استخراج ویژگی و بررسی کارآیی روش‌های کاهش بُعد در زمینهٔ تحلیل احساس

راضیه برادران[1]* و عفت گلپر رابوکی[2]
[1]گروه مهندسی کامپیوتر و فناوری اطلاعات، دانشگاه قم، قم، ایران
[2]گروه ریاضی، دانشگاه قم، قم، ایران

چکیده

امروزه با فراگیرشدن دسترسی به اینترنت و به‌خصوص شبکه‌های اجتماعی، امکان به‌اشتراک‌گذاری عقاید و نظرات کاربران فراهم شده است. از سوی دیگر تحلیل احساس و عقاید افراد می‌تواند نقش به‌سزایی در تصمیم‌گیری سازمان‌ها و تولیدکنندگان داشته باشد. ازاین‌رو وظیفهٔ تحلیل احساس و یا عقیده‌کاوی به زمینهٔ پژوهشی مهمی در حوزه پردازش زبان طبیعی تبدیل شده است. یکی از چالش‌های استفاده از شیوه‌های یادگیری ماشینی در حوزهٔ پردازش زبان طبیعی، انتخاب و استخراج ویژگی‌های مناسب از میان تعداد زیاد ویژگی‌های اولیه برای دست‌یابی به مدلی با صحت مطلوب است. در این پژوهش دو روش فشرده‌سازی براساس تجزیه‌های ماتریسی SVD و NMF و یک روش بر اساس شبکه‌های عصبی برای استخراج ویژگی‌های مؤثرتر و با تعداد کمتر در زمینه تحلیل احساس در مجموعه‌داده نظرات به زبان فارسی مورد استفاده و تأثیر سطح فشرده‌سازی و اندازهٔ مجموعه‌داده در صحت مدل‌های ایجادشده مورد ارزیابی قرارگرفته شده است. بررسی‌ها نشان می‌دهد که فشرده‌سازی نه‌تنها از بار محاسباتی و زمانی ایجاد مدل کم می‌کند، بلکه می‌تواند صحت مدل را نیز افزایش دهد. بر طبق نتایج پیاده‌سازی، فشرده‌سازی ویژگی‌ها از ۷۷۰۰ ویژگی اولیه به دوهزار ویژگی با استفاده از شبکهٔ عصبی، نه‌تنها باعث کاهش هزینه محاسسباتی و فضای ذخیره‌سازی می‌شود، بلکه می‌تواند صحت مدل را از ٪ ۷۷/۰۵ به ٪ ۷۷/۸۵ افزایش دهد. از سوی دیگر در مجموعه داده کوچک با استفاده از روش SVD نتایج بهتری به‌دست می‌آید و با تعداد ویژگی دوهزار می‌توان به صحت ٪ ۶۳/۹۲ در مقابل ٪ ۶۳/۵۷ دست پیدا کرد؛ هم‌چنین آزمایش‌ها حاکی از آن است که فشرده‌سازی با استفاده از شبکه عصبی در صورت بزرگی مجموعه‌داده برای ابعاد پایین مجموعه ویژگی، بسیار بهتر از سایر روش‌ها عمل می‌کند. به‌طوری‌که تنها با یکصد ویژگی استخراج‌شده با استفاده از فشرده‌ساز شبکهٔ عصبی از ۷۷۰۰ ویژگی اولیه می‌توان به صحت قابل قبول ٪ ۷۴/۴۶ در مقابل صحت اولیه ٪ ۷۷/۰۵ با ۷۷۰۰ ویژگی دست یافت.

واژگان کلیدی: پردازش زبان طبیعی، تحلیل احساس، کدگذار خودکار، تجزیه مقدار تکین، تجزیه نامنفی ماتریس

Feature Extraction and Efficiency Comparison Using Dimension Reduction Methods in Sentiment Analysis Context

Razieh Baradaran*[1] & Effat Golpar Raboky[2]
[1]Department of Computer Engineering and Information Technology, University of Qom, Qom, Iran
[2]Department of Mathematics, University of Qom, Qom, Iran

Abstract

Nowadays, users can share their ideas and opinions with widespread access to the Internet and especially social networks. On the other hand, the analysis of people's feelings and ideas can play a significant role in the decision making of organizations and producers. Hence, sentiment analysis or opinion mining is an

* Corresponding author
*نویسندهٔ عهده‌دار مکاتبات

● تاریخ ارسال مقاله: ۱۳۹۶/۴/۳۱ ● تاریخ پذیرش: ۱۳۹۸/۳/۲۹ ● تاریخ انتشار: ۱۳۹۸/۱۰/۰۷ ● نوع مطالعه: پژوهشی

[DOI: 10.29252/jsdp.16.3.88]

important field in natural language processing. One of the most common ways to solve such problems is machine learning methods, which creates a model for mapping features to the desired output. One challenge of using machine learning methods in NLP fields is feature selection and extraction among a large number of early features to achieve models with high accuracy. In fact, the high number of features not only cause computational and temporal problems but also have undesirable effects on model accuracy.

Studies show that different methods have been used for feature extraction or selection. Some of these methods are based on selecting important features from feature sets such as Principal Component Analysis (PCA) based methods. Some other methods map original features to new ones with less dimensions but with the same semantic relations like neural networks. For example, sparse feature vectors can be converted to dense embedding vectors using neural network-based methods. Some others use feature set clustering methods and extract less dimension features set like NMF based methods. In this paper, we compare the performance of three methods from these different classes in different dataset sizes.

In this study, we use two compression methods using Singular Value Decomposition (SVD) that is based on selecting more important attributes and non-Negative Matrix Factorization (NMF) that is based on clustering early features and one Auto-Encoder based method which convert early features to new feature set with the same semantic relations. We compare these methods performance in extracting more effective and fewer features on sentiment analysis task in the Persian dataset. Also, the impact of the compression level and dataset size on the accuracy of the model has been evaluated. Studies show that compression not only reduces computational and time costs but can also increase the accuracy of the model.

For experimental analysis, we use the Sentipers dataset that contains more than 19000 samples of user opinions about digital products and sample representation is done with bag-of-words vectors. The size of bag-of-words vectors or feature vectors is very large because it is the same as vocabulary size. We set up our experiment with 4 sub-datasets with different sizes and show the effect of different compression performance on various compression levels (feature count) based on the size of dataset size.

According to experiment results of classification with SVM, feature compression using the neural network from 7700 to 2000 features not only increases the speed of processing and reduces storage costs but also increases the accuracy of the model from 77.05% to 77.85% in the largest dataset contains about 19000 samples. Also in the small dataset, the SVD approach can generate better results and by 2000 features from 7700 original features can obtain 63.92 % accuracy compared to 63.57 % early accuracy.

Furthermore, the results indicate that compression based on neural network in large dataset with low dimension feature sets is much better than other approaches, so that with only 100 features extracted by neural network-based auto-encoder, the system achieves acceptable 74.46% accuracy against SVD accuracy 67.15% and NMF accuracy 64.09% and the base model accuracy 77.05% with 7700 features.

