https://doi.org/10.22131/sepehr.2023.552846.2867

# استخراج عملکرد محل از محتواهای متنی کاربرتولید با استفاده از روش‌های یادگیری ماشین

مینا کریمی [۱]  محمدسعدی مسگری [۲]

تاریخ دریافت مقاله: ۱۴۰۱/۰۳/۰۳  تاریخ پذیرش مقاله: ۱۴۰۱/۱۰/۰۳

*********

**چکیده**

**امروزه با افزایش روزافزون استفاده کاربران از شبکه‌های اجتماعی، اطلاعات مکانی مردم‌گستر رشد چشمگیری داشته است. از میان انواع اطلاعات، محتواهای متنی کاربرتولید غالباً در ساختار مشخصی به اشتراک گذاشته نمی‌شوند. یکی از ویژگی‌های عمده این نوع اطلاعات محل‌مبنا بودن آن‌ها است. محل‌های مورد گفتگوی بشر معمولاً همراه با ابهام و وابسته به بافت است. عملکرد محل یا به عبارتی عمده فعالیت‌هایی که افراد در یک محل انجام می‌دهند، به عنوان یک بافت در توصیفات محل، ازجمله ویژگی‌های عمده و متمایزکننده محل است. هدف این تحقیق استخراج عملکرد محل با استفاده از تحلیل محتواهای متنی کاربرتولید به اشتراک گذاشته شده توسط کاربران است. به این منظور ابتدا محل‌ها و نظرات کاربران در مورد محل‌ها در وبگاه TripAdvisor به عنوان محتواهای متنی، جمع‌آوری شده، سپس از روش‌های مختلف پردازش زبان طبیعی به منظور آماده‌سازی و پیش‌پردازش داده‌ها استفاده می‌شود. در ادامه برای هر دیدگاه کاربر یک مجموعه واژگان با استفاده از مقادیر TF-IDF به عنوان مقادیر بردار ویژگی ساخته می‌شود. سپس در یک روش نظارت‌شده این مقادیر به همراه عملکرد محل‌ها به عنوان ورودی به یک طبقه‌بندی‌کننده لجستیک رگرسیون به منظور آموزش مدل داده شده و با استفاده از آن عملکرد محل بر روی داده‌های آزمایشی پیش‌بینی شده است. نتایج ارزیابی روش از طریق محاسبه ماتریس درهم ریختگی نشان می‌دهد، صحت کلی روش پیشنهادی در حدود ۹۶ درصد است که رقم قابل توجهی است. همچنین بیشترین دقت و امتیاز F1 برای محل‌های سرو خوراکی است، در حالی‌که اقامتگاه‌ها به دلیل شباهت عملکردی به هتل‌ها کمترین دقت و امتیاز F1 را دارند ولی با این وجود نتایج آن‌ها نیز قابل اطمینان و رضایت‌بخش است.**

**واژه‌های کلیدی: محل، عملکرد محل، محتواهای کاربرتولید، پردازش زبان طبیعی، یادگیری ماشین، متن**

*********

۱- دانشجوی دکتری، گروه مهندسی سیستم اطلاعات مکانی، دانشکده مهندسی نقشه‌برداری، دانشگاه صنعتی خواجه‌نصیرالدین طوسی، تهران، ایران (نویسنده مسئول)
minakarimi@email.kntu.ac.ir

۲- استادیار، گروه مهندسی سیستم اطلاعات مکانی، دانشکده مهندسی نقشه‌برداری - دانشگاه صنعتی خواجه‌نصیرالدین طوسی، تهران، ایران
mesgari@kntu.ac.ir

## ۱- مقدمه

در علوم اطلاعات مکانی همواره سعی شده است اطلاعات مکانی با دقت هرچه بیشتر و در قالب فضا[1] و مختصات ارائه شوند. در حالی‌که استدلال، رفتار و درک انسان‌ها براساس محل[2] است نه فضا ***(Goodchild, 2015)***. فضا مفهومی انتزاعی است که می‌تواند در یک محیط مبتنی بر رایانه صوری‌سازی شود، در حالی‌که محل با تجربیات انسانی جهان مرتبط است ***(Couclelis, 1992)***. محل‌های مورد گفتگوی بشر معمولاً همراه با ابهام و وابسته به بافت[3] است. امروزه به‌کارگیری جای نام[4] به‌جای مختصات متداول شده است و سرویس‌های مختلفی از قبیل سرویس‌های مسیریابی نظیر مپکوئست[5]، مداخل مکانی[6] ***(Hill, 2000)*** و موتورهای جستجوی فضایی معنایی ***(SPRIT)***[7] ***(Purves et al., 2007)*** به این منظور طراحی و اجرا شده است.

در دهه‌های اخیر سرعت شهرنشینی افزایش یافته که به شهرها امکان می‌دهد طیف گسترده‌ای از عملکردها[8] و فعالیت‌های انسانی را دربرداشته باشند. این امور به کاوش در سطوح زمین و فضای شهری محدود نمی‌شوند و شامل مفاهیم انسان‌گرا مانند مناطق[9] ***(Hartshorne, 1969)*** و محل‌ها ***(Tuan, 1979)*** نیز می‌شوند. با توسعه شبکه‌های اجتماعی در سال‌های اخیر، اطلاعات مکانی مردم‌گستر به طرز چشمگیری درحال رشد است.

