکلات مطرح‌شده با ترکیب روش مبتنی بر Tf-Idf و رویکردهای توزیع‌شدهٔ[1] مبتنی بر تعبیهٔ واژه‌ها[2] (روش FastText)، سعی در بهبود دقت و صحت ارزیابی تشخیص پیام‌های درخواستی شد.

تا امروز کارهای زیادی در زمینهٔ تشخیص پیام‌های درخواست در پیام‌رسان‌ها برای زبان فارسی انجام نشده‌است. چهارچوب روش پیشنهادی در این پژوهش کلی است و به‌راحتی می‌توان آن را برای سایر شبکه‌های اجتماعی برخط گسترش داد. بخش دوم به کارهای گذشته اختصاص یافته‌است. در بخش سوم روش پیشنهادی ارائه شد و در بخش چهارم به ارزیابی مدل پیشنهادی پرداخته و درنهایت در بخش پنجم نتیجه‌گیری ارائه شده‌است.

۲- پیشینهٔ پژوهش

با توجه به هدف این پژوهش که ایجاد سامانه‌ای برای هوشمندسازی کسب‌وکارها برای زبان فارسی است، ابتدا مروری بر روی کارهای انجام‌شده در حوزهٔ هوشمندی کسب‌وکار صورت پذیرفت. از آنجا که می‌خواستیم با بررسی روش‌های موجود و کارهای پیشین، به طراحی سامانه‌ای کارآمد برای تشخیص پیام‌های درخواست در زبان فارسی بپردازیم، یک رویکرد کلی برای طراحی سامانه انتخاب و در ادامه به مرور اجمالی پژوهش‌های انجام‌شده در زبان انگلیسی و سپس پژوهش‌های زبان فارسی پرداخته شد.

یکی از حوزه‌های اصلی در زمینهٔ هوشمندی کسب‌وکار، شناسایی موضوعات اصلی مرتبط با آن است. نگاش در سال ۲۰۰۴ هشت موضوع پردازش دادهٔ برخط[3]، انبار داده، مصورسازی[4]، داده‌کاوی[5]، بازاریابی ارتباط با مشتریان[6]، سامانه‌های تصمیم‌یار[7]، سامانه اطلاعاتی سازمانی[8]، مدیریت دانش[9]، سامانه اطلاعاتی جغرافیایی[10] را به‌عنوان موضوعات اصلی هوشمندی کسب‌وکار معرفی کرد[۱۴]. چن و همکاران در سال ۲۰۱۲ در پژوهش خود با عنوان «تحلیل و هوشمندی کسب‌وکار: از کلان‌داده تا اثرات بزرگ»، فرصت‌ها و چالش‌های حوزهٔ هوشمندی کسب‌وکار را تحلیل کردند و پنج حوزهٔ تحلیل داده‌های کلان[11]، تحلیل وب[12]، تحلیل شبکه[13]، تحلیل متن و تحلیل موبایل را موضوعات اصلی هوشمندی کسب‌وکار معرفی کردند[۱۵].

کیاو و همکاران در سال ۲۰۲۳ سامانه‌های تجارت الکترونیکی را پرکاربردترین ابزارهای دیجیتالی معرفی کردند؛ همچنین بیان کردند هوشمندسازی این سامانه‌ها و متن‌کاوی[14] به بهبود سیاست‌های بازاریابی الکترونیکی از طریق پیش‌بینی هدف مشتری، تصمیم‌گیری برای تبلیغات و توصیه به مشتری کمک می‌کند[۴]. در پژوهشی دیگر مورو و همکاران در سال ۲۰۱۵ داده‌کاوی، انباره داده و سامانه‌های پشتیبانی از تصمیم را به‌عنوان زیرحوزه‌های مرتبط با هوشمندی کسب‌وکار معرفی کردند که به‌منظور انجام این امر از روش‌های متن‌کاوی استخراج زیرحوزه‌های مرتبط با هوشمندی کسب‌وکار استفاده کردند[۵]. یانگ و همکاران در سال ۲۰۲۲ تجزیه‌وتحلیل احساسات در متن و متن‌کاوی را برای کارهای تجاری دارای اهمیت دانستند و آن‌را مورد بررسی قرار دادند؛ آن‌ها به این نتیجه رسیدند که تجزیه‌وتحلیل احساسات[15] و متن‌کاوی از اخبار دارای ارزش اقتصادی بالایی است[۶].

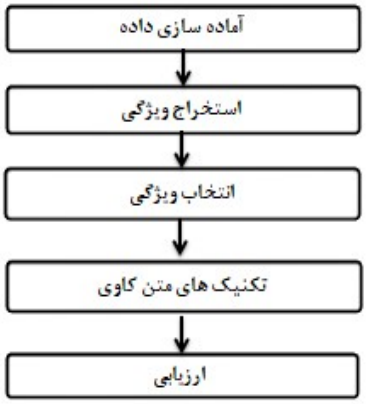
همان‌طورکه بیان شد، متن‌کاوی یکی از حوزه‌های اثرگذار برای هوشمندسازی کسب‌وکارهاست. در این پژوهش از روش‌های متن‌کاوی جهت ارائهٔ مدلی برای پیش‌بینی حوزه‌های مرتبط با هوشمندی کسب‌وکار استفاده شد. متن‌کاوی شامل سه مؤلفهٔ اصلی بازیابی اطلاعات[16]، پردازش اطلاعات[17] و یک‌پارچگی اطلاعات[18] است[۷]. فرایند کلی متن‌کاوی در شکل (۱) نشان داده شده‌است.

۱. آماده‌سازی داده‌ها: شامل تقسیم جملات به واژه‌ها، حذف ایست‌واژه‌ها و ریشه‌یابی واژه‌ها.
۲. استخراج ویژگی: تبدیل متن به کیسه‌واژه‌ها[19] یا مدل‌های برداری برای تحلیل.
۳. انتخاب ویژگی: حذف ویژگی‌های غیرمرتبط برای افزایش دقت و سرعت.
۴. تکنیک‌های متن‌کاوی: استفاده از روش‌های طبقه‌بندی، خوشه‌بندی، کشف موضوع و خلاصه‌سازی بر اساس هدف متن‌کاوی.

[1] Distributed
[2] Word Embedding
[3] On-Line Data Processing (OLAP)
[4] Business intelligence tools and visualization
[5] Data mining
[6] Customer Relationship Management (CRM)
[7] Decision Support Systems (DSS)
[8] Organizational behavior
[9] Process management
[10] Geographic Information Systems (GIS)
[11] (Big) data analytics
[12] Web analytics
[13] Network analytics
[14] Text Mining
[15] Sentiment Analysis
[16] Information retrieval
[17] Information processing
[18] Information integration
[19] Bag-of-Words

[Downloaded from jsdp.rcisp.ac.ir on 2026-09-05]
[DOI: 10.61186/jsdp.22.1.39]

۵. ارزیابی: سنجش نتایج با معیارهایی مانند دقت (accuracy)، دقت مثبت (precision)، بازخوانی (recall) و f-score.



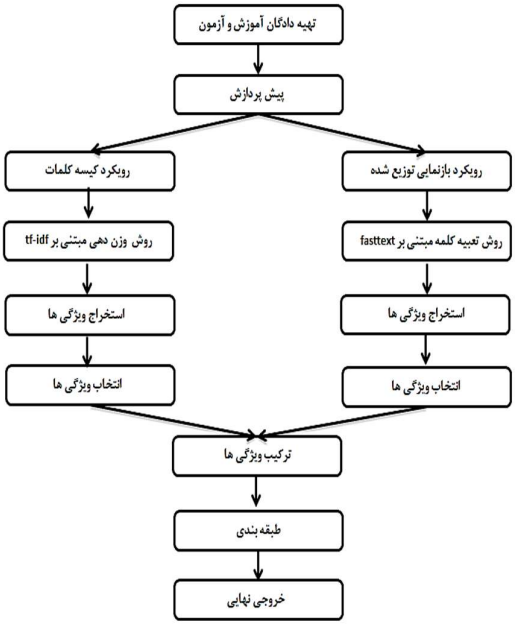
(شکل-۱): فرایند متن‌کاوی [۲و۸]
(Figure-1): Text-mining process