Keywords: Natural Language Processing, Sentiment Analysis, Opinion Mining, Auto-Encoder, Singular Value Decomposition, Nonnegative Matrix Factorization

۱- مقدمه

عقاید، تصمیم‌ها و انتخاب‌های افراد بسیار متأثر از نظرات و عقاید دیگر افراد است؛ ازاین‌رو، اغلب قبل از تصمیم‌گیری در مورد چیزی، نظرات دیگران را نسبت به آن موضوع جستجو می‌کنیم. این مسأله نه‌تنها در مورد فرد بلکه در سطح سازمان نیز صادق است [19]؛ بنابراین «آن‌چه مردم فکر می‌کنند» بخش مهمی از اطلاعات تأثیرگذار در فرایند تصمیم‌گیری است [23].

با فراگیرشدن دسترسی به اینترنت و به‌طور خاص‌تر شبکه‌های اجتماعی امکان ثبت و دست‌یابی به حجم زیادی از نظرات افراد در مورد مسائل گوناگون فراهم شده است. ازاین‌رو تحلیل احساس و یا عقیده‌کاوی به زمینه پژوهشی مهمی در حوزه پردازش زبان طبیعی تبدیل شده است.

تحلیل احساس یا عقیده‌کاوی (اغلب این دو مفهوم به یک معنا به‌کار برده می‌شوند هر چند در اصل تفاوت‌هایی نیز دارند) به بررسی نظرات کاربران در مورد محصولات، خدمات، سازمان‌ها، افراد، رویدادهاو دیگر موارد می‌پردازد؛ درواقع می‌توان تحلیل احساس را یک وظیفه طبقه‌بندی در نظر گرفت که نظرات را به سه ردهٔ احساس مثبت، احساس منفی و احساس بی‌طرف نسبت به موضوع مورد نظر طبقه‌بندی می‌کند. وظیفه طبقه‌بندی در تحلیل احساس به سه سطح طبقه‌بندی تقسیم می‌شود. تحلیل احساس در سطح سند، در سطح جمله و در سطح مؤلفه. تحلیل احساس در سطح سند، کل سند را به‌عنوان یک واحد اطلاعاتی در نظر گرفته و احساس مثبت، منفی و یا بی‌طرف به آن نسبت می‌دهد. تحلیل احساس در سطح جمله، به طبقه‌بندی احساس جملات می‌پردازد و تحلیل احساس در سطح مؤلفه برچسب احساسی هر موجودیت را تعیین می‌کند [20]؛ برای مثال تحلیل احساس می‌تواند در حوزه‌های تجاری برای تصمیم‌گیری در مورد یک محصول [16] و یا در حوزه‌های پزشکی [5] و به‌طورکلی در تمام حوزه‌هایی که بهره‌گیری از نظرات دیگران می‌تواند در تصمیم‌گیری دخالت داشته باشد، مفید باشد.

درکل شیوه‌های طبقه‌بندی احساس به شیوه‌های

[DOI: 10.29252/jsdp.16.3.88]

یادگیری ماشینی، شیوه‌های مبتنی بر لغت‌نامه و شیوه‌های ترکیبی تقسیم‌بندی می‌شود. شیوه‌های یادگیری ماشینی از الگوریتم‌های یادگیری ماشینی معروف مانند ماشین بردار پشتیبان و ویژگی‌های زبان‌شناختی استفاده می‌کنند. شیوه‌های مبتنی بر لغت‌نامه متکی بر لغت‌نامه احساسی هستند که مجموعه‌ای از موارد[1] احساسی به‌همراه رتبهٔ احساسی آن‌ها است [4-3]. شیوه ترکیبی نیز از هر دو روش به‌صورت ترکیبی استفاده می‌کند [6].

از مهم‌ترین مزایای شیوه‌های یادگیری ماشینی ایجاد مدل‌های مستقل از دامنه است [29]. شیوه‌های یادگیری ماشینی در حوزه پردازش زبان طبیعی به‌طور کلی و به‌طور خاص در حوزه تحلیل احساس از ویژگی‌هایی استفاده می‌کنند که شامل واژگان یا ترکیبی از واژگان پیکره متنی هستند و مدل‌های آماری را به این پدیده‌های زبانی نگاشت می‌دهند [7].

به‌دلیل بزرگ‌بودن اندازهٔ لغت‌نامه و یا تعداد لغات متمایز پیکره متنی، یکی از چالش‌های مهم در حوزه پردازش زبان طبیعی وجود مجموعه ویژگی‌های با طول زیاد و تُنُک است که منجر به مشکلات محاسباتی و زمانی می‌شود [12].

هم‌چنین این تعداد ویژگی زیاد درصورت عدم دسترسی به تعداد بالای نمونه‌های آموزشی مستعد مشکل عدم تعمیم‌دادن مناسب مدل ایجادشده و باعث محدودشدن و انطباق بیش از حد به نمونه‌های آموزشی و توانایی پایین مدل در تخمین درست نمونه‌های نادیده و با اصطلاح «بیش‌برازش[2]» می‌شود. یک شیوه برای غلبه بر این مشکل تولید مجموعه ویژگی با ابعاد پایین از مجموعه ویژگی‌های اولیه است. در این صورت نتایج حاصل از اعمال الگوریتم‌های یادگیری بر مجموعه ویژگی با ابعاد کم، پایدارتر و قابل اعتمادتر از مدل ایجادشده بر اساس مجموعه ویژگی اولیه است.

برای تولید مجموعه ویژگی با ابعاد کم یک شیوه طراحی ویژگی‌ها بر اساس نظر متخصص و با اطلاع از دانش زبان‌شناختی است که درواقع تولید ویژگی به‌طور دستی است. چنین راه‌کاری با وجود کارایی بالای مدل ایجادشده از مشکل زمان‌بربودن تولید ویژگی‌ها و عدم انطباق همان ویژگی‌ها با دامنه جدید رنج می‌برد که لزوم تولید ویژگی‌های با ابعاد پایین را به‌صورت خودکار ایجاب می‌کند [7].

یکی از روش‌های تولید خودکار ویژگی‌های با ابعاد کم فشرده‌سازی مجموعه ویژگی‌های اولیه است. استفاده از روش‌های فشرده‌سازی و کاهش ابعاد[3] می‌تواند در کم‌کردن تعداد ویژگی‌ها و حذف ویژگی‌های نامرتبط و حذف نوفه مفید باشد و مشکلات یادشده را تا حد زیادی مرتفع کند.

