یک کاربرد از رویکرد تحلیل توپولوژیکی داده در طبقه‌بندی اشعار فارسی

نیره الیاسی[1*]، مهدی حسینی مقدم[2]

استادیار دانشکدهٔ علوم ریاضی و کامپیوتر، دانشگاه خوارزمی، تهران، ایران[1*]

کارشناس ارشد مهندسی داده، آکسفورد، انگلستان[2]

چکیده

تحلیل توپولوژیکی داده شاخه‌ای بدیع و به‌سرعت در حال رشد در علوم داده است که مجموعه‌ای از ابزارهای هندسی و توپولوژیکی را برای استخراج ویژگی‌های مرتبط از دادهٔ پیچیدهٔ بُعد بالا فراهم می‌کند. در این مقاله، دو روش از بهترین‌های تحلیل توپولوژیکی داده؛ یعنی هموlوژی ماندگار و نگاشت‌گر به‌منظور طبقه‌بندی اشعار دو تن از بهترین شعرای ایران؛ یعنی فردوسی و حافظ، به‌کار گرفته می‌شود. در این پژوهش از روش‌شناسی تحلیل متن سنتی فراتر رفته تا کارآمدی و بهینه‌بودن تحلیل توپولوژیکی داده در زمینهٔ متن‌کاوی و انتساب نویسنده، نشان داده شود. نکتهٔ نوآورانهٔ این مقاله استفاده از تحلیل توپولوژیکی داده در انتساب نویسنده است که پیش‌تر نیز سابقه نداشته است؛ همچنین قابلیت بصری‌سازی نتایج با استفاده از نمودارهای پایا، بارکد و نگاشت‌گر که منحصر به تحلیل توپولوژیکی داده است و به‌سادگی قابل تفسیر توسط هر خواننده‌ای است، این امید را می‌دهد که از این به بعد تحلیل توپولوژیکی داده به‌عنوان یک دریچه جهت رمزگشایی در ادبیات و علوم‌انسانی بتواند مورداستفاده قرار گیرد.

واژگان کلیدی: تحلیل توپولوژیکی داده، هموlوژی ماندگار، نگاشت‌گر، اشعار فارسی.

An application of the topological data analysis approach in Persian poetry classification

Naiereh Elyasi[1], Mehdi Hosseini Moghadam[2*]

Assistant Professor, Faculty of Mathematical sciences and Computer, Kharazmi university, Tehran, Iran[1*]

NLP/LLM(Team lead), Oxford university, Oxford, UK[2]

Abstract

This research delves into authorship attribution through an avant-garde lens, employing Topological Data Analysis (TDA) as a potent instrument to unravel intricate patterns within classical Persian poetry. The focal point of this study is the distinguished works of Ferdowsi and Hafez, two preeminent Persian poets, exploring the latent structures in their verses through the lenses of Persistent Homology and Mapper a pair of TDA methodologies. The discernment between Non-Semantic and Semantic authorship attribution methodologies lays the groundwork, elucidating the significance of capturing structural nuances in textual data. The main focus of this investigation revolves around the deployment of Persistent Homology a cutting-edge technique that transcends traditional text analysis methodologies. It operates in high-dimensional spaces, extracting topological features, and rendering them comprehensible through persistent diagrams. This paper meticulously unpacks the mathematical underpinnings of Persistent Homology, providing a stepwise exposition of its application, focusing on Homology, Simplicial Complex, and Group Theory. These foundational elements converge to empower extracting meaningful topological signatures from the poetic corpus. In tandem, Mapper, another TDA tool, unfolds as a pivotal player in this explorative journey. This algorithmic entity facilitates dimensionality reduction and simplicial complex construction to portray an accurate depiction of the intrinsic topological architecture residing in the dataset. The intricacies of Mapper's workflow from filter function selection to binning and clustering are meticulously detailed, forming a coherent narrative of its operational dynamics. Transitioning from theoretical discourse to practical implementation, this research adopts a case study approach, weaving Ferdowsi and Hafez's poetic masterpieces into the TDA tapestry. Beyond the mere application of algorithms, the study delves into the

* نویسندهٔ عهده‌دار مکاتبات

* Corresponding author

• تاریخ ارسال مقاله: ۱۴۰۲/۶/۱۴ • تاریخ پذیرش: ۱۴۰۴/۳/۲۱ • تاریخ انتشار: ۱۴۰۴/۹/۲۸ • نوع مطالعه: کاربردی

[DOI: 10.66224/jsdp.22.3.77]

realm of accuracy assessments, subjecting the Mapper algorithm to rigorous tests, and gauging the precision of its poem classifications within identified clusters. An additional layer of complexity unfolds as the research embraces semantic clustering, elucidating thematic resonances embedded within the verses. The results borne out of this meticulous exploration not only underscore the efficacy of TDA methodologies in unveiling the intricate structures of Persian poetry but also offer a nuanced perspective on their interpretability and utility in the realm of authorship attribution. The poetic narrative, with its semantic richness and structural subtleties, emerges as a fertile ground for the application of TDA, pushing the boundaries of text classification methodologies. This research, therefore, contributes significantly to the evolving discourse on the intersection of literature and data science, offering a profound understanding of how TDA can be wielded as a transformative lens to decipher the profound threads of authorial expression.

Keywords: Topological data analysis, Persistent Homology, Mapper, Persian poems.

۱-مقدمه

۱-۱-انتساب نویسنده

پیشرفت فناوری و استفاده از اینترنت و شبکه‌های اجتماعی موجب تولید حجم انبوهی از اطلاعات مبتنی بر متن، از جمله اخبار، پیام‌ها و نظرها شده‌است که برای تحلیل این حجم عظیم از داده‌های بدون ساختار باید آن‌ها را طبقه‌بندی کرد. فرایند تخصیص برچسب به برخی متون بر اساس محتوای آن‌ها طبقه‌بندی متن نامیده می‌شود. انتساب نویسنده یکی از شاخه‌های اصلی طبقه‌بندی متن است که سعی می‌کند نویسندهٔ متن را بر اساس محتوای آن شناسایی کند. موضوع انتساب نویسنده از دیرباز موردتوجه بوده و همچنان موضوعی محبوب است. به‌دلیل پیشرفت در رایانه‌های دیجیتال، این حوزه در دههٔ گذشته پیشرفت‌های سریعی را تجربه کرده‌است و می‌توان انتساب نویسندگی را به دو روش اصلی تقسیم کرد:

- غیرمعنایی: این روش سعی می‌کند نویسنده را بر اساس طول واژگان و جملات و واژگان به‌کاررفته در متنی مشخص شناسایی کند[۱ و۲].
- معنایی: این روش ساختار زبان را بر اساس تحلیل معنایی آن در نظر می‌گیرد [۳، ۴و ۵].

در این مقاله سعی شده‌است تا انتساب نویسندگی بر اساس روشی جدید به‌نام تحلیل توپولوژیکی داده انجام شود.

۱-۲- تحلیل توپولوژیکی داده

در دههٔ اخیر با حجم روزافزون داده‌ها و پیشرفت در فناوری، با دنیای پربار اطلاعات روبه‌روییم؛ به‌منظور ادراک بهتر از آنچه در دنیای ما اتفاق می‌افتد، باید این حجم عظیم داده را جمع‌آوری و به کمک علم داده تجزیه‌وتحلیل کرد. یکی از مشکلات اصلی در علم داده، پرداختن به داده‌های با ابعاد بالا و تبدیل آن‌ها به داده‌هایی با ابعاد کمتر به‌منظور تسهیل در تجزیه‌وتحلیل است. تجزیه‌وتحلیل داده‌های توپولوژیکی(TDA) یکی از شاخه‌های جدید و به‌سرعت درحال‌رشد علم داده است که سعی در تحلیل داده‌ها با مطالعهٔ ارتباط توپولوژیکی آن‌ها و همچنین کاهش ابعاد داده‌ها دارد[۶]. این شاخه بر دو شاخهٔ بسیار مهم «آمار» و «توپولوژی جبری» استوار است. باتوجه‌به مقاوم‌بودن روش‌های تحلیل توپولوژیکی داده در برابر نوفه، این شاخه می‌تواند برخی از مشکلات جدی در علم داده را حل کند. هدف تحلیل توپولوژیکی داده، کاهش ابعاد در داده‌های با ابعاد بالا و همچنین تحلیل ساختار یا شکل توپولوژیکی داده‌ها و درنهایت خوشه‌بندی داده‌های پیچیده است؛ ازآنجا که ماهیت داده‌ها تصادفی است، این رویکرد بر اساس مفاهیم آماری و ترکیبیاتی توسعه یافته است؛ همچنین TDA[1]، روش‌های داده‌کاوی نوآورانه‌ای را ارائه می‌دهد که می‌تواند کارایی روش‌های یادگیری ماشین را بهبود بخشد.

تحلیل توپولوژیکی داده دو روش اصلی دارد: «همولوژی ماندگار» و «نگاشت‌گر». در همولوژی ماندگار، پالایشی از هم‌بافت‌های سادکی ساخته می‌شود و سپس ساختارهای توپولوژیکی اصلی داده‌ها استخراج می‌شود. برای شهود بهتر همولوژی ماندگار، برخی نمودارها مانند «نمودار ماندگار»، «بارکد» و «چشم‌انداز پایا» برای نشان‌دادن ویژگی‌های توپولوژیکی اصلی داده‌ها ابداع شدند. روش همولوژی ماندگار پیش‌تر در تحلیل موج فشار پالس [۶]، تحلیل تصاویر سه‌بعدی [۸ و ۹]، در رفتار رأی‌گیری [۱۰]، شبکه‌های مغز [۱۱]، تحلیل تصویر[۱۲]، برنامه‌ریزی مسیر [۱۳] استفاده شده‌است. ایدهٔ پشت الگوریتم نگاشت‌گر این است که یک مجموعه‌داده با ابعاد بالا و نوفه را به شیء ترکیبیاتی (هم‌بافت سادکی) کاهش دهد. چنین شیء سعی می‌کند شکل اصلی داده‌های با ابعاد بالا را در خود حفظ کند. الگوریتم نگاشت‌گر پیش‌تر برای داده‌های بیماران مبتلا به سرطان پستان [۱۴]، بازنمایی متن برای پردازش زبان طبیعی [۱۵]، متن‌کاوی[۱۵، ۱۶، ۱۷ و۱۸]، تشخیص موضوع در توییتر[۱۹] و داده‌های بالینی استفاده شده‌است[۴، ۲۰، ۲۱، ۲۲، ۲۳و ۲۴].