کارهای زیادی در حوزهٔ متن‌کاوی بر روی شبکه‌های اجتماعی برای کسب‌وکارهای الکترونیکی مورد بررسی قرار گرفته‌است؛ برای مثال گریکو و پولی در سال ۲۰۱۹ در شبکهٔ اجتماعی توییتر تأثیر پیام‌ها در مدیریت برند را مورد بررسی قرار دادند[۹]. آخوندی و همکاران در سال ۲۰۱۸ نیز مدیریت زنجیرهٔ تأمین برند تلفن هوشمند را بررسی کردند[۱۰]. در شبکهٔ اجتماعی فیس‌بوک، هی و همکاران در سال ۲۰۱۳ به‌منظور کمک به مشاغل در استفاده از دانش رسانه‌های اجتماعی برای تصمیم‌گیری، استخراج متن را در صفحات فیس‌بوک و توییتر بر روی بزرگ‌ترین زنجیره در صنعت پیتزای ایالات متحده اعمال کردند[۱۱]. یانگ و همکاران در سال ۲۰۲۲ در مطالعهٔ خود تجزیه‌وتحلیل احساسات در مسائل تجاری را با متن‌کاوی مورد بررسی قرار دادند و اعلام کردند چالشی حل نشده‌است. آن‌ها این پژوهش را در دو مرحلهٔ پیش‌پردازش و طبقه‌بندی احساسات انجام دادند. در مرحلهٔ نخست کاهش داده‌های اضافه را مورد بررسی قرار دادند و در مرحلهٔ دوم یک مدل تحلیل متن مبتنی بر شبکه‌های عصبی کانولوشن برای تجزیه‌وتحلیل اخبار در سطح جمله انجام دادند[۶]. سوما و همکاران در سال ۲۰۱۹ رویکرد شبکه‌های عصبی عمیق را با استفاده از RNN و LSTM برای آموزش، داده‌های آرشیو اخبار تامپسون رویترز را از سال‌های ۲۰۰۳ تا ۲۰۱۲ و برای آزمایش پیش‌بینی مدل خود از سال ۲۰۱۳ استفاده کردند[۱۲]. موهان و همکاران در سال ۲۰۱۹ دقت پیش‌بینی سهام را با جمع‌آوری مقدار زیادی از داده‌های سری زمانی و تجزیه‌وتحلیل آن با استفاده از یک مدل یادگیری عمیق در مورد مقالات خبری بهبود دادند[۱۳]. پیک و همکاران (۱۳۹۷) در زمینهٔ تجارت الکترونیکی روی داده‌های استخراج‌شده از تلگرام کار کردند. آن‌ها از مدل مارکوف در این پژوهش برای طبقه‌بندی متن استفاده کردند[۲]. محمدی و همکاران (۱۳۹۸) برای تشخیص پیام‌های درخواست، سه شبکهٔ عمیق را بر روی داده‌های تلگرام مورد بررسی قرار دادند[۱۵].

۳- روش پیشنهادی

شبکه‌های اجتماعی و نرم‌افزارهای پیام‌رسان منبع عظیمی از اطلاعات‌اند. امروزه شبکه‌های اجتماعی بخشی از زندگی روزمرهٔ افراد شده‌اند و کاربران می‌توانند در گروه‌ها با ارسال پیام، نیازهای خود را مطرح کنند. پیام‌ها به واسطهٔ در دسترس‌بودن همگانی شبکه‌های اجتماعی، عامیانه و دارای غلط‌های املایی هستند.

هدف از پژوهش حاضر بهبود عملکرد تشخیص پیام‌های درخواست است. برای هر پیام خروجی یک برچسب «درخواست» و «عدم درخواست» در نظر گرفته شد و از روش‌های متن‌کاوی جهت ارائهٔ مدلی برای پیش‌بینی حوزه‌های مرتبط با هوشمندی کسب‌وکار استفاده شد. مراحل انجام این پژوهش در شکل (۲) نمایش داده شده‌است.



(شکل-۲): چهارچوب مدل پیشنهادی
(Figure-2): The framework of the proposed model

از آنجا که می‌خواهیم در حین حفظ ارزش واژه‌ها، با غلط‌های املایی و عامیانه‌بودن پیام‌ها مقابله کنیم برای پیاده‌سازی سامانه از رویکرد کیسه‌واژه‌ها برای حفظ ارزش واژه‌ها و رویکرد بازنمایی توزیع‌شده برای مقابله با غلط‌های املایی استفاده شد که در ادامه به توضیح جزئیات هر یک می‌پردازیم.

[Downloaded from jsdp.rcisp.ac.ir on 2026-09-05]
[DOI: 10.61186/jsdp.22.1.39]

۳-۱- تهیهٔ داده‌های آموزش و آزمون

نخستین مرحله آماده‌سازی داده‌هاست. داده‌های مورد استفاده در این پژوهش شامل پیام‌های منتشرشده در شبکهٔ اجتماعی تلگرام به زبان فارسی است. با توجه به اینکه هدف، تشخیص درخواست در سطح جمله است، تنها پیام‌هایی با طول دو تا سیصد نماد در داده‌ها وجود دارد. داده‌های مورد استفاده در این پژوهش شامل ۸۵۷۴۱ پیام منتشرشده در شبکهٔ اجتماعی تلگرام است. نمونه‌ای از داده‌های خام در جدول (۱) نشان داده شده‌است.

(جدول-۱): نمونه‌ای از داده‌های خام این پژوهش و برچسب درخواست و عدم درخواست آن‌ها

(Table-1): An example of the raw data of this research and their request and non-request label

شناسهٔ کاربر	متن پیام	برچسب پیام
۱۳۶۱۰۲ ۳۰۹۶	"قاتق آجیل جنس سیلور با قیمت مناسب میخام هر کی داره پی وی"	درخواست
۱۰۸۵۸۱ ۶۰۵۰	♥ افتخار ما رضایت مشتریست ,جنس با کیفیت قیمت مناسب ,ارزان و با کیفیت خریدن هنر است	عدم درخواست

۳-۲-پیش‌پردازش

پیش‌پردازش مهم‌ترین گام در متن‌کاوی و پردازش زبان طبیعی است. در حوزهٔ متن‌کاوی، مرحلهٔ پیش‌پردازش در جهت استخراج بهتر دانش از داده‌های متنی استفاده شد. پیام‌ها در شبکه‌های اجتماعی بدون ساختارند و بیشتر شامل داده‌های اضافی زیادی از جمله شکلک‌ها، واژه‌های غیر انگلیسی و تاریخ هستند. این داده‌های اضافی به متن‌کاوی کمک نمی‌کنند و می‌توانند حذف شوند؛ بنابراین در روند اجرای روش پیشنهادی برای آماده‌سازی متن، عملیاتی از قبیل حذف داده‌های اضافی مورد نیاز است. نمونه‌ای از داده‌های اضافی شامل موارد زیر است:

حذف پیوندها، منشن‌ها و هشتگ‌ها: از آنجا که هدف پژوهش ما تشخیص پیام‌های درخواست بود، پیوندها و تگ‌ها تأثیری در کار ما نداشتند و وجود چنین نمادهایی در متون گیج‌کننده و ارزش لغات بااهمیت را کاهش می‌داد؛ بنابراین حذف شدند.

۱. **حذف اعداد و حروف غیرفارسی:** از آنجا که هدف پژوهش ما کار بر روی داده‌های فارسی است، کلیهٔ اعداد و حروف غیر فارسی از متن پیام‌ها پاک شدند؛ زیرا وجود آن‌ها تأثیر مثبتی در فرایند طبقه‌بندی ندارد.

۲. **حذف کلیهٔ علامت‌ها و نمادهای خاص و شکلک‌ها:** ساده‌ترین روش تشخیص حالت یک نویسنده، مشاهدهٔ شکلک‌هایی است که فرد در متن به‌کار برده‌است. با این شکلک‌های بسیار ساده می‌توان مفاهیم احساسی پیچیده‌ای را انتقال داد. با وجود اینکه شکلک‌ها در متن‌کاوی می‌توانند گاهی بسیار مؤثر باشند، اما به دلایل زیر نادیده گرفته شدند:

- استفادهٔ نادرست کاربر: همهٔ افراد از شکلک‌ها برای بیان منظور خود استفاده نمی‌کنند. این امکان وجود دارد که افراد برحسب عادت، در پایان متن شکلک به‌کار ببرند.
- کاربرد بسیار کم شکلک‌ها: در متن‌های جدی شکلک‌ها کاربرد بسیار کمی دارند؛ همچنین در بحث تجزیه‌وتحلیل احساسات که از حوزه‌های متن‌کاوی است، گاهی شکلک‌ها از پیام‌ها حذف می‌شوند؛ برای مثال کاربران هیجان‌زده یا بسیار عصبانی ممکن است اصلاً از شکلک‌ها استفاده نکنند. مطابق با پژوهش پارک و همکاران در سال ۲۰۱۳ میزان استفاده از شکلک‌ها در محیط‌های رسمی کمتر از ده درصد است. در این پژوهش تمام نمادها و شکلک‌های تلگرام و تلفن همراه از متون پیام حذف شد که برای این کار کلیهٔ نمادها و شکلک‌ها در یک فایل آماده ذخیره و از کلیه پیام‌ها حذف شد.