یکی از ویژگی‌های عمده این نوع اطلاعات محل‌مبنا بودن آن‌ها است ***(Purves, Winter, & Kuhn, 2019)***. با این وجود اطلاعات به‌دست آمده از شبکه‌های اجتماعی اغلب دید کامل و روشنی نسبت به مفهوم مکان و اطلاعات مکانی ندارند و اطلاعات مکانی بین پدیده‌ها، کاربری‌ها، نقاط موردعلاقه (POI)[10]، سیستم حمل و نقل و غیره را که در شکل‌گیری و عملکرد محل[11] مهم و دخیل هستند، لحاظ نمی‌کنند. این امر در نهایت توان و قدرت آن‌ها را در کار با مفهوم محل محدود می‌کند. از سوی دیگر در سیستم اطلاعات مکانی (GIS)[12] مفهوم محل به خوبی دیده نشده است. در چنین شرایطی برای تحلیل داده‌های محل‌مبنا نیاز به تعریف محل در سیستم‌های اطلاعات مکانی احساس می‌شود. این تعریف شامل دو دسته کلی فضا و معنا است. فضا به مختصات جغرافیایی محل اشاره دارد، در حالی‌که معنا به درک و مفهوم ارائه شده توسط یک محل می‌پردازد ***(Purves, Winter, & Kuhn, 2019; Papadakis, Gao, & Baryannis, 2019)***.

در این حالت، GIS باید بتواند با استفاده از دانش و داده در مورد فعالیت و تجربه انسان، عملکرد و فضا را با هم مرتبط کند. به بیان ساده‌تر GIS باید بتواند عملکرد محل‌ها را کشف کند ***(Papadakis, Gao, & Baryannis, 2019)*** که الزاماً به سادگی و وضوح در داده‌های ذخیره شده وجود ندارد. هدف اصلی این مقاله استخراج عملکرد محل از محتواهای متنی کاربر تولید نظیر نظرات کاربران در وبگاه‌های سفراست. برای دستیابی به این هدف از روش‌های مختلف تحلیل متن و پردازش زبان طبیعی (NLP)[13] به منظور طبقه‌بندی نظرات کاربران درمورد محل استفاده می‌شود.

امروزه به‌کارگیری روش‌های یادگیری ماشین به منظور طبقه‌بندی متون رشد چشمگیری داشته است. از آنجایی‌که عملکرد هر محل در این تحقیق مشخص است، در یک روش نظارت شده با استفاده از الگوریتم لجستیک رگرسیون، عملکرد محل‌ها پیش‌بینی می‌شود. درنهایت نتایج طبقه‌بندی ارزیابی می‌شود.

این تحقیق به دنبال استخراج عملکرد محل است که با کاربری متفاوت است. برای نمونه امروزه افراد، خانه‌ها و ویلاهای خود را به عنوان اقامتگاه در اختیار مسافران

1- Space
2- Place
3- Context
4- Placename
5- MapQuest
6- Gazetteers
7- SpatiallyAwareInformationRetrivalon Internet
8- Functionality
9- Regions
10- Point of Interest
11- Place Functionality
12- Geospatial Information System
13- Natural Language Processing

و گردشگران قرار می‌دهند. بنابراین، این محل‌ها عملکرد اقامتی دارند. در حالی‌که کاربری آن‌ها به عنوان مسکونی ثبت شده است. همچنین کاربری پاساژهای خرید تجاری است، در حالی‌که می‌توانند عملکردهای متفاوتی داشته باشند. برای نمونه مجتمع کوروش عملکرد خرید دارد، با این وجود به دلیل وجود یک سینمای بزرگ، می‌تواند عملکرد تفریحی و سرگرمی داشته باشد. یا حتی گاهی پاساژگردی و نه صرفاً خرید به عنوان عملکرد تفریحی محسوب می‌شود. این مجتمع همچنین به دلیل داشتن کافی‌شاپ و رستوران عملکرد خوردنی دارد و گاهی افراد قرار ملاقات‌ها و دیدارهایشان را در این مجتمع می‌گذارند.

برای ورود به مروری بر تحقیقات پیشین ذکر این نکته ضروری است که؛ مکالمه در مورد مکان‌ها با یک کامپیوتر طیف وسیعی از چالش‌ها را برای هوش مصنوعی فعلی ایجاد می‌کند. محتوای ارتباطات به شدت به دیدگاه‌ها و ترجیحات شخصی افراد درگیر در گفتگو بستگی دارد و ممکن است حاوی سطوح بالایی از ابهام باشد که تنها از طریق اطلاعات متقابل و بافت محلی قابل حل است ***(Winter et al., 2021)***.

رلف (۱۹۷۶) محل را به عنوان الگویی منحصربه‌فرد از عوارض فیزیکی، ظواهر، فعالیت‌ها و عملکردها محدود می‌کند. کیفیت منحصربه‌فرد آن شامل قدرت تمرکز روی اهداف، تجربیات و اقدامات انسان در بعد فضایی است. همچنین بیان می‌کند محل براساس تجربه محیطی که به نوبه خود ناشی از بافت فضایی[1] محل است، ساخته می‌شود و از این طریق به ارتباط نزدیک محل و مکان می‌پردازد ***(Relph, 1976)***.

وینتر و همکاران (۲۰۲۱) چالش‌هایی که برای گفتگو درباره محل‌ها وجود دارد، از جمله ماهیت مفهومی گسترده‌تر محل‌ها در مقیاس‌های جغرافیایی، انعطاف‌پذیری، ابهام و ابهام زبان، وابستگی به بافت تأثیرگذار بر زمینه‌سازی و ابهام‌زدایی زبان، و تفسیر روابط بخشی از و سایر روابط فضایی را بررسی کرده و کارهای نوظهور را خلاصه می‌کنند تا به بررسی روش‌هایی که رایانه‌ها را قادر می‌سازند تا به شریک گفتگوی انسان‌ها در مکان‌ها و مکان‌ها تبدیل شوند، بپردازند ***(Winter et al., 2021)***.

مُکنیک (۲۰۲۲) به مروری روی بحث‌های فعلی و جهت‌های آتی در مورد چگونگی گسترش پارادایم فضایی فرض شده فعلی به سمت اطلاعات محلی می‌پردازد. به این ترتیب، رویکردهای ممکن و چشم‌اندازهای آتی و همچنین محدودیت‌های نظریه‌های اطلاعات فضایی و سیستم‌های اطلاعات فضایی بررسی شده است ***(Mocnik, 2022)***. همچنین در سمپوزیم اخیر علوم اطلاعات مبتنی بر محل، تحقیقات اخیر در این حوزه بررسی شده است ***(Mocnik&Westerholt, 2022)***.