در این نوشتار، ابتدا برخی از پیشینه‌های ریاضی «نظریهٔ گروه»، «هم‌بافت سادکی» و «همولوژی» مرور می‌شوند؛ سپس الگوریتم «همولوژی ماندگار» معرفی و این روش برای انتساب نویسندگی بر مجموعه‌داده‌های اشعار به کار گرفته و نتایج را تحلیل می‌شوند؛ سپس یک روش جدید در تحلیل توپولوژیکی داده؛ یعنی نگاشت‌گر معرفی می‌شود و برای انتساب نویسنده بر مجموعه‌داده‌های اشعار اعمال و نتیجهٔ آن به‌عنوان هم‌بافت سادکی که فعل‌وانفعالی است و می‌توان آن را با استفاده از آمار

[1] تحلیل توپولوژیکی داده

[DOI: 10.66224/jsdp.22.3.77]

به روش‌های مختلف کمی‌سازی کرد، تجزیه‌وتحلیل می‌شود. در بخش آخر، نتایج به‌کارگیری هر دو الگوریتم روی مجموعهٔ اشعار مقایسه می‌شوند.

بخش نوآورانهٔ مقاله، مربوط به این واقعیت است که تابه‌حال تحلیل توپولوژیکی داده در بخش انتساب نویسنده به‌کار نرفته است و این روش با دقت خوبی این کار را انجام می‌دهد.

۲- پیش‌نیازها

مطالب این بخش‌ها همگی استانداردند و می‌توان آن‌ها را در مرجع [۱] فارسی یافت.

۲-۱-نظریهٔ گروه

تعریف ۱. یک گروه یک جفت $(G,*)$ است که $*$ یک عملیات دوتایی در مجموعهٔG است به‌همراه یک عضو همانی e از مجموعه که در روابط زیر صدق می‌کنند. برای هر $a, b, c \in G$ داشته باشیم:

- $(a * b) * c = a * (b * c)$
- $e * a = a * e = a$
- برای هر a یک عضو منحصربه‌فرد a^{-1} وجود دارد که $a * a^{-1} = e$.

تعریف ۲. گروه G را آبلی نامیم هرگاه برای $a, b, \in G$ دل‌خواه داشته باشیم: $a * b = b * a$.

تعریف ۳. زیرمجموعهٔ H را یک زیرگروه$(G,*)$ می‌نامیم هرگاه داشته باشیم $(H,*)$ خودش یک گروه باشد.

تعریف ۴. برای دو گروه $(G,*)$ و $(G,*')$، نگاشت $\varphi: G \to G'$ را یک هوموورفیسم گروه نامیم هرگاه

$$\varphi(a * b) = \varphi(a) *' \varphi(b) \qquad (۱)$$

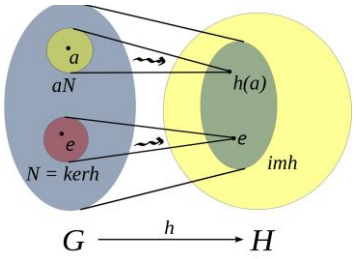
تعریف ۵. برای هوموورفیسم $\varphi: G \to G'$ هستهٔ φ را با $\ker \varphi$ نمایش می‌دهیم و تعریف می‌کنیم:

$$\ker \varphi = \{a \in G: \varphi(a) = e'\} \qquad (۲)$$

و تصویر φ را با $Im\ \varphi$ نمایش می‌دهیم و تعریف می‌کنیم:

$$Im\ \varphi = \{\varphi(a): a \in G\} \qquad (۳)$$

در زیر این دو مفهوم تصویری نمایش داده شده‌اند.



(شکل-۱):تصویر و هستهٔ هوموورفیسم
(Figure-1): Image and kernel of homomorphism

تعریف ۶. برای زیرگروه H از G و $a \in G$ مجموعهٔ $aH = \{ah: h \in H\}$ را یک هم‌دستهٔ H می‌نامیم.

مجموعهٔ تمام هم‌دسته‌های H تشکیل یک گروه می‌دهند که آن را با $\frac{G}{H}$ نمایش می‌دهیم و گروه خارج قسمتی می‌نامیم.

۲-۲-هم‌بافت سادکی و همولوژی

تعریف ۷. برای بردارهای $u_1, u_2, \dots, u_n \in \mathbb{R}^n$ و ضرایب $\lambda_1, \lambda_2, \dots, \lambda_n \in \mathbb{R}$ که $\sum_i \lambda_i = 1$ ترکیب $\lambda_1 u_1 + \lambda_2 u_2 + \cdots \_ \lambda_n u_n$ را یک ترکیب آفین از بردارها نامیم؛ اگر به‌علاوهٔ هر λ_i نامنفی باشد ترکیب فوق را یک ترکیب محدب نامیم.

تعریف ۸. پوش محدب[1] هر مجموعه‌ای از بردارهای $u_1, u_2, \dots, u_n \in R^n$، مجموعهٔ تمام ترکیبات محدب از این بردارها است.

تعریف ۹. یک $k-$ سادک σ پوش محدب $k+1$ نقطه به‌طور آفین مستقل $u_0, \dots, u_k$ است. پس هر ۰-سادک یک رأس و ۱- سادک یک یال و ۲- سادک یک مثلث و ۳- سادک یک چهاروجهی است.

تعریف ۱۰. حال یک اجتماع خاص از چند سادک باید تعریف شود که به آن هم‌بافت سادکی می‌گویند. تعریف شهودی K هم‌بافت سادکی این است که اگر یک سادک در K باشد، تمام وجوه آن نیز باید در K باشد؛ علاوه‌بر این، سادک‌ها باید در وجه‌ها به‌هم بچسبند یا جدا باشند.

تعریف ۱۱. وجه یک سادک σ تولیدشده توسط $\{u_i\}_{i=0}^{k}$، پوش محدب تولیدشده توسط زیرمجموعه‌ای از مجموعهٔ فوق است و اگر زیرمجموعه کل مجموعهٔ فوق نباشد به آن وجه سره می‌گویند.

تعریف ۱۲. یک هم‌بافت سادکی K، مجموعه‌ای متناهی از سادک‌ها است که اگر یک سادک در هم‌بافت سادکی باشد هر وجه آن نیز در هم‌بافت سادکی است و برای دو سادک $\sigma, \tau \in K$ داشته باشیم $\sigma \cap \tau$ یا تهی و یا وجهی از هر دو σ, τ است؛ به‌عبارت ساده هم‌بافت سادکی مجموعه‌ای از بلوک‌های سازنده‌اش یعنی سادک‌ها است که فقط می‌توانند از یک وجه که خودش نیز سادک است به‌هم بچسبند.

تعریف ۱۳. گروه p- زنجیر از یک هم‌بافت سادکی K که با $(C_p(K), +)$ نشان داده می‌شود، گروه آبلی آزاد از p- سادک‌های جهت‌دار$[\sigma]$ است که در آن داریم $[\sigma] = -[\tau]$ اگر و فقط اگر $\sigma = \tau$ ، اما جهت‌های متفاوتی دارند. عنصری از $C_p(K)$، یک p- زنجیر نامیده می‌شود که با $\sum_i \lambda_i [\sigma_i]$ که در آن $\lambda_i \in \mathbb{Z}$ و $\sigma_i \in K$ نشان داده می‌شود.

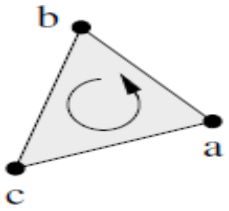
تعریف ۱۴. فرض کنید K یک هم‌بافت سادکی و $\sigma \in K$ به‌صورت $\sigma = [v_0, \dots, v_k]$ باشد، هوموورفیسم مرز $\partial_k: C_k(K) \to C_{k-1}(K)$ به‌صورت زیر تعریف می‌شود:

$$\partial_k(\sigma) = \sum_i (-1)^i [v_0, \dots, \widehat{v_i}, \dots, v_k] \qquad (۴)$$

[1] Convex hull

[DOI: 10.66224/jsdp.22.3.77]

که $\widehat{v_i}$ نشان می‌دهد v_i در دنباله حذف شده‌است.



(شکل-۲): ۲- سادک [a, b, c]
(Figure-2): simplex[a,b,c]

مثال ۱۵. مرز مثلث جهت‌دار در شکل (۲) به‌شرح زیر است:

$$\partial[a,b,c] = [b,c] - [a,c] + [a,b] = [b,c] + [c,a] + [a,b] \quad (۵)$$

تعریف ۱۶. k-امین گروه دورها $Z_k = ker\partial_k$ است. یک زنجیر که عضوی از Z_k است، یک k- دور نامیده می‌شود. k-امین گروه مرزی همان $B_k = Im\partial_{k+1}$ است. به زنجیر d که عضوی از B_k است، k-امین مرز می‌گویند.

لم ۱۷. لم بنیادی همولوژی: برای هر $(p+1)$ -زنجیر d داریم $\partial_p\partial_{p+1}\, d = 0$.

از آخرین لم، می‌دانیم که گروه مرزها زیرگروهی از گروه دورها را تشکیل می‌دهند و بعد می‌توانیم گروه خارج‌قسمتی از آن‌ها را تشکیل دهیم؛ به عبارت دیگر، می‌توانیم گروه دورها را به کلاس‌هایی از دورهایی تقسیم کنیم که از نظر مرزی با یکدیگر تفاوت دارند. این مقدمات به مفهوم گروه‌های همولوژی و بعد آن‌ها منجر می‌شود که اکنون آن‌ها تعریف می‌شوند و مورد بحث قرار می‌گیرند.

تعریف ۱۸. p-امین گروه همولوژی با H_p نمایش داده و تعریف می‌شود. $H_p = \frac{Z_p}{B_p}$ بعد این گروه با β_p نمایش داده و عدد بتی p-ام نامیده می‌شود.

تعریف ۱۹. هم‌بافت ویتوریس-ریپس[1] با قطر ϵ هم‌بافت سادکی است که به‌صورت زیر تعریف می‌شود:

$$VR(\epsilon) = \{\sigma:\ diam(\sigma) \le \epsilon\} \quad (۶)$$

که جایی که قطر σ ($diam(\sigma)$) بزرگترین فاصلهٔ ممکن از نقاط در σ است.

تعریف ۲۰. دنباله‌ای از زیرمجموعه‌های سادهٔ تو در توی $\emptyset \subseteq \cdots K_0 \subseteq K_1 \subseteq K_2 \subseteq \cdots$ یک پالایش نامیده می‌شود.

تعریف ۲۱. پالایش $\emptyset \subseteq K_0 \subseteq K_1 \subseteq K_2 \subseteq \cdots$ را در نظر بگیرید.

به‌طور طبیعی دنباله‌ای از هوموفیسم‌های تولیدشده به‌صورت زیر را خواهیم داشت:

$$H_p(K_0) \longrightarrow H_p(K_1) \longrightarrow H_p(K_2) \longrightarrow \cdots$$

ما می‌گوییم که p-امین کلاس همولوژی $[c]$ در K_i متولد می‌شود، اگر $[c] \in K_i$ اما $[c] \notin K_{i-1}$، و می‌گوییم $[c]$در K_i می‌میرد، اگر $[c] \in K_i$ اما $[c] \notin K_{i-1}$.

[1] Vietoris-Rips

۳- همولوژی ماندگار

همولوژی ماندگار سعی می‌کند گروه‌های همولوژی و حفره‌ها را با کمک پالایش‌ها پیدا و ردیابی کند. در این بخش سعی می‌شود مفهوم و الگوریتم همولوژی ماندگار توضیح داده شود و سپس در طبقه‌بندی متن اعمال می‌شود. یک هم‌بافت سادکی پالایش‌شده با توابع مرزی آن، هم‌بافت پایا نامیده می‌شود. برای یک دنبالهٔ صعودی از اعداد حقیقی مثبت، به یک هم‌بافت پایا می‌رسیم.

هدف پژوهش حاضر تجزیه‌وتحلیل خواص توپولوژیکی مجموعه‌داده‌های ابر نقطه‌ای با تجزیه‌وتحلیل هم‌بافت پایای آن است. برخی از نمایش‌های معمولی برای همولوژی ماندگار عبارت‌اند از: بارکد، نمودار پایدار و چشم‌انداز پایا. یک بارکد هر مولد پایا (کلاس‌هایی که p-امین گروه همولوژی را تولید می‌کنند) را با یک خط افقی نشان می‌دهد که از نخستین مرحلهٔ پالایش در جایی که آن ظاهر می‌شود شروع می‌شود و به مرحله‌ای از پالایش که در آن ناپدید می‌شود، پایان می‌یابد. یک نمایش بصری همولوژی ماندگار، نمودار پایدار است که برای هر مولد یک نقطه با مؤلفهٔ x به‌عنوان زمان تولد مؤلفهٔ y به‌عنوان زمان مرگ، رسم می‌کند.

تعریف ۲۲. p -نمودار ماندگاری D از یک پالایش $\emptyset \subseteq K_0 \subseteq K_1 \subseteq K_2 \subseteq \cdots$ به‌صورت زیر تعریف می‌شود.

فرض کنید μ_p^{ij} تعداد کلاس‌های p- بعدی مستقل باشد که در مرحلهٔ K_i متولد می‌شوند و در K_j می‌میرند، آن‌گاه D با رسم مجموعه‌ای از نقاط (i,j) با ضریب μ_p^{ij} به‌دست می‌آید که در آن قطر ناحیه با ضریب بی‌نهایت هم به نمودار اضافه می‌شود.

برای مقایسهٔ دو نمودار ماندگار، معیارهایی تعریف شده‌است که دو مورد از مهم‌ترین آن‌ها فاصلهٔ تنگنا (bottleneck) و واسرشتاین (Wasserstein) است.

تعریف ۲۳. فرض می‌شود D_1, D_2 دو نمودار ماندگار و B مجموعهٔ همهٔ توابع دوسویی $\varphi: D_1 \to D_2$ باشد. اگر $\|.\|_\infty$ نرم سوپریمم باشد؛ سپس فاصلهٔ تنگنا بین دو نمودار ماندگاری D_1, D_2 که با $W_\infty(D_1, D_2)$ نشان داده شده‌است به‌صورت زیر تعریف می‌شود.

$$W_\infty(D_1, D_2) = \inf_{\varphi \in B} \operatorname{Sup}_{x \in D_1} \|x - \varphi(x)\|_\infty \quad (۷)$$

تعریف ۲۴. فرض کنید D_1, D_2 دو نمودار ماندگار و B مجموعهٔ همهٔ توابع دوسویی $\varphi: D_1 \to D_2$ باشد. اگر $\|.\|_\infty$ نرم سوپریمم باشد؛ سپس فاصلهٔ تنگنا بین دو نمودار ماندگاری D_1, D_2 که با $W_p(D_1, D_2)$ نشان داده شده‌است به‌صورت زیر تعریف می‌شود.

$$W_p(D_1, D_2) = \left[\inf_{\varphi \in B} \sum_{x \in D_1} \|x - \varphi(x)\|_\infty^p\right]^{\frac{1}{p}} \quad (۸)$$

از آنجا که تجزیه‌وتحلیل اطلاعات مربوط به گروه‌ها و حفره‌های همولوژی بسیار دشوار است، می‌توانیم از یک

[DOI: 10.66224/jsdp.22.3.77]

روش بصری‌سازی به نام «بارکد» استفاده کنیم، ایده به شرح زیر است:

اگر حفره‌ای در ε_{t_1} ظاهر شد، نمودار شروع به رسم خطی می‌کنیم که در آن شروع خط در ε_{t_1} در محور x باشد و اگر در ε_{t_2} ناپدید شد، کشیدن خط را متوقف می‌کنیم و انتهای خط در ε_{t_2} خواهد بود.

چشم‌انداز پایا روش دیگری است که در [۲۵] برای تجسم هموولوژی ماندگار معرفی شده‌است.

تعریف ۲۵. چشم‌انداز پایا تابع $\lambda: \mathbb{N} \times \mathbb{R} \to \overline{\mathbb{R}}$ است که در آن $\overline{\mathbb{R}}$ نشان‌دهندهٔ اعداد حقیقی توسعه‌یافتهٔ $[-\infty, \infty]$ است. تعریف دیگر ممکن است آن را دنباله‌ای از توابع $\lambda_k: \mathbb{R} \to \mathbb{R}$ در نظر بگیرد که در آن $\lambda_k(t) = \lambda(k,t)$.

$$(9)\ \lambda_k(t) = sup\{m \geq 0 : \beta^{(t-m),(t+m)} \geq k\}$$

که در آن $\beta^{i,j}$ بعد گروه $\frac{H_i}{H_j}$ است.

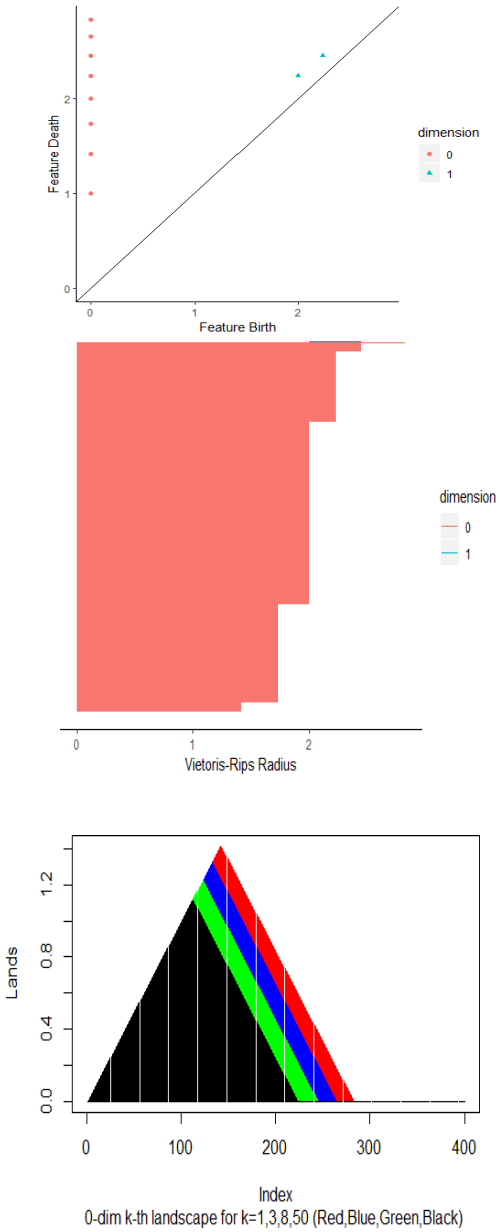
نمودار چشم‌انداز پایا اعداد بتی پایا و غیرپایا را نشان می‌دهد؛ برای مثال بیشینهٔ نمودار چشم‌انداز پایا نشان‌دهندهٔ پایدارترین عدد بتی است.

۳-۱-الگوریتم هموولوژی ماندگار

فرض کنید $\mathbb{P}$ یک دادهٔ ابر نقطه‌ای باشد. ابتدا هم‌بافت ویتوریس-ریپس برای $\mathbb{P}$ به‌صورت زیر ساخته می‌شود: دنبالهٔ صعودی از اعداد حقیقی مثبت $\varepsilon_1, \varepsilon_2, \ldots$ را در نظر بگیرید؛ سپس پوششی از دایره‌هایی با مرکز نقاط $\mathbb{P}$ و قطر ε_1 ساخته می‌شود؛ بنابراین تعداد نقاط داده در ابرداده، با دایره‌ها برابر است. در مرحلهٔ بعد بین مرکز هر دو دایره که تقاطع دارند، یک یال کشیده می‌شود و لذا یک هم‌بافت سادکی $VR(\varepsilon_1)$ ایجاد شده‌است. همین فرایند برای همه $i = 1,2,3, \ldots$ انجام می‌شود؛ درنتیجه یک پالایش از هم‌بافت‌های $VR(\varepsilon_i)$ موجود است؛ از آنجایی که تجزیه‌وتحلیل اطلاعات مربوط به گروه‌ها و حفره‌های هموولوژی بسیار دشوار است، می‌توان از یک نمودار شهودی به نام بارکد استفاده کرد، ایده به این صورت است که اگر حفره‌ای در ε_{t_1} ظاهر شد، بارکد شروع به رسم پاره‌خطی می‌کند که مبدأ خط در ε_{t_1} در محور x است و اگر در ε_{t_2} ناپدید شد، رسم پاره‌خط را متوقف می‌کند و انتهای پاره‌خط در ε_{t_2} خواهد بود.

۳-۳- نتایج پیاده‌سازی الگوریتم هموولوژی ماندگار بر اشعار حافظ و فردوسی

هموولوژی ماندگار در بسته‌های نرم‌افزار R مانند "TDA" و "TDAstatus" پیاده‌سازی شده‌است. در این مقاله از این دو بسته برای طبقه‌بندی متن استفاده شده‌است. در این بخش از داده‌های متنی بخشی از (اشعار) دو شاعر ایرانی (حافظ و فردوسی) استفاده شده‌است. مجموعه‌داده‌ها از «شاهنامه» و «غزلیات حافظ» جمع‌آوری شد که شامل حدود هشت‌هزار مصرع از هر کتاب است. پس از پیش‌پردازش، داده‌ها به الگوریتم tf-idf داده می‌شود تا ماتریس نظیر داده متن ایجاد شود؛ سپس ماتریس به‌عنوان ورودی الگوریتم هموولوژی ماندگار وارد می‌شود. در ادامه نمودار ماندگار، بارکد و چشم‌انداز پایا که برای یک نمونه از اشعار فردوسی شامل هزار مصراع رسم‌شده را در شکل (۳) می‌توان دید.



(شکل- ۳): به‌ترتیب نمودارهای نخست تا سوم نمودار ماندگار، بارکد و چشم‌انداز پایا را برای هزار مصراع فردوسی نشان می‌دهد

(Figure-3): Diagrams from the top to the bottom respectively show the persistent diagram, barcode and landscape for 1000 hemistich of Ferdowsi.

(جدول-۱):۰-فاصله‌ها و ۱-فاصله‌های واسرشتاین بین بخش‌های متناظر اشعار فردوسی و حافظ

(Table-1): Wasserstein 0-distances and 1-distances between corresponding parts of Ferdowsi and Hafez poems

بخش‌های مختلف اشعار	۰-فاصله	۱-فاصله
فردوسی ۱ و حافظ ۱	۴۴۹.۴۲۷۴	۰.۶۸۸۱۷۸۸
فردوسی ۲ و حافظ ۲	۸۲.۱۶۰۹۸	۰.۲۹۲۸۳۹۹
فردوسی ۳ و حافظ ۳	۸۲.۸۴۳۰۸	۳.۷۶۵۸۰۸
فردوسی ۴ و حافظ ۴	۵۵۳.۳۸۴۵	۲.۰۷۸۴۱۸
فردوسی ۵ و حافظ ۵	۱۱۸.۲۳۱۶	۳.۳۱۲۱۱۲
فردوسی ۶ و حافظ ۶	۱۴۱.۳۲۱	۰.۵۲۴۸۲۵۴

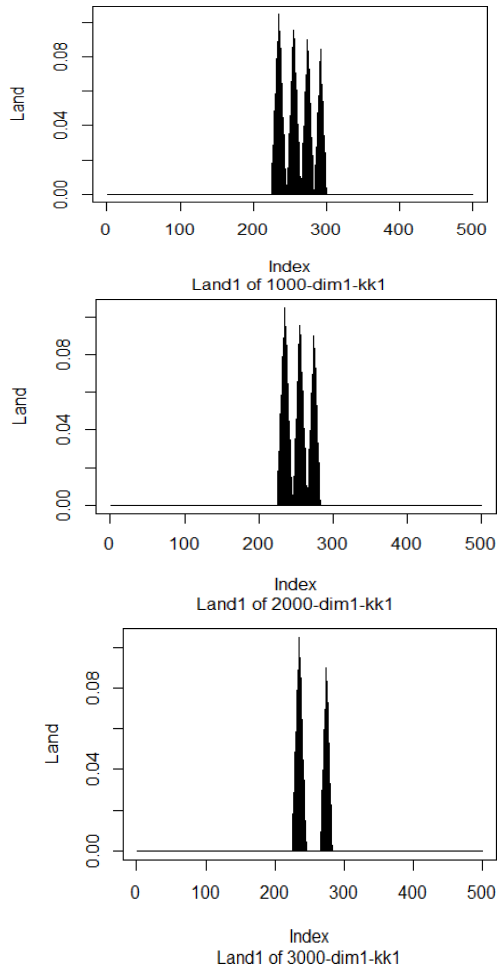
[DOI: 10.66224/jsdp.22.3.77]

فردوسی ۷ و حافظ ۷	۷۴.۲۸۲۹۶	۳.۲۸۲۴۶۸
فردوسی ۸ و حافظ ۸	۳۸۷.۲۶۲۸	۰.۴۵۰۳۸۷۱

(جدول-۲): ۰-فاصله‌ها و ۱-فاصله‌های واسرشتاین بین بخش‌های متوالی اشعار فردوسی

(Table-2): Wasserstein 0-distances and 1-distances between consecutive parts of the Ferdowsi poems

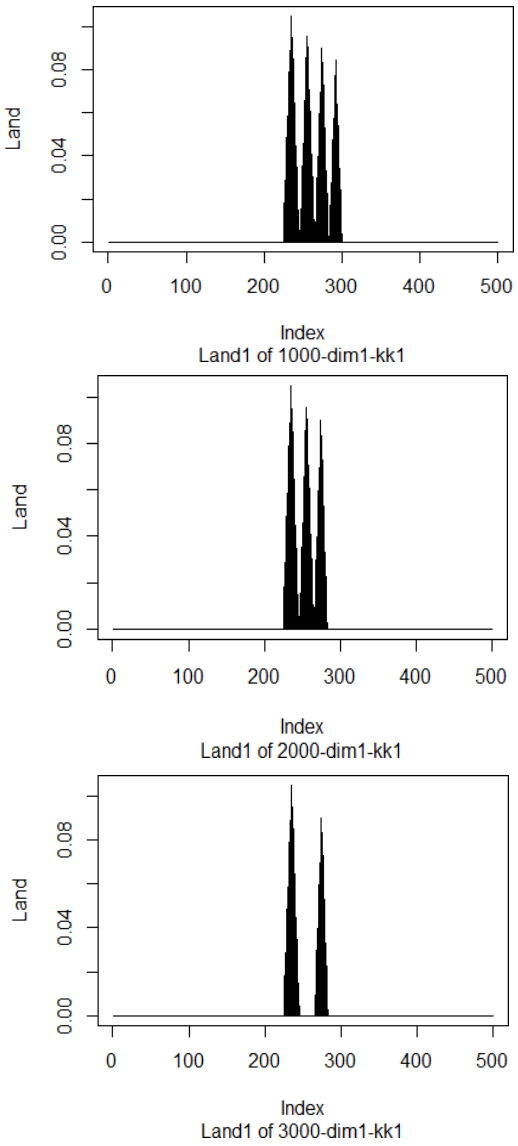
بخشهای مختلف اشعار	۰-فاصله	۱-فاصله
فردوسی ۲ و حافظ ۲	۱۴۶.۱۰۸۵	۰.۳۶۳۴۲۰
فردوسی ۳ و حافظ ۳	۵.۶۴۰۱۸	۳.۴۰۵۹۳۲
فردوسی ۴ و حافظ ۴	۳۱۰.۸۶۹۲	۱.۹۴۶۳۱۹
فردوسی ۵ و حافظ ۵	۲۷۸.۳۹۷۳	۱.۴۲۷۱۵۵
فردوسی ۶ و حافظ ۶	۳۸.۹۳۹۸۴	۳.۰۰۳۳۸۸
فردوسی ۷ و حافظ ۷	۱۴.۵۸۰۷	۳.۲۶۶۱۴۵
فردوسی ۸ و حافظ ۸	۱۶.۷۷۷۹۸	۳.۳۲۲۸۲۴



(شکل-۴): سه تصویر بالا چشم‌انداز پایای سه بخش نخست از اشعار حافظ را نشان می‌دهد.

(Figure-4): Three top images show the landscapes of the first three parts of Hafez's poems

همچنین هشت‌هزار مصراع حافظ به هشت قسمت با طول هزار تقسیم شده؛ سپس نمودار ماندگار و نخستین چشم‌انداز پایا هر قسمت نیز محاسبه شده‌است؛ درنهایت چشم‌انداز میانگین این قسمت‌ها ترسیم شده و برای هشت‌هزار مصراع «فردوسی» نیز همین موارد انجام شده‌است. در مرحلهٔ آخر، فاصله‌های واسرشتاین بین نمودارهای پایای بخش‌های متناظر اشعار «حافظ» و «فردوسی» محاسبه شده‌است. این نتایج در ادامه به‌شرح زیر آورده می‌شود: نتایج محاسبه ۰-فاصله‌ها و ۱-فاصله‌های واسرشتاین را برای بخش‌های مختلف اشعار حافظ و فردوسی که شمارهٔ بخشی از اشعار بعد از شاعر آن مشهود است در جدول (۱) آمده‌است؛ همچنین فاصلهٔ واسرشتاین بین بخش‌های متوالی از اشعار فردوسی در جدول (۲) در ادامه خواهد آمد.



(شکل-۵): سه تصویر بالا چشم‌انداز پایای سه بخش نخست از اشعار فردوسی را نشان می‌دهد.

(Figure-5): Three top images show the landscapes of the first three parts of Ferdowsi's poems.

۴- الگوریتم نگاشت‌گر

الگوریتم نگاشت‌گر توسط سینگ[1]، ممولی[2] و کارلسون[3] [۲۶] به‌عنوان ابزاری هندسی برای تحلیل و بصری‌سازی مجموعه‌داده‌ها معرفی شد. هدف نگاشت‌گر کاهش بُعد مجموعه‌داده، همچنین فشرده‌کردن مجموعه‌داده به یک

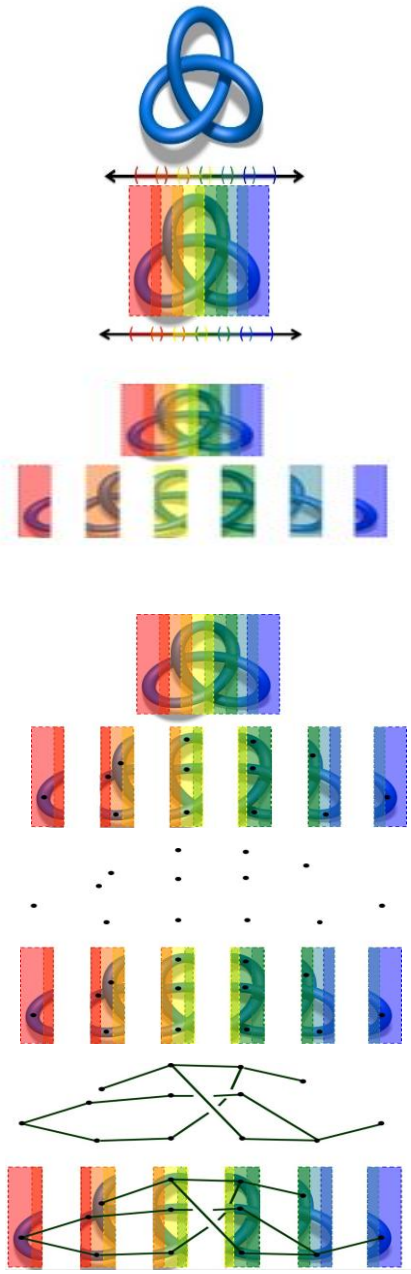
[1] Singh
[2] Mémoli
[3] Carlsson

[DOI: 10.66224/jsdp.22.3.77]

هم‌بافت سادکی که در حالت ساده گراف است که تجزیه‌وتحلیل ساختار توپولوژیکی یا شکل داده‌ها را آسان کند. در اینجا جدولی از نمادها برای توضیح ریاضیات روش نگاشت‌گر معرفی می‌شود.

(جدول-۳): جدول نمادهای روش نگاشت‌گر
(Table- 3): Table of symbols of mapper

نماد	توضیحات
$\mathbb{X}$	فضای زمینهٔ دادهٔ ابری $(\mathbb{R}^n)$
$\mathbb{p}$	دادهٔ ابری نقطه‌ای $(\mathbb{p} \subseteq \mathbb{x})$
$\mathbb{Y}$	فضای پارامتر که به‌طور معمول $\mathbb{Y} = \mathbb{R}$ است
$\mathbb{F}$	تابع پالایش $(\mathbb{F}: \mathbb{P} \subseteq \mathbb{x} \to \mathbb{y})$
Γ	برد تابع پالایش تحدید به $\mathbb{P}$
$\mathbb{U}$	یک پوشش $\mathbb{p}(\mathbb{p} \subseteq \cup U_\alpha)$
$\mathcal{E}$	یک پوشش از زیر بازه‌های Γ که هم‌پوشانی دارند



(شکل-۶): ایدهٔ شهودی نگاشت‌گر
(Figure-6): Intuitive idea of mapper

خلاصهٔ الگوریتم نگاشت‌گر با توجه به نمادهای بالا و دانش توپولوژی و آمار به شرح زیر است:

- ابتدا باید با تابع پالایش مناسب $\mathbb{F}: \mathbb{P} \subseteq \mathbb{x} \to \mathbb{y}$ شروع کرد.
- سپس برد $\mathbb{F}$ و تحدیدش به $\mathbb{P}$ را یافته و آن را Γ نامید.

پس از آن، لازم است Γ به زیر بازه‌های $\mathcal{E}$ تقسیم شود تا در مرحلهٔ بعد یک پوشش برای $\mathbb{p}$ از طریق $\mathbb{F}^{-1}$ ایجاد شود.

-برای هر $\varepsilon_i \in \mathcal{E}$ مجموعهٔ زیر ساخته می‌شود:

$$X_i = \{x: \mathbb{F}(x) \in \varepsilon_i\}$$
$$(\mathbb{P} \subseteq \cup X_i)$$

- نقاط هر عضو X_i از $\mathbb{U}$ با یک متریک مناسب خوشه‌بندی شود؛ بنابراین برای هر X_i، خوشه‌های X_{ij} ایجاد می‌شود.
- هر خوشه X_{ij} به‌عنوان یک رأس در نظر گرفته می‌شود و وقتی $X_{ij} \cap X_{rs} \neq \emptyset$ بین دو رأس X_{ij} و X_{rs} یک یال کشیده می‌شود.

ایدهٔ شهودی پشت نگاشت‌گر در شکل (۶) نشان داده شده‌است و می‌توان آن را به‌صورت زیر توضیح داد: فرض کنید یک دادهٔ ابرنقطه‌ای وجود دارد که یک شکل را نشان می‌دهد؛ برای مثال یک گره. ابتدا کل داده‌ها بر روی یک سیستم مختصات با ابعاد کمتر به‌منظور کاهش پیچیدگی از طریق کاهش ابعاد تصویر می‌شود (در اینجا داده‌ها از روی گره به محور $\boldsymbol{x}$ تصویر می‌شود)؛ سپس فضای پارامتر (محور$\boldsymbol{x}$) به چند بازه با درصد هم‌پوشانی مفروض تقسیم می‌شود. بعد، با گرفتن تصویر وارون این بازه‌ها تحت $\mathbb{F}$، داده‌ها در نواحی دارای هم‌پوشانی قرار می‌گیرند؛ سپس از الگوریتم‌های خوشه‌بندی به‌منظور طبقه‌بندی نقاط هر ناحیه به چند خوشه استفاده می‌شود؛ درنهایت، می‌توان هم‌بافت سادگی خود را ایجاد کرد.

۴-۱-الگوریتم برخط تحلیل‌دادهٔ نگاشت‌گر بر اساس TDA و مجموعه‌دادهٔ ما

پیاده‌سازی الگوریتم نگاشت‌گر در حال حاضر در بسته‌های پایتون مانند "Mapper" و "Kepler Mapper" در دسترس است. در این مقاله از "Kepler Mapper" در کنار سایر بسته‌های پایتون استفاده شده‌است؛ همچنین از داده‌های متنی (اشعار) دو شاعر ایرانی حافظ و فردوسی استفاده شده‌است.

مجموعه‌داده‌ها از «شاهنامه» (کتاب حماسی فردوسی) و «غزلیات حافظ» که شامل حدود نه‌هزار مصراع مختلف (از اشعار حماسی گرفته تا عاشقانه) از هر دو کتاب است. ابتدا به پیش‌پردازش داده پرداخته می‌شود که شامل مراحل زیر است:

۱-پاک‌سازی اولیه (cleaning)

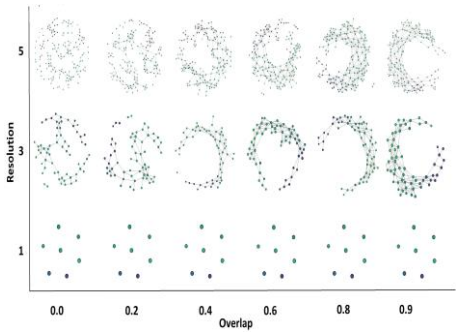
- حذف علائم نگارشی مانند نقطه، ویرگول، گیومه؛
- حذف نویسه‌های غیرمجاز یا اعداد؛

[DOI: 10.66224/jsdp.22.3.77]

• تبدیل تمام نویسه‌ها به فرم استاندارد (برای مثال «ی» به «ی»، «ک» به «ک»)؛

۲-توکن‌سازی (Tokenization) تقسیم هر مصراع به واژگان با استفاده از ابزارهای استاندارد مانند Hazm یا Parsivar.

۳-حذف ایست‌واژه‌ها (Stopword Removal)



(شکل-۷): مقایسهٔ وضوح‌های مختلف و درصد هم‌پوشانی: همان‌طور که از جدول بالا مشخص است، هرچه وضوح بیشتری داشته باشیم، نمودار حاصل بهتر طبقه‌بندی می‌شود و هر چه درصد هم‌پوشانی بیشتر باشد، نمودار حاصل فشرده‌تر می‌شود

(Figure-7): Comparison of different resolutions and percentage of overlap: As it is clear from the table above, the more resolution we have, the resulting graph is better classified, and the higher the percentage of overlap, the more compact the resulting graph is

استفاده از فهرست ایست‌واژه‌های رایج فارسی (مانند «از»، «به»، «در»، «که»، «برای»). این واژگان به‌دلیل فراوانی زیاد و بار معنایی پایین، برحسب معمول حذف می‌شوند تا نویسه‌های معنادار باقی بمانند.

۴-نمایش برداری داده‌ها (Vectorization)

دو رویکرد در مقاله مورد استفاده قرار گرفته است:

• TF-IDF: تولید ماتریس وزن‌دهی بر اساس فراوانی-معکوس واژگان در مصراع‌ها.

• BERT Embedding: استفاده از مدل پیش‌آموزش‌یافته برای استخراج بردارهای معنایی از مصراع‌ها.

۵-استانداردسازی (Normalization)

• در روش TF-IDF، مقادیر بردارها نرمال شده‌اند (برای مثال واحد برداری)

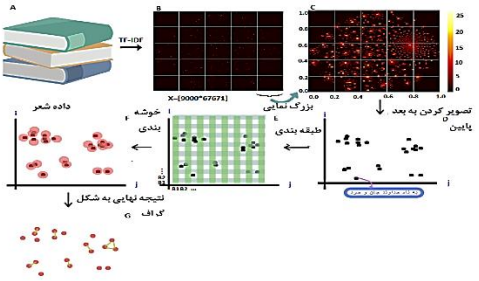
• در روش BERT، پس از تولید بردارها، از MinMaxScaler برای مقیاس‌بندی استفاده شده‌است.

۶- کاهش ابعاد (Dimensionality Reduction)

استفاده از الگوریتم‌هایی مانند UMAP و TruncatedSVD برای کاهش ابعاد بردارهای TF-IDF یا BERT به‌منظور تحلیل با Mapper و Homology.

۴-۱-۱- روش TF-IDF

در این بخش برای تبدیل داده به یک ماتریس به شکل $[]_{67671\times 9000}$(#واژه در پیکره × # مصرع)، از "TF-IDF" و سپس از "truncatedSVD" و "t-SNE" به‌عنوان توابع پالایش استفاده شد؛ همچنین وضوح و درصد هم‌پوشانی مناسب برای نواحی فضای پارامتر R2 انتخاب شد. (شکل (۷) وضوح و مقادیر مختلف هم‌پوشانی را نشان می‌دهد. نگاشت‌گر درنتیجهٔ وضوح بالاتر، پارتیشن‌بندی و طبقه‌بندی بهتر داده‌ها و درصد هم‌پوشانی بالاتر نمودار فشرده‌تری را خواهد داد.)؛ درنهایت، نواحی موجود در مجموعه‌داده را با «خوشه‌بندی تجمعی[1]» موجود در بستهٔ «sklearn» با متریک شباهت کسینوسی و پیوند کامل[2]، خوشه‌بندی می‌کنیم. تمامی این مراحل در شکل (۸) بصری‌سازی شده‌است.



(شکل-۸): بصری‌سازی پیاده‌کردن نگاشت‌گر

(Figure-8): Visualization of the implementation of the mapper

۴-۱-۲- روش BERT Embedding

در این روش، از ترکیب چندین روش مدرن یادگیری ماشین و پردازش زبان طبیعی برای تحلیل و نمایش داده‌های متنی استفاده شده‌است. توضیحات این مراحل به‌صورت جامع و به‌ویژه با تمرکز بر بخش استخراج تعبیه‌ها (embeddings) با استفاده از مدل BERT در ادامه آمده‌است. مدل BERT یکی از معروف‌ترین مدل‌های ازپیش آموزش‌داده‌شده در پردازش زبان طبیعی (NLP) است که توانایی بسیار بالایی در استخراج ویژگی‌های معنایی از متن دارد. در این بخش، به‌کمک BERT تعبیه‌های معنایی برای هر نمونهٔ متنی استخراج می‌شود. روند اجرای این مرحله به‌شرح زیر است؛ در ابتدا، توکنایزر و مدل BERT با نام all-MiniLM-L12-v2 که نسخه‌ای کوچک‌تر و بهینه‌شده از BERT است، بارگذاری می‌شود. این مدل به‌صورت خودکار از مخزن مدل‌های آماده فراخوانی می‌شود و نیازی به تنظیمات پیچیدهٔ اولیه ندارد. توکنایزر، هر متن ورودی را به قالبی عددی تبدیل می‌کند که مدل BERT قادر به درک آن باشد. این تبدیل شامل جداسازی واژگان، تبدیل آن‌ها به شاخص‌های عددی و تنظیم طول ورودی‌ها (پدینگ و برش) می‌شود تا با ابعاد ورودی مدل BERT هم‌خوانی داشته باشد.

برای هر متن در مجموعه‌داده، توکنایزر ورودی را به توکن‌های مورد نیاز مدل تبدیل می‌کند و نتیجه را به‌عنوان ورودی به مدل BERT می‌فرستد؛ مدل BERT

[1] Agglomerative Clustering
[2] Complete linkage

[DOI: 10.66224/jsdp.22.3.77]

سپس با استفاده از توکن‌ها، عملیات پردازش را آغاز می‌کند. این مدل به‌صورت عمیق و لایه‌بندی شده‌است و در هر لایه به‌صورت گام‌به‌گام، ویژگی‌های پیچیده‌تر زبانی و معنایی را از ورودی استخراج می‌کند.

از خروجی لایهٔ آخر مدل BERT، بردار تعبیهٔ نهایی هر واژه به‌دست می‌آید، اما برای نمایندگی کل متن، از میانگین بردارهای این واژگان استفاده می‌شود. این روش که به‌عنوان «میانگین تعبیه‌ها» شناخته می‌شود، اطلاعات کل متن را به یک بردار واحد با ابعاد ثابت تبدیل می‌کند. این بردار نهایی، یک تعبیهٔ معنایی برای هر نمونهٔ متنی است و ویژگی‌های زبانی و معنایی مهم متن را در خود دارد.

تعبیه‌های BERT به‌طور معمول در ابعاد بالا و پیچیده‌اند و استفادهٔ مستقیم از آن‌ها برای خوشه‌بندی یا تجسم داده‌ها ممکن است به محاسبات سنگین منجر شود؛ به همین دلیل، از روش UMAP (Uniform Manifold Approximation and Projection) برای کاهش ابعاد این بردارها استفاده می‌شود. این الگوریتم به‌گونه‌ای طراحی شده‌است که با کاهش ابعاد، تا حد ممکن ساختار و روابط موجود در داده‌ها را حفظ کند. در این کد، ابعاد تعبیه‌ها به پنج کاهش می‌یابد تا پردازش‌های بعدی سریع‌تر و کارآمدتر انجام شوند. در مرحلهٔ بعد نقشه‌برداری و خوشه‌بندی با KeplerMapper به‌صورت زیر انجام می‌شود:

۱-انتخاب تابع فرافکنی و خوشه‌بندی

۲-پوشش‌دهی و ساخت گراف

۳-تجسم گراف نهایی و نمایش تعاملی

داده‌های کاهش‌یافته با استفاده از روش استانداردسازی MinMax در محدوده‌ای مشخص فرافکنی می‌شوند تا به‌شکل یکنواخت‌تری برای نگاشت‌گر آماده شوند. این کار با استفاده از MinMaxScaler انجام می‌شود. برای خوشه‌بندی، از AgglomerativeClustering استفاده شده‌است که نوعی الگوریتم خوشه‌بندی سلسله‌مراتبی است و داده‌ها را در پنج خوشهٔ مجزا گروه‌بندی می‌کند. کلاس Cover با تعیین یک هم‌پوشانی سی‌درصدی بین بخش‌های داده‌ها، پوششی مناسب برای نگاشت‌گر ایجاد می‌کند. این پوشش‌دهی به نگاشت‌گر اجازه می‌دهد تا خوشه‌ها و زیرخوشه‌های داده‌ها را به‌صورت گراف نمایش دهد که در آن هر گره نشان‌دهندهٔ یک خوشه یا زیرخوشه از داده‌های متنی است؛ درنهایت، با توجه به ویژگی‌های هر خوشه، راهنماهای متنی (tooltip) سفارشی ساخته می‌شوند که شامل محتوای خلاصه‌ای از هر نمونهٔ متنی است و در کنار گراف برای کاربر قابل مشاهده است. این راهنماها به شکل HTML طراحی شده‌اند تا کاربر بتواند محتوای هر خوشه را به‌صورت تعاملی مشاهده کند. گراف نهایی با تمام ویژگی‌های مذکور ذخیره و آماده استفاده می‌شود. در این روش ترکیبی، از قدرت مدل BERT برای استخراج ویژگی‌های معنایی، از UMAP برای کاهش پیچیدگی داده‌ها و از Mapper برای نمایش توپولوژی داده‌ها استفاده شده‌است تا بتوان به‌صورت بصری به تحلیل داده‌های متنی پرداخت.

تبدیل متن، به‌ویژه شعر، به داده‌های توپولوژیکی برای تحلیل، فرایندی چندمرحله‌ای است که ساختار معنایی و روابط پنهان در متن را به‌صورت هندسی نمایش می‌دهد. در ادامه، جزئیات کامل این فرایند از ابتدا تا انتها توضیح داده می‌شود. در ابتدا، متن خام (شعر) باید پیش‌پردازش شود تا برای تحلیل‌های بعدی آماده باشد، پس از آن باید متن به واحدهای معنادار (مانند غزل‌های منفرد در دیوان حافظ) تقسیم شود؛ سپس نرمال‌سازی متن یعنی حذف یا استانداردسازی علائم نگارشی، تبدیل حروف به شکل استاندارد و حذف فاصله‌های اضافی مد نظر قرار می‌گیرد و همچنین باید به حذف واژگان ایست[1] یعنی واژگان پرتکرار و کم‌اهمیت مانند حروف اضافه و ربط پرداخت.

ریشه‌یابی یا لم‌سازی؛ یعنی تبدیل واژگان به شکل پایهٔ آن‌ها (در شعر کلاسیک فارسی این مرحله می‌تواند پیچیده باشد) نیز لازم است. حال به مرحلهٔ دوم؛ یعنی استخراج ویژگی و تولید تعبیه (Embedding) می‌رسیم که در آن، متن پیش‌پردازش شده به بردارهای عددی تبدیل می‌شود که نشان‌دهندهٔ معنا و محتوای متن هستند. مدل SBERT چندزبانه که در روش ما استفاده شد، یک شبکهٔ عصبی عمیق است که برای تولید تعبیه‌های متنی آموزش دیده‌است. این مدل، از معماری ترانسفورمر با توجه دوطرفه[2] استفاده می‌کند که روابط بین واژگان را در تمام متن در نظر می‌گیرد؛ سپس هر متن (غزل) از مدل SBERT عبور داده می‌شود و یک بردار با ابعاد بالا (به‌طور معمول ۳۸۴، ۵۱۲، یا ۷۶۸ بعدی) تولید می‌کند.

این بردارها در یک فضای چندبعدی قرار می‌گیرند که در آن متون با معانی مشابه به یکدیگر نزدیک‌ترند؛ برای مثال، برای هر غزل حافظ، یک بردار از بعد فضای تعبیه تولید می‌شود. به‌طور معمول تعبیه‌های تولیدشده دارای ابعاد بسیار بالا هستند که مصورسازی و تحلیل آن‌ها را دشوار می‌سازد؛ بنابراین، لازم است روش‌های کاهش ابعاد مانند تحلیل مؤلفه‌های اصلی (PCA) استفاده شوند. این روش ماتریس کوواریانس داده‌ها را محاسبه می‌کند و از

[1] Stop words
[2] bidirectional attention

[DOI: 10.66224/jsdp.22.3.77]

بردارهای ویژه با بزرگترین مقادیر ویژه برای نگاشت داده‌ها به فضایی با ابعاد کمتر استفاده می‌کند؛ همچنین می‌توان از روش‌های غیرخطی مانند t-SNE یا UMAP که روابط غیرخطی بین داده‌ها را بهتر حفظ می‌کنند، نیز برای کاهش ابعاد استفاده کرد.

برای داده‌های شعری ما، به‌طور معمول ابعاد از صدها به دو یا سه بُعد کاهش می‌یابد تا مصورسازی امکان‌پذیر شود؛ درنهایت در این پژوهش به ایجاد نمایش توپولوژیک باKepler Mapper که یک الگوریتم تحلیل داده‌های توپولوژیک است و ساختار توپولوژیک داده‌ها را از طریق تقسیم‌بندی و خوشه‌بندی کشف می‌کند، پرداخته می‌شود. ابتدا باید لنز مناسب اختیار کرد، لنز[1] تابعی است که داده‌ها را به یک فضای پارامتری ساده‌تر می‌نگارد. در مورد دادهٔ شعر، لنز می‌تواند PCA باشد؛ سپس برای ایجاد پوشش[2] فضای پارامتری به مجموعه‌ای از هم‌پوشانی‌ها[3] تقسیم می‌شود؛ هر هم‌پوشانی یک پالایه ایجاد می‌کند که نقاط داده را بر اساس مقادیر لنز انتخاب می‌کند. مرحلهٔ بعد خوشه‌بندی مجموعه‌های پوشش، با یکی از الگوریتم‌های خوشه‌بندی مانند DBSCAN است تا خوشه‌های محلی شناسایی شوند. مرحلهٔ آخر ساخت هم‌بافت سادکی یا گراف است که در آن خوشه‌ها به‌عنوان رئوس گراف درنظر گرفته می‌شوند. دو رأس اگر دارای نقاط دادهٔ مشترک باشند، با یک یال به‌هم متصل می‌شوند؛ برای مثال، در مورد غزلیات حافظ هر رأس می‌تواند مجموعه‌ای از غزل‌های با ویژگی‌های مشابه باشد و یال‌ها نشان‌دهندهٔ شباهت‌های بین گروه‌های مختلف غزل‌ها هستند. یک نمونهٔ تفسیر و تحلیل نمایش توپولوژیک نهایی، بینش‌هایی دربارهٔ ساختار و روابط در داده‌های شعری ارائه می‌دهد؛ برای مثال رئوس نشان‌دهندهٔ گروه‌های غزل‌ها با مضامین، سبک یا ویژگی‌های زبانی مشابه و یال‌ها نیز بیان‌کنندهٔ ارتباطات بین گروه‌های مختلف، که می‌تواند نشان‌دهندهٔ انتقال‌های تدریجی در سبک یا مضمون باشد، هستند. شکل کلی گراف می‌تواند ساختارهای کلان‌تر مانند دوره‌های مختلف شاعری یا تکامل سبک را نشان دهد. نقاط انشعاب می‌تواند محل‌هایی که مسیرهای مختلف سبکی یا مضمونی از هم جدا می‌شوند را روایت کند. به‌عنوان یک مثال کاربردی غزل زیر از حافظ را در نظر بگیرید:

«دلم ز صومعه بگرفت و خرقه سالوس /کجاست دیر مغان و شراب ناب کجَ» ابتدا متن نرمال‌سازی شده و به‌عنوان یک واحد کامل در نظر گرفته می‌شود؛ سپس از مدل SBERT عبور داده می‌شود و یک بردار ۵۱۲ بعدی تولید می‌کند که مفاهیم معنایی مانند «شراب»، «صومعه»، «خرقه» و مضمون «ریاستیزی» را کدگذاری می‌کند. بردار ۵۱۲ بعدی به یک بعد پایین‌تر (برای مثال دو بعد) کاهش می‌یابد تا بتوان آن را مصور کرد. در گراف توپولوژیک ساخته شده جایی که خوشهٔ حاوی این غزل به خوشه‌های دیگر با مضامین «عرفانی» یا «انتقادی» متصل می‌شود. این تحلیل توپولوژیک به پژوهش‌گر اجازه می‌دهد الگوهای پنهان و ساختارهای معنایی را در مجموعهٔ شعر شناسایی کند که با روش‌های سنتی تحلیل متن قابل تشخیص نیستند. در این پژوهش دو آزمون دقت، روی نمودار شکل خود بررسی شد؛ ابتدا کل نمودار به سه خوشه تقسیم شد. («حافظ»، «فردوسی»، «هردوشاعر»)، که در خوشهٔ «حافظ» گره‌هایی وجود دارد که درصد بالایی از اشعار حافظ را شامل می‌شود؛ به همین ترتیب خوشهٔ «فردوسی» نیز دارای گره‌هایی است که درصد بالای اشعار فردوسی را در بر می‌گیرد و در خوشهٔ «هر دوشاعر» به‌طور تقریبی به یک اندازه هر دو نوع شعر وجود دارد؛ سپس بررسی می‌شود که در حقیقت چند درصد از اشعار خوشهٔ حافظ متعلق به اشعار حافظ است و همین روش برای خوشه‌های دیگر نیز استفاده می‌شود. برای این کار به‌سادگی تعداد اشعار حافظ در هر گره در خوشهٔ «حافظ» بر تعداد تمام اشعار در همان خوشه تقسیم و همین آزمایش برای خوشه‌های دیگر نیز انجام می‌شود؛ برای مثال برای خوشهٔ «حافظ» محاسبات زیر انجام می‌شود؛ درصد دقت برابر است با تقسیم (تعداد اشعار حافظ در هر گره خوشه) به (تعداد تمام اشعار در کل خوشه)؛ بنابراین اگر درصد دقت α برای یک خوشه وجود داشته باشد به این معنی است که α درصد از اشعار آن خوشه به‌درستی برچسب‌گذاری شده‌است.

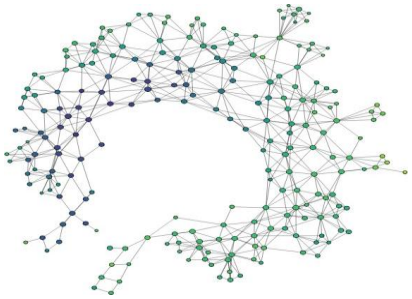
در شکل (۹) پس از بررسی نخستین آزمون روی هر خوشه، نتایج زیر حاصل شد: درصد دقت برای خوشهٔ «حافظ» هشتاد درصد، برای خوشهٔ «فردوسی» درصد دقت حدود ۹۴درصد و برای خوشهٔ «هر دو» در هر خوشه درصد دقت برای اشعار حافظ چهل درصد و برای اشعار فردوسی شصت درصد بوده‌است؛ بنابراین برای خوشهٔ «حافظ» می‌توان گفت که هشتاد درصد اشعار این خوشه دارای برچسب درستی هستند و به همین ترتیب برای خوشه‌های دیگر. در گام بعدی سعی شد برخی از بخش‌های نمودار بر اساس معنایی به چند خوشه تقسیم شوند. برای تجزیه‌وتحلیل بصری هر خوشه، متن در هر خوشه به‌عنوان یک ابر واژه نشان داده شد که تنها در یک نگاه به‌راحتی قابل درک است.

در تصویر شکل (۱۰.و) آزمون دوم روی شکل نمودار خود بررسی شد، چند خوشه انتخاب؛ سپس بررسی شدند که آیا از نظر معنایی با یکدیگر مرتبط‌اند، پنج خوشه به‌طور تصادفی از طیف رنگ‌های متضاد خروجی نگاشت‌گر («قرمز»، «آبی»،

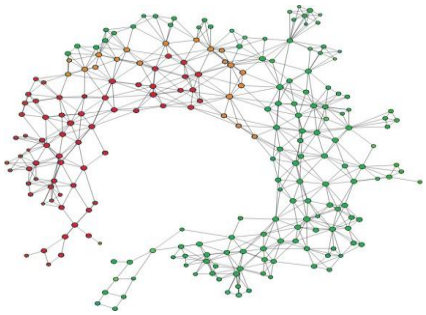
[1] Lens
[2] Cover
[3] overlapping sets

[DOI: 10.66224/jsdp.22.3.77]

«سیاه»، «زرد» و «نارنجی») همان‌طور که در قسمت (و) شکل (۱۰) ظاهر شده‌اند، انتخاب می‌شوند و سپس ابر واژگان رسم شدند: الف) ابر واژهٔ «سیاه» است و به‌دلیل فراوانی برخی واژگان مشخص است که متعلق به اشعار حافظ است و در تحلیل معنایی در مجموعهٔ اشعار عاشقانه خواهد بود؛ ب) ابر واژهٔ «قرمز» است و به‌دلیل فراوانی واژگان در ابر واژه می‌توان آن را از نظر معنایی مربوط به پرستش خدا دانست؛ ج) ابر واژهٔ «زرد»؛ د) ابر واژهٔ برای «نارنجی» و ه) ابر واژهٔ «آبی» است. مقایسهٔ نتایج گراف‌های حاصل از روش نگاشت‌گر، می‌تواند دوباره قابل اطمینان‌بودن این روش در خوشه‌بندی بر اساس معنا و ماهیت واژگان نویسنده و درنهایت دقت خوب این روش در انتساب نویسنده را تأیید کند.



(شکل-۹.الف): گراف اصلی
(Figure-9.a): original graph



(شکل-۹.ب): خوشه‌بندی‌شده برای نخستین آزمون
(Figure-9.b): clustered for first test

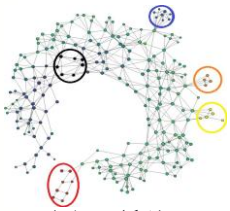
(شکل -۱۰.الف)
(Figure-10.a)

(شکل -۱۰.ب)
(Figure- 10.b)

(شکل -۱۰.ج)
(Figure- ۱۰.c)

(شکل -۱۰.د)
(Figure- ۱۰.d)

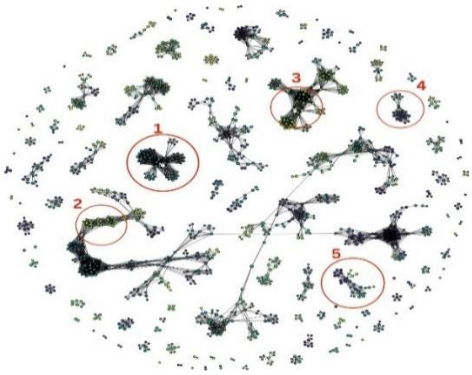
(شکل -۱۰.ه)
(Figure -۱۰.e)



(شکل - ۱۰.و)
(Figure -۱۰.f)

کما اینکه کاربر با دانش نسبی در ادبیات و شعر با مقایسهٔ واژگان خوشه‌بندی‌شده می‌تواند تا حدودی حدس بزند که کدام اشعار مربوط به کدام شاعر است؛ همچنین قابلیت بصری‌سازی این روش در دسته‌بندی اشعار پژوهش‌گران را امیدوار می‌کند که این روش می‌تواند در پردازش متن‌های متنوع ادبی و انتساب نویسنده کارا و موفق باشد.

۲-۴- نتایج روش BERT Embedding



(شکل-۱۱): خوشه‌بندی با روش نگاشت‌گر به همراه BERT Embedding
(Figure–11): clustring by keppler mapper and BERT Embedding

در این بخش خروجی نگاشت‌گر به‌همراه روش BERT Embedding در شکل (۱۱) دیده می‌شود که در ادامه جدولی از نمونه اشعار گراف این شکل نیز آورده می‌شود.

(جدول -۴): جدول شامل نمونه‌ای از اشعار حافظ و فردوسی در گراف
(Table-4): containing poems of Hafez and Ferdowsi

شماره	متن
۱	در حضرت کریم تمنا چه حاجت است اظهار احتیاج خود آن جا چه حاجت است احباب حاضرند به اعدا چه حاجت است می‌دانست وظیفه تقاضا چه حاجت است حسد چه می‌بری ای سست نظم بر حافظ
۲	یکی آتشی بر شده تابناک یکی مرکزی تیره بود و سیاه تویی خویشتن را به بازی مدار بود تا بود هم بدین یک نهاد دل از تیرگی‌ها بدین آب شوی که یزدان به آتش بسوزد تنش نهادش به سر بر یکی تیره ترگ
۳	هشیوار دیوانه خواند ورا همان خویش بیگانه داند ورا همی بر شد آتش فرود آمد آب پذیرنده هوش و رای و خرد

[DOI: 10.66224/jsdp.22.3.77]

شماره	متن
	مگر مردمی خیره خوانی همی همی خواند خواننده بر هر کسی
۴	حافظ چه شد ار عاشق و رند است و نظرباز حافظ گمشده را با غمت ای یار عزیز
۵	پیداست نگارا که بلند است جنابت که خواجه خاتم جم یاوه کرد و بازنجست دست از سر آبی که جهان جمله سراب است

دوباره با مقایسهٔ نتایج به‌دست‌آمده می‌بینیم که ترکیب روش‌های پیش‌پردازش جدید با نگاشت‌گر به میزان زیادی کارایی آن را افزایش می‌دهد. پس با قدرت بصری‌سازی نگاشت‌گر می‌توان نتایج دقیق‌تر را به‌شکل بهتری نیز نمایش داد.

۵- مقایسهٔ نتایج با برخی از بهترین روش‌های سنتی و مدرن یادگیری ماشین برای طبقه‌بندی متون

در ادامه جداولی برای مقایسهٔ دقت نتایج به‌دست‌آمده توسط Keppler mapper به‌همراه پیش‌پردازش Sbert با دو روش یادگیری ماشین برای طبقه‌بندی متون یعنی جنگل تصادفی (سنتی) و XGBoost (مدرن) آورده شده‌است. XGBoost یک چهارچوب بهینه‌سازی‌شده برای الگوریتم‌های درخت تصمیم تقویت شده‌است که در سال ۲۰۱۶ توسط تیان چی و همکاران معرفی شد[۲۷]. این روش بر اساس اصول تقویت گرادیان عمل می‌کند، اما با بهینه‌سازی‌های قابل توجهی در سرعت و عملکرد همراه است.

در XGBoost، مدل‌های ضعیف‌تر (به‌طور معمول درختان تصمیم) به‌صورت متوالی ساخته می‌شوند تا خطاهای مدل‌های پیشین را اصلاح کنند. این رویکرد با استفاده از یک تابع هدف منظم‌سازی‌شده که شامل یک جزء خطا و یک جزء پیچیدگی است، به کاهش بیش‌برازش کمک می‌کند.

ویژگی‌های اصلی XGBoost عبارت است از:

۱. منظم‌سازی[1]: کنترل پیچیدگی مدل برای جلوگیری از بیش‌برازش.

۲. هرس درخت[2]: استفاده از الگوریتم‌های هرس عمق-نخست برای حذف شاخه‌های غیرمفید.

۳. پردازش موازی[3]: استفاده از چندین هستهٔ پردازنده برای سرعت‌بخشیدن به آموزش.

۴. مدیریت داده‌های گمشده: توانایی مدیریت خودکار داده‌های گمشده.

۵. بهینه‌سازی سیستمی: استفادهٔ بهینه از حافظه و منابع محاسباتی.

در زمینهٔ پردازش زبان طبیعی، XGBoost می‌تواند از بردارهای ویژگی استخراج‌شده به‌وسیلهٔ مدل‌های زبانی پیشرفته مانند SBERT بهرهٔ مؤثر ببرد و با استفاده از این اطلاعات معنایی، طبقه‌بندی دقیقی از متون ارائه دهد. در کاربردهای متنی مانند طبقه‌بندی غزلیات حافظ، XGBoost می‌تواند با استفاده از بردارهای معنایی تولیدشده به‌وسیلهٔ SBERT، الگوهای پیچیده در متون را شناسایی کند. سرعت بالای این الگوریتم و مصرف پایین حافظه آن را برای کار با مجموعه‌های متنی بزرگ مناسب می‌سازد. روش XGBoost در مقایسه با روش‌های سنتی‌تر مانند جنگل تصادفی، عملکرد بهتری در دقت و سرعت از خود نشان می‌دهد و به همین دلیل در سال‌های اخیر برای کاربردهای پردازش زبان طبیعی بسیار محبوب شده‌است.

(جدول-۵): مقایسهٔ دقت طبقه‌بندی دستهٔ ۰-غزل‌های کوتاه با دو الگوریتم سنتی طبقه‌بندی متن و روش پژوهش حاضر در آخر

(Table-5): Accuracy comparison of classification of 0-sonnets by two of traditional text mining algorithm and our method at last

روش طبقه‌بندی در دسته؛ ۰-غزل‌های کوتاه	F۱ معیار (وزن‌دار)	دقت داخلی (precision)	فراخوانی (Recall)
جنگل تصادفی (Forest Random)	۰.۸۹۹۷	۰.۹۰۱۴	۰.۹۰۰
الگوریتم XGBoost	۸۰۰۰,۰	۸۰۰۰,۰	۰.۸۰۰۰
SBERT با نگاشت Kepler	۰.۸۹۹۷	۰.۹۰۱۴	۰.۹۰۰

(جدول-۶): مقایسهٔ دقت طبقه‌بندی دستهٔ ۱-غزل‌های متوسط با دو الگوریتم سنتی طبقه‌بندی متن و روش ما در آخر

(Table-6): Accuracy comparison of classification of 0-sonnets by two of traditional text mining algorithm and our method at last

روش طبقه‌بندی در دستهٔ ۱-غزل‌های متوسط	دقت داخلی (precision)	فراخوانی (Recall)	F۱ معیار
جنگل تصادفی (Random Forest)	۰.۸۸	۰.۹۴	۰.۹۱
الگوریتم XGBoost	۰.۸۱	۰.۸۱	۰.۸۱
SBERT با نگاشت Kepler	۰.۸۸	۰.۹۴	۰.۹۱

(جدول-۷): مقایسهٔ دقت طبقه‌بندی همه اشعار با دو الگوریتم سنتی طبقه‌بندی متن و روش ما در آخر

(Table-7): Accuracy comparison of classification of all of poems by two of traditional text mining algorithm and our method at last

روش طبقه‌بندی	دقت داخلی (precision)	فراخوانی (Recall)	F۱ معیار
جنگل تصادفی (Random Forest)	۰.۹۲	۰.۸۶	۰.۸۹
الگوریتم XGBoost	۰.۷۹	۰.۷۹	۰.۷۹
SBERT با نگاشت Kepler	۰.۹۲	۰.۸۶	۰.۸۹

[1] Regularization
[2] Tree Pruning
[3] Parallel Processing

[DOI: 10.66224/jsdp.22.3.77]

۶- نتیجه‌گیری و کارهای آینده

ابعاد تعبیه‌های برت بسیار بالا است و استفادهٔ مستقیم از آن‌ها منجر به پیچیدگی محاسباتی می‌شود. عدد پنج به‌صورت تجربی انتخاب شد تا ضمن حفظ ساختار معنایی داده‌ها، الگوریتم نگاشت‌گر عملکرد بهینه داشته باشد. در کارهای آینده، تحلیل حساسیت با ابعاد مختلف برای انتخاب بهینه‌تر انجام خواهد شد؛ همچنین بر داده‌های متنی دیگر نیز الگوریتم می‌تواند پیاده‌سازی شود.

۷-مراجع 7-References

[۱] الیاسی نیره، تیموری حسین، پاک‌نیت سروش، *مقدمه‌ای بر تحلیل توپولوژیکی داده: نظریه و رویکرد*، انتشارات دانشگاه تفرش، ۱۴۰۰.

[1] Eliasi, N., Teimouri, H., Pakniat, S., *An Introduction to Topological Data Analysis: Theory and Approach*, Tafresh University Press, 1400.

[2] Holmes, D. "Authorship Attribution", *Computers and the Humanities*, 28:87-106. Kluwer Academic Publishers, 1995

[3] Malyutov, M.B. "Authorship Attribution of Texts: a Review", *Proceedings of the program "Information transfer"* held in ZIF. University of Bielefeld, Germany, 17 pages, 2004.

[4] Chaski, C. "Who's at the Keyword? Authorship Attribution in Digital Evidence Investigations", *International Journal of Digital Evidence*, Volume 4, Issue 1 , 2005.

[5] Diederich, J., Kindermann, J., Leopold, E., Paas, G. "Authorship Attribution with Support Vector Machines", *Applied Intelligence*, 19(1): pp.109-123, 2003.

[6] Stamatatos, E., Fakotakis, N., Kokkinakis, G. "Computer-Based Authorship Attribution Without Lexical Measures", *Computers and the Humanities*, 35: pp. 193-214, Kluwer Academic Publishers, 2001.

[7] Carlsson, G, *Topology and data*, Bull, Amer, Math, Soc, 46, pp. 255-308, 2009.

[8] Emrani, S., Saponas, T., Morris, D., and Krim, H. A. "Novel Framework for Pulse Pressure Wave Analysis Using Persistent Homology", *IEEE Signal Processing Letters*, Vol. 22, No. 11, 2015.

[9] Günther, D., Reininghaus, J., Hotz, I., and Wagner, H. "Memory Efficient Computation of Persistent Homology for 3D Images using Discrete Morse Theory", 24th SIBGRAPI Conference on Graphics, *Patterns and Images*, 28-31, 2011.

[10] Lum, P. Y., Singh. G, Lehman, A., Ishkanov. T., Vejdemo-Johansson, M., Alagappan, M., Carlsson, J., Carlsson, G. "Extracting insights from the shape of complex data using topology", *Sci*. Rep, 3, 1236 (2013).

[11] Lum, P. Y., Lehmann, L., Singh, G., Ishkhanov, T., Vejdemo- Johansson, M., Carlsson , G., "The topology of politics: voting connectivity in the US House of Representatives", KTH, School of Computer Science and Communication (CSC), *Computer Vision and Active Perception*, CVAP, 2012.

[12] Lee, H., Ghung, M. K., Kang, H., Kim, B., Lee, D. "Discriminative persistent homology of Brain networks", *IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pp. 841-844, 2011.

[13] Edelsbrunner, H. "Persistent Homology In Image processing", *Lecture Notes in Computer Science book series*, LNCS, vol. 7877, pp. 182-183, 2013.

[14] Bhattacharya, S., Ghrist, R., and Kumar, V. "Persistent Homology for Path Planning in Uncertain Environments", *IEEE TRANSAC- TIONS ON ROBOTICS*, VOL. 31, NO. 3, pp. 1-13, JUNE 2015.

[15] Nicolau, M., Levine, A. J., Carlsson, G. "Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival", *Proc Natl Acad Sci U S A*, 108 (17), pp.7265-70(2011).

[16] Zhu, X. "Persistent Homology: An Introduction and a New Text Representation for Natural Language Processing", *Twenty-Third International Joint Conference on Artificial Intelligence*, 2013.

[17] Elyasi, N., Hosseini Moghadam, M. "An introduction to a new text classification and visualization for natural language processing using topological data analysis", *arXiv preprint arXiv*:1906.01726, (2019), https://arxiv.org/pdf/1906.01726 .

[18] Gholizadeh, S., Seyeditabari, A., and Zadrozny, W. "Topological Signature of 19th Century Novelists: Persistent Homology in Text Mining", *big data and cognitive computing*, 2(4), 2018, doi:10.3390/bdcc2040033.

[19] NILSSON, D., EKGREN, A., "Topology and Word Spaces", *Bachelors Thesis at CSC*.

[20] Torres P., Hromic H., Heravi B. "Topic Detection in Twitter Using Topology Data Analysis", *ICWE 2015 Workshops*, LNCS 9396, pp. 186-197, 2015.

[21] Romano, D., Nicolau, M., Quintin, E. M., Mazaika, P. K., Light- body, A. A., Hazlett, H. C., Piven, J., Carlsson, G., Reiss, A. L. "Topological methods reveal high and low functioning neuro- phenotypes within fragile X syndrome", *Human Brain Mapping*, 35, 9, pp. 4904-4915, 2014.

[22] Rizvi, A., Camara, P., Kandror, E., Roberts, T., Schieren, I. et al. "Single-cell topological RNA-seq analysis reveals insights into cellular differentiation and development", *Nature Biotechnology*, 35(6), pp. 551-560, 2017.

[23] Mihaela, E., Sardiu, M. E., Joshua, M. Gilmore., Groppe, B., Florens, L., Michael, P. Washburn, "Identification of Topological Network Modules in Perturbed Protein Interaction Networks", *Scientific Reports*, 7(1), pp. 1-13, 2017.

[24] Jessica, L., Nielson, Jesse Paquette, Aiwen, W. Liu, Cristian F. Guandique, C., Amy. Tovar et al. "Topological data analysis for discovery in preclinical spinal

[DOI: 10.66224/jsdp.22.3.77]

cord injury and traumatic brain injury", *Nature Communications*, Oct 14, 2015.
[25] Saggar, M., Sporns, O., Gonzalez-Castillo, J., Bandettini, P., Carlsson, G. et al. "Towards a new approach to reveal dynamical organization of the brain using topological data analysis", *Nature Communications*, 9 (1), Apr 11, 2018.
[26] Bubenik, P., Dotko, P. "A persistence landscapes toolbox for topological statistics", *J. Symbolic Computation*, 78, pp. 91-114 (2016).
[27] Singh, G., Mémoli, F., Carlsson, G. "Topological methods for the analysis of high dimensional data sets and 3d object recognition", *In Eurographics Symposium on Point-Based Graphics* (eds Botsch, M., Pajarola, R.) (The Eurographics Association), 2007.
[28] Chen, T., Guestrin, C., "Xgboost: A scalable tree boosting system", *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, San Francisco, CA, pp. 785-794, USA 2016.

نیره الیاسی استادیار دانشکدهٔ علوم ریاضی و کامپیوتر دانشگاه خوارزمی است. ایشان مدرک دکتری تخصصی ریاضی خود را در رشتهٔ دکتری پیوسته ریاضی از دانشگاه صنعتی امیرکبیر در سال ۱۳۸۹ دریافت کرد. زمینه‌های پژوهشی ایشان مدل‌سازی ریاضی نظریه‌های فیزیک کلاسیک و نسبیت، تحلیل توپولوژیکی داده، ابزارهای هندسی و توپولوژیکی یادگیری ماشین و یادگیری ژرف است.
نشانی رایانامهٔ ایشان عبارت است از:
elyasi82@khu.ac.ir

مهدی حسینی مقدم کارشناسی خود را در دانشکدهٔ علوم ریاضی و کامپیوتر دانشگاه خوارزمی در سال ۱۳۹۷به پایان رساند و در همان سال در رشتهٔ کارشناسی‌ارشد علوم داده در دانشکدهٔ مهندسی کامپیوتر دانشگاه خوارزمی پذیرفته شد. هم‌اکنون در کشور انگلستان به‌عنوان متخصص علوم داده مشغول به‌کار است. زمینه‌های کاری ایشان هوش مصنوعی، پردازش زبان طبیعی، یادگیری ژرف و تحلیل توپولوژیکی داده‌است.
نشانی رایانامهٔ ایشان عبارت است از:
m.h.moghadam1996@gmail.com

[DOI: 10.66224/jsdp.22.3.77]