۳. **حذف حروف اضافه:** حذف واژه‌های اضافی نظیر حروف اضافه، حروف ربط و غیره در مضمون کلی متن تأثیر نمی‌گذارند و باعث خلاصه‌سازی متن می‌شود؛ به همین دلیل می‌توان آن‌ها را از متن پیام‌ها حذف کرد. برای این کار از یک فرهنگ لغت شامل حروف ربط و اضافه استفاده شد و این حروف از متن حذف شدند.

۴. **ریشه‌یابی واژه‌ها:** در پردازش زبان طبیعی گاهی در متون یک واژه می‌تواند به اشکال مختلفی (شکل اسم مفرد و جمع) ظاهر شود؛ درحالی‌که معنای یکسان دارند و دارای ریشهٔ مشترک هستند[۱۷]؛ برای مثال سه واژهٔ کتابها، کتاب‌ها و کتاب ها دارای ریشهٔ مشترک کتاب هستند و برای بررسی و محاسبهٔ احساس نهفته در پیام اضافی هستند؛ بنابراین ضروری است تا تمام واژه‌ها را به واژهٔ ریشهٔ خود تبدیل کنیم. در این مرحله واژه‌ها را مجزا بررسی و ریشهٔ واژه‌ها را جایگزین می‌کنیم. تقی‌زاده و همکاران در سال ۱۳۹۱ یک متد جدید برای ریشه‌یابی واژه‌ها در زبان فارسی معرفی کردند[۴]. در این پژوهش برای ریشه‌یابی واژه‌ها در متون پیام‌ها از این رویکرد استفاده شد. شکل (۳) نمای کلی مراحل پیش‌پردازش در این پژوهش را نشان می‌دهد.

۳-۳-رویکرد کیسه‌واژه‌ها

نخستین مرحله برای درک متن به وسیلهٔ ماشین، بازنمایی متن است. رایج‌ترین مدل بازنمایی متن روش کیسه‌واژه‌ها است[۱۹]. برای ساخت کیسه‌واژه‌ها از مجموعه‌داده‌های

[Downloaded from jsdp.rcisp.ac.ir on 2026-09-05]

[DOI: 10.61186/jsdp.22.1.39]

آموزشی استفاده می‌شود. در این مدل واژگان یا جمله‌ها به‌عنوان کیسه‌واژه نمایش داده می‌شوند و تکرار آن‌ها مورد بررسی قرار می‌گیرد[۲۰]. یکی از رویکردهای رایج مبتنی بر کیسه‌واژه‌ها برای استخراج ویژگی‌ها روش Tf-Idf است.

(شکل-۳): نمای کلی پیش‌پردازش [۲]
(Figure-3): Overview of preprocessing

۳-۳-۱- روش Tf-Idf

قبل از آنکه قادر به اجرای الگوریتم‌های یادگیری ماشین و یادگیری عمیق روی مجموعه‌ای از پیام‌ها باشیم باید بتوان پیام‌ها را به‌عنوان بردارهای دوبه‌دو مقایسه کرد. در این پژوهش از روش فراوانی واژه در مقابل فراوانی سند استفاده شد. مقادیر هر متغیر نسبت به سند با استفاده از روش Tf-Idf محاسبه شد. این روش امتیاز میزان ارتباط واژهٔ T در متن D را طبق فرمول (۱) محاسبه می‌کند.

(۱)

Score (D, T) = Term Frequency (T, D) * log (N / Doc Frequency (T))

Term Frequency: تعداد تکرار واژه در یک پیام را محاسبه می‌کند.

N: تعداد کل پیام‌های موجود در مجموعهٔ هدف است.

Doc Frequency: تعداد تکرار واژه در کل پیام‌های مجموعه است.

به بیانی دیگر این روش، واژه‌ها را بر اساس اهمیت وزن‌دهی می‌کند که فراوانی واژه، نسبت تعداد تکرار یک واژه در پیام به تعداد واژه‌های موجود در پیام است. در مقابل فراوانی پیام لگاریتم، نسبت تعداد پیام‌ها به تعداد پیام‌هایی است که واژهٔ موردنظر را دارد. این روش باعث می‌شود تا واژگانی که در مجموعهٔ کمتری از پیام‌ها هستند، وزن بیشتری و واژگانی که در اغلب پیام‌ها موجودند، وزن کمتری پیدا کنند. به این ترتیب واژه‌های با وزن کمتر را می‌توان حذف کرد که در این پژوهش تعداد واژه‌های استخراج‌شده به سیصد متغیر رسید. به بیانی دیگر روش tf-idf وزن واژگانی را که همیشه فرکانس بالایی دارند، متعادل می‌کند[۲۱]. این روش دارای مزایا و معایبی است که به‌شرح زیر است:

مزایا:

- راهی آسان برای استخراج واژگان کلیدی در یک متن و سنجش منحصربه‌فرد بودن؛
- متعادل‌کردن وزن واژگانی که فرکانس بالایی دارند؛
- آسان‌بودن جهت محاسبه

معایب:

- معنا و مفهوم واژه‌ها را در ظر نمی‌گیرد؛
- ماتریس ویژگی به‌دست‌آمده در این روش، بزرگ و خلوت[1] است.

فرایند بردارسازی کامل به پیکرهٔ متنی وابسته است؛ یعنی واژه‌های یکسان در متن‌های مختلف، بردارهای متفاوتی خواهند داشت.

۳-۳-۲-انتخاب ویژگی

انتخاب ویژگی فرایندی است که در آن ویژگی‌های مرتبط شناسایی و ویژگی‌های غیر مرتبط و تکراری کاهش می‌یابد. هدف از این فرایند مشاهدهٔ ویژگی‌هایی است که مسئله را به‌خوبی و با حداقل ویژگی تشریح می‌کنند[۲۲]. این کار مزایای مختلفی دارد:

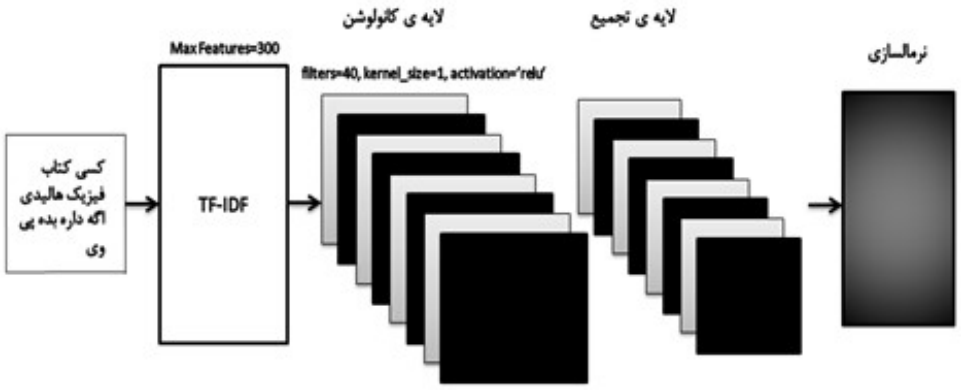
- بهبود کارایی و افزایش دقت الگوریتم‌ها
- درک داده و کسب دانش
- کاهش مجموعه ویژگی‌ها
- افزایش سرعت

در این پژوهش بعد از استخراج ویژگی، از شبکه‌های کانولوشن عمیق[2] برای انتخاب ویژگی استفاده شد. شبکهٔ عصبی کانولوشن یا به‌اختصار CNN در ابتدا از مطالعات قشر بینایی گربه الهام گرفته شد[۲۳]. شبکهٔ عصبی کانولوشن، شبکه‌های عصبی پیش‌خوری[3] هستند که هر نرون در یک لایه، ورودی را از نرون‌های مجاور لایهٔ پیشین دریافت می‌کند. همسایگی‌ها این امکان را به CNNها می‌دهند که الگوهای پیچیده‌تر را به یک روش سلسله‌مراتبی با ترکیب ویژگی‌های سطح پایین ابتدایی و ویژگی‌های سطح بالا بازشناسایی کنند. در متن‌کاوی ویژگی‌های مرتبهٔ بالا (n-gram) می‌توانند از ویژگی‌های مرتبهٔ پایین ساخته شوند. در این صورت ترتیب نه در سطح متن، بلکه در سطح محلی ضروری است. در شبکه‌های اجتماعی، پیام‌های کاربران از

[1] Sparse
[2] Convolutional Neural Network (CNN)
[3] Feed forward neural networks

[Downloaded from jsdp.rcisp.ac.ir on 2026-09-05]
[DOI: 10.61186/jsdp.22.1.39]

نوع عامیانه است و ترتیب واژه‌ها نباید درنظر گرفته شود؛ برای مثال دو پیام «کتاب می‌خواهم» و «کتاب می‌فروشم» را در نظر بگیرید؛ برای تشخیص درخواست از پیام‌های ارسال‌شده از سمت کاربران اهمیت ندارد که «کتاب می‌خواهم» در ابتدا و یا در پایان متن آورده شده باشد؛ بنابراین از شبکه‌های CNN برای مدل



(شکل -۴): معماری مدل نخست پیشنهادی
(Figure-4): Architecture of the first proposed model

پیشنهادی استفاده شد. در شکل (۴) معماری روش پیشنهادی آورده شده‌است. در معماری پیشنهادی تنها از سه لایه کانولوشن، تجمیع[1] و نرمال‌سازی[2] علاوه‌بر لایهٔ رمزنگاری[3] استفاده شد.

لایهٔ کانولوشن: لایهٔ کانولوشن، عملیاتی خطی است که به دنبال آن تبدیل غیر خطی می‌آید. عملیات خطی درواقع ضرب هر نمونه از یک پنجرهٔ یک‌بعدی بر روی متن ورودی است که این عملیات توسط پالایه[4] اعمال می‌شود. پالایه به‌عنوان یک ماتریس از پارامترها نمایش داده می‌شود.

لایهٔ تجمیع: در این لایه خروجی‌ها با هم پیوست می‌شوند.

لایهٔ نرمال‌سازی: از این لایه برای تسریع هم‌گرایی و آموزش بهتر مدل استفاده می‌شود.

در این پژوهش معماری‌های مختلف شبکه‌های عمیق مورد بررسی قرار گرفت. نتایج برای هر معماری جداگانه بررسی و بهترین نتیجه برای مدل پیشنهادی استفاده شد.

۳-۴- رویکرد بازنمایی توزیع‌شده

بازنمایی توزیع‌شدهٔ واژگان یا با نام مستعار تعبیهٔ واژگان از یک مجموعه‌دادهٔ آموزشی مانند ویکی‌پدیا به گونه‌ای آموخته می‌شود که واژه‌های دارای معنی مشابه نزدیک به یکدیگر باشند[۲۱]. در این رویکردها شباهت معنایی بین واژه‌ها حفظ می‌شود؛ به عبارتی دیگر، با بردارهای به‌دست‌آمده می‌توان معنای واژه‌ها را تشخیص داد و میزان شباهت واژه‌های مختلف را با یکدیگر به‌دست آورد. یکی از این رویکردهای استخراج ویژگی، روش FastText است[۲۴].

۳-۴-۱-روش FastText

این روش نخستین‌بار توسط Piotr Bojanowski، در سال ۲۰۱۶ در آزمایشگاه پژوهشی هوش مصنوعی Facebook توسعه داده شد. FastText بر روی شانزده میلیارد توکن مورد آموزش قرار گرفت؛ به این ترتیب، این مدل‌ها واژه‌های زیادی را پوشش داده و برای هر واژه بردار ایجاد می‌کنند[۲۵]. این روش زمان آموزش را نسبت به سایر مدل‌های تعبیهٔ واژه‌ها تا حد زیادی کوتاه و در عین حال اثر طبقه‌بندی را حفظ می‌کند. روش FastText از رویکرد مبتنی بر n-gram برای ساخت مدل زبانی استفاده می‌کند. هر گاه FastText با یک واژهٔ جدید و ناشناخته که پیش از این در دایرهٔ لغات خود نداشت، برخورد کند آن واژه را به n-gram تجزیه می‌کند؛ سپس بردارهای مربوطه را جمع می‌کند تا بردار واژهٔ اصلی را به‌دست آورد. از جمله مزایای روش FastText، می‌توان موارد زیر را نام برد:

- کمک به حل مشکلات خارج از دایرهٔ لغات مانند غلط املایی
- سرعت بسیار بالا در آموزش و استنتاج نتایج

در این پژوهش همانند بخش قبل بعد از استخراج ویژگی با استفاده از روش FastText معماری‌های مختلف شبکه‌های عمیق مورد بررسی قرار گرفت. نتایج برای هر معماری به‌صورت جداگانه بررسی و بهترین نتیجه برای مدل پیشنهادی استفاده شد. معماری پیشنهادی در مدل دوم در شکل (۵) آورده شده‌است. در معماری پیشنهادی تنها از سه لایه کانولوشن تجمیع و نرمال‌سازی علاوه‌بر لایهٔ رمزنگاری[5] استفاده شد.

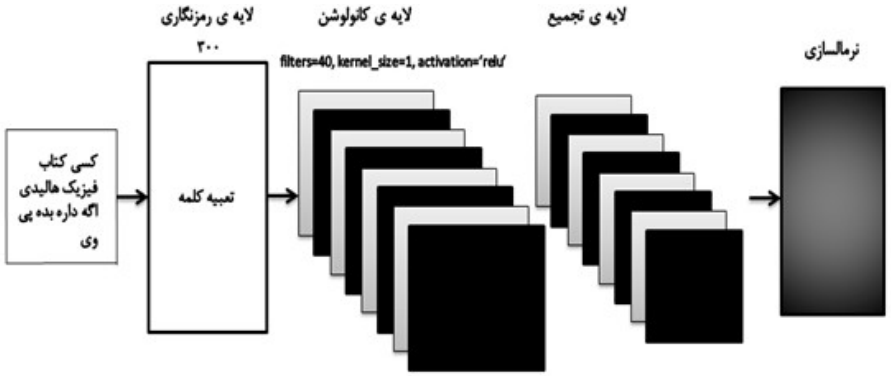
۳-۵- طبقه‌بندی

برای ترکیب دو مدل به‌دست‌آمده، در این مرحله ویژگی‌های استخراج‌شده از هر دو مدل به یک بردار

[1] Pooling
[2] Batch Normalization
[3] Embedding
[4] Filter
[5] Embedding

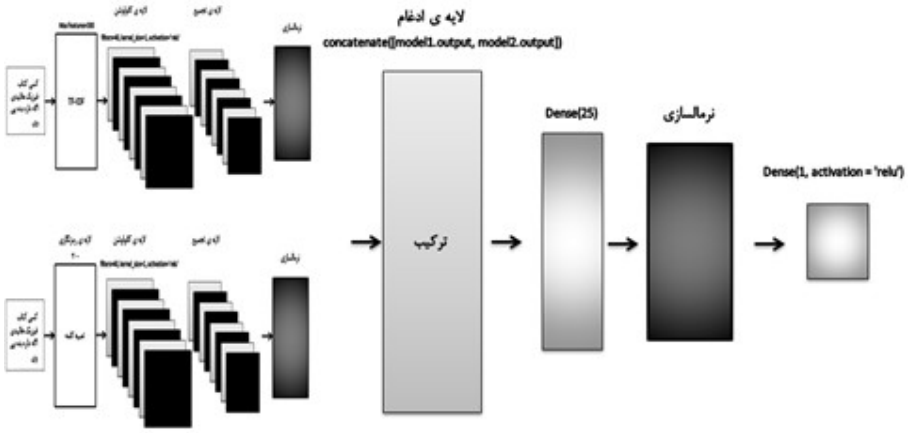
[Downloaded from jsdp.rcisp.ac.ir on 2026-09-05]
[DOI: 10.61186/jsdp.22.1.39]

نگاشت داده شد؛ سپس برای طبقه‌بندی یک شبکهٔ عصبی عمیق مورد استفاده قرار گرفت. معماری نهایی مدل پیشنهادی در شکل (۶) آورده شده‌است.



(شکل -۵): معماری مدل دوم پیشنهادی

(Figure-5): The architecture of the second proposed model



(شکل-۶): معماری مدل پیشنهادی

(Figure-6): Architecture of the proposed model

همچنین برچسب‌هایی که برای این پیام‌ها توسط مدل تخمین زده شد با برچسب واقعی پیام‌ها مورد مقایسه قرار گرفت. در این پژوهش نسبت پیام‌های آموزش به تست هشتاد به بیست است. در این پژوهش چهار معیار اصلی در تجزیه‌وتحلیل متن درنظر گرفته شد. نسبت پیام‌هایی که به درستی طبقه‌بندی شده‌اند به تعداد کل پیام‌ها، accuracy نامیده می‌شود. دو معیار دیگری که برای مقایسهٔ الگوریتم‌های طبقه‌بندی مورد استفاده قرار می‌گیرد، precision و recall است. یکی از معیارهای مهم در حوزهٔ متن‌کاوی precision است. این معیار کسری از پیام‌های بازیابی‌شده‌ای را که مرتبطاند نشان می‌دهد؛ همچنین معیار recall نشان‌دهندهٔ کسری از پیام‌های مربوط بازیابی شده‌است. معیار کاربردی دیگر f-score است. این معیار میانگین هارمونیک دو معیار precision و recall است و در کل میزان تناسب مدل با داده‌ها را نشان می‌دهد.

(۲)

$$precission = \frac{(relevant \cap retrieved)}{retried}$$

(۳)

$$recall = \frac{(relevant \cap retrieved)}{relevant}$$

(۴)

$$F - measure = 2 \times \left(\frac{precision \times recall}{precision + recall}\right)$$

۴- ارزیابی

در این بخش ابتدا مراحل انجام کار توضیح داده، سپس تأثیر مرحلهٔ پیش‌پردازش بررسی می‌شود. در پایان نتایج ارزیابی مدل پیشنهادی با استفاده از رویکردهای مبتنی بر یادگیری ماشین[1] و رویکردهای مبتنی بر یادگیری عمیق[2] ارائه و همچنین نتایج مدل پیشنهادی پژوهش حاضر نسبت به کارهای پیشین ارزیابی می‌شود.

[1] Machine Learning
[2] Deep Learning

[Downloaded from jsdp.rcisp.ac.ir on 2026-09-05]
[DOI: 10.61186/jsdp.22.1.39]

۴-۱-مراحل انجام کار

مجموعه‌داده‌های مورد استفاده در این پژوهش شامل ۸۵۷۴۱ پیام درج‌شده توسط کاربران در گروه‌های مختلف است. پیش‌پردازش نخستین گام در فرایند متن‌کاوی است. پس از مراحل پیش‌پردازش داده‌های مورد استفاده از تعداد ۸۵۷۴۱ پیام به ۷۶۷۸۷ کاهش یافت. تعداد پیام‌های مورد استفاده در این پژوهش در جدول (۲) نشان داده شده‌است. همانند کارهای پیشین از این تعداد به‌صورت تصادفی پیام‌ها به نسبت هشتاد به بیست برای داده‌های آموزشی و آزمون تقسیم شد.

(جدول-۲): تعداد پیام‌های مورد استفاده در این پژوهش
(Table-2): The number of messages used in this study

تعداد کل پیام‌ها	۷۶۷۸۷
برچسب «درخواست»	۳۹۲۱۸
برچسب «عدم درخواست»	۳۷۵۶۹

در مرحلهٔ دوم پس از پیش‌پردازش در مدل نخست پس از تبدیل داده‌های متنی که ساختارنیافته‌اند به داده‌های ساختاریافته با استفاده از روش منتخب در استخراج ویژگی یعنی tf-idf، واژگان جهت استفاده در مدل‌سازی استخراج شدند. پس از این مرحله واژگان کلیدی استخراج‌شده برای کاهش ویژگی، شبکهٔ عصبی عمیق CNN مورد استفاده قرار گرفت. در مدل دوم برای تعبیهٔ واژه‌ها از روش استخراج ویژگی FastText و برای کاهش ویژگی شبکهٔ عصبی عمیق از CNN استفاده شد. پس از این مراحل برای افزایش کارایی و استفاده از فواید هر دو شبکه این دو مدل ترکیب شدند و برای طبقه‌بندی، شبکهٔ عصبی عمیق مورد استفاده قرار گرفت.

۴-۲- پیش پردازش

در این بخش تأثیر مرحلهٔ پیش‌پردازش بررسی شد. در جدول (۳) مدل پیشنهادی بدون مرحلهٔ پیش‌پردازش مورد ارزیابی قرار گرفت. نتیجهٔ ارزیابی مدل طراحی‌شده همان طور که انتظار می‌رفت نسبت به مدل بدون مرحلهٔ پیش‌پردازش بالا بود.

(جدول-۳): تأثیر بخش پیش‌پردازش دادگان در سه حالت مجزا
(Table-3): The effect of data preprocessing in three different modes

	A	P	R	F1
مدل بدون پیش-پردازش	۰.۸۹۲۴	۰.۸۹۱۶	۰.۸۹۱۵	۰.۸۹۱۵
مدل پیشنهادی	۰.۹۰۶۰	۰.۹۰۶۱	۰.۹۰۶۰	۰.۹۰۶۰

۴-۳- آزمایش‌ها

در این بخش به ارائه و بررسی نتایج آزمایش روی بخش‌های مختلف روش پیشنهادشده می‌پردازیم. جهت ارزیابی هر یک از مدل‌ها، از چهار معیار Accuracy، Precision، Recall و F-score استفاده شد.

۴-۳-۱-رویکردهای مبتنی بر یادگیری ماشین

در این مرحله بعد از استخراج ویژگی با استفاده از روش tf-idf برای طبقه‌بندی از روش‌های یادگیری ماشین استفاده شد. دو نوع عمدهٔ روش‌های یادگیری ماشین، یادگیری بدون نظارت[1] و یادگیری با نظارت[2] است. در روش با نظارت داده‌ها شامل برچسب‌اند و با استفاده از این داده‌های برچسب‌گذاری‌شده مدل مورد نظر آموزش داده می‌شود؛ اما در روش‌های بدون نظارت داده‌ها برچسبی ندارند. در این پژوهش از برخی از روش‌های یادگیری ماشین با نظارت نظیر ماشین بردار پشتیبان[3]، نایوبیز[4] و جنگل تصادفی[5] استفاده شد. در ادامه توضیح مختصری از هر یک بیان شده‌است.

- **ماشین بردار پشتیبان:** هدف از این طبقه‌بندی، تعیین جداکننده در فضای جست‌وجو است که می‌تواند بهترین طبقه‌های مختلف را جدا کند. بهترین جداکننده بین کلاس‌ها باید بزرگ‌ترین فاصلهٔ طبیعی را با هر یک از نقاط داده‌ای داشته باشد. این روش داده‌ها را تحلیل می‌کند و مرزهای تصمیم‌گیری را با داشتن ابر صفحات تعریف می‌کند. در جدول (۴) استفاده از الگوریتم ماشین بردار پشتیبان برای طبقه‌بندی به‌ازای تخمین‌سازهای[6] مختلف آورده شده‌است. نتایج نشان داد در linear می‌توان به بهترین دقت در داده‌های خود دست یافت.
- **نایوبیز:** این روش یکی از رایج‌ترین و ساده‌ترین روش‌های طبقه‌بندی احتمالاتی است. این روش احتمال پیشین یک طبقه‌بندی را بر اساس توزیع واژه‌ها در پیام محاسبه می‌کند. در این روش از قضیهٔ بیز برای محاسبه و پیش‌بینی احتمال یک ویژگی مشخص‌شده متعلق به یک طبقه‌بندی خاص استفاده می‌کند. این روش بر پایهٔ احتمالات است و ویژگی‌ها مستقل از هم‌اند. در جدول (۵) استفاده از الگوریتم نایوبیز برای طبقه‌بندی به‌ازای مقادیر مختلف alpha آورده شده‌است. نتایج نشان می‌دهد در alpha=0.2 می‌توان به بهترین دقت رسید.
- **جنگل تصادفی:** جنگل تصادفی مجموعه‌ای از درخت‌های تصمیم است. درخت تصمیم، فضای دادهٔ

[1] Unsupervised Learning
[2] Supervised Learning
[3] Support Vector Machines (SVM)
[4] Naive Bayes
[5] Random Forest
[6] Estimators

[Downloaded from jsdp.rcisp.ac.ir on 2026-09-05]
[DOI: 10.61186/jsdp.22.1.39]

آموزشی را سلسله‌مراتبی تقسیم می‌کند که در آن شرط بر روی مقدار ویژگی‌ها برای تقسیم داده‌ها استفاده می‌شود. شرط پیش‌فرض، وجود یا نبود یک یا چند واژه است؛ تقسیم فضای داده به‌صورت بازگشتی انجام می‌شود تا گره‌های برگ حاوی حداقل تعداد رکوردهای لازم برای طبقه‌بندی باشد. جنگل تصادفی یک الگوریتم ترکیبی است که از مجموعه‌ای از درخت‌های تصمیم برای مقابله با بیش‌برازش[1] استفاده می‌کند. الگوریتم درخت تصادفی روی نمونه‌هایی از داده درختان تصمیم می‌سازد و سپس از هر کدام پیش‌بینی می‌گیرد؛ درنهایت با رأی‌گیری بهترین راه‌حل را انتخاب می‌کند. در جدول (۶) استفاده از الگوریتم جنگل تصادفی برای طبقه‌بندی آورده شده‌است.

(جدول-۴): بررسی مقادیر مختلف kernel در الگوریتم ماشین بردار پشتیبان

(Table-4): Examining different kernel values in support vector machine algorithm

	A	P	R	F1
Linear Support Vector Classification	0.8767	0.8769	0.8768	0.8767
Sigmoid Support Vector Classification	0.7839	0.7839	0.7838	0.7838
Nu-Support Vector Classification	0.8609	0.8612	0.8611	0.8609

(جدول-۵): بررسی مقادیر مختلف پارامتر alpha در الگوریتم نایوبیز

(Table-5): Examining different values of alpha parameter in Newby's algorithm

	A	P	R	F1
alpha=0	0.7601	0.7611	0.7606	0.7601
alpha=0.2	0.7677	0.7679	0.7678	0.7677
alpha=0.4	0.7577	0.7580	0.7579	0.7577
alpha=0.6	0.7622	0.7625	0.7622	0.7621
alpha=0.8	0.7598	0.7600	0.7598	0.7597
alpha=1	0.7610	0.7612	0.7611	0.7610
alpha=2	0.7587	0.7590	0.7588	0.7586
alpha=4	0.7564	0.7567	0.7566	0.7564
alpha=6	0.7546	0.7550	0.7549	0.7546
alpha=8	0.7666	0.7667	0.7666	0.7665

(جدول-۶): بررسی معیارهای ارزیابی در روش جنگل تصادفی

(Table-6): Examining evaluation criteria in random forest method

	A	P	R	F1
Random Forest	0.8505	0.8539	0.8515	0.8503

۴-۳-۲-رویکردهای مبتنی بر یادگیری عمیق

یادگیری عمیق به‌عنوان یک روش یادگیری بسیار قدرتمند ظاهر شده‌است. شبکه‌های مبتنی بر یادگیری عمیق چندین لایه از ویژگی‌ها را یاد می‌گیرند؛ سپس نتایجی را تولید می‌کنند. یادگیری عمیق در بسیاری از دامنه‌های کاربردی مانند بینایی ماشین و تشخیص صوت به موفقیت بسیار خوبی دست یافته است و این شبکه‌ها در متن‌کاوی نیز محبوب شده‌اند. در این پژوهش برخی از روش‌های یادگیری عمیق نظیر شبکه‌های عصبی بازگشتی[2] و شبکه‌های کانولوشن مورد بررسی قرار گرفت. در هر کدام از شبکه‌ها سه معماری بررسی شده‌است. در معماری سوم تنظیم و انتخاب بهترین پارامترها انجام گرفته‌است. در هر قسمت مقدار epochs برابر پانزده و مقدار batch-size برابر ۱۲۸ درنظر گرفته شد. در این قسمت به توضیح مختصر از هر کدام و سپس ارزیابی آن‌ها می‌پردازیم.

- **شبکه‌های عصبی بازگشتی:** شبکه‌های عصبی بازگشتی یا به‌اختصار RNN نوعی از شبکه‌های عصبی هستند. در این شبکه‌ها ارتباط بین نرون‌ها به صورت حلقهٔ جهت‌دار است. این شبکه‌ها به طور اختصاصی برای کار بر روی دنباله‌ها به وجود آمدند[۲۶]. شبکه‌های RNN از حافظهٔ داخلی خود برای پردازش دنباله‌ای از ورودی‌ها استفاده می‌کنند. یک RNN را می‌توان همانند یک زنجیره از لایه‌های عصبی ساده که برخی پارامترها را به اشتراک می‌گذارند دید. به بیانی دیگر یک کار مشترک را در هر گام، اما با ورودی‌های متفاوت اجرا می‌کنند. یک نوع خاص شبکه‌های RNN، شبکه‌های حافظهٔ کوتاه‌مدت طولانی[3] یا به‌اختصار LSTM است. درعمل هنگام انتخاب RNNها از واحد LSTM استفاده می‌کنند؛ زیرا این واحدها به صورتی طراحی شده‌اند که مسئلهٔ انفجار یا ناپدیدشدن گرادیان[4] را حل می‌کنند و اطلاعات را در بازه زمانی طولانی‌تری به خاطر می‌سپارند[۲۷]؛ به بیانی دیگر واحد LSTM، میزان اطلاعاتی که از گذشته بایستی به خاطر سپرده شوند، میزان اطلاعاتی که از ورودی جاری باید در حافظه نوشته شود و میزان اطلاعاتی را که باید به گام زمانی بعدی و لایه‌های بالاتر داده شود، می‌تواند کنترل کند. در این قسمت ابتدا استخراج ویژگی با روش tf-idf انجام، سپس از شبکه LSTM برای طبقه‌بندی استفاده شد. در جدول (۷) سه معماری مختلف مورد بررسی قرار گرفت. جزئیات هر کدام از معماری‌ها در شکل (۷) نمایش داده شده‌است. معماری سوم بهترین نتیجه را از خود نشان داد.
- **شبکه‌های کانولوشن:** سرعت رشد شبکهٔ کانولوشنی بسیار زیاد است؛ به‌طوری‌که در مدت کوتاهی، در بسیاری از زمینه‌های بینایی رایانه مانند شناسایی عمل انسان، تشخیص اشیا، شناسایی چهره و ردیابی به نتایج

[1] Overfitting

[2] Recurrent Neural Network (RNN)
[3] Long-Short Term Memory (LSTM)
[4] Vanishing and Exploding gradients

[Downloaded from jsdp.rcisp.ac.ir on 2026-09-05]
[DOI: 10.61186/jsdp.22.1.39]

بسیار خوبی رسیده‌است. با سیطره بر بینایی رایانه، شبکهٔ کانولوشنی در سایر زمینه‌های هوش مصنوعی مانند پردازش زبان و گفتار نیز وارد شد و نتایج خوبی به‌دست آمد. در این پژوهش همانند بخش پیشین بعد از استخراج ویژگی با روش tf-idf از شبکه‌های کانولوشن استفاده می‌شود.

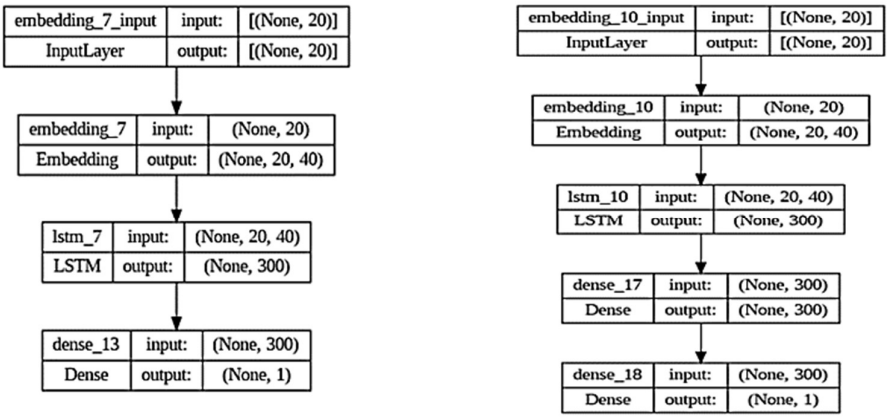
شبکه‌های کانولوشن در جدول (۸) به ازای معماری‌های مختلف بررسی شد. جزئیات هر کدام از معماری‌ها در شکل (۸) نمایش داده شده‌است.

(جدول-۷): بررسی چند معماری مختلف برای مدل نخست

(Table-7): Examining several different architectures for the first model

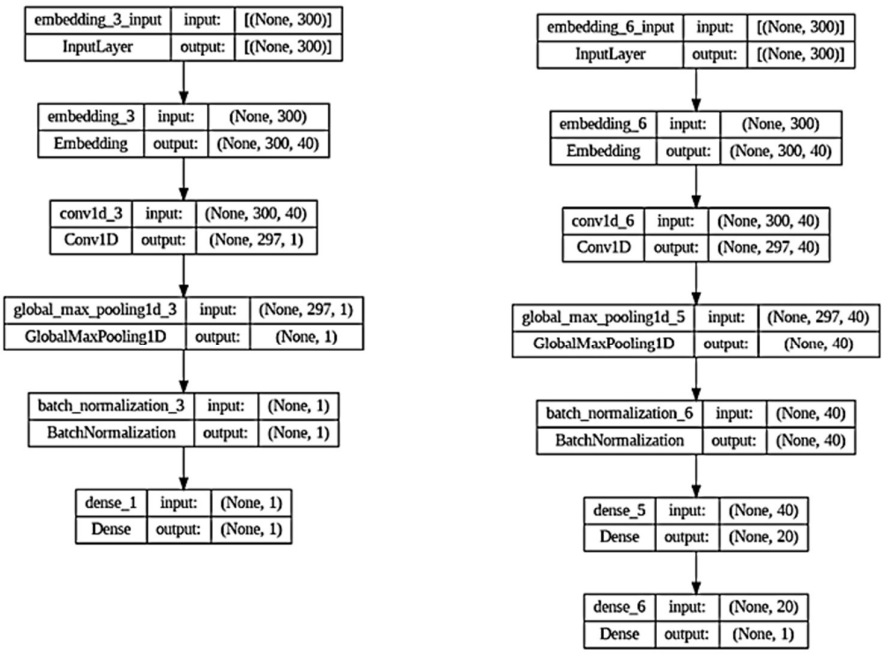
	A
معماری نخست	۰.۵۶۱۹
معماری دوم	۰.۶۸۱۸
معماری پیشنهادی	۰.۷۹۳۴

در هر دو نمونه معماری آورده‌شده تابع فعال‌ساز relu مورد استفاده قرار گرفت. در هر قسمت مقدار epochs با سعی و خطا برابر ده به دست آمد. مقدار batch_size برابر با ۱۲۸ درنظر گرفته شد.



(شکل-۷): دو معماری دیگر ارزیابی‌شده با مدل نخست (معماری نخست سمت راست، معماری دوم سمت چپ)

(Figure-7): Two other architectures evaluated with the first model (the first architecture on the right, the second architecture on the left)



(شکل-۸): دو معماری دیگر ارزیابی‌شده با مدل دوم (معماری نخست سمت راست، معماری دوم سمت چپ)

(Figure-8): Two other architectures evaluated with the first model (the first architecture on the right, the second architecture on the left)

[Downloaded from jsdp.rcisp.ac.ir on 2026-09-05]

[DOI: 10.61186/jsdp.22.1.39]

(جدول-۸): بررسی چند معماری مختلف با مدل دوم

(Table-8): Examining several different architectures with the second model

	A
معماری نخست	۰.۴۹۱۰
معماری دوم	۰.۵۹۲۱
معماری پیشنهادی	۰.۸۹۹۳

۴-۳-۳- ارزیابی مدل پیشنهادی

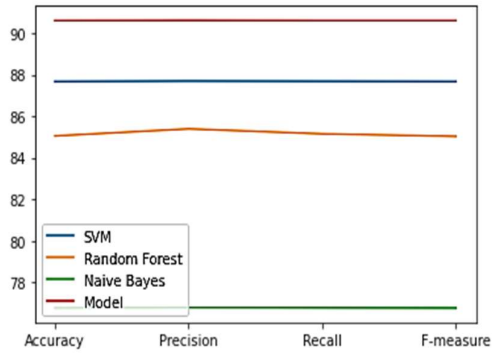
در این بخش مدل پیشنهادی با رویکردهای مبتنی بر یادگیری ماشین و مقالات منتشرشده مورد بررسی قرار گرفت. پس از تحلیل پارامترهای مختلف در هر سه الگوریتم یادگیری ماشین، بهترین نتایج در جدول (۹) با مدل پیشنهادی مقایسه شد؛ همان طور که از نتایج ارزیابی نمایان است در روش پیشنهادی بهترین نتیجه به‌دست آمد.

(جدول-۹): ارزیابی الگوریتم‌های طبقه‌بندی با مدل پیشنهادی

(Table-9): Evaluation of classification algorithms with the proposed model

	A	P	R	F1
SVM	0.8767	0.8769	0.8768	0.8767
Naive Bayes	0.7677	0.7679	0.7678	0.7677
Random Forest	0.8505	0.8539	0.8515	0.8503
Our Model	0.9060	0.9061	0.9060	0.9060

مطابق با شکل (۹) روش پیشنهادی نسبت به دیگر رویکردهای مبتنی بر یادگیری ماشین به معیارهای ارزیابی بالاتری دست یافت.



(شکل-۹): مقایسهٔ روش پیشنهادی با روش‌های مبتنی بر یادگیری ماشین

(Figure-9): Comparison of the proposed method with methods based on machine learning

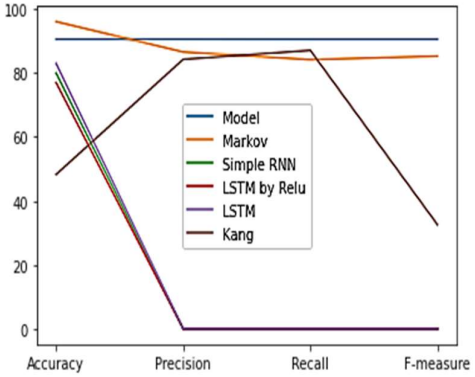
مقالات محدودی در این حوزه بر روی شبکه‌های اجتماعی به زبان فارسی منتشر شده‌است. تا کنون سه مقاله بر روی داده‌های استفاده‌شده در این پژوهش، منتشر شده‌است. در جدول (۱۰) معیارهای ارزیابی مدل پیشنهادی نسبت به کارهای گذشته نشان داده شده‌است. نتایج بیان‌کنندهٔ رسیدن به معیارهای ارزیابی بالاتر نسبت به کارهای گذشته بود.

(جدول-۱۰): مقایسه روش پیشنهادی با سایر روش‌ها

(Table-10): Comparison of the proposed method with other methods

	A	P	R	F1
Our Model	۰.۹۰۶۰	۰.۹۰۶۱	۰.۹۰۶۰	۰.۹۰۶۰
Simple RNN [3]	۰.۸۰	گزارش نشده	گزارش نشده	گزارش نشده
by Relu [3] LSTM	۰.۷۷	گزارش نشده	گزارش نشده	گزارش نشده
LSTM [3]	۰.۸۳	گزارش نشده	گزارش نشده	گزارش نشده
Markov [2]	۰.۹۶۱	۰.۸۶۶	۰.۸۴۲	۰.۸۵۳
Kang [2]	۰.۴۸۳	۰.۸۴۳	۰.۸۷۱	۰.۳۲۶

مطابق با شکل (۱۰) روش پیشنهادی در سه معیار precision، recall و f-score بالاترین مقادیر را از خود نشان داد.



(شکل-۱۰): مقایسه روش پیشنهادی با روش های پیشین

(Figure-10): Comparison of the proposed method with previous methods

۵- نتیجه‌گیری

با پیشرفت فناوری شبکهٔ اطلاع‌رسانی جهانی و رشد آن، حجم عظیمی از داده‌های موجود در در این بستر برای کاربران شبکهٔ رایانه‌ای جهانی مبادلهٔ اطلاعات وجود دارد و داده‌های زیادی نیز تولید می‌شود. محبوبیت شبکه‌های اجتماعی مانند توییتر، فیس‌بوک و تلگرام به‌سرعت روبه‌افزایش است؛ زیرا این امکان را به مردم می‌دهند تا نظرات خود را در مورد موضوعات به اشتراک بگذارند و ابراز، با جوامع مختلف گفت‌وگو یا پیام‌هایی را در سراسر جهان ارسال کنند. شبکه‌های اجتماعی به‌عنوان منبع مهمی برای تجزیه‌وتحلیل، از نظر ماهیت ساختاریافته نیستند؛ بنابراین به پردازش‌هایی نظیر طبقه‌بندی یا خوشه‌بندی نیاز دارند تا اطلاعات معنی‌دار برای استفاده‌های بعدی را ارائه دهند. تجزیه‌وتحلیل پیام‌های شبکه‌های اجتماعی هدف این پژوهش بود؛ برای انجام این کار ابتدا از داده‌های موجود در دو مقالهٔ منتشرشده استفاده شد؛ سپس در مرحلهٔ پیش‌پردازش به روش‌های

[Downloaded from jsdp.rcisp.ac.ir on 2026-09-05]

[DOI: 10.61186/jsdp.22.1.39]

مختلف سعی در بهبود دقت کار شد. از آنجا که پژوهش حاضر بر روی داده‌های شبکه‌های اجتماعی و چالش اصلی این پژوهش وجود پیام‌های عامیانه و غلط املایی فراوان بود؛ بنابراین مدل پیشنهادی در حین حفظ ارزش واژه‌ها باید با این چالش نیز مقابله کند. به همین سبب دقت در روش‌های مبتنی بر کیسه‌واژه‌ها با وجود محاسبهٔ ارزش واژه‌ها در پیام‌ها پایین است؛ ازاین‌رو در این پژوهش روش‌های ترکیبی ارائه شد. بعد از مرحلهٔ پیش‌پردازش ترکیبی از روش‌های مبتنی بر کیسه‌واژه‌ها مانند tf-idf و روش‌های مبتنی بر بازنمایی توزیع‌شده همچون FastText استفاده شد و عملکرد مدل پیشنهادی بر روی داده‌های شبکه‌های اجتماعی با کارهای پیشین مقایسه شد. نتایج نشان‌دهندهٔ دقت و صحت بالاتر مدل پیشنهادی نسبت به کارهای پیشین بود.

6-References ۶-مراجع

[۱] خبرگزاری جمهوری اسلامی (۲۰۲۱) ، آخرین وضعیت شبکه‌های اجتماعی، https://www.irna.ir/news/

[1] https://www.irna.ir/news/ (2021)

[۲] آ. پیک، م. زارع چاهوکی، م. آقاصرام، «تشخیص پیام‌های درخواست در پیام‌رسان تلگرام مبتنی بر اثربخشی کاهش وضعیت ها در مدل مخفی مارکوف»، *دومین کنفرانس بین‌المللی پژوهش‌های کاربردی در علوم برق و کامپیوتر*، ۱۳۹۷.

[2] A. peyk, M. Zare Chahooki, "Detection of request messages in Telegram messenger based on the effectiveness of reducing situations in the hidden Markov model", *Second International Conference on Applied Research in Electrical and Computer Science*, 2019.

[۳] ع. محمدی، م. رضاییان، «تعیین قطبیت نظرات کاربران و تشخیص درخواست ها با کمک تکنیک‌های یادگیری عمیق در تلگرام»، *چهارمین کنفرانس ملی تحقیقات کاربردی در مهندسی برق، مکانیک*، کامپیوتر و فناوری اطلاعات، ۱۳۹۷.

[3] A. Mohammadi, M. Rezaee, "Determining the polarity of users' opinions and recognizing requests with the help of deep learning techniques in Telegram", *Fourth National Conference on Applied Research in Electrical Engineering, Mechanics, Computer and Information Technology*, 2019.

[۴] م. دیانتی، م. صدرالدینی، ا. راسخ، ح. تقی‌زاده، «روشی مستقل از زبان جهت ریشه‌یابی کلمات با استفاده از معیار شباهت»، *یازدهمین کنفرانس سراسری سیستم های هوشمند*، ۱۳۹۱.

[4] M. Dianati, M. Sadredini, "A language-independent method for rooting words using similarity criteria", *11th Iranian Conference on Intelligent Systems*, 2013.

[5] S. Moro, P. Cortez, P. Rita, "Business intelligence in banking: A literature analysis from 2002 to 2013 using text mining and latent Dirichlet allocation", *Expert Systems with Applications*, vol. 42, no. 3, pp. 1314-24, 2015.

[6] M. Yang, B. Jiang, Y. Wang, T. Hao, Y. Liu, "News text mining-based business sentiment analysis and its significance in economy", *Frontiers in Psychology*, vol. 13, pp. 1-7 , 2022.

[7] H. Hassani, C. Beneki, S. Unger, MT. Mazinani, "Text mining in big data analytics", *Big Data and Cognitive Computing*, vol. 4, no. 1, 2020.

[8] A. Gasparetto, M. Marcuzzo, A. Zangari, A. Albarelli, "A survey on text classification algorithms: From text to predictions", *Information*, vol. 13, no. 2, pp. 83, 2023.

[9] F. Greco, A. Polli, "Emotional Text Mining: Customer profiling in brand management", *International Journal of Information Management* , vol. 51, pp. 101934, 2020.

[10] A. Akundi, B. Tseng, J. Wu, E. Smith, "Text mining to understand the influence of social media applications on smartphone supply chain", *Procedia Computer Science*, vol. 140, pp. 87-94, 2018.

[11] W. He, S. Zha, L. Li, "Social media competitive analysis and text mining: A case study in the pizza industry", *International journal of information management*, vol. 33, no. 3, pp. 464-72, 2013.

[12] W. Souma, I. Vodenska, H. Aoyama, "Enhanced news sentiment analysis using deep learning methods", *Journal of Computational Social Science*, vol. 2, no. 1, pp. 33-46, 2019.

[13] S. Mohan, S. Mullapudi, S. Sammeta, "Stock price prediction using news sentiment analysis", *In2019 IEEE fifth international conference on big data computing service and applications (BigDataService)*, pp. 205-208, 2019.

[14] S. Negash, "Business intelligence", *Communications of the association for information systems*, vol. 13, no. 15, pp. 177-195, 2004.

[15] H. Chen, RHL. Chiang, VC. Storey, "Business intelligence and analytics: From big data to big impact", *MIS quarterly*, vol. , no. 1, pp. 1165-1188, 2012.

[16] J. Park, V. Barash, C. Fink, M. Cha, "Emoticon style: Interpreting differences in emoticons across cultures", *In Proceedings of the international AAAI conference on web and social media*, vol. 7, no. 1, pp. 466-475, 2013.

[17] K. Spirovski, E. Stevanoska, A. Kulakov, "Comparison of different model's performances in task of document classification", *In Proceedings of the 8th International Conference on Web Intelligence, Mining and Semantics, Novi Sad, Serbia*, pp. 1-12, 2018.

[18] KS. Kyaw, P. Tepsongkroh, C. Thongkamkaew, "Business Intelligent Framework Using Sentiment Analysis for Smart Digital Marketing in the E-Commerce Era", *Asia Social*, vol. 16, no. 3, pp. e252965-e252965, 2023.

[19] D. Yan, K. Li, S. Gu, L. Yang, "Network-based bag-of-words model for text classification", *IEEE Access*, vol. 8, pp. 82641-82652, 2020.

[20] WA. Qader, MM. Ameen, "An overview of bag of words; importance, implementation, applications, and challenges", *In2019 international engineering conference (IEC)* , pp. 200-204, 2019.

[Downloaded from jsdp.rcisp.ac.ir on 2026-09-05]

[DOI: 10.61186/jsdp.22.1.39]

[21] C. Niu, W. Zhang, S. Byna, Y. Chen, "Kv2vec: A Distributed Representation Method for Key-value Pairs from Metadata Attributes", *IEEE High Performance Extreme Computing Conference (HPEC)*, pp. 1-7, 2022.
[22] J. Cai, J. Luo, S. Wang, S. Yang, "Feature selection in machine learning: A new perspective", *Neurocomputing*, vol. 300, pp. 70-79, 2023.
[23] DH. Hubel, TN. Wiesel, "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex", *The Journal of physiology*, vol. 160, no. 1, pp. 106-154, 1962.
[24] A. Joulin, E. Grave, P. Bojanowski, M. Douze, "Fasttext. zip: Compressing text classification models", *arXiv preprint*, 2016.
[25] P. Bojanowski, E. Grave, A. Joulin, "Enriching word vectors with subword information", *Transactions of the association for computational linguistics*, vol. 5, pp. 135-146, 2017.
[26] JL. Elman, "Finding structure in time", *Cognitive Science*, vol. 14, no. 2, pp. 179-211, 1990.
[27] S. Hochreiter, J. Schmidhuber, "Long short-term memory", *Neural computation*, vol. 9, no. 8, pp. 1735-1780, 1997.

پردیس مرادبیکی دانشجوی دکترای دانشکدهٔ مهندسی برق و کامپیوتر دانشگاه صنعتی اصفهان در رشتهٔ نرم‌افزار و الگوریتم است و دانش‌آموختهٔ مقطع کارشناسی‌ارشد در همان رشته از دانشگاه یزد است. وی به حوزهٔ پژوهشی پردازش زبان طبیعی علاقه‌مند است.

نشانی رایانامهٔ ایشان عبارت است از:
p.moradbeiki@ec.iut.ac.ir

علیرضا بصیری کارشناسی خود را در سال ۸۷ در گرایش مهندسی فناوری اطلاعات از دانشگاه اصفهان و مقطع کارشناسی‌ارشد و دکترا را نیز در همان رشته از دانشگاه تهران در سال‌های ۸۹ و ۹۶ به پایان رساند. از سال ۱۳۹۷ نیز به عضویت هیئت علمی دانشکدهٔ مهندسی برق و کامپیوتر دانشگاه صنعتی اصفهان درآمد. از سال ۱۴۰۱ تاکنون نیز رئیس مرکز فناوری اطلاعات دانشگاه صنعتی اصفهان است.

نشانی رایانامهٔ ایشان عبارت است از:
Basiri@iut.ac.ir

[Downloaded from jsdp.rcisp.ac.ir on 2026-09-05]

[DOI: 10.61186/jsdp.22.1.39]