در این پژوهش که به وظیفه تحلیل احساس به‌صورت خودکار در متون فارسی پرداخته شده است. ابتدا با استفاده از روش‌های فشرده‌سازی کلاسیک مانند تجزیه مقدار تکین[4] (SVD) و تجزیه نامنفی[5] (NMF) ماتریس‌ها و روش‌های نوین مانند روش‌های مبتنی بر شبکه عصبی و به‌طور خاص کدگذار خودکار[6]، فشرده‌سازی مجموعه ویژگی‌ها انجام شده و سپس طبقه‌بندی احساس صورت گرفته است. به عبارت دیگر در این پژوهش کارایی هریک از این روش‌های فشرده‌سازی را در مسأله تحلیل احساس بررسی کرده‌ایم. نتایج به‌دست‌آمده حاکی از آن است که به‌طور کلی فشرده‌سازی نه‌تنها می‌تواند زمان و هزینه محاسباتی را کاهش، بلکه می‌تواند کارایی وظیفه طبقه‌بندی را نیز افزایش دهد؛ علاوه‌براین به مقایسه کارایی روش‌های فشرده‌سازی یادشده نسبت به اندازه‌های مختلف مجموعه داده آموزشی پرداخته شده است و نشان داده‌ایم که با زیادشدن تعداد نمونه‌های آموزشی فشرده‌سازی با استفاده از کدگذار خودکار نتایج بهتری را حاصل می‌کند.

به‌طور خلاصه کارهای صورت‌گرفته در این پژوهش عبارتند از:

- ایجاد یک سامانه تحلیل احساس در زبان فارسی به‌صورت خودکارو بدون نیاز به تعریف مجموعه ویژگی‌ها با استفاده از متون نظرات به‌صورت خام و مبتنی بر فشرده‌سازی مجموعه ویژگی‌ها؛
- ارائهٔ دسته‌بندی روش‌های فشرده‌سازی و مقایسه کارایی شیوه‌های مختلف فشرده‌سازی مجموعه ویژگی‌ها برای تحلیل احساس با توجه به اندازه مجموعه داده؛

ادامه این مقاله به‌صورت زیر سازماندهی شده است. در بخش ۲ به مروری بر کارهای صورت‌گرفته در این زمینه پرداخته شده است. بخش ۳ روش پیشنهادی مورد بررسی قرار می‌گیرد و نتایج حاصل از ارزیابی‌ها را در بخش ۴ خواهید دید؛ درنهایت بخش ۵ به نتیجه‌گیری و جهت‌گیری‌های آینده اختصاص یافته است.

[1] terms
[2] Overfitting
[3] Dimension Reduction
[4] Singular Value Decomposition
[5] Nonnegative Matrix factorization
[6] Autoencoder

[DOI: 10.29252/jsdp.16.3.88]

۲- مروری بر کارهای انجام گرفته

در سال‌های اخیر پژوهش‌های بسیاری در زمینه تحلیل احساس صورت گرفته است. از این میان برخی از این پژوهش‌ها به حوزه انتخاب ویژگی برای وظیفه تحلیل احساس اختصاص یافته‌اند که ما در اینجا به بررسی آنها می‌پردازیم.

روش‌های گوناگونی برای استخراج و یا انتخاب ویژگی وجود دارد؛ برخی از این روش‌ها به انتخاب ویژگی‌های مهم‌تر از میان ویژگی‌های موجود می‌پردازند؛ مانند بهره اطلاعاتی[1]، اطلاعات دوسویه[2]، تحلیل مؤلفه‌های اساسی[3] (PCA) یا SVD (در اینجا PCA همان کاری را انجام می‌دهد که SVD انجام می‌دهد [27]) و غیره. از این میان روش PCA از محبوبیت بیشتری برخوردار بوده است و در حوزه‌های دیگر نیز مانند پردازش تصاویر [2] به کار گرفته شده است. برای نمونه می‌توان به [10] اشاره کرد. این پژوهش نظرکاوی بر روی نظرات مربوط به فیلم را انجام می‌دهد. در این کار از PCA برای کاهش ابعاد مجموعه ویژگی‌ها استفاده شده و نتایج طبقه‌بندی با استفاده از الگوریتم‌های بیزی ساده[4] و Linear Vector Quantization (LVQ) مقایسه شده که البته نتایج به‌دست‌آمده با حالتی که از PCA استفاده نشود مقایسه نشده است. در پژوهش دیگری که مربوط به کاوش نظرات مربوط به دوربین دیجیتال از تارنمای آمازون است از PCA برای کم‌کردن ابعاد و از ترکیبی از طبقه بندها برای کاوش نظرات استفاده و نتایج حاصل از آن با رگرسیون لجستیک[5] و SVM مقایسه شده است [26].

در پژوهش دیگری در سال ۲۰۱۶ چند الگوریتم استخراج ویژگی شامل PCA و استخراج ویژگی مبتنی بر درخت تصمیم و استخراج ویژگی مبتنی بر جنگل تصمیم به‌کار برده شده و نتایج اعمال چند الگوریتم طبقه‌بندی شامل Cart، بیزی ساده، LVQ با هم مقایسه و نشان داده شده است که استخراج ویژگی مبتنی بر جنگل تصمیم بهتر عمل می‌کند؛ ولی دراین کار نرخ‌های گوناگون فشرده‌سازی در PCA به‌کار برده نشده است [11].

در [30] نیز سه روش استخراج ویژگی شامل PCA، تحلیل معنایی پنهان[6] (LSA) پنهان و تطابق تصادفی[7] (RP) در زمینه تحلیل احساس مقایسه شده است که البته این هر سه روش در دستهٔ نخست روش‌ها طبق دسته‌بندی ارائه‌شده در این مقاله قرار می‌گیرند.

برخی دیگر از روش‌های استخراج ویژگی به‌طور خودکار ویژگی‌های اولیه را به ویژگی‌های جدید با ابعاد کمتر نگاشت می‌دهند؛ به‌طوری که ویژگی‌های جدید روابط معنایی ویژگی‌های قبلی را حفظ کند. برای مثال می‌توان به تبدیل بردار ویژگی‌های اولیه به بردار ویژگی‌های تعبیه‌شده با ابعاد کمتر اشاره کرد که از شبکهٔ عصبی برای تولید بردار با ابعاد کمتر و حافظ اطلاعات و روابط معنایی استفاده می‌شود [21-22].

به‌عنوان نمونه در پژوهش [25] از کدگذار خودکار مبتنی بر شبکهٔ عصبی، استفاده شده است. در این مقاله برای استخراج ویژگی در وظیفه تحلیل احساس از ترکیبی از مدل‌ها مبتنی بر کدگذار خودکار استفاده و نشان داده است که کدگذار خودکار در این زمینه بهبود ایجاد می‌کند؛ البته در این کار از دو مجموعه داده استفاده شده که یکی شامل حدود ۱۰۵۰۰ و دیگری پنجاه‌هزار نمونه آموزشی است که تاحدودی مقادیر قابل توجهی است.

برخی دیگر از روش‌ها از خوشه‌بندی مجموعه ویژگی‌های اولیه استفاده می‌کنند و مجموعه ویژگی‌های با ابعاد کمتر به‌دست می‌آورند [17]. در این کار از NMF و لغت نامه برای تحلیل احساس استفاده شده است که درواقع با NMF ویژگی‌ها خوشه‌بندی شده و خوشه‌ها به‌عنوان ویژگی‌های جدید در نظر گرفته می‌شوند. در این کار NMF با چندین الگوریتم خوشه‌بندی مشابه دیگر مقایسه شده است.