استفاده از نظریه قابلیت محیط[2] ***(Gibson, 1977)***یکی دیگر از روش‌هایی است که برای مدل‌سازی محل در علوم اطلاعات مکانی به کار گرفته شده است ***(Jordan et al., 1998)***. قابلیت محیط به امکاناتی اشاره دارد که اشیاء برای افراد فراهم می‌کنند. برای مثال، شیء «پله» قابلیت «بالا رفتن» را برای افراد فراهم می‌آورد. به این ترتیب، می‌توان محل را به‌عنوان یکی از اشیایی در نظر گرفت که قابلیت‌هایی برای انسان فراهم می‌کنند.

پاپاداکیس و همکاران (۲۰۱۶)یک مدل هستی‌شناسی از مفهوم محل براساس عملکرد[3] آن‌ها ارائه کرده‌اند ***(Papadakis, Resch, & Blaschke, 2016)***. این مدل الگوهای مکانی محل‌ها را استخراج می‌کند و فضا را با عملکرد ترکیب می‌کند. شایان ذکر است که عملکرد تنها زیرمجموعه‌ای از بافت‌های فضایی را بیان می‌کند. محل‌ها اگر با سایر بافت‌های فضایی نظیر احساسات، تجربیات، اهداف حضور و رویدادها در محل همراه باشند، می‌توانند پیچیده‌تر شوند ***(Papadakis & Blaschke, 2017)***. تفاوت‌های عملکردی محل‌های جغرافیایی یا قابلیت محل به‌عنوان بعد اساسی برای تعیین مشخصات محل‌های جغرافیایی درنظر گرفته می‌شود ***(Alazzawi, Abdelmoty, & Jones, 2012)***.

1- Spatial Context

2- AffordanceTheory

3- Function

بخش قابل‌توجهی از تحقیقات پیشین، به داده‌های حرکتی افراد مانند جریان‌های رفت‌وآمد یا خطوط سیر تاکسی‌ها، داده‌های کارت‌های هوشمند اتوبوس و مترو، سوابق اجاره دوچرخه‌های عمومی به منظور شناسایی مناطق عملکردی وابسته است *(Fan et al., 2015; T. Zhou et al., 2020; Tao et al., 2019)*. یکی از نقاط ضعف موارد یاد شده این است که به یک نوع اطلاعات متکی هستند. این امر آن‌ها را به نمایشی تک بعدی محدود می‌کند که ممکن است برای توصیف ماهیت چندجانبه مفهومی نظیر منطقه عملکردی کافی نباشد. همچنین، نتایج ممکن است تحت‌تأثیر در دسترس بودن، کیفیت و پوشش داده‌ها قرار گیرد زیرا به منابع دادهای وابسته است که ممکن است به صورت گسترده در دسترس نباشند (مثل خط سیر تاکسی).

همانطور که توسط سو و همکاران(۲۰۱۷)اشاره شده است، این مسئله در داده‌های مکانی داوطلبانه (VGI)[1] بیشتر مشهود است *(Su et al., 2017)*.

برای مقابله با این مسائل، تحقیقات به‌طور فزاینده‌ای بر روی ترکیب انواع داده‌های مختلف متمرکز می‌شوند. ترکیب داده‌های سرویس‌های مبتنی بر موقعیت (LBS)[2] با داده‌های نقاط موردعلاقه به دلیل موقعیت‌یابی با دقت بالا، ویژگی تعامل‌پذیری و حجم زیاد داده‌ها، تأثیر به‌سزایی در شناسایی الگوهای مکانی دارد *(H. Han, Yu, & Long, 2015)*.

در این راستا نولاس و همکاران(۲۰۱۱) از فراوانی اعلام حضورها[3] در هر دسته محل در شبکه اجتماعی فوراسکوئر[4] استفاده کردند تا مشخص کنند کدام دسته‌ها در مناطق خاصی رایج‌تر هستند *(Noulas et al., 2011)*. به‌طور مشابه ژو و ژانگ(۲۰۱۶)داده‌های توئیترو فوراسکوئر را برای استخراج توزیع فضایی فعالیت‌های افراد (مانند غذا و رستوران‌ها، فروشگاه‌ها و خدمات، فضای باز و تفریحی) و برای تجزیه و تحلیل مکانی–زمانی در تعیین شدت نقاط داغ[5] به کار بردند. همچنین از مدل طبقه‌بندی شده در یک سیستم بلادرنگ برای استخراج عملکردهای مختلف شهری استفاده شده است *(X. Zhou & Zhang, 2016)*. در هیچ یک از این تحقیقات از داده‌های متنی به منظور استخراج عملکرد محل استفاده نشده است.

در تحقیق وسردنی و همکاران (۲۰۱۶)از ۱۴ دانشجو خواسته شد سه محل را توصیف کنند *(Vasardani, Tomko, & Winter, 2016)*. سپس براساس ۱۵ ویژگی ساختاری الکساندر که یک کل را مشخص می‌کند *(Alexander, 2002)*، توصیف از محل را دسته‌بندی کنند. از بین ویژگی‌هایی که در مجموعه الکساندر نیامده ولی غالباً در توصیفات به آن اشاره شده است، قابلیت محل است. بنابراین خصوصیات عملکردی می‌تواند خصوصیات ساختاری را در صوری‌سازی محل تکمیل کند *(Vasardani, Tomko, & Winter, 2016)*.

از دیدگاه خوری و همکاران (۲۰۰۶) بیشتر اطلاعات معنایی یک جمله در فعالیت توصیف‌شده توسط آن قرار دارد. بنابراین در روش آن‌ها آن واژگانی از متن که ارتباط‌شان یک فعالیت را نشان می‌دهد، استخراج می‌شود *(Khoury, Karray, & Kamel, 2006)*. تفاوت‌های عملکردی محل‌های جغرافیایی به لحاظ فعالیت‌هایی که فرد ممکن است در یک محل انجام دهد یا قابلیت محل به عنوان بعد اساسی برای تعیین مشخصات محل‌های جغرافیایی درنظر گرفته می‌شود *(Alazzawi, Abdelmoty, & Jones, 2012)*.

امروزه با گسترش استفاده کاربران از شبکه‌های اجتماعی محتواهای کاربرتولید رشد چشمگیر و قابل ملاحظه‌ای داشته‌اند. از میان انواع محتواهای کاربرتولید، محتواهای متنی غالباً ساختارنیافته هستند و بسیار به بافت و کاربرد بستگی دارند. تحقیقات مختلفی از محتواهای متنی کاربرتولید نظیر توصیفات محل و نظرات کاربران به منظور استخراج اطلاعات محل استفاده کرده‌اند.

العزاوی و همکاران (۲۰۱۲) با استفاده از تحلیل متن و استخراج افعال حرکتی ترکیب شده با اسامی در مجموعه

1- Volunteered Geographic Information
2- Location-Based Service
3- Check-ins
4- Forsquare
5- Hotspots

داده دیبی‌پدیا[1]، نوع فعالیت‌ها و سرویس‌های ممکن همراه با نوع محل‌ها را استخراج نمودند (*Alazzawi, Abdelmoty, & Jones, 2012*).

ادامز و یانوویچ(۲۰۱۵) متن غیرساختار یافته از مقالات ویکی‌پدیا و دیبی‌پدیا را به‌منظور استخراج امضاهای[2] موضوعی با استفاده از مدل LDA که می‌توانند نوع محل همراه شده با هر مقاله را توصیف کنند، به کار بردند (*Adams & Janowicz, 2015*).

هوبل و همکاران(۲۰۱۶) از NLP روی نظرات کاربران در TripAdvisor[3] برای مرکز تاریخی وین، به منظور یافتن نام‌های مرکب مرتبط به عوارض جغرافیایی، استفاده کردند. سپس این نام‌ها و برچسب‌های OSM[4] برای آموزش یک طبقه‌بندی‌کننده بیزبرای کشف نواحی شناختی استفاده شدند (*Hobel, Fogliaroni, & Frank, 2016*).

ادامز و مکنزی(۲۰۱۳) با به‌کارگیری LDA بر روی نوشته‌های وبگاه‌های سفر از سایت travelblog[5]، مضامین مربوط به محل‌های سراسر دنیا را شناسایی و با استفاده از داده‌های VGI شباهت بین محل‌ها را محاسبه کردند (*Adams & McKenzie, 2013*).

ذکر این نکته ضروری است که ویژگی مهم این تحقیق تمرکز بر روی تحلیل محتواهای متنی است. اگرچه محتواهای متنی در تحقیقات زیاد استفاده شده است، ولی در هیچ یک از این تحقیقات، عملکرد محل از طریق تحلیل محتواهای متنی کاربرتولید استخراج نشده است. همچنین در تحقیقات پیشین غالباً عملکرد محل از داده‌های حرکتی و یا POI استخراج شده است، در حالی‌که در این تحقیق به دنبال استخراج عملکرد محل از نظرات کاربران در مورد محل هستیم.همچنین در این تحقیق از الگوریتم‌های یادگیری ماشین نظیر لجستیک رگرسیون در استخراج عملکرد محل از محتواهای متنی استفاده شده است.

1- DBPedia
2- Signatures
3- https://www.tripadvisor.com
4- Open Street Map
5- https://www.travelblog.org

## ۲- مواد و روش تحقیق

در این بخش ابتدا به معرفی داده‌های تحقیق و منطقه مورد مطالعه پرداخته شده و سپس روش پیشنهادی تحقیق به منظور استخراج عملکرد محل از نظرات کاربران ارائه شده است.

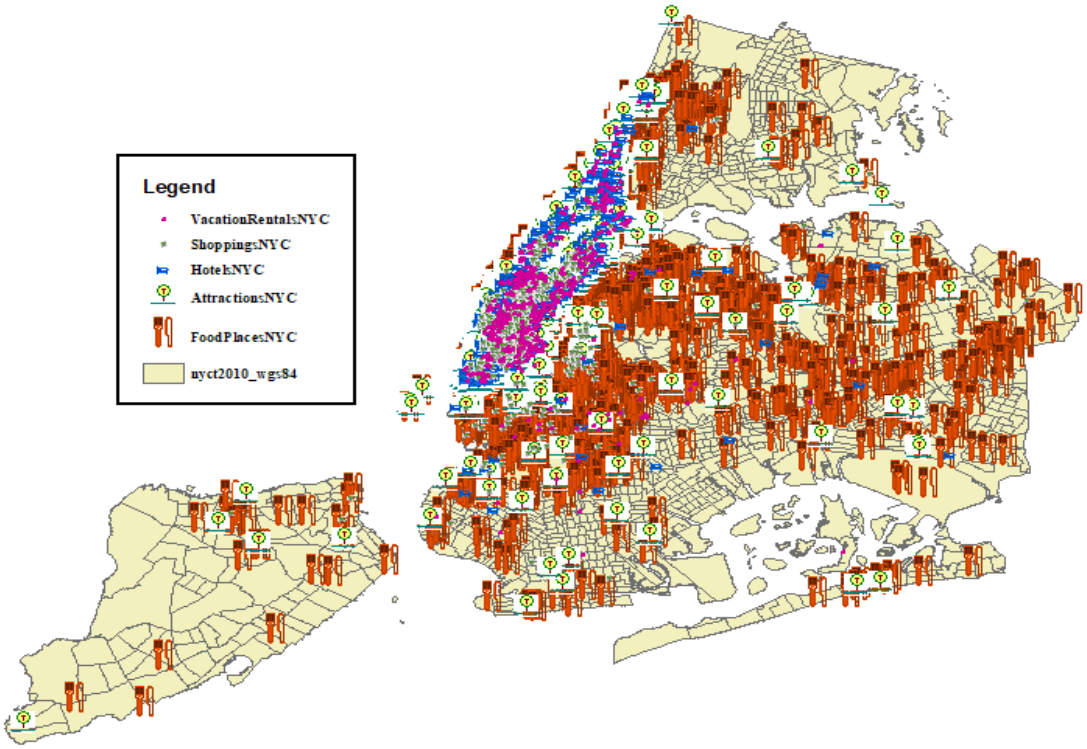
### ۲-۱- داده‌ها

اطلاعات مرتبط با محل‌ها در تحقیق حاضر از وبگاه TripAdvisor از طریق خزیدن در وب[6] با استفاده از زبان برنامه‌نویسی پایتون و از طریق Ajax query و کتابخانه‌های BeautifulSoup و Selenium استخراج شده است. ابتدا برای هر نوع محل، تمامی محل‌های مرتبط با آن دسته شامل شناسه، نام، نوع، زیر نوع، امتیاز و طول و عرض جغرافیایی محل استخراج شده است و در یک فایل csv ذخیره شده است. سپس برای هر محل حداکثر ۱۰۰۰ دیدگاه آخر از نظرات کاربران به زبان انگلیسی با استفاده از خزیدن مجدد در وب استخراج شده است. همراه با متن هر دیدگاه، شناسه، عنوان و امتیاز دیدگاه نیز ذخیره شده است. در ادامه تمامی این فایل‌های csv ادغام شده و نهایتاً دو فایل یکی برای محل‌ها و یکی برای نظرات کاربران درمورد محل‌ها ایجاد شده است. این دو فایل با استفاده از شناسه محل در فایل محل‌ها و فایل نظرات به ترتیب به عنوان کلید اصلی و کلید فرعی به یکدیگر مرتبط می‌شوند. این داده‌ها در اکتبر ۲۰۲۰ در دسترس بوده‌اند.

### ۲-۲- منطقه مورد مطالعه

شهر نیویورک ایالات متحده آمریکا به عنوان منطقه مورد مطالعه انتخاب شده است. نیویورک به عنوان پرجمعیت‌ترین و پرتراکم‌ترین، و وسیع‌ترین شهر ایالات متحده آمریکا و وسیع‌ترین منطقه شهری در جهان است (*DecennialCensus P.L, 2020*) این شهر از پنج ناحیه شامل برانکس، بروکلین، منهتن، کوئینز و استتن آیلند تشکیل شده‌است و یکی از

6- Web Crawling




**نگاره۱: منطقه مورد مطالعه، شهر نیویورک**

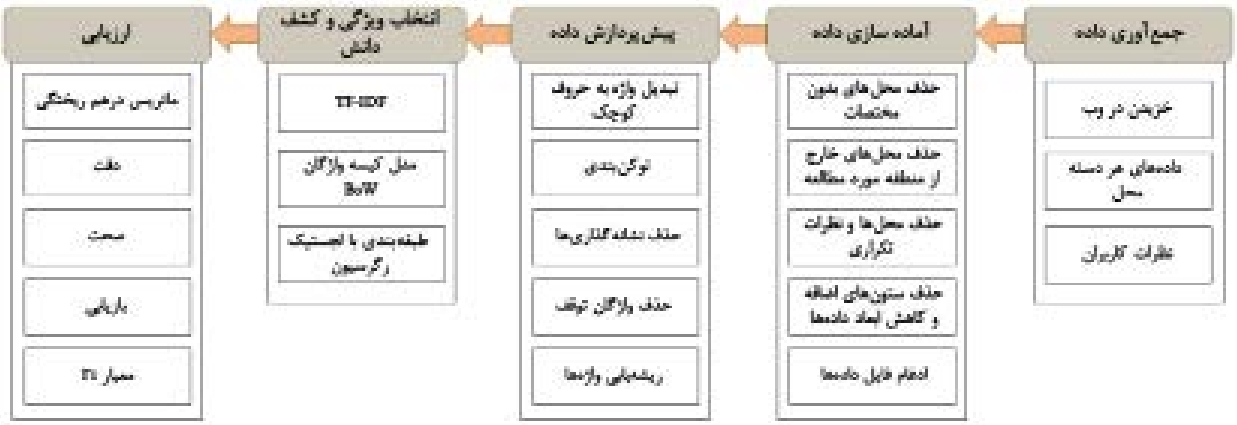
قطب‌های فرهنگی، سیاسی و اقتصادی جهان است. عده‌ای به‌دلیل اهمیت سیاسی، اقتصادی و فرهنگی منحصربه‌فرد نیویورک، این شهر را "پایتخت جهان" نامیده‌اند (*Top 8. Citiesby GDP, 2018*). در نگاره (۱) منطقه مورد مطالعه و محل‌های جمع‌آوری‌شده در این منطقه نشان داده شده است.

## ۲-۳- روش پیشنهادی

در NLP تلاش می‌شود همان‌گونه که مفاهیم زبان طبیعی توسط انسان تجزیه‌وتحلیل می‌شود، برای کامپیوتر نیز قابل فهم باشد. درمورد متون غیرساختاریافته یا نیمه‌ساخت‌یافته؛ ابتدا باید با روش‌هایی، آن‌ها را ساختارمند کرد و سپس از این روش‌ها برای استخراج اطلاعات و دانش استفاده کرد. برای دستیابی به هدف تحقیق چارچوب کلی روش پیشنهادی به‌طور خلاصه در نگاره(۲) ارائه شده است.

در اولین مرحله از این تحقیق، اطلاعات استخراج شده در دو فایل مجزا (یک فایل مرتبط با محل‌ها و یک فایل مرتبط با نظرات کاربران) با فرمت csv ذخیره شده است. با درنظر گرفتن هدف اصلی این تحقیق، مزیت این داده‌ها نسبت به داده‌های محل‌مبنای دیگر، عدم وابستگی آن‌ها به توصیفات رسمی از محل و دسترسی گسترده به آن‌ها است. نگاره (۳) ابرواژگان استخراج شده از تمامی نظرات جمع‌آوری شده کاربران را نشان می‌دهد.

با وجود دقت مکانی بالای اطلاعات، به دلیل ماهیت کاربرمحور بودن آن‌ها ممکن است در اطلاعات توصیفی یا مکانی محل عدم صحت وجود داشته و دارای خطا و اشتباه در ثبت موقعیت و یا نوع دسته‌بندی محل باشند. همچنین اشتباهات تایپی جزء اجتناب‌ناپذیر این نوع از داده‌ها است. بنابراین در مرحله بعد، داده‌های استخراج‌شده نیازمند آماده‌سازی هستند. به این منظور ابتدا محل‌های بدون مختصات از مجموعه داده حذف شده و سپس با استفاده از شیپ فایل نیویورک، محل‌های خارج از منطقه مورد مطالعه حذف شده‌اند. همچنین با حذف فیلدهای




**نگاره۲: چارچوب پیشنهادی تحقیق**


**نگاره۳: ابرواژگان ایجادشده از نظرات کاربران درمورد محل‌ها**

**جدول۱: تعداد محل‌ها و نظرات کاربران در هر دسته محل**

| نوع محل | محل‌های دیدنی | محل‌های سرو خوراکی | هتل‌ها | مراکز خرید | اقامتگاه‌ها | مجموع |
|---|---|---|---|---|---|---|
| تعداد محل‌ها | ۱۲۰۳ | ۱۳۲۸۲ | ۸۴۲ | ۱۰۴۹ | ۱۰۸۰ | ۱۷٤٥٦ |
| تعداد نظرات | ۹٥٦٦۱ | ۳۲٥۷۷٤ | ۱٥۷٦۲٥ | ۳۱٤۰٤ | ۳۳۳٤ | ٦۱۳۷۹۸ |

نامرتبط، ابعاد و حجم داده‌ها کاهش و فقط داده‌های مرتبط بررسی و تحلیل شده‌اند. در ادامه موارد تکراری و یا محل‌هایی که نوع آن‌ها نامشخص است، حذف شده‌اند. در جدول (۱) تعداد محل‌ها و نظرات کاربران به تفکیک هر دسته محل پس از آماده‌سازی ارائه شده است. آماده‌سازی و پیش‌پردازش داده‌ها مهم‌ترین و زمان‌برترین بخش فرآیند داده کاوی است. حدود ۸۰ درصد فرآیند صرف آماده‌سازی و پیش‌پردازش داده‌ها شده *(J. Han &Kamber, 2001)* که موجب افزایش کارآمدی تحلیل متن می‌شود. به منظور پیش‌بینی عملکرد محل در یک دیدگاه،

در مرحله سوم باید متن نظرات پیش‌پردازش شود. به این منظور از کتابخانه NLTK استفاده شده است. ابتدا متن هر دیدگاه به حروف کوچک تبدیل شده و توکن‌بندی می‌شود. سپس نشانه‌گذاری‌ها و واژگان توقف[1] حذف

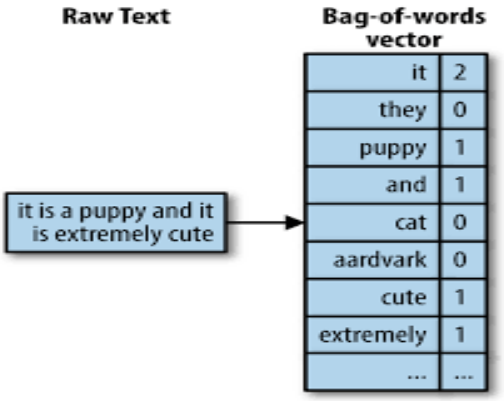
1- Stopwords

می‌شود. در ادامه با تبدیل هر واژه به شکل اصلی آن از طریق فرآیند ریشه‌یابی[1]، ریشه هر واژه مشخص می‌گردد. به این ترتیب هر دیدگاه کاربر به رشته‌ای از واژگان که به فرم اصلی درآمده‌اند، تبدیل می‌شود.

در مرحله چهارم ویژگی‌های[2] مورد نظر انتخاب و بردار ویژگی‌ها ایجاد می‌شود. از جمله روش‌های انتخاب ویژگی عبارتند از: مدل مجموعه واژگان (BoW)[3]، مدل تبدیل واژه به بردار (Word2Vec) و مدل تبدیل سند به بردار (Doc2Vec).

در این تحقیق از مدل BoW استفاده شده است. مدل مجموعه واژگان که اغلب به عنوان مدل فضای برداری دیده می‌شود، فقط برای واژه‌ها و فرکانس وقوع آن‌ها محاسبه می‌شود. با استفاده از مجموعه واژگان استخراج شده برای هر سند (یا متن که در این تحقیق دیدگاه کاربر است)، یک بردار ویژگی ایجاد می‌شود. هر ویژگی یک واژه (اصطلاح)[4] و مقادیر ویژگی، وزن آن است. وزن اصطلاح می‌تواند مقادیر دودویی، فراوانی اصطلاح (TF)[5] و یا فراوانی اصطلاح- معکوس فراوانی سند (TF-IDF)[6] باشد. در این مقاله، مقادیر ویژگی با استفاده از روش TF-IDF وزن‌دار شده‌اند.

BoW ترتیب و تعامل واژه‌ها را نادیده می‌گیرد و با هر واژه به عنوان یک ویژگی منحصر به فرد رفتار می‌کند. همچنین ساختار نحوی را نادیده می‌گیرد، با این حال نتایج مناسبی را برای آنچه برخی کاربردهای وابسته به نحو در نظر می‌گیرند، ارائه می‌دهد. این مشاهده نشان می‌دهد که نمایش‌های ساده، هنگامی که با مقادیر زیادی از داده‌ها همراه شوند، ممکن است به خوبی یا بهتر از نمایش‌های پیچیده‌تر عمل کنند. در نگاره (٤) مدل BoW در ایجاد بردار ویژگی نمایش داده شده است.




**نگاره٤: نمایشی از مدل مجموعه واژگان در ایجاد بردار ویژگی**

در مرحله پنجم، باید به کشف دانش پرداخت. انوع مختلفی از روش‌های یادگیری ماشین در کارهای NLP استفاده می‌شود که غالباً به دو دسته کلی نظارت‌شده و نظارت‌نشده تقسیم می‌شوند. یادگیری نظارت شده عموماً برای کارهای طبقه‌بندی و یادگیری نظارت‌نشده غالباً برای کارهای خوشه‌بندی به کار برده می‌شود. روش‌های مختلفی به منظور طبقه‌بندی متون ارائه شده است. روش‌های یادگیری ماشین از جمله پرکاربردترین این روش‌ها هستند. از آنجایی‌که هدف این تحقیق طبقه‌بندی نظرات کاربران در استخراج عملکرد محل است، در ادامه از روش لجستیک رگرسیون، به منظور طبقه‌بندی نظرات کاربران استفاده شده است. ذکر این نکته ضروری است که نه روش یادگیری ماشین و همچنین روش یادگیری عمیق برت به منظور طبقه‌بندی با استفاده از روش پیشنهادی، پیاده‌سازی شده و از نظر معیارهای رایج ارزیابی روش‌های یادگیری ماشین، مقایسه شده است. روش طبقه‌بندی لجستیک رگرسیون بهترین کارایی را در توازن بین دقت و زمان اجرای الگوریتم دارد. از این‌رو در تحقیق حاضر از الگوریتم لجستیک رگرسیون استفاده شده است. در نهایت در مرحله ششم کارایی روش پیشنهادی بررسی و نتایج با استفاده از معیارهای ارزیابی روش‌های یادگیری ماشین نظیر ماتریس در هم‌ریختگی[7] تفسیر و ارزیابی شد. با توجه به اینکه در این تحقیق به نوعی طبقه‌بندی انجام می‌شود، ارزیابی عملکرد معمولاً

1- Stemming and/or Lemmatization
2- Features
3- Bag-of-Words
4- Term
5- Term Frequency
6- Term Frequency-Inverse Document Frequency
7- ConfusionMatrix

براساس صحت پیش‌بینی با استفاده از مجموعه داده‌های آزمایش سنجیده می‌شود. با استفاده از این مجموعه داده‌ها مقادیر مثبت درست (TP)[1] (تعداد آیتم‌هایی که به درستی در کلاس مثبت طبقه‌بندی شده‌اند)، منفی درست (TN)[2] (تعداد آیتم‌هایی که به‌درستی در کلاس منفی طبقه‌بندی شده‌اند)، مثبت نادرست (FP)[3] (تعداد آیتم‌هایی که به صورت نادرست در کلاس مثبت طبقه‌بندی شده‌اند)، و منفی نادرست (FN)[4] (تعداد آیتم‌هایی که به صورت نادرست در کلاس منفی طبقه‌بندی شده‌اند) برای انجام ارزیابی عملکرد روش محاسبه می‌شوند. معمولاًدر ارزیابی روش‌های یادگیری ماشین معیارهای زیر درنظر گرفته می‌شود:

• **تشکیل ماتریس درهم ریختگی:** به ماتریسی گفته می‌شود که عملکرد الگوریتم‌های مربوطه را نشان می‌دهند. معمولاً چنین نمایشی برای الگوریتم‌های یادگیری نظارت‌شده استفاده می‌شود، اگر چه در یادگیری نظارت‌نشده نیز با عنوان ماتریس تطابق کاربرد دارد. این ماتریس یک ماتریس مربعی N* N است که N همان تعداد کلاس‌های ما در دسته‌بندی کننده می‌باشد. هر ستون از ماتریس، نمونه‌ای از مقدار پیش‌بینی شده را نشان می‌دهد. درصورتی که هر سطر نمونه‌ای واقعی (درست) را دربر دارد. در نگاره (۵) این ماتریس نشان داده شده است.

| | | برچسب پیش‌بینی شده | |
|---|---|---|---|
| | | مثبت | منفی |
| برچسب شناخته شده | مثبت | TP | FN |
| | منفی | FP | TN |

**نگاره۵: ماتریس درهم ریختگی**

ماتریس درهم ریختگی فقط برای کارهای طبقه‌بندی استفاده می‌شود، و به همین دلیل نمی‌تواند در مدل‌های رگرسیون یا سایر مدل‌های غیرطبقه‌بندی استفاده شود. ماتریس درهم ریختگی، نتایج حاصل از طبقه‌بندی را براساس اطلاعات واقعی موجود، نمایش می‌دهد. حال براساس این مقادیر می‌توان معیارهای مختلف ارزیابی دسته‌بندی و اندازه‌گیری دقت را تعریف کرد.

• **صحت**[5]: صحت به این معناست که مدل تا چه اندازه خروجی را درست پیش‌بینی می‌کند. در حقیقت بیانگر میزان الگوهایی است که درست پیش‌بینی شده‌اند. این معیار به صورت رابطه (۱) تعریف می‌شود:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$ رابطه (۱)

• **دقت**[6]: وقتی که مدل نتیجه را مثبت[7] پیش‌بینی می‌کند، این نتیجه تا چه اندازه درست است؟ در حقیقت نسبت نمونه‌های مرتبط در میان نمونه‌های بازیابی شده است. این معیار در رابطه (۲) تعریف شده است:

$$precision = \frac{TP}{TP + FP}$$ رابطه (۲)

• **بازیابی[8] یا حساسیت[9] (یا نرخ مثبت واقعی)**: این معیار بیانگر این است که چه نسبتی از تشخیص‌های مثبت واقعی به درستی شناسایی شده‌اند؟ زمانی که ارزش منفی‌های اشتباه[10] بالا باشد، معیار بازیابی، معیار مناسبی خواهد بود. در رابطه (۳) نحوه محاسبه این معیار نشان داده شده است:

$$recall = sensitivity = TPR = \frac{TP}{TP + FN}$$ رابطه (۳)

• **امتیاز F1**[11]: مطابق رابطه (٤)، این معیار دقت و بازیابی را با هم در نظر می‌گیرد. معیار F1 در بهترین حالت، یک و در بدترین حالت صفر است. در واقع این ارزیابی شامل سطح معینی از مصالحه بین نرخ TP و منفی درست TN و بین بازیابی و دقت است. دقت بالا و بازیابی پایین می‌تواند صحت خوبی از روش ارائه کند.

1- True Positive
2- True Negative
3- False Positive
4- False Negative

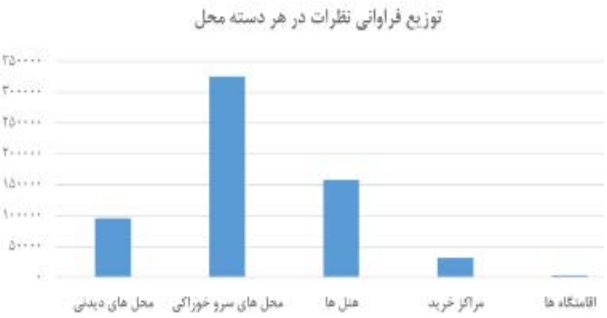
5- Accuracy
6- Precision
7- Positive
8- Recall
9- Sensitivity
10- False Negatives
11- F1-Score

$$F1=\frac{2\times(precision\times recall)}{2TP+FP+FN}=\frac{2TP}{2TP+FP+FN}$$

رابطه (٤)

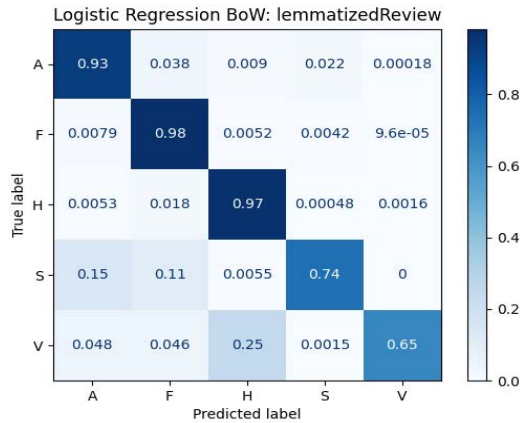
## ۳- ارزیابی نتایج

نگاره (٦) فراوانی توزیع داده‌ها در هر دسته محل را نشان می‌دهد. همانطور که مشخص است، مجموعه داده استخراج شده از وبگاه TripAdvisor به شدت نامتوازن است، به‌طوری که نظرات مربوط به محل‌های سرو خوراکی بیش از نیمی از کل داده‌ها را به خود اختصاص داده است، در حالی‌که نظرات مربوط به اقامتگاه‌ها در حدود ۰/۵٪ از کل داده‌ها را شامل می‌شود. به منظور مقابله با تأثیر نامتوازن بودن کلاس‌ها، به نمونه‌های مثبت وزن بیشتری تخصیص داده شده است.




**نگاره٦: توزیع فراوانی نظرات کاربران در هر دسته محل**

از زبان برنامه‌نویسی پایتون و کتابخانه‌های مختلف NLP برای پیاده‌سازی روش پیشنهادی استفاده شده است.کتابخانه NLTK به عنوان یکی از قدرتمندترین کتابخانه‌ها در زمینه پیش‌پردازش متون، برای آماده‌سازی و پیش‌پردازش نظرات کاربران به‌کار گرفته شده است. برای نمونه از واژگان توقف تعریف شده در این کتابخانه به منظور حذف واژگان توقف موجود در نظرات کاربران استفاده شده است. همچنین واژه‌ها با استفاده از این کتابخانه به فرم اصلی و ریشه تبدیل شده‌اند. از کتابخانه Gensim برای ایجاد مجموعه واژگان استفاده شده است. کتابخانه Scikit-Learn نیز به منظور ایجاد و آموزش طبقه‌بندی‌کننده و همچنین محاسبه معیارهای ارزیابی به‌کار گرفته شده است. به‌صورت تصادفی ۷۵٪ مجموعه داده برای آموزش مدل و ۲۵٪ برای آزمایش نتایج درنظر گرفته شده است. روش طبقه‌بندی پیشنهادی یک روش نظارت شده است، به این صورت که متن نظرات کاربران پس از پیش‌پردازش و همچنین برچسب محل‌ها به عنوان ورودی به منظور آموزش یک طبقه‌بندی‌کننده (الگوریتم لجستیک رگرسیون) داده می‌شود. سپس برای داده‌های آزمایشی[1]، برای هر نظر برچسب محل مرتبط با آن که همان عملکرد محل است، پیش‌بینی می‌شود. در نهایت لازم است کارایی روش پیشنهادی بررسی و ارزیابی گردد. این معیارهای ارزیابی برای داده‌های آزمایشی ارائه شده است. پس از پیش‌بینی عملکرد محل‌ها با استفاده از روش پیشنهادی، به منظور محاسبه ماتریس درهم ریختگی از کتابخانه Scikit-Learn استفاده شده است. نگاره (۷) ماتریس درهