Karafan

Quarterly Scientific Journal of Technical and Vocational University

Autumn 2023, Vol. 20, No. 3, p. 531-551

Journal Homepage: https://karafan.tvu.ac.ir/?lang=en

:10.48301/KSSA.2023.386361.2458

TVU

Presenting a Framework for Intelligent Sentiment Analysis Using a Novel Method of feature Combination and Meta-initiative in particle Swarm Optimization

Mahdi Basiri[1*], Framerz Fathnejad[2]

[1]Assistant Professor of the Department of Knowledge Management, Faculty of Social Sciences, Command University and Aja.

[2]Assistant Professor of Mathematics Department, Faculty of Basic, Electronic Campus of Azad University.

ARTICLE INFO

Article Type:
Original Research

Received: 03.07.2023
Revised: 07.24.2023
Accepted: 08.30.2023

Keyword:
Feature Combination Method
Intelligent Analysis
Emotions
Group Optimization of Particles

***Corresponding Author:**
Mahdi Basiri
Email: *basiri60@gmail.com*

ABSTRACT

Today, with the increase in the use of the internet, people have turned to the internet to buy their products or to learn about various topics. There are a large number of virtual pages where users post their opinions on various topics. A large amount of data exists which extracting useful information from is a costly and time-consuming task. Opinion mining is the process of intelligent analysis of the sentiments of users who have expressed their opinions in relation to a specific topic with the capability of extract them. The machine learning method is one of the most optimal and efficient methods for extracting knowledge from users' opinions on the offered products. In these methods, the training data of a system is given to classify user opinions. One of the most important classification steps is data reduction. By using the new feature combination method, the set of extracted features can be reduced to a greater extent than the feature selection method, which leads to a subset of useful information with a much smaller volume and higher recognition power. In this research, particle group optimization algorithm was used to optimize the combination of features. To evaluate the proposed method, MATLAB software was used to evaluate the proposed method, and experiments were conducted on four data sets. The results of the research showed that the use of the feature combination method increased the efficiency of classification and reduced the effect of this increase in the decrease of the efficiency of the classifier.

©2023 the authors. Published by Technical and Vocational University, Tehran, Iran. This article is an open-access article distributed under the terms and conditions of the Creative Commons Attribution-Noncommercial 4.0 International (CC BY-NC License) (https://creativecommons.org/licenses/by-nc/4.0/)

E-ISSN: 2538-4430
ISSN: 2382-9796

EXTENDED ABSTRACT

Introduction

Today, with the increase in the use of the internet, people turn to the internet to buy their products or to learn about various topics. There are a large number of virtual pages where users post their opinions on various topics, and extracting useful information from this large amount of data is a costly and time-consuming task. Opinion mining is the process of intelligent analysis of the sentiments of users who have expressed their opinions in relation to a specific topic and has the capability to extract them. Opinion mining is a new and emerging research field that uses data mining and natural language processing to recover information and discover knowledge from text. The goal of opinion mining is to enable computers to recognize and express emotions. A view or behavior that is based on emotions instead of logic is called emotion. Therefore, Kawi opinion is also known as sentiment analysis. Business organizations have spent a great deal of money on consultants and researchers to know the feelings and opinions of customers about their products. Similarly, people are interested in other people's opinions about products, services and topics to find the best choices.

Methodology

This type of research is quite easy to gather through web forums, blogs, discussion groups, and comment boxes. The machine learning method is one of the most optimal and efficient methods for extracting knowledge from users' opinions about the offered products. In these methods, the training data of a system is given to classify user opinions. One of the most important classification steps is data reduction. By using the new feature combination method, the set of extracted features can be reduced more than the feature selection method, which leads to a subset of useful information with a much smaller volume and higher recognition power.

Results and discussion

The present research was analytical-experimental in terms of its practical purpose and its implementation method. In this research, to increase the speed and accuracy of classification, data reduction methods using the feature combination method were used. In this method, after the feature selection stage, attempts were made to find combinations of features that lead to building a stronger feature. In addition, fold cross-validation was carried out on the data sets to minimize the effect of changes in the sets. Out of the total 100% of data already in the system, 90% of the data was considered as training data and the remaining 10% was presented to the algorithm as a test subject to knowing in which group the 10% of the data was classified. After presenting this 10% of the data to the algorithm, it can be compared with the real grouping to check the accuracy of the classification.

After finishing the work of this 10% that was assigned to the algorithm as a test, another 10% of the total data was selected for the next stage test and the remaining 90% was assigned to the training data. To obtain the final accuracy result of the presented method, this process was carried out in ten steps with different 10% and the average was obtained.

In this research, a particle group optimization algorithm was used to optimize the combination of features. MATLAB software was used to measure the proposed method, and experiments were conducted on four data sets. The results of the research showed that the use of the feature combination method increases the efficiency of classification and reduces the effect of this increase in the drop of classifier efficiency.

In this research, a new method for the classification of emotions based on the combination of features is presented. The presented framework is based on two new perspectives of feature selection from the mass of data and a combination of selected features. This new method turns the set of features into a subset of stronger features, which reduces the cost, and time and increases the accuracy of information classification.

In the current research, a new method for the classification of emotions based on the combination of features is presented. The presented framework is based on two new perspectives of feature selection from the mass of data and a combination of selected features. This new method turns the set of features into a subset of stronger features, which reduces the cost, and time and increases the accuracy of information classification.

One of the strengths of using the feature selection and feature combination methods is the consensus of the potentials of the defining features of the text (words) and the reduction of features to increase the efficiency of the classification accuracy, which can be cited in the results obtained between the various methods.

Conclusion

In the present research, experiments were conducted to evaluate the proposed method and the experimental results were analyzed using three classification techniques (SVM, NB, KNN) and data reduction (feature selection). It is also concluded from the obtained results that the use of the feature combination method increases the efficiency of classification and reduces the effect of this increase in the loss of efficiency of the classifier. The point that should be kept in mind in future research is that the greater the relationship between the words chosen to combine the features, the greater the improvement in the accuracy of the system. Using the technique of combining with a hierarchical structure and choosing words according to this structure can help this goal.

Finally, extracting and discovering knowledge from a very large amount of data requires a great deal of time and money. Therefore, we must use methods to reduce the data size. Data reduction techniques can be applied without losing the correctness of information and final results. By reducing data in different stages of data mining processing, it can bring simplicity to the presented model so that it is more understandable.

کارفن
فصلنامه علمی دانشگاه فنی و حرفه‌ای
پاییز ۱۴۰۲، دوره ۲۰، شماره ۳، ۵۵۱-۵۳۱
آدرس نشریه: https://karafan.tvu.ac.ir/
doi:10.48301/KSSA.2023.386361.2458

ارائه چارچوبی برای تحلیل هوشمند احساسات با استفاده از روش جدید ترکیب ویژگی و فرا ابتکاری در بهینه‌سازی گروهی ذرات

مهدی بصیری[1*]، فرامرز فتح‌نژاد[2]

۱- استادیار، گروه مدیریت دانش، دانشکده علوم اجتماعی، دانشگاه فرماندهی و ستاد آجا.

۲- استادیار، گروه ریاضی، دانشکده علوم پایه، دانشگاه آزاد واحد الکترونیکی.

اطلاعات مقاله

نوع مقاله: مقاله پژوهشی

دریافت مقاله: ۱۴۰۱/۱۲/۱۶
بازنگری مقاله: ۱۴۰۲/۰۵/۰۲
پذیرش مقاله: ۱۴۰۲/۰۶/۰۸

کلید واژگان:
روش ترکیب ویژگی
تحلیل هوشمند
احساسات
بهینه‌سازی گروهی ذرات

***نویسنده مسئول:** مهدی بصیری
پست الکترونیکی:
basiri60@gmail.com

چکیده

امروزه با افزایش زمینهٔ استفاده از اینترنت، افراد برای خرید محصولات خود و یا اطلاع از موضوعات مختلف به اینترنت مراجعه می‌نمایند. تعداد زیادی از صفحات مجازی وجود دارند که کاربران نظرات خود را در مورد موضوعات مختلف درج می‌کنند، در نتیجه حجم زیادی از داده‌ها وجود دارد که استخراج اطلاعات سودمند از آنها کار پرهزینه و زمانبری است. نظرکاوی[1] فرآیند تحلیل هوشمند احساسات کاربرانی است که نظرات خود را در ارتباط با یک موضوع مشخص طرح نموده و قابلیت استخراج دارند. روش یادگیری ماشین یکی از بهینه‌ترین و کارآمدترین روش‌ها برای استخراج دانش از میان نظرات کاربران دربارهٔ محصولات ارائه شده می‌باشد. در این روش‌ها، داده‌های آموزشی یک سیستم جهت طبقه‌بندی نظرات کاربران داده می‌شود. یکی از مهمترین مراحل طبقه‌بندی، کاهش داده می‌باشد. با به‌کارگیری روش جدید ترکیب ویژگی[2] می‌توان مجموعهٔ ویژگی‌های استخراج شده را بیشتر از روش انتخاب ویژگی کاهش داد، که به یک زیر مجموعه‌ای از اطلاعات مفید با حجم بسیار کمتر و میزان قدرت تشخیص بالاتر رسید. در این تحقیق از الگوریتم بهینه‌سازی گروهی ذرات جهت بهینه کردن ترکیب ویژگی‌ها استفاده شده است. برای سنجش روش پیشنهاد شده از نرم‌افزار متلب استفاده شده است که بر روی چهار مجموعهٔ داده، آزمایش‌هایی صورت گرفت. نتایج تحقیق نشان داد که استفاده از روش ترکیب ویژگی، کارایی طبقه بندی را افزایش داده و از تأثیر این افزایش در افت کارایی دسته‌بندی کننده می‌کاهد.

[1] Opinion Mining
[2] Feature Unionization

شاپای الکترونیکی: ۲۵۳۸-۴۴۳۰
شاپای چاپی: ۲۳۸۲-۹۷۹۶

©2023 the authors. Published by Technical and Vocational University, Tehran, Iran. This article is an open-access article distributed under the terms and conditions of the Creative Commons Attribution-Noncommercial 4.0 International (CC BY-NC License) (https://creativecommons.org/licenses/by-nc/4.0/)

مقدمه

در عرصهٔ صنعت و تولید سازمان‌هایی که محصولات تازه ارائه می‌کنند، نظرات و احساسات مشتریانشان در خصوص محصولات، برایشان مهم و قابل توجه می‌باشد و کارشناسان متخصص در متن کاوی اطلاعات به دست آمده را مورد تجزیه تحلیل قرار می‌دهند؛ برای مطالعهٔ دیدگاه‌ها و عقاید مشتریان همواره از نظرات آنان با آغوش باز استقبال می‌کنند. در عین حال، کاربران یا مشتریان نیز می‌خواهند قبل از خرید محصولات از نظرات و عقاید دیگر کاربران مطلع گردند و بر اساس این نتایج به انتخاب محصول مورد نظر خود اقدام نمایند.

در چنین شرایطی و با رقابتی‌تر شدن محیط فعالیت کسب و کارها، رویکرد مشتری محوری و توجه بر نیازهای مشتری در سطح جهانی، نقش حیاتی در رشد و توسعهٔ کسب و کارها ایفا می‌نماید [۱]. نظرکاوی برخط روز به روز حساس‌تر و مجبوب‌تر می‌شود و اطلاعات به‌دست آمده هم برای تولیدکنندگان و هم برای مشتریان جویای محصولات ارزش روز افزون‌تری دارد. با این حال، انجام این فرایند به صورت دستی دشوار و زمان‌بر است. تولید کنندگان با حجم انبوهی از نظرات کاربران سر و کار دارند و این امر موجب شده‌است تا راهکاری کارآمدی جهت سهولت دریافت این نظرات از کاربران سیستم باشند، در این راستا از خودکار سازی فرایند دریافت نظرات استفاده نمودند که به کاربران این امکان را می‌دهد که نظرات خود را به صورت فرم متن، تصویر، صوت یا دادهٔ ویدئویی بیان نمایند [۱].

نظرکاوی[1] یکی از روش‌های استخراج تحلیل هوشمند احساسات کاربرانی است که نظرات خود را در باب یک موضوع مطرح شده، نوشته‌اند. به عبارت بهتر نظرکاوی زمینهٔ مطالعاتی است که سعی می‌کند احساسات، رفتار، نظرات و تحلیل افراد مختلف را نسبت به یک موجودیت و ویژگی‌های آن بیان کند؛ نظرات اشخاص دیگر می‌تواند در خصوص محصولی که استفاده کرده یا خریده‌اند بسیار مهم باشد. بر این اساس مقاله پیش‌رو به دنبال ایجاد یک سیستم طبقه‌بندی شده از نظرات افراد به منظور کارایی در استفاده از روش‌های کاهش داده و تسریع فرایند طبقه‌بندی داده‌ها می‌باشد [۲].

مبانی نظری

نظرکاوی

نظرکاوی یا تحلیل احساسات یک زمینهٔ تحقیقاتی جدید و نو ظهور است که با استفاده از داده کاوی و پردازش زبان طبیعی بازیابی اطلاعات و کشف دانش از متن صورت می‌گیرد. هدف نظرکاوی این است که رایانه را قادر سازیم بتواند احساسات را تشخیص دهد و بیان کند. دید یا رفتاری که بر اساس احساسات به جای منطق باشد، احساس گفته می‌شود. بنابراین نظرکاوی به تحلیل احساسات نیز معروف است [۳].

تحلیل احساسات[2]

تحلیل احساس موضوعی است که به تحلیل با جزییات نظرکاوی می‌پردازد در نظرکاوی ما به دنبال طبقه‌بندی نظرات به دو گروه مثبت و منفی هستیم ولی در تحلیل احساسات به جزئیات بیشتری پرداخته می‌شود. بدین شکل که احساسات مثبت و منفی را نیز به گروه‌های کوچک‌تر تقسیم می‌کنیم. حس منفی شامل تنفر، ترس، عصبانیت، بی‌حرکتی و حس مثبت به اعتماد و خوشحالی تقسیم می‌شود. در برخی موارد حس بی‌طرفانه قسم سوم کار است و شامل تعجب کردن و مشارکت می‌شود. در تحقیقات اخیر در زمینهٔ نظرکاوی تحلیل احساسات بیشتر مورد توجه بوده است[۴].

[1] Opinion Mining
[2] Sentiment Analysis

روش‌های مختلف انتخاب ویژگی

با توجه به پیچیدگی مسائل و تغییر مداوم آنها در دنیای واقعی از سال ۲۰۰۰ به بعد شاهد ارائه الگوریتم‌های فرا ابتکاری جدید جهت حل مسائل مختلف هستیم. جهت حل مسأله انتخاب ویژگی با استفاده از الگوریتم‌های فرا ابتکاری، هر جواب ممکن مسأله به صورت رشته‌ای از صفر و یک تعریف می‌شود. طول رشته مساوی تعداد کل ویژگی‌ها بوده و مقدار صفر برای هر بیت نشان‌دهنده عدم انتخاب و مقدار یک بیانگر انتخاب ویژگی مربوطه می‌باشد. الگوریتم‌های فرا ابتکاری از تابع شایستگی برای ارزیابی و جستجو استفاده نموده و به علت هوش جمعی قادر به کشف جواب می‌باشند [۵].

جدول ۱. تعدادی از جدیدترین الگوریتم‌های فراابتکاری.

ردیف	عناون مقاله	نویسندگان	سال ارائه	مرجع
۱	انتخاب ویژگی‌ها و الگوریتم دسته‌میگو پیشرفته برای خوشه‌بندی اسناد متنی	ابولیکا[1]	۲۰۱۹	[۶]
۲	الگوریتم بهینه‌سازی نهنگ برای انتخاب ویژگی	مفرجا و میرجلیلی[2]	۲۰۱۸	[۷]
۳	انتخاب ویژگی‌ها با استفاده از الگوریتم بهینه‌سازی جنگل	قایمی، فیضی درخشی	۲۰۱۶	[۸]
۴	انتخاب ویژگی با استفاده از الگوریتم جمعیت مورچگان	کاشف و نظام آبادی‌پور	۲۰۱۵	[۹]
۵	انتخاب زیر مجموعه ویژگی با رویکرد بهینه‌سازی گرگ خاکستری	اماری[3] و همکاران	۲۰۱۵	[۳]

مروری بر روش بهینه‌سازی گروه ذرات

روش بهینه‌سازی گروه ذرات در اواسط دهۀ ۱۹۹۰ توسط کندی و ابرهارت اختراع گردید در این روش، حرکت گروه پرندگان به عنوان بخشی از مطالعۀ اجتماعی شناختی که به پژوهش در مورد تصور هوش جمعی در جوامع زیستی می‌پردازد، شبیه‌سازی می‌گردد. در روش PSO، مجموعۀ راه‌حل‌های تصادفاً انتخاب شده (گروه اولیه) در فضای طراحی در جهت نیل به راه حل بهینه در میان تعدادی تکرار(حرکت) بر اساس مقدار زیاد اطلاعات موجود در مورد فضای طراحی منتشر می‌شود که باهم تلفیق شده و کلیۀ اعضای گروه از آن بهره می‌برند. روش PSO از توانایی دسته‌های پرندگان، دستۀ ماهی‌ها و گلۀ جانوران برای سازش با محیط، یافتن منابع سرشار غذایی و دوری از شکارچیان (صیادان) با اجرای شیوۀ تقسیم اطلاعات الهام گرفته و یک حسن تکاملی دارا می‌باشد.

در حرکت دسته‌جمعی پرندگان همیشه حرکت تک‌تک پرندگان گرایش به سمت پرندۀ هدایت‌گر(سردسته) دارد. اگر به طور اتفاقی پرندۀ هدایت‌گر انحراف کوچکی از مسیر اصلی داشته باشد، همۀ پرندگان این انحراف کوچکی را همراهی می‌کنند. سردسته بهترین موقعیت را دارد. یافتن بهترین و بهینه‌ترین نقطه برای هر ذره مطلوب است و هر ذره برای یافتن بهترین نقطه همیشه در حال تغییر به وضعیت مطلوب‌تر است. و بنا به جابه‌جا شدن قطعاً سرعت دارد.

الگوریتم PSO نیز مشابه سایر الگوریتم جمعیتی از میان مجموعه پاسخ‌های ممکن به دنبال پاسخ بهینه می‌گردد و این جستجو تا زمانی ادامه دارد که شرایط پایان الگوریتم وجود نداشته باشد. در اینجا هر x پاسخ ذره نمایش داده شده است (این تشابه مانند کروموزم در الگوریتم ژنتیک است).

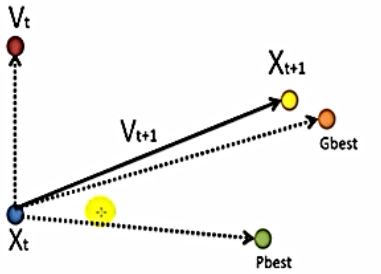
برای انجام کار باید به تشابه دیگر الگوریتم‌ها یک جمعیت اولیه داشته‌باشیم. در اینجا ضمانت حرکت به سمت ناحیه بهینه معادلۀ سرعت ذرات می‌باشد. در این معادلۀ سه عضو اصلی به شرح ذیل می‌باشند:

- **سرعت**

[1] Abualigah
[2] Mafarja & Mirjalili
[3] Emary

- **مؤلفه شناختی pbest**: بهترین حالات ممکن برای ذره
- **مؤلفه جمعی gbest**: بهترین ذره از میان انبوه ذرات که تا به‌حال نمایان شده است

برای پیاده‌سازی مجازی عکس‌العمل‌های تک‌تک ذره‌ها از بهترین ذرهٔ محلی (بهترین حالتی را که یک ذره تاکنون داشته است و یا داخل یک همسایگی مشخص است) (Personal Best) (و یا ممکن است بهترین ذره عمومی (Global Best)) از میان حجم انبوه ذرات بهترین ذره انتخاب شود) باشد.



Xt: حالت قبلی (فعلی) ذره است

Xt+۱: حالت کنونی (آینده) است

اگر ما به حالت قبلی (فعلی) مقدار عددی سرعت آینده را اضافه کنیم به حالت کنونی (آینده) می‌رسیم

Vt+۱ سرعت کنونی از میزان شانسی سه مؤلفه (سرعت قبلی (فعلی) و gbest و pbest) به‌دست می‌آید.

Pnew = Pold + Vnew

Vnew=Vold+C1*R1*(Plocal best -Pold)+C2*R2*(Pglobal best -Pold)

C۱ و C۲ مقادیر عددی ثابت و همیشه مثبت و R۱ و R۲ اعداد شانسی هستند که حتماً در بازهٔ [۰,۱] تولید می‌شوند.

پارامتری به نام وزن اینرسی برای قابلیت بهتر جستجو، به صورت ضریبی در پارامتر سرعت الگوریتم اضافه می‌گردد:

Vnew=W*Vold+C1*R1*(Plocal best-Pold)+C2*R2*(Pglobal best-Pold)

وزن اینرسی میزان تأثیر سرعت ذرات در گام قبل را بر سرعت فعلی تعیین می‌نماید. به‌صورتی که مقادیر بزرگ از وزن اینرسی موجب افزایش بهبود قابلیت جستجوی عمومی الگوریتم می‌شود و فضای بیشتری مورد بررسی قرار می‌گیرد. حال آن‌که با مقادیر کوچک وزن اینرسی فضای مورد بررسی محدود شده و جستجو در این فضای محدود شده صورت می‌گیرد. به همین دلیل، معمولاً الگوریتم با مقدار بزرگی از وزن اینرسی شروع به حرکت می‌کند که سبب جستجوی گستردهٔ فضا در ابتدای اجرای الگوریتم شده و این وزن به مرور در طول زمان کاهش می‌یابد که سبب تمرکز جستجو در فضای کوچک در گام‌های پایانی می‌شود.

تمامی الگوریتم‌ها از دو مکانیسم تنوع و تمرکز استفاده می‌کنند: در مکانیسم تنوع ابتدا الگوریتم فضای بیشتری از فضای جستجو را مورد جستجو قرار دهد و هرچه به انتهای الگوریتم نزدیک‌تر می‌شویم میزان تنوع را کم می‌کنیم و فرایند جستجو را به سمت ناحیهٔ بهینه متمرکز می‌کنیم.

w (وزن اینرسی) در واقع بیان‌کنندهٔ میزان تنوع است که در ابتدای الگوریتم زیاد و در مراحل انتهایی الگوریتم میزان آن کم و کمتر خواهد شد.

در مراحل اولیه ذرات به‌صورت تصادفی در سرتاسر فضای جستجو مقداردهی می‌شوند که این موقعیت‌های اولیه به‌دست آمده از مقداردهی در فضای جستجو به‌عنوان بهترین تجربهٔ شخصی ذرات نیز شناخته می‌شوند.

در قدم بعدی از میان ذرات موجود بهترین ذره انتخاب و به عنوان بهترین پاسخ (gbest) شناخته می‌شود. سپس گروه ذرات تا زمانی که شرایط پایان انتهایی الگوریتم تحقق یابد در فضای جستجو حرکت می‌نمایند. حرکت ذرات در فضای جستجو شامل اعمال معادلهٔ سرعت به گروه ذرات می‌باشد که موقعیت هر ذره براساس معادلهٔ اعمال شده بر آن تغییر می‌کند.

مقدار برازش جدید حاصل از ذره، با مقدار بهترین پاسخ (pbest) ذره مقایسه می‌گردد، اگر موقعیت جدید به دست آمده ذره دارای برازش بهتری باشد، این موقعیت جدید جایگزین موقعیت pbest می‌شود.

کارهای مرتبط

حسینی و سادات نواب [۱۰] در تحقیقی با عنوان «ارائه رویکرد ترکیبی مبتنی بر الگوریتم بهینه سازی ذرات و الگوریتم جستجوی گرانشی برای انتخاب ویژگی» به دنبال ارائه روشی برای تحلیل و دسته‌بندی حجم زیاد داده‌های افراد در شبکه‌های اجتماعی مجازی بوده است. نتایج تحقیق وی نشان داد که الگوریتم ترکیبی پیشنهادی محققان دقت بیشتری را نسبت به الگوریتم جستجوی گرانشی نشان می دهد.

محمدی و ناظمی [۲] در پژوهشی با عنوان تجزیه و تحلیل احساسات در سطح ویژگی محصول و مبتنی بر جنسیت کاربران به بررسی احساسات مشتریان دو شرکت اپل و سامسونگ در شبکهٔ اجتماعی توئیتر پرداخته است. نتایج تحقیق وی نشان داد محبوبیت ویژگی‌های مختلف محصول بین کاربران مرد و زن متفاوت بوده و بر اساس این نتایج، صاحبان کسب و کار می‌توانند اقدام به تولید محصولاتی با مرکز جنسیت افراد نمایند.

محمدی و خلج [۱۱] در مقالهٔ خود با عنوان ارائه مدلی برای عقیده کاوی در سطح ویژگی برای نظرات کاربران هتل‌ها، یک روش ترکیبی و جدید بر اساس یک رویکرد رایج در تحلیل احساسات، استفاده از واژگان و الگوریتم ژنتیک برای تولید ویژگی‌هایی برای طبقه‌بندی بار احساسی نظرات، ارائه شده است. بدین صورت که دو روش ساخت فهرست واژگان یکی با استفاده از روش‌های آماری و دیگری با استفاده از الگوریتم ژنتیک ارائه گردیده است. واژگان موردالاشاره با فرهنگ واژگان احساس عمومی و استاندارد لیو بینگ آمیخته می‌شوند. نتایج نشان می‌دهد روش پیشنهادی از روش‌های پایه براساس واژه نامه‌های احساسی روی این مجموعه داده بهتر عمل کرده و معیارهای ارزیابی صحت، دقت، بازخوانی بالا بوده است.

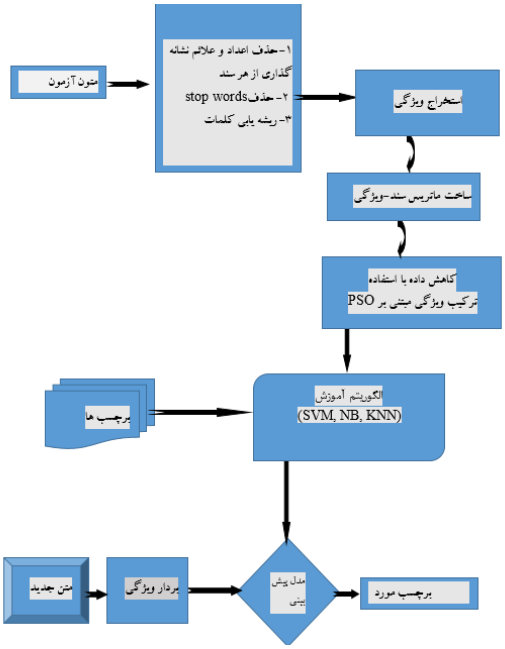
ابراهیم و وانگ[1] [۱۲] در سال ۲۰۱۹ از لغت نامهٔ سنتی استرنت برای تعیین گرایش، قدرت و شدت احساسات توئیت‌های کاربران در مورد ۵ برند خرده فروشی برخط در انگلستان استفاده کردند که به هر توئیت یک نمره در بازه (+۵ و – ۵) اختصاص می‌دهد. نتایج این پژوهش نشان داد که استفاده از این لغت‌نامه به دلیل آن‌که از قوانین زبانی شامل منفی کننده‌ها و شکلک‌ها برای محاسبهٔ قدرت و شدت قطبیت نظرات استفاده می‌کند، برای توئیتر که شامل داده‌های کوتاه و غیررسمی است به نحوی جواب می‌دهد.

نظیم شا و راجسواری[2] [۱۳] در تحقیقی به بررسی نقش پیشرفت‌های فناوری اطلاعات با تمرکز بر هوش مصنوعی در ردیابی پنج حس انسانی اشاره نموده‌اند. نتایج تحقیق آنها نشان داد که حس بینایی، شنوایی، چشایی، بویایی و لامسه قابلیت ردیابی توسط هوش مصنوعی را دارند و در نتیجه ارتباط بهتری بین محصول و مشتری برقرار می‌گردد.

[1] Ibrahim & Wang
[2] Nazim Sha & Rajeswari

روش‌شناسی

تحقیق پیش‌رو از نظر هدف کاربردی و روش اجرای آن توصیفی–تحلیلی است. در این پژوهش برای افزایش سرعت و دقت طبقه‌بندی از روش‌های کاهش داده به روش ترکیب ویژگی، بهره گرفته شده است. در این روش پس از مرحلهٔ انتخاب ویژگی‌های سعی در یافتن ترکیباتی از ویژگی‌ها که منجر به ساختن یک ویژگی قوی‌تر می‌گردد که از روش PSO جهت جستجو استفاده می‌شود. شکل ۱ نحوهٔ گردش کار و مدل پیشنهادی را نشان می‌دهد.



شکل ۱. مدل پیشنهادی تحقیق.

تجزیه و تحلیل داده‌ها

پیش پردازش‌های زبانی

در مرحلهٔ پیش پردازش برای این‌که بتوانیم یک الگوریتم وزن‌دهی داشته باشیم. در مرحلهٔ پیش پردازش برای انجام پردازش با مراجعه به یک ماتریس از محورهای اسناد در ماتریسی به نام سند- کلمه پیش پردازش با مراجعه به این ماتریس فعالیت‌های معلوم، معمول و مشترکی بر روی مدارک از لحاظ زبانی انجام شود.

حذف کلمات زائد

در جملات به کلماتی زائد گفته می شود که واسط یا رابط جملات هستند و تکرارشان در جملات زیاد است، به مانند: «و»، «گر»، «یا»، «the»، «or»، این کلمات با این‌که بار معنایی ندارند و و نقششان در جملات فقط رابط و واسط می‌باشند؛ به دلیل نداشتن بار معنایی در مرحلهٔ پیش‌پردازش زبان طبیعی حذف می‌گردند. برای حذف کلمات رابط باید

یک فرهنگ لغاتی یا لیستی تهیه نمود و هرکجا در متن به این کلمات رسیدیم با مراجعه به لیست یا فرهنگ لغات از متن حذف گردد. در زبان انگلیسی چندین نمونه از این فرهنگ لغات تدارک دیده شده است که تقریباً شامل میانگین ۵۰۰ کلمه است.

ریشه‌یابی کلمات

ریشه‌یابی یعنی کاهش دادن لغات به ریشه‌های آنها گفته می‌شود. بنابراین«computing» و «compute» و «computer» به «compute» که ریشهٔ اصلی است کاهش می‌یابند. الگوریتم ریشه‌یاب «مارتین پورتر » جزء مهم‌ترین و معروف‌ترین الگوریتم‌های ریشه‌یاب است.

استخراج ویژگی‌ها

استخراج ویژگی به فرایندی گفته می‌شود که با استفاده از الگوریتم‌های کاهش داده و آماری می‌توان ویژگی‌های مهم و ارزشمند داده‌های ورودی را به‌دست آورد. در این پژوهش عملیات داده‌کاوی بر روی داده‌های متنی انجام می‌شود که با هدف تحلیل هوشمند احساسات توسط سیستم‌های رایانه‌ای از روی متون صورت می‌گیرد در نتیجه کلمات مهم به عنوان ویژگی در نظر گرفته می‌شود و استخراج می‌گردد.

تشکیل ماتریس نهایی سند-لغت

در مرحلهٔ تشکیل ماتریس نهایی پس از مشخص شدن ویژگی‌ها یک ماتریس (صفر و یک) را تشکیل می‌دهیم. که در این ماتریس ستون‌های آن بر اساس ویژگی‌های استخراج شده از اطلاعات و هر سطر آن نشان‌دهندهٔ یک سند یا الگو می‌باشد. از این ماتریس (صفر و یک) برای آموزش سیستم تشخیص استفاده خواهیم کرد.

کاهش داده

در کل کاهش داده دو مرحله صورت می‌گیرد: نمونه و سطح ویژگی. در این تحقیق برای وزن دادن به داده‌ها از دو روش ترکیب ویژگی و انتخاب ویژگی توسط الگوریتم IG1 استفاده می‌کنیم. الگوریتم IG در واقع داده‌های نامناسب را حذف می‌کند و سپس ویژگی‌های به‌دست آمده را ترکیب می‌نماید که این عملیات موجب تشکیل یک زیر مجموعه از ویژگی‌های کاهش‌یافته و قوی‌تر می‌شود، با هدف افزایش کارایی، سرعت و دقت بالاتر در یادگیری می‌شود.

در این پژوهش روش ترکیب ویژگی بدون استفاده از تکنیک جستجو و استفاده از تکنیک جستجو PSO جهت بهینه‌سازی ترکیبات ویژگی‌ها با یگدیگر مقایسه شده است.

- FU (ترکیب ویژگی ساده)
- FU-PSO (ترکیب ویژگی مبتنی بر PSO)

انتخاب ویژگی (Feature Selection)

حذف ویژگی‌های نامرتبط و تکراری یکی از راه‌کارهای انتخاب ویژگی می‌باشد. یک مجموعه کاهش یافته و مؤثرتر و قابل قبول‌تر از ویژگی‌ها می‌سازد. یکی از بهترین و پرکاربردترین این تکنیک‌ها IG می‌باشد که این روش نیز مورد استفاده قرار گرفته است.

[1] Information Gain

$$IG = P(t_k, c_i) \log \frac{P(t_k, c_i)}{P(t_k).P(c_i)} + P(\bar{t}_k, c_i) \log \frac{P(\bar{t}_k, c_i)}{P(\bar{t}_k).P(c_i)}$$

ترکیب ویژگی (Feature Unionization)

هر روزه داده‌های فراوانی توسط افراد دربارهٔ محصولات و خدمات در سایت‌های رسانه‌ای و اجتماعی ایجاد می‌شود، به دلیل انبوه داده‌های ایجاد شده این داده‌ها بدون ساختار و پویا و دارای نویز می‌باشند در سایت‌های رسانه‌ای و اجتماعی هیچ مرکزیتی برای داده‌های ایجاد شده وجود ندارد و موجب شده این داده ها به صورت پراکنده در همه جا پخش شوند. تجزیه و تحلیل داده‌های به‌دست آمده و تغییر شکل آنها و همچنین تکنیک‌هایی که در مرحلهٔ جمع آوری داده‌ها استفاده می‌شوند، مناسب برای داده‌هایی با حجم پایین و متوسط می‌باشند. داده‌کاوی، داده‌هایی با حجم بسیار بالا نیازمند صرف زمان زیادی است پس باید از روش‌هایی برای کاهش اندازه داده‌ها استفاده کنیم. تکنیک‌هایی که برای کاهش داده‌ها استفاده می‌شوند نباید باعث از دست دادن درستی داده‌ها شود و بدون به خطر انداختن نتایج نهایی استخراج دانش وارد عمل شوند. کنکاش بر روی داده‌های کمتر سرعت بیشتر و کارایی بالاتری دارد. در نتیجه عملیات پردازش بر روی داده‌هایی با حجم کمتر موجب تسریع در عملیات و کارایی بهینه‌تر خواهد شد که با کاهش داده‌ها در مراحل مختلف استخراج دانش و سادگی مدل پیشنهادی و نمایش اطلاعات خواهد شد که مدل قابل درک بهتری خواهد بود. تسریع در فرایند داده‌کاوی، افزایش دقت و کارایی و ساده‌سازی داده‌ها از اهداف کلی روش‌های کاهش داده‌ها تلقی می‌شوند. اما دستیابی هم‌زمان به تمام این اهداف شدنی نیست.

لذا نگهداری این اهداف در سطح مطلوب و برقراری توازن بین آنها روشی مناسب و امکان‌پذیر به نظر می‌رسد. در این پژوهش کاهش اطلاعات در دو مرحله صورت می‌گیرد که عبارتند از: کاهش صفات خاصه (ستون) و کاهش نمونه‌ها (سطر) [۱۴].

یکی از بهینه‌ترین روش‌های کاهش اطلاعات در سطح صفات خاصه انتخاب ویژگی (Feature Selection) می‌باشد [۱۵].

این روش با حذف ویژگی‌های نامرتبط و زائد، یک زیرمجموعهٔ بهینه از ویژگی‌ها می‌سازد، به‌طوری‌که باعث افزایش کارایی طبقه‌بندی می‌گردد. اگرچه روش انتخاب ویژگی، یک زیرمجموعهٔ کاهش یافته از ویژگی‌ها با قدرت تشخیص بالاتر، جهت طبقه‌بندی را پیدا می‌کند، ولی این زیرمجموعه با حفظ قدرت تشخیص همچنان قابل کاهش می‌باشد.

برای طبقه‌بندی متون، کلمات هم سطح (هم‌معنی) به عنوان ویژگی‌های مختلف در طبقه‌بندی‌های مختلف قرار می‌گیرند. برای مثال در طبقه‌بندی متون کلمات «خوب» و «عالی» به عنوان دو ویژگی نزدیک به‌هم لحاظ می‌شوند که اگر ترکیب گردند مسلماً قابلیت ساخت یک ویژگی قوی‌تر خواهد شد. جرقهٔ اصلی ترکیب ویژگی‌ها از اینجا آمد که ترکیب ویژگی‌های نزدیک به‌هم منجر به تولید ویژگی‌های بهتر می‌شود. در این روش، انتخاب ویژگی‌ها یک فرآیند جستجو در میان داده‌ها می‌باشد که در فرآیند جستجو هدف اصلی یافتن بهترین ترکیب از ویژگی‌ها می‌باشد.

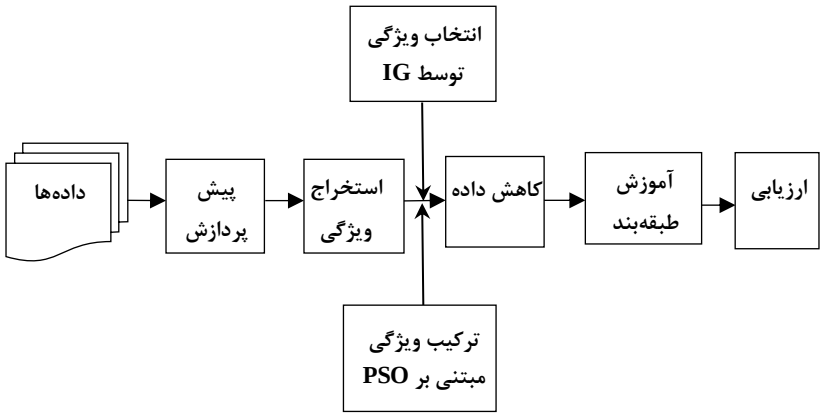
آموزش دسته‌بندی کننده

به منظور ایجاد یک تناظر بین مجموعه‌ای از کلاس‌های از پیش تعریف شده و اسناد متنی نیازمند دسته‌بندی متون می‌باشیم. در واقع این روش برای یافتن رابطه‌ای بین یک مجموعه از متون $D=\{di, \ldots, d2, d1\}$ با مجموعه‌ای از موضوعات $C=\{cj, \ldots, c2, c1\}$ می‌باشد. از اسناد متنی (کلاس‌های معین) برای طبقه‌بندی یک مجموعهٔ آموزشی استفاده می‌شود. با استفاده از مجموعهٔ آموزشی، مدل طبقه‌بندی ارائه می‌شود کلاس سند جدید را مشخص می‌سازد. برای سنجش میزان کارایی مدل طبقه‌بندی شده نیاز به یک مجموعه از داده‌ها برای تست نیازمندیم که این داده‌ها مستقل از مجموعهٔ داده‌های آموزشی در نظر گرفته می‌شود. برچسب‌های تخمین زده شده با برچسب واقعی اسناد

مقایسه می‌شود. روش‌های ارائه شده در زمینهٔ دسته‌بندی متون عبارتند از: درخت‌های تصمیم‌گیری، روش نزدیک‌ترین همسایه، روش بیزین ساده، روش شبکه‌های عصبی، ماشین‌های بردار پشتیبان (SVM) و غیره .

در این پژوهش از روش ماشین‌های بردار پشتیبان استفاده می‌شود.

دقت عبارت است از: نسبت اسنادی که به درستی طبقه‌بندی شده‌اند به تعداد کل اسناد [۱۶].



شکل ۲. چارچوب روش پیشنهادی.

ارزیابی طبقه‌بندی

در این تحقیق برای ارزیابی نتایج طبقه‌بندی از معیارهای میزان کاهش داده و میزان دقت (Accuracy) استفاده می‌کنیم.

$$accuracy = \frac{Tn}{N}$$

در فرمول نوشته شده Tn تعداد داده‌های درست تشخیص داده شده نسبت به کل داده‌ها(N) می‌باشد .

و درفرمول زیر Dn تعداد نمونه‌های انتخاب شده (ویژگی‌ها) به کل داده‌ها (N) است.

$$r = \frac{Dn}{N}$$

پیاده‌سازی

داده‌های آموزشی پس از پاک‌سازی و استخراج ویژگی‌ها و ترکیب آنها منجر به ساخت یک زیر مجموعه‌ای از داده آموزشی با کیفیت بالاتر می‌گردد. سپس از این مجموعه داده جهت آموزش طبقه‌بندی استفاده می‌گردد. برای اجرای الگوریتم پیشنهادی بر روی مجموعه‌ای از داده‌هایی که خروجی نهایی آن ماتریس وزن‌دهی شدهٔ سند-لغت می‌باشد، احتیاج به پیمودن گام‌هایی در پیش پردازش مجموعهٔ داده‌های به‌منظور آماده‌سازی آن برای اجرای الگوریتم مورد نظر می‌باشد. در این قسمت، اشاره‌ای اجمالی به گام‌هایی که برای انجام این عملیات باید پیموده شوند خواهیم داشت.

پیش‌پردازش

پیش‌پردازش داده‌ها با انجام عملیاتی مانند: حذف علائم نشانه‌گذاری‌شده، حذف اعداد، حذف حروف انگلیسی از هر سند از میان داده‌های موجود بر روی متون جمع‌آوری شده، برای استخراج ویژگی‌ها صورت خواهد گرفت.

حذف کلمات زائد[1]

در فرایند حذف کلمات زائد نمی‌توان از محصولات آماده استفاده نمود زیرا موجب پیچیدگی خروجی آنها می‌شود. به‌همین منظور برای حذف کلمات بی‌ارزش از دیکشنری کلمات زائد استفاده می‌کنیم.

ریشه‌یابی

فرایند ریشه‌یابی به وسیلهٔ ماتریس‌های از قبل تولید شده صورت می‌گیرد که در این مرحله کلمات به ریشهٔ خود بر می‌گردند. با این کار از افزونگی ویژگی‌ها جلوگیری می‌شود.

استخراج ویژگی

یکی از متداول‌ترین تکنیک‌های استخراج ویژگی در متن کاوی الگوریتم BOW می‌باشد. الگوریتم BOW به گونه‌ای عمل می‌کند که هر کلمه را به عنوان یک ویژگی در نظر می‌گیرد و یک ماتریس سند- ویژگی باینری به آن اختصاص می‌دهد و هر سطر آن معادل یک متن یا سند و هر ستون آن معادل یک کلمه می‌باشد. در ماتریس تشکیل شده برای هر سند که دارای کلمه یا ویژگی‌های برای آن است مقدار یک و در غیر این‌صورت مقدار صفر قرار می‌دهد.

انتخاب ویژگی

در فرایند انتخاب ویژگی از تکنیک الگوریتم IG استفاده می‌کنیم که این الگوریتم در نرم‌افزار متلب به‌صورت یک تابع پیاده‌سازی خواهد شد. در این پژوهش انتخاب ویژگی و ترکیب ویژگی به صورت مجزا با یکدیگر مقایسه شده‌اند.

ترکیب ویژگی

در اینجا با ذکر یک مثال ترکیب ویژگی را نشان می‌دهیم. فرض کنید ۱۰ داده با ۷ ویژگی در دو کلاس مختلف وجود دارد (جدول۲).

جدول ۲. مجموعه داده‌های فرضی.

D	F۱	F۲	F۳	F۴	F۵	F۶	F۷	C
۱	۱	۰	۱	۰	۱	۰	۰	۱
۲	۱	۰	۰	۱	۱	۰	۰	۱
۳	۱	۰	۰	۱	۱	۰	۰	۱
۴	۰	۱	۱	۰	۰	۰	۰	۱
۵	۰	۱	۱	۰	۰	۰	۰	۱
۶	۰	۱	۰	۱	۰	۰	۱	۱-

[1] Stop words

D	F۱	F۲	F۳	F۴	F۵	F۶	F۷	C
۷	۰	۱	۰	۱	۰	۰	۱	۱-
۸	۰	۱	۰	۱	۱	۱	۰	۱-
۹	۰	۱	۰	۱	۱	۱	۰	۱-
۱۰	۰	۱	۰	۱	۱	۱	۰	۱-

همان‌طور که می‌دانیم با استفاده از روش انتخاب ویژگی می‌توان ویژگی‌های نامرتبط و زائد را حذف کرد. بنابراین در جدول ۴-۱ $F1=\overline{F2}$ و $F3=\overline{F4}$ و (F۲ , F۴) یا (F۱ , F۳) را توسط روش انتخاب ویژگی می‌توان حذف کرد. یکی دیگر از ویژگی‌های نامرتبط F5 است که می‌تواند حذف شود. بنابراین زیر مجموعهٔ بهینه زیر می‌تواند ساخته شود (جدول ۳).

جدول۳. زیر مجموعه داده‌های انتخاب شده.

D	F۱	F۳	F۶	F۷	C
۱	۱	۱	۰	۰	۱
۲	۱	۰	۰	۰	۱
۳	۱	۰	۰	۰	۱
۴	۰	۱	۰	۰	۱
۵	۰	۱	۰	۰	۱
۶	۰	۰	۰	۱	۱-
۷	۰	۰	۰	۱	۱-
۸	۰	۰	۱	۰	۱-
۹	۰	۰	۱	۰	۱-
۱۰	۰	۰	۱	۰	۱-

اگرچه این مجموعه داده به خوبی توسط روش انتخاب ویژگی کاهش داده شده است، ولی همچنان قابل کاهش می‌باشد. اجتماع ویژگی‌های (F۱, F۳) یا (F۶, F۷) می‌تواند یک ویژگی پر قدرت‌تر نسبت به کلاس C بسازد. در واقع ویژگی‌هایی را که اجتماعشان منجر به ساخت یک ویژگی قوی‌تر می‌شوند ترکیب می‌کنیم. به‌عنوان مثال اجتماع F۶ و F۱ نمی‌تواند بر اساس هر معیاری، ویژگی خوبی بسازد در حالی‌که اجتماع F۳ و F۱ ویژگی قوی‌تری را می‌سازد. بنابراین می‌توان توسط روش جدید ترکیب ویژگی زیرمجموعه ویژگی بهینه و قوی‌تری ساخت (جدول ۴). همان‌طور که مشاهده می‌شود دو ویژگی F۱ و F۲ بسیار قوی‌تر شده‌اند.

جدول۴. زیر مجموعه داده‌های ترکیب شده.

D	(F۱∪F۳)F′1	(F۶∪F۷)F′2	C
۱	۱	۰	۱
۲	۱	۰	۱
۳	۱	۰	۱
۴	۱	۰	۱

D	(F۱∪F۳)F′1	(F۶∪F۷)F′2	C
۵	۱	۰	۱
۶	۰	۱	۱-
۷	۰	۱	۱-
۸	۰	۱	۱-
۹	۰	۱	۱-
۱۰	۰	۱	۱-

در این مرحله، ویژگی‌ها با استفاده از معیار ارزیابی بر اساس کیفیت‌شان مرتب شده و سپس از بالا شروع به ترکیب ویژگی‌ها می‌کنیم. ویژگی‌ها دو به دو با یکدیگر با استفاده از عمگر OR ترکیب می‌شوند که منجر به ساختن یک ویژگی می‌شود، این ترکیب زمانی صورت می‌گیرد که این ویژگی جدید توسط معیار ارزیابی، ارزش‌گذاری شده و از هر دو ویژگی قبلی قوی‌تر شده باشد. بنابراین ویژگی‌هایی که جهت ترکیب مناسب هستند با یکدیگر ترکیب شده و منجر به ساخت یک زیر مجموعهٔ کاهش یافته و قوی‌تری می‌شود. بر اساس ایده پیشنهاد شده، می‌توان ویژگی‌های به دست آمده را با هم ترکیب کرد و یک ویژگی قوی‌تری ایجاد نمود. در این روش، انتخاب ویژگی‌ها با استفاده از جستجو در میان داده‌ها است که در هدف استخراج و کشف بهترین ترکیب از ویژگی‌ها می‌باشد. این روش ترکیب ویژگی‌ها را به کمک معیارهایی (مانند IG) وزن‌دهی و به‌صورت نزولی مرتب می‌کند. سپس شروع به ترکیب ویژگی‌ها می‌نماید. در واقع اولین ترکیب را به‌دست می‌آورد که لزوماً بهترین ترکیب نیست. در اینجا این روش را با FU نشان می‌دهیم. ولی در روش پیشنهادی، بهترین ترکیبات توسط الگوریتم PSO پیدا می‌شود که در زیر توضیح داده شده است (FU_PSO).

به طور کلی اگر بخواهیم ساختار تشکیل شده را با زبان ساده‌ای بیان‌کنیم، می‌توان گفت: تمام ذرات به دنبال بهترین نقطه می‌گردند، در هر بار که حرکت انجام می‌شود ذرات تابع برازش خود را محاسبه می‌کنند (طبق فرمول زیر) هر ذره‌ای که بهترین برازش (Fitness function) را داشته باشد (یعنی به پاسخ نزدیک‌تر است) به دیگران اطلاع می‌دهد و دیگر ذرات به سمت او حرکت می‌کنند. این حرکت تا زمانی ادامه پیدا می‌کند که همهٔ ذرات در بهترین نقطه در کنار یکدیگر جمع شوند. که در اینجا بهترین ترکیبات ویژگی‌ها می‌باشد. در حل یک مسئله با PSO در ابتدا باید فرمت ذرات و تابع برازش مشخص شوند.

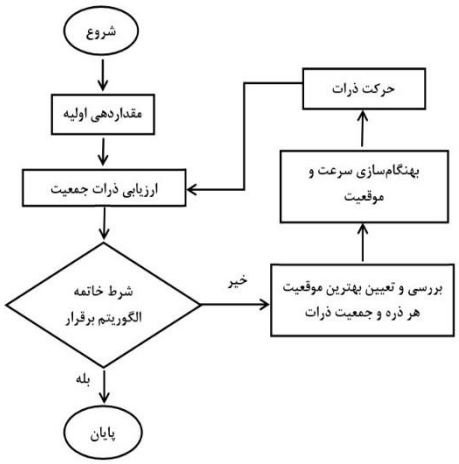
فرمت ذرات

با فرض این‌که تعداد ویژگی‌ها n باشد و قرار باشد به m ویژگی توسط FU کاهش داده شوند. در اینجا هدف آن است که چه ویژگی‌هایی با هم ترکیب می‌شوند. بنابراین تعداد هر ذره در PSO به تعداد n می‌باشد و مقادیر آنها یک عدد بین ۱ تا m می‌باشد. ویژگی‌هایی که هم شماره باشند با جهت ساختن یک ویژگی واحد با یکدیگر ترکیب می‌گردند.

جدول ۵. فرم ذرات ویژگی‌ها.

F۱	F۲	F۳	F۴	F۵	F۶	F۷		Fn
۶	۴	۶	۱	۴	۶	۲	.....	۱

به عنوان مثال در اینجا ویژگی‌های (۱،۳و۶) با یکدیگر ترکیب و یک ویژگی واحد می‌سازند.
در نهایت بهترین ذره که نشان‌دهندهٔ بهترین ترکیبات هستند توسط PSO پیدا می‌شود.



شکل ۳. روش کار الگوریتم PSO.

تابع برازش

تابع برازش هر ذره را بر اساس کیفیت آن ارزش‌گذاری می‌کند. در اینجا به ازای هر ذره ترکیبات را توسط اپراتور OR انجام داده و سپس برای هر ویژگی(m ویژگی) مقدار IG را محاسبه و با یکدیگر جمع می‌کنیم. مقدار این تابع برای آن ذره‌ای که ترکیبات بهتری را نشان می‌دهد بیشتر می‌باشد.

$$Fitness\_function = \sum_{i=1}^{m} IG(Fi)$$

تابع دیگر استفاده از طبقه‌بندی جهت ارزیابی کیفیت ذره مورد نظر بر اساس معیار دقت می‌باشد.

$$Fitness\_function = acc(F)$$

دسته‌بندی (طبقه‌بندی)

در این پژوهش از سه الگوریتم (SVM۱، NB۲ و KNN۳) استفاده شده است. NB یک روش تقسیم‌بندی احتمالی ساده است که بر اساس به‌کارگیری فرضیه‌های مستقل و قضیه بیز داده‌های آموزشی را انتخاب می‌کند. برای ارائه الگوریتم نایوبیز نیاز به تخمین‌های پارامتر تکراری پیچیده نیست و به آسانی برای مجموعه داده‌های متنی بزرگ به کار می‌رود.

الگوریتم SVM یک روشی برای استخراج دانش از میان داده‌های انبوه و ایجاد مدل‌های ریاضی پیچیده است. با استفاده از الگوریتم SVM می‌توان مدل‌سازی غیر خطی را انجام داد. الگوریتم SVM بهتر از الگوریتم KNN است زیرا می‌تواند شاخص‌های اصلی داده‌ها را حفظ کند. طبقه‌بندی الگوریتم KNN یکی از جزئی‌ترین و ساده‌ترین متدها برای طبقه‌بندی داده است. تشکیل یک دیتاست در الگوریتم KNN به وسیله رأی اکثریت همسایگان آن تقسیم

[1] Support vector machine
[2] Naïve Bayesian
[3] K Nearest Neighbors

بندی می‌شود که k=۱ را به‌عنوان یک برچسب برای نزدیکترین همسایه در نظر می‌گیرد. در این پژوهش k=۳ در نظر گرفته شده است.

نتایج آزمایشی

آزمایش‌ها

کلیهٔ آزمایش‌ها روی سیستمی با سیستم‌عامل Windows ۸ و سی پی یو ۳.۰GHz و رم ۸.۰۰ GB اجرا شده است. در این پژوهش روش پیشنهادی را به کمک نرم‌افزار متلب R۲۰۱۳a پیاده‌سازی کرده‌ایم.

مجموعهٔ داده‌ها

در این پژوهش برای مرحلهٔ آزمایش نیاز به مجموعه‌ای از داده‌های متنی داریم که داده‌های متنی زیادی وجود دارند که در این تحقیق به‌دلیل نیاز به مجموعهٔ داده‌های متنی بزرگ از وب‌سایت Amazon مجموعه داده‌ها استخراج شده و مورد آزمایش قرار گرفته و براساس ویژگی‌های استخراج شده طبقه‌بندی می‌شوند. در جدول ۶ چهار نوع داده که شامل: دی وی دی – کتاب – الکترونیک و آشپزخانه به همراه تعداد ویژگی‌های‌شان مورد آزمایش قرار می‌گیرند (ویژگی‌های داده‌ها دارای مقادیر صفر و یک می‌باشند یعنی اگر ویژگی وجود داشته باشد مقدار یک را خواهد داشت و در غیر این‌صورت مقدار صفر را خواهد داشت).

در واقع سه الگوریتم (SVM، NB و KNN) بر روی مجموعهٔ داده‌های جمع‌آوری‌شده پیاده‌سازی شده است و بخش‌بندی مجموعهٔ داده‌ها و الگوریتم‌ها به‌عنوان یک اصل در نظر می‌گیریم. در بعضی مواقع برای رسیدن به اطمینان خاطر بیشتر مجبور به تغییر بخش‌بندی اولیه می‌باشیم.

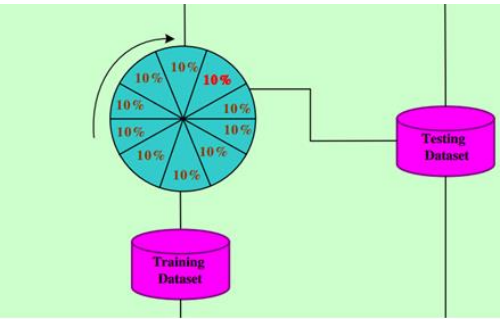
جدول ۶. داده‌های آزمایشی.

نام طبقه	تعداد Instance	تعداد Feature
دی وی دی	۲۰۰۰	۴۶۵۷
کتاب	۲۰۰۰	۸۴۵۷
الکترونیک	۲۰۰۰	۴۲۱۶
آشپزخانه	۲۰۰۰	۳۷۳۸

روش ارزیابی

در این پژوهش اعتبارسنجی Fold cross-validation بر روی مجموعهٔ داده‌ها برای به‌حداقل رساندن تأثیر تغییر در مجموعه‌ها صورت گرفت. از مجموع ۱۰۰٪ داده‌هایی که از قبل در سیستم بوده است ما هر بار ۹۰ ٪ از داده‌ها را به عنوان داده‌های آموزشی در نظر می‌گیریم و ۱۰درصد باقی مانده را به عنوان تست به الگوریتم می‌دهیم منوط به این که می‌دانیم ۱۰درصد از داده‌ها در کدام گروهها طبقه‌بندی می‌شوند. پس از دادن این ۱۰درصد از داده‌ها به الگوریتم‌مان می‌توانیم با گروه‌بندی واقعی مقایسه‌اش کنیم تا میزان درستی طبقه‌بندی را بررسی کنیم. پس از اتمام کار با این ۱۰درصد که به‌عنوان تست به الگوریتم تخصیص داده بودیم، یک ۱۰ درصد دیگر از کل داده‌ها را برای تست مرحلهٔ بعدی برمی‌داریم و ۹۰درصد باقی مانده را به دادهٔ آموزشی اختصاص می‌دهیم. برای به‌دست آوردن نتیجهٔ دقت نهایی روش

ارائه شده این فرایند را در ده مرحله با ۱۰٪های مختلف انجام می‌دهیم و میانگین می‌گیریم مراحل انجام شده را مورد بررسی قرار می دهیم.



شکل ۴. اعتبار سنجی.

نتایج آزمایشی

مرحلهٔ بعد از انتخاب ویژگی‌ها از میان داده‌های انبوه و کشف نمونه‌های قابل استناد در سیستم وگروه‌بندی کلمات، پیاده‌سازی روش‌های پیشنهادی و مقایسهٔ الگوریتم‌های گروه‌بندی می‌باشد. این آزمایش‌ها شامل: استفاده از سه روش گروه‌بندی اطلاعات با استفاده از الگوریتم‌های (SVM، NB و KNN) و ترکیب ویژگی مبتنی بر بهینه‌سازی گروه ذرات (FU-PSO) و روش‌های انتخاب ویژگی(FS) می‌باشد که نتایج به‌دست آمده در جدول شماره ۵ و شکل شماره ۴ مشاهده می‌شود. با مشاهدهٔ جدول می‌توان بهبود عملکرد سیستم را بعد از استفاده از روش پیشنهاد شده به‌ویژه با استفاده از الگوریتم SVM به عنوان گروه‌بندی، مشاهده نمود.

در جدول شماره ۵ نتیجهٔ عکس‌العمل یادگیرنده‌های پایه را روی ۴ گروه از داده‌ها مختلف پیاده کرده و دقت گروه‌بندی و تعداد کاهش داده‌ها نشان داده شده است. شکل شماره ۴دقت هر سه گروه‌بندی در حالت استفاده از راه‌حل‌های انتخاب ویژگی(FS) و ترکیب ویژگی مبتنی بر بهینه‌سازی ذرات (FU-PSO) را نشان می‌دهد. طبق اسناد به دست آمده از نتایج می‌توان بهبود عملکرد گروه‌بندی را مشاهده نمود. افزایش بهبود و کارایی گزارش شده حدوداً ۳٪ است که این بهبود و کارایی ارائه شده در SVM بیشتر از دیگر گروه‌بندهای می‌باشد و روش پیشنهادی کارایی بهینه‌تری جهت گروه‌بندی کلمات در مقایسه با استفاده از انتخاب ویژگی و نمونه به‌صورت جداگانه را دارد. طبق نتایج به دست آمده میزان دقت افزایش‌یافته و با کاهش داده‌ها موجب سرعت بخشیدن مرحلهٔ یادگیری شده است.

به عنوان مثال در مجموعهٔ داده دی وی دی، دقت گروه بندی با استفاده از SVM از ۰.۷۸ به ۰.۸۱۱ افزایش داشته است در حالی‌که تعداد ویژگی‌ها از ۴۶۵۷ به ۱۶۱ عدد کاهش یافته است. البته با توجه به نتایج به دست آمده از جداول و شکل‌ها می‌توان بهبود در دیگر طبقه‌بندی‌ها را مشاهده نمود، که تایید کننده کارایی استفاده از روش پیشنهاد شده می‌باشد.

جدول۷. میزان دقت گروه‌بندهای مختلف برای داده‌های دی وی دی.

تکنیک	بدون استفاده از کاهش داده		FS		FU-PSO	
	دقت	Fe#	دقت	Fe#	دقت	Fe#
SVM	۰.۷۶۷	۴۶۵۷	۰.۷۸۰	۵۰۰	۰.۸۱۱	۱۶۱
NB	۰.۷۵۵	۴۶۵۷	۰.۷۶۰	۵۰۰	۰.۷۸۵	۱۶۰

تکنیک	بدون استفاده از کاهش داده		FS		FU-PSO	
	دقت	Fe#	دقت	Fe#	دقت	Fe#
KNN	۰.۶۴۰	۴۶۵۷	۰.۶۷۷	۵۰۰	۰.۷۱۰	۱۶۴

جدول ۸. میزان دقت گروه بندهای مختلف برای داده‌های کتاب.

تکنیک	بدون استفاده از کاهش داده		FS		FU-PSO	
	دقت	Fe#	دقت	Fe#	دقت	Fe#
SVM	۰.۷۶۲	۸۴۵۷	۰.۷۸۰	۵۰۰	۰.۸۲۱	۲۰۰
NB	۰.۷۸۲	۴۶۵۷	۰.۷۹۲	۳۰۰۰	۰.۷۹۵	۹۶۰
KNN	۰.۶۴۷	۴۶۵۷	۰.۶۷۷	۳۰۰۰	۰.۷۱۲	۹۵۵

جدول ۹. میزان دقت گروه‌بندی داده‌های الکترونیک.

تکنیک	بدون استفاده از کاهش داده		FS		FU-PSO	
	دقت	Fe#	دقت	Fe#	دقت	Fe#
SVM	۰.۷۷۲	۴۲۱۶	۰.۷۸۵	۱۵۰۰	۰.۸۰۳	۵۰۰
NB	۰.۸۰۲	۴۲۱۶	۰.۸۱۰	۱۰۰۰	۰.۸۳۵	۳۵۰
KNN	۰.۷۲۵	۴۲۱۶	۰.۷۳۲	۱۰۰۰	۰.۷۵۳	۳۳۰

جدول ۱۰. گروه‌بندهای مختلف برای داده‌های یک آشپزخانه.

تکنیک	بدون استفاده از کاهش داده		FS		FU-PSO	
	دقت	Fe#	دقت	Fe#	دقت	Fe#
SVM	۰.۸۲۰	۳۷۳۸	۰.۸۲۲	۵۰۰	۰.۸۴۳	۲۰۰
NB	۰.۷۹۵	۳۷۳۸	۰.۸۰۵	۲۰۰۰	۰.۸۳۲	۷۰۰
KNN	۰.۷۲۰	۳۷۳۸	۰.۷۲۲	۲۰۰۰	۰.۷۴۷	۷۵۰

گروه‌بندی بر روی همه مجموعهٔ داده‌ها می‌باشد که با توجه به نمودار می‌توان مشاهده نمود که الگوریتم SVMبالاترین دقت و قدرت طبقه‌بندی بیشتری را نسبت به دو الگوریتم دیگر دارد. نتایج بررسی‌ها گویای بهبود عملکرد گروه‌بند ناشی از بهره‌جویی ترکیب ویژگی مبتنی بر روش بهینه‌سازی ذرات در زمینهٔ طبقه‌بندی هوشمند احساسات است. در کل باید گفت که به‌کارگیری این راه حل با توجه به کاهش داده در سطح ویژگی تأثیر قابل توجهی در افزایش کارایی گروه‌بندی دارد. در ضمن باید به تأثیر ترکیب ویژگی با الگوریتم KNN و تأثیر موفق آن که دلیل حساسیت این الگوریتم به داده‌های نامناسب می‌باشد، برای طبقه‌بندی بهینه و مؤثرتر دقت شود.

نتیجه‌گیری

در این مقاله یک روش جدید جهت طبقه‌بندی احساسات مبتنی بر ترکیب ویژگی‌ها ارائه شده است. چارچوب ارائه شده بر اساس دو دیدگاه جدید انتخاب ویژگی از میان انبوه داده‌ها و ترکیب ویژگی‌های انتخاب شده است. این روش جدید مجموعهٔ ویژگی‌ها را به یک زیر مجموعه از ویژگی‌های قوی‌تر تبدیل می‌کند که باعث کاهش هزینه، زمان و افزایش دقت طبقه‌بندی اطلاعات می‌شود. نتایج این بخش با نتایج حاصل از تحقیق محمدی و خلج [۱۱] تطابق دارد.

یکی از نقاط قوت استفاده از روش‌های انتخاب ویژگی و ترکیب ویژگی، اجماع پتانسیل‌های ویژگی‌های تعریف کننده متن (کلمات) و کاهش ویژگی‌ها در جهت افزایش کارایی دقت طبقه‌بندی است، که در نتایج به‌دست آمده میان روش‌های مختلف، قابل استناد می‌باشد. نتایج این بخش نیز با یافته‌های ابراهیم و وانگ [۱۲] مطابقت دارد.

در این تحقیق آزمایش‌هایی جهت ارزیابی روش پیشنهادی انجام شد و نتایج آزمایشی با استفاده از سه تکنیک طبقه‌بندی (SVM, NB, KNN) و کاهش داده (انتخاب ویژگی‌ها) بررسی شد. همچنین از نتایج به‌دست آمده استنباط می‌شود که استفاده از روش ترکیب ویژگی، کارایی طبقه‌بندی را افزایش داده و از تأثیر این افزایش در افت کارایی دسته‌بندی‌کننده می‌کاهد. نکته‌ای که باید در کارهای آتی در نظر داشت، این است که هر چه ارتباط بین لغات انتخاب شده جهت ترکیب ویژگی‌ها بیشتر باشد، دقت سیستم بهبود بیشتری خواهد یافت. استفاده از تکنیک ترکیب با ساختار سلسلهٔ مراتبی و انتخاب لغات با توجه به این ساختار می‌تواند به این هدف کمک کند.

نهایتاً این‌که استخراج و کشف دانش از داده‌هایی با حجم بسیار بالا نیازمند هزینه و زمان زیادی است. بنابراین مستلزم این است که از روش‌هایی برای کاهش اندازه داده‌ها استفاده نمائیم. تکنیک‌های کاهش داده می‌توانند بدون از دست دادن درستی اطلاعات و نتایج نهایی وارد عمل شوند که با کاهش داده‌ها در مراحل مختلف پردازش داده‌کاوی اطلاعات می‌توان سادگی مدل ارائه شده را به همراه داشته باشد به‌طوری که مدل ارائه شده قابل فهم‌تر خواهد بود.

References

[1] Hosseinian, A. H., Teimourpour, B., & Jamali Hondori, B. (2019). A Hybrid Algorithm for Detecting Communities of Social Networks based on the Modularity Density Criterion. *Business Intelligence Management Studies*, *8*(29), 61-86. https://doi.org/10.22054/ims.2019.10376

[2] Mohammadi, S., & Nazemi, E. (2021). Sentiment Analysis at the Product Feature Level and Based on Users Gender. *Business Intelligence Management Studies*, *10*(37), 267-296. https://doi.org/10.22054/ims.2021.52110.1723

[3] Emary, E., Zawbaa, H. M., Grosan, C., & Hassenian, A. E. (2015). Feature Subset Selection Approach by Gray-Wolf Optimization. In A. Abraham, P. Krömer, & V. Snasel (Eds.), Afro-European Conference for Industrial Advancement. Springer International Publishing. https://doi.org/10.1007/978-3-319-13572-4_1

[4] Athar, A., Butt, W. H., Anwar, M. W., Latif, M., & Azam, F. (2017, February 24-26). *Exploring the Ensemble of Classifiers for Sentimental Analysis: A Systematic Literature Review*. Proceedings of the 9th International Conference on Machine Learning and Computing, Singapore, Asia. https://doi.org/10.1145/3055635.3056601

[5] Khadem, M., Toloie Eshlaghy, A., & Fathi Hafshejani, K. (2023). Introducing a new meta-heuristic algorithm to solve the feature selection problem. *Journal of Future Studies Management*, *33*(3), 16-27. https://doi.org/10.30495/jmfr.2022.55572.2264

[6] Abualigah, L. M. Q. (2019). *Feature selection and enhanced krill herd algorithm for text document clustering*. Springer Cham. https://doi.org/10.1007/978-3-030-10674-4

[7] Mafarja, M., & Mirjalili, S. (2018). Whale optimization approaches for wrapper feature selection. *Applied Soft Computing*, *62*, 441-453. https://doi.org/10.1016/j.asoc.2017.11.006

[8] Ghaemi, M., & Feizi-Derakhshi, M-R. (2016). Feature selection using Forest Optimization Algorithm. *Pattern Recognition*, *60*, 121-129. https://doi.org/10.1016/j.patcog.2016.05.012

[9] Kashef, S., & Nezamabadi-pour, H. (2015). An advanced ACO algorithm for feature subset selection. *Neurocomputing*, *147*, 271-279. https://doi.org/10.1016/j.neucom.2014.06.067

[10] Hosseini, M., & Navabi, M. S. (2023). Hybrid PSO-GSA based approach for feature selection. *Journal of Industrial Engineering and Management Studies*, *10*(1), 1-15. https://jiems.icms.ac.ir/article_166460.html

[11] Mohammadi, S., & Khalaj, E. (2021). Presenting the model for opinion mining at the document feature level for hotel users' reviews. *Journal of Information and Communication Technology*, *13*(49-50), 85-102. http://rimag.ricest.ac.ir/fa/Article/16379

[12] Ibrahim, N. F., & Wang, X. (2019). Decoding the sentiment dynamics of online retailing customers: Time series analysis of social media. *Computers in Human Behavior*, *96*, 32-45. https://doi.org/10.1016/j.chb.2019.02.004

[13] Nazim Sha, S., & Rajeswari, M. (2019, February 26-28). *Creating a Brand Value and Consumer Satisfaction in E-Commerce Business Using Artificial Intelligence.* Proceedings of International Conference on Sustainable Computing in Science, Technology and Management Amity University Rajasthan, Jaipur-India. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3351618

[14] Khan, F. H., Qamar, U., & Bashir, S. (2016). SentiMI: Introducing point-wise mutual information with SentiWordNet to improve sentiment polarity detection. *Applied Soft Computing*, *39*, 140-153. https://doi.org/10.1016/j.asoc.2015.11.016

[15] Dehghani Ashkazari, S., Derhami, V., Zare Bidoki, A. M., & Basiri, M. E. (2020). Persian Opinion Mining based on Transfer Learning. *Tabriz Journal Of Electrical Engineering*, *50*(3), 1215-1224. https://tjee.tabrizu.ac.ir/article_11360.html?lang=en

[16] Yousefpour, A., Ibrahim, R., & Hamed, H. N. A. (2017). Ordinal-based and frequency-based integration of feature selection methods for sentiment analysis. *Expert Systems with Applications*, *75*, 80-93. https://doi.org/10.1016/j.eswa.2017.01.009