همان‌طور که بیان شد، پژوهش‌های انجام‌گرفته تاکنون مقایسه‌ای بین روش‌های دسته‌های مختلف انجام نداده‌اند. هم‌چنین اندازه مجموعه داده می‌تواند یک عامل تأثیرگذار در میزان صحت روش‌های گوناگون باشد. در این پژوهش هر سه نوع روش استخراج ویژگی (استخراج ویژگی مبتنی بر SVD، استخراج ویژگی مبتنی بر کدگذار خودکار، استخراج ویژگی مبتنی بر خوشه‌بندی NMF) به‌کار برده شده و کارایی آن‌ها در اندازه‌های مختلف مجموعه داده بررسی و مقایسه شده است.

۳- روش‌های استخراج ویژگی

در هر مجموعه داده که به‌صورت ماتریس ذخیره می‌شود، اطلاعات زیادی پنهان شده است. روش کاهش بُعد یک روش قدرتمند برای پیداکردن ساختار پنهان داده است. از مزایای

[1] Information Gain
[2] Mutual information
[3] Principal Component Analysis
[4] Naive Bayes
[5] Logistic Regression
[6] Latent Semantic Analysis
[7] Random Projection

[DOI: 10.29252/jsdp.16.3.88]

این روش کاهش حافظه و حجم محاسبات، تشخیص نوفه و کشف روابط پنهان در ماتریس اولیه است. این روش در مرحله پیش‌پردازش اطلاعات انجام می‌شود که یک تقریب با رتبه پایین از داده‌های با بعد بالا استخراج می‌کند.

در این پژوهش سه الگوریتم فشرده‌سازی برای کاهش ابعاد ویژگی‌ها و استخراج ویژگی‌های جدید در نظر گرفته شده است. این الگوریتم‌ها به یک دسته از روش‌های استخراج ویژگی مطرح‌شده در بخش ۲، روش‌های مبتنی بر استخراج ویژگی‌های مهم‌تر، روش‌های مبتنی بر استخراج بردار ویژگی‌های تعبیه‌شده و روش‌های مبتنی بر خوشه‌بندی ویژگی‌ها، تعلق دارد.

در این مقاله سه شیوه فشرده‌سازی به‌منظور استخراج ویژگی‌های جدید و با ابعاد کمتر مورد بررسی قرار گرفته و نتایج حاصل مقایسه شده است:

- استخراج ویژگی مبتنی بر تجزیه مقدار تکین
- استخراج ویژگی مبتنی بر کدگذار خودکار
- و استخراج ویژگی مبتنی بر تجزیه نامنفی ماتریس‌ها که در ادامه هریک از این روش‌ها توضیح داده شده است..

۳-۱- تجزیه مقدار تکین

SVD یکی از مهم‌ترین و پرکاربردترین روش‌ها در تجزیه و فشرده‌سازی ماتریس‌ها به حساب می‌آید که می‌تواند بر روی هر ماتریس با مقادیر حقیقی اعمال شود.

تجزیه مقدار تکین ماتریس A را به حاصل‌ضرب سه ماتریس به‌صورت $A=UDV^T$ تجزیه می‌کند، به‌طوری‌که U و V ماتریس‌های متعامد و D یک ماتریس قطری با مقادیر مثبت حقیقی است که به‌صورت نزولی مرتب شده‌اند.

در اغلب کاربردها ماتریس داده بسیار بزرگ است و لازم است ماتریس A را با یک ماتریس رتبه پایین‌تر مانند B تقریب بزنیم، به‌طوری‌که ماتریس B تقریب خوبی برای A باشد. از مهم‌ترین ویژگی‌های SVD که به‌طور گسترده در فشرده‌سازی اطلاعات به‌کار می‌رود خاصیت بهترین تقریب‌کنندگی آن است .طبق قضیه بهترین تقریب‌کننده Eckart-Young، SVD بهترین تقریب از رتبه دلخواه K را برای ماتریس اولیه ارایه می‌دهد. فرض کنید $A=UDV^T$. حال $A_k=U_kD_kV_k^T$ در نظر بگیرید به‌طوری‌که U_k و V_k به‌ترتیب شامل k ستون اول ماتریس‌های U و V و D_k ماتریس قطری k در k است که شامل k سطر و ستون نخست D است. در این صورت A_k بهترین تقریب با رتبه بیشینهٔ k برای ماتریس A است؛ یعنی $\min\| A-B \|_F=\| A-A_k \|_F$. برای ملاحظات بیشتر به [9] مراجعه کنید.

۳-۲- تجزیه نامنفی ماتریس‌ها

هر چند تجزیه SVD بهترین تقریب رتبه پایین را تولید می‌کند؛ ولی عناصر ماتریس‌های حاصل از تجزیه ممکن است منفی باشند؛ اما در برخی از کاربردها مانند صوت و تصویر ماهیت داده نامنفی است . تجزیه نامنفی یکی از موضوعات به‌روز در زمینه جبر خطی است که برای کاهش بُعد، تحلیل داده‌ها، تشخیص الگو، کشف اطلاعات معنی‌دار در داده‌ها و خوشه‌بندی و پیداکردن نماینده‌ای کارآمد برای داده‌ها و طبقه‌بندی به کار می‌رود [8].

در تجزیه نامنفی برای ماتریس نامنفی $A \in R^{m\times n}$ و یک عدد صحیح $k < min\{m \text{ و } n\}$، ماتریس‌های نامنفی $W \in R^{m\times k}$ و $H \in R^{k\times n}$ مورد نظر است، به‌طوری‌که تابع زیر را کمینه کند:

$$F(W,H)=\frac{\|A-WH\|_F^2}{2} \qquad (۱)$$

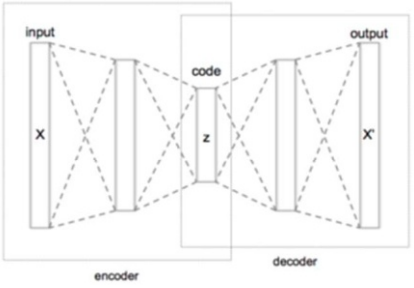
حاصل ضرب WH را یک تجزیه نامنفی A می‌نامیم، W ماتریس پایه و H ماتریس ضرایب نام دارد [15]. ستون‌های ماتریس W برای خوشه‌بندی داده‌ها مورد استفاده قرار می‌گیرد، به این دلیل که ستون‌های W فضای ستون‌های ماتریس A را می‌سازند.. در اینجا اگر W، k ستون داشته باشد، k خوشه برای خوشه‌بندی در نظر گرفته می‌شود. در NMF هر ستون (هر سند) از ماتریس A در یک خوشه قرار دارد. در این روش سند j ام در خوشه i قرار دارد اگر درایه h_{ij} در H بزرگترین درایه h_j باشد. [13]. درواقع با خوشه‌بندی داده‌ها به‌وسیله NMF ، ماتریس W که ماتریس خوشه‌ها است به جای ماتریس A به‌کار برده می‌شود.

۳-۳- کدگذار خودکار

کدگذار خودکار یک نوع معماری شبکه عصبی شامل لایه‌های ورودی، خروجی و دست‌کم یک لایهٔ مخفی است (شکل (۱)). لایه ورودی و خروجی یکسان هستند و درواقع شبکه مقادیر وزن‌ها و لایه مخفی را به‌گونه‌ای یاد می‌گیرد که خروجی معادل ورودی به‌دست آید. هدف یک کدگذار خودکار یادگیری یک بازنمایی از یک مجموعه داده است که به‌طورمعمول به هدف کاهش ابعاد انجام می‌گیرد؛ به‌طوری‌که این بازنمایی تا حد امکان حافظ اطلاعات معنایی داده اصلی باشد [14, 18].

[DOI: 10.29252/jsdp.16.3.88]

(شکل-۱): معماری کلی یک کدگذار خودکار [28]
(Figure-1): global architecture of Autoencoder [28]



۴- روش پیشنهادی

در این پژوهش سه الگوریتم فشرده‌سازی یادشده در بخش قبل، به مجموعه‌داده در اندازه‌های مختلف و در سطوح فشرده‌سازی مختلف اعمال شده و مقایسه‌ای از نتایج حاصل ارائه شده است؛ علاوه‌براین یک مدل نیز بدون اعمال فشرده‌سازی و با همان ویژگی‌های اولیه ایجاد شده که از آن به‌عنوان مدل پایه نام‌برده شده است.

در این بخش به بررسی و تحلیل آزمایش‌های انجام‌گرفته و نتایج حاصل از آن می‌پردازیم. در ابتدا مجموعه‌دادهٔ به‌کار رفته معرفی و در ادامه روال انجام کار بیشتر توضیح داده می‌شود.

۴-۱- داده

در این پژوهش از پیکره تحلیل احساس سنتی پرس [1] استفاده شده است. این مجموعه‌داده شامل نظرات کاربران در مورد محصولات دیجیتال است که از تارنمای دیجی‌کالا جمع‌آوری و در فرمت XML سازماندهی شده است. یکی از ویژگی‌های این پیکره آن است که شامل جملاتی از فارسی به هر دو صورت رسمی و غیررسمی (محاوره‌ای) است. در این پیکره بار معنایی جملات به‌صورت نمره‌دهی در یک بازه شامل {+۲و+۱و۰و-۱و-۲} است که به یک جمله بر حسب میزان مثبت یا منفی احساس آن نسبت داده می‌شود.

در ادامه چند مثال از نمونه‌های آموزشی در مجموعه داده آورده شده است:

نمره -۲: این گوشی فاجعه است و به‌طورکامل ناامیدم کرد.

نمره ۰: این گوشی رو ماه پیش از دیجی کالا خریداری کردم.

نمره +۱: مصرف انرژی گوشی خوبه در مجموع ازش راضیم.

پیکره سنتی پرس شامل مجموعه‌ای از فایل‌هاست که هر فایل شامل نظرات کاربران در مورد محصول خاصی است. جدول (۱) حاوی اطلاعات مربوط به تعداد فایل‌های هر گروه محصول در این پیکره است.

(جدول-۱): تعداد فایل‌های مربوط به هر نوع محصول در پیکره
(Table-1): the number of files for each product in corpus

نوع محصول	تعداد
تلفن همراه	73
دوربین دیجیتال	64
دوربین فیلم‌برداری	30
تبلت	30
نوت بوک	15
پخش‌کننده موزیک	12
چاپگر	12
تجهیزات مربوط به رایانه	11
تلویزیون	11
کنسول بازی	6
پویشگر	5

این پیکره درمجموع شامل ۲۹۹ فایل محصول و درکل ۱۹۲۹۸ جمله حاوی نظرات کاربران است؛ البته نسخه اولیه آن تعداد نظرات کمتری را شامل می‌شود که به‌طورتقریبی برابر با ۸۸۵۰ نظر است.

۴-۲- پیش‌پردازش

فاز پیش‌پردازش یکی از مراحل اصلی هر پژوهش داده‌کاوی است که نقش بسیاری در کارایی مدل ایجادشده دارد. این پژوهش نیز از این قایده مستثنی نیست.

در این گام ابتدا جملات حاوی نظر به همراه برچسب معنایی مربوطه از فایل‌های پیکره استخراج و سپس قطعه‌بندی[1]، ریشه‌یابی[2] و حذف ایست‌واژه‌ها[3] انجام شد. قطعه‌بندی و ریشه‌یابی با استفاده از کتابخانه هضم[4] که یک کتابخانه رایگان برای پردازش زبان فارسی در پایتون است انجام شد. و برای حذف ایست‌واژه‌ها یک فهرست از واژهای مورد نظر در این کار تهیه و از نمونه‌ها حذف شد. ریشه هریک از واژگان پیکره به‌عنوان ویژگی برای مدل ایجادشده لحاظ شده و بنابراین برای هریک از جملات پیکره به‌عنوان یک نمونه بردار ویژگی ایجاد شده است؛ به‌طوری‌که به‌ازای هر کلمه درصورتی‌که آن کلمه در جمله حضور داشته باشد درایه مربوطه مقدار می‌گیرد در غیر این صورت صفر است؛ درنهایت ماتریس سندواژه برای کل نمونه‌های مجموعه‌داده موجود ایجاد شده است.

در این کار برای سادگی بیشتر، مسأله رده‌بندی به صورت یک مسأله سه رده شامل ردهٔ احساس مثبت احساس منفی و احساس بی‌طرف در نظر گرفته شده است.

در این آزمایش، هشتاد درصد مجموعه‌داده به‌عنوان مجموعه آموزشی و بیست‌درصد به‌عنوان مجموعه آزمون در نظر گرفته شده است.

[1] tokenization
[2] stemming
[3] Stop words
[4] http://www.sobhe.ir/hazm/

[DOI: 10.29252/jsdp.16.3.88]

۳-۴- وزن‌دهی به ویژگی‌ها

در طبقه‌بندی متن به‌طور معمول از وزن‌دهی «فراوانی واژه عکس فراوانی سند[1]» (tf-idf) استفاده می‌شود. اما در طبقه‌بندی احساس Pang و دیگران به این نکته اشاره کرده است که تعداد تکرار واژگان تأثیری در احساس کلی و مثبت و یا منفی‌بودن آن ندارد [24]؛ بنابراین از وزن‌دهی به ویژگی‌ها به‌صورت دودویی به معنای حضور (یک) یا عدم حضور (صفر) ویژگی (واژه) در نمونه (جمله) مربوطه استفاده شده و درنهایت بردار ویژگی برای هر نمونه به‌صورت مجموعه‌ای از صفر و یک حاصل شده است.

۴-۴- سطح فشردگی ویژگی‌ها

سطح فشردگی، اندازه بردار ویژگی حاصل از اعمال الگوریتم فشرده‌سازی است که توسط کاربر تعیین می‌شود. به‌عنوان مثال سطح فشردگی یکصد به این معناست که اندازهٔ بردار ویژگی برابر یکصد است.

۵-۴- پیاده‌سازی

در این پژوهش از چهار اندازهٔ مختلف مجموعه‌داده اصلی استفاده شده که اطلاعات مربوط به آنها در جدول (۲) نشان داده شده است. درواقع نسخه اولیه مجموعه‌داده به‌عنوان مجموعه داده یک و نسخه نهایی به‌عنوان مجموعه‌داده چهار محسوب می‌شود. سه روش فشرده‌سازی در ابعاد فشرده‌سازی مختلف به هر کدام اعمال و سپس الگوریتم طبقه‌بندی SVM با استفاده از زبان برنامه‌نویسی پایتون پیاده‌سازی شده است. هم‌چنین طبقه‌بندی بدون اعمال فشرده‌سازی نیز انجام شده و از مدل ایجادشده به‌عنوان مدل پایه استفاده شد. آمار مربوط به مجموعه‌داده‌ها و نتایج حاصل در جدول‌های (۲ تا ۵) نشان داده شده است.

(جدول-۲): تعداد نظرات و ویژگی‌های مجموعه‌داده‌های به‌کاررفته
(Table-2): the number of comments and feature size in used datasets

عنوان	اندازه مجموعه داده (تعداد نظرات)	تعداد ویژگی
مجموعه داده ۱	2853	3911
مجموعه داده ۲	6159	5066
مجموعه داده ۳	8850	6010
مجموعه داده ۴	19298	7706

[1] term frequency – inverse document frequency

(جدول-۳): نتایج ارزیابی روش‌ها در مجموعه‌دادهٔ یک
(Table-3): evaluation results of the methods in 1 Dataset

صحت (٪)			
مدل پایه	63.57		
سطح فشردگی (تعداد ویژگی)	SVD	NMF	Autoencoder
100	57.79	55.34	55.17
500	58.67	56.22	61.12
1000	59.89	61.65	60.24
2000	63.92	56.39	63.57
	63.92	52.19	61.12

(جدول-۴): نتایج ارزیابی روش‌ها در مجموعه‌دادهٔ دو
(Table-4): evaluation results of the methods in 2Dataset

صحت (٪)			
مدل پایه	68.29		
سطح فشردگی (تعداد ویژگی)	SVD	NMF	Autoencoder
100	63.39	54.79	66.56
500	65.42	63.55	68.67
1000	65.99	67.45	69.89
2000	67.13	67.56	70.05
3000	69.32	67.37	70.13
4000	68.26	64.29	69.56
5000	68.26	62.74	68.99

(جدول-۵): نتایج ارزیابی روش‌ها در مجموعه‌دادهٔ سه
(Table-5): evaluation results of the methods in 3 Dataset

صحت (٪)			
مدل پایه	72.03		
سطح فشردگی (تعداد ویژگی)	SVD	NMF	Autoencoder
100	67.46	63.78	67.63
500	68.70	68.93	69.66
1000	70.90	70.90	71.58
2000	71.41	71.75	71.75
3000	71.58	71.41	70.58
4000	71.92	70.39	60.62
5000	72.03	68.98	70.79

همان‌طورکه مشاهده می‌شود، زمانی که اندازه مجموعه‌داده کوچک باشد، روش SVD و کدگذار خودکار تفاوت چندانی ندارند و در برخی سطوح فشرده‌سازی SVD و

[DOI: 10.29252/jsdp.16.3.88]

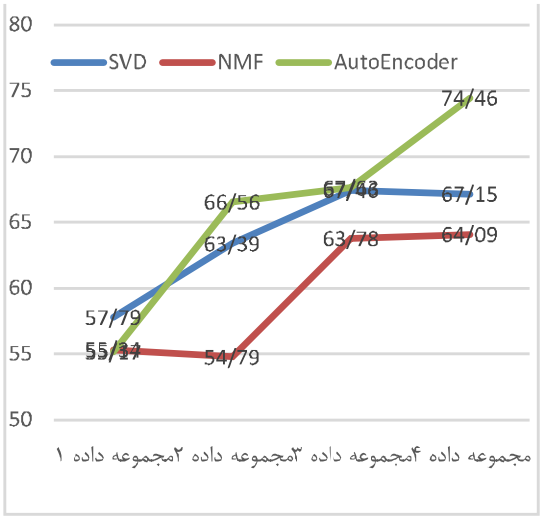
در برخی، کدگذار خودکار بهتر عمل می‌کند. از میان دو روش SVD و NMF در اغلب موارد SVD نتایج بهتری دارد. با افزایش اندازهٔ مجموعه‌داده، کدگذار خودکار نسبت به روش‌های دیگر در اغلب موارد نتایج بهتری را نشان می‌دهد به‌خصوص در سطوح فشرده‌سازی پایین این تفاوت بیشتر مشهود است. به‌عنوان مثال در مجموعه‌دادهٔ چهار با سطح فشردگی صد، کدگذار خودکار با هفت درصد اختلاف در صحت بهتر از SVD عمل می‌کند (شکل‌های (۲) و (۳)). این نکته بسیار ارزشمند است که می‌توان با تعداد ویژگی بسیار کم در مقایسه با ویژگی‌های اولیه (برای نمونه صد در مقابل ۷۷۰۶ در مجموعه‌داده چهار) به صحتی قابل قبول و نزدیک به صحت مدل پایه (در اینجا ۷۴/۴۶ درصد برای ویژگی‌های با سطح فشردگی صد در مقابل ۷۷/۰۵ درصد برای مدل پایه با کل ویژگی‌ها) دست یافت. به‌نظر می‌رسد این نتایج حاکی از ماهیت شبکه عصبی در مدل‌کردن داده با حجم بالاست. به‌طوری‌که وجود دادهٔ بسیار برای دست‌یابی به مدلی با صحت بالا ضروری است. علت این امر وجود پارامترهای زیاد شبکهٔ عصبی برای یادگیری مدل است که نیاز به مجموعه‌دادهٔ بسیار دارد؛ در صورتی‌که اندازهٔ مجموعه‌داده کوچک باشد، مدل‌های سنتی به‌طورمعمول نتایج بهتری را از خود نشان می‌دهند و نتایج آزمایش‌ها ما نیز چنین مطلبی را تأیید می‌کند.

نکته دیگر این‌که در بسیاری موارد با فشرده‌سازی می‌توان به صحتی بالاتر از صحت مدل پایه دست یافت که این به خاصیت حذف داده‌های غیر ضروری در فرایند فشرده‌سازی برمی‌گردد. به‌عنوان مثال در مجموعه‌دادهٔ چهار با سطح فشرده‌سازی دوهزار (تعداد ویژگی) به صحت ۷۷/۸۵ درصد دست یافتیم که صحت مدل پایه با ۷۷۰۶ ویژگی ۷۷/۰۵ درصد است.

(جدول-۶): نتایج ارزیابی روش‌ها در مجموعه‌دادهٔ چهار

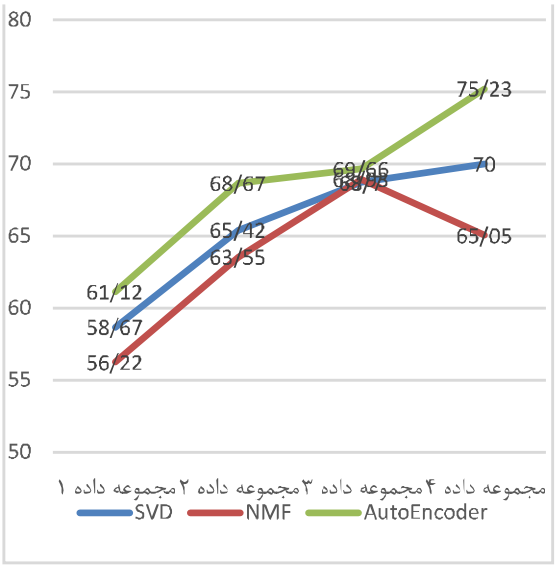
(Table-6): evaluation results of the methods in4 Dataset

صحت (٪)			
مدل پایه	77.05		
سطح فشردگی (تعداد ویژگی)	SVD	NMF	Autoencoder
100	67.15	64.09	74.46
500	70.00	67.05	75.23
1000	73.34	69.87	75.91
2000	75.26	72.90	77.85
3000	76.71	73.34	77.07
4000	76.58	73.55	77.23
5000	76.79	74.25	77.49
6000	76.97	74.71	77.33



(شکل-۲): صحت مدل‌ها-سطح فشردگی ۱۰۰

(Figure-2): the model Accuracy-compression level 100



(شکل-۳): صحت مدل ها-سطح فشردگی ۵۰۰

(Figure-3): the model Accuracy-compression level 500

۵- نتیجه‌گیری

تحلیل احساس یکی از وظایف پرطرفدار در حوزهٔ پردازش زبان طبیعی است که تاکنون پژوهش‌های بسیاری در این زمینه به‌ویژه در زبان انگلیسی انجام شده است. از جمله شیوه‌های رایج در حل این‌گونه مسائل شیوه‌های یادگیری ماشینی است که به ایجاد مدلی برای نگاشت ویژگی‌ها به خروجی مطلوب مبادرت می‌کند. بنابراین انتخاب ویژگی‌های مناسب نقش مهمی در تولید مدل با صحت بالا دارد.

یکی از چالش های این حوزه وجود تعداد ویژگی‌های بالاست که نه‌تنها باعث مشکلات محاسباتی و زمانی می‌شود بلکه در صحت مدل نیز تأثیر نامطلوب دارد.

در این پژوهش از سه شیوهٔ کاهش بُعد برای استخراج

[DOI: 10.29252/jsdp.16.3.88]

و فشرده‌سازی ویژگی‌ها استفاده شده است. این شیوه‌ها عبارتند از تجزیه مقدار تکین که مبتنی بر انتخاب ویژگی‌های مهم‌تر، تجزیه نامنفی ماتریس که مبتنی بر خوشه‌بندی ویژگی‌های اولیه است و کدگذار خودکار که ویژگی‌های اولیه را به یک مجموعه ویژگی‌های جدید با ابعاد کوچک‌تر و حافظ روابط معنایی ویژگی‌های اولیه تبدیل می کند. ما ویژگی‌های تولیدی از این روش‌ها را در سطوح فشردگی مختلف و در اندازه‌های مختلف مجموعه داده فارسی با یکدیگر مقایسه کردیم.

نتایج حاکی از کارایی بهتر روش کدگذار خودکار در صورت بالابودن اندازه مجموعه‌داده به‌خصوص در سطوح فشردگی پایین است؛ به‌طوری‌که در مجموعه‌دادهٔ چهار با مجموعه ویژگی با اندازه صد که با استفاده از کد گذار خودکار به‌دست‌آمده، می‌توان به صحتی (۷۴/۴۶) نزدیک به صحت مدل ایجادی با کل ویژگی‌ها (۷۷/۰۵) که حدود ۷۷۰۰ ویژگی است، دست یافت. از سوی دیگر در اندازه پایین مجموعه‌داده، SVD می‌تواند روش پایدارتری محسوب شود. هم‌چنین با اعمال شیوه‌های فشرده‌سازی با اینکه تعداد ویژگی کاهش می‌یابد، صحت مدل می‌تواند افزایش یابد که این ناشی از حذف ویژگی‌های غیر ضروری و مستعد خطاست. بهترین نتیجه طبقه‌بندی با الگوریتم SVM و بر اساس معیار صحت، مربوط به مجموعه داده با اندازهٔ ۱۹۲۹۸ و مجموعه ویژگی تولیدی توسط کدگذار خودکار با سطح فشردگی دوهزار است که برابر با ۷۷/۸۵ درصد است.

۶- مراجع 6- References

[۱] حسینی، پدرام، احمدیان رمکی، علی، ملکی، حسن، انواری، منصوره، میرروشندل، ابولقاسم، "پیکرهٔ فارسی تحلیل احساس سِنتی پِرس"، سومین همایش ملی زبان‌شناسی رایانشی، تهران، دانشگاه صنعتی شریف، ۱۳۹۳.

[1] P. Hosseini, A. Ahmadian-Ramaki, H. Maleki, M. Anvari and A. Mirroshandel, "Sentipers: A sentiment analysis corpus for Persian", in *3th National Conference on Linguistics*, Tehran: Sharif University of Technology, 2015.

[2] شاهدوستی، حمید رضا، قاسمیان، حسن، "استفاده از تبدیل PCA مکانی جهت ادغام تصاویر چند طیفی و تک رنگ"، پردازش علائم و داده‌ها، دوره ۱۰، شماره ۱، صفحات ۶۹-۷۸، ۱۳۹۲.

[2] H. Ghassemian, and H. R. Shahdoosti. "Multispectral and Panchromatic image fusion using Spatial PCA", in *Signal and Data Processing*, vol. 10, pp. 78-69, 2013.

[3] عسکریان، احسان، کاهانی، محسن، شریفی، شهلا، "حس‌نگار: شبکه واژگان حسی فارسی"، پردازش علائم و داده‌ها. دوره ۱۵، شماره ۱، صفحات ۷۱-۸۶، ۱۳۹۷.

[3] E. Asgarian, M. Kahani, and S. Sharifi, "HesNegar: Persian Sentiment WordNet", *Signal and Data Processing*, vol. 15, pp. 71-86, 2018.

[4] نجف‌زاده، محسن، راحتی قوچانی، سعید، قائمی، رضا، "یک چارچوب نیمه‌نظارتی مبتنی بر لغت‌نامه وفقی خودساخت جهت تحلیل نظرات فارسی"، پردازش علائم و داده‌ها، دوره ۱۵، شماره ۲، صفحات ۸۹-۱۰۲، ۱۳۹۷.

[4] M. Najafzadeh, S. Rahati Quchani, R. Ghaemi, "A Semi-supervised Framework Based on Self-constructed Adaptive Lexicon for Persian Sentiment Analysis", *Signal and Data Processing*, vol. 15, pp. 89-102, 2018.

[5] نوفرستی، سمیرا، شمس فرد، مهرنوش، "ساخت نیمه‌خودکار یک پیکره از نظرات غیر مستقیم در دامنه دارو و به کارگیری آن در تعیین قطبیت نظرات"، مجله پردازش علائم و داده‌ها، شماره ۲، صفحات ۳۵-۴۲، ۱۳۹۵.

[5] S. Noferesti, and M. Shamsfard. "Automatic building a corpus and exploiting it for polarity classification of indirect opinions about drugs", *Signal and Data Processing*, vol 2, pp. 35-42, 2017.

[6] D. Ankitkumar, R. Badre, and M. Kinikar, "A Survey on Sentiment Analysis and Opinion Mining", *International Journal of Innovative Research in Computer and Communication Engineering*, vol. 2, no. 11, November 2014.

[7] J. Blitzer, "Dimensionality Reduction for Language, A Survey of Dimensionality Reduction Techniques for Natural Language", 2008, [Online]. Available:http://john.blitzer.com/papers/wpc2.pdf . [Accessed: 10 July 2017].

[8] M. Chu, F. Diele, R. Plemmons, and S. Ragni, "Optimality, Computation and Interpretation of NonNegative Matrix Factorizations", October 2014. Available: http://users.wfu.edu/plemmons/papers/chu_ple.pdf. [Accessed: 10 July 2017].

[9] G. Golub, and C. V. Loan, Matrix computation, 3th ed. Baltimore, Maryland: JHU Press, 1989.

[10] J. Jotheeswaran, B. MadhuSudhanan, and R.Loganathan, "Feature Reduction using Principal Component Analysis for Opinion Mining", *International Journal of Computer Science and Telecommunications*, vol. 3, no. 5, pp. 118-121, May 2012.

[DOI: 10.29252/jsdp.16.3.88]

[11] J. Jotheeswaran, and S.Koteeswaran, "Feature Selection using Random Forest method for Sentiment Analysis", *Indian Journal of Science and Technology*, vol. 9, no. 3, pp. 1-7, January 2016.

[12] E. Keogh, and A. Mueen, "Curse of dimensionality", In: Encyclopedia of Machine Learning, Springer, pp. 257–258, 2010.

[13] J. Kim, and H. Park, "Sparse nonnegative matrix factorization for clustering", *Technical Report CSE Technical Reports*, GTCSE-08-01, Georgia Institute of Technology, 2008.

[14] D.P. Kingma, and M. Welling, "Auto-Encoding Variational Bayes", Cornell University Library, ArXiv: 1312.6114, December 2013.

[15] D. D. Lee, and H. Sebastian Seung, "Algorithms for Non-Negative Matrix Factorization", *Advances in Neural Information Processing Systems*, vol. 13, pp. 556-562, 2001.

[16] TS. Lee, BC. Shia, and CL. Huh, "Social Media Sentimental Analysis in Exhibition's Visitor Engagement Prediction", *American Journal of Industrial and Business Management*, vol. 06, pp. 392-400. March 2016.

[17] T. Li, Y. Zhang, and V. Sindhwani, "A non-negative matrix tri-factorization approach to sentiment classification with lexical prior knowledge", in *Proceedings of ACL-IJCNLP*, 2009, pp. 244–252.

[18] C. Y. Cheng, J. W Liou, D. R Liou, "Autoencoder for Words", *Neurocomputing*, vol. 139, pp. 84–96, September 2014.

[19] B. Liu,"Sentiment Analysis and Opinion Mining", *Synthesis lectures on human language technologies*, vol. 5. no. 1, pp. 1-167, 2012.

[20] W. Medhat, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications: A survey", *Ain Shams Engineering Journal*, vol. 5, no. 4, pp. 1093-1113, December 2014.

[21] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space", ICLR, 2013.

[22] T. Mikolov, M. Karafiát, L. Burget, J. Cernocký, and S. Khudanpur, "Recurrent neural network based language model," in *INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association*, 2010, pp. 1045–1048.

[23] B. Pang, L. Lee, "Opinion mining and sentiment analysis", Foundations and Trends in Information Retrieval, vol. 2, no. 1-2, pp. 1-135, 2008.

[24] B. Pang, L. Lee, S. Vaithyanathan, "Thumbs up? Sentiment classification using machine learning techniques", in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 79–86, 2002.

[25] W. Rong, Y. Nie, Y. Ouyang, B. Peng, and Z. Xiong, "Auto-encoder Based Bagging Architecture for Sentiment Analysis", *Journal of Visual Languages and Computing*, vol. 25, pp. 840-849, 2014.

[26] G. Vinodhini, and RM. Chandrasekaran, "Opinion mining using principal component analysis based ensemble model for e-commerce application", CSI Transactions on ICT, vol. 2, pp. 169–179, November 2014.

[27] M. E. Wall, A. Rechtsteiner, and L. M. Rocho, "Singular Value Decomposition and Principal Component Analysis", chapter 5 in A Practical Approach to Microarray Data Analysis Kluwer Academic Publishers, Boston, MA, 91-109, 2003.

[28] Wikipedia-Autoencoder, [Online]. Available: https://en.wikipedia.org/wiki/Autoencoder. [Accessed: 10 July 2017].

[29] Y. Yoshida, T. Hirao, T. Iwata, M. Nagata, and Y. Matsumoto, "Transfer learning for multiple-domain sentiment analysis identifying domain dependent/independent word polarity." in *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence*, 2011.

[30] N. Zainuddin,A. Selamat and R. Ibrahim, "Hybrid Sentiment Classification on Twitter Aspect-Based Sentiment Analysis," *Applied Intelligence*, vol. 48, no. 5, pp. 1218-1232, May 2018.

راضیه برادران مدرک کارشناسی خود را در رشته مهندسی کامپیوتر گرایش نرم‌افزار از دانشگاه قم در سال ۸۹ دریافت کرد. کارشناسی ارشد را نیز در همان دانشگاه در رشتهٔ مهندسی فناوری اطلاعات گرایش تجارت الکترونیک در سال ۹۱ به پایان رساند. هم‌اکنون نیز در رشته فناوری اطلاعات گرایش چندرسانه‌ای در دوره دکترای همان دانشگاه مشغول به تحصیل است. زمینه‌های پژوهشی مورد علاقهٔ وی پردازش زبان طبیعی، یادگیری ماشینی و داده کاوی است.

نشانی رایانامه ایشان عبارت است از:

r_baradaran_stu@yahoo.com

عفت گلپر رابوکی مدرک کارشناسی خود را در رشتهٔ ریاضی کاربردی گرایش کامپیوتر در سال ۷۳ از دانشگاه صنعتی امیرکبیر و مدرک کارشناسی ارشد و دکترا را به‌ترتیب در سال ۷۵ و ۹۰ در رشته ریاضی کاربردی از دانشگاه صنعتی شریف دریافت کرد. ایشان از سال ۷۹ عضو هیأت علمی دانشگاه قم و زمینه‌های مورد علاقه‌شان، گرافیک رایانه‌ای، پردازش تصویر و داده‌کاوی است.

نشانی رایانامه ایشان عبارت است از:

g.raboky@qom.ac.ir

[DOI: 10.29252/jsdp.16.3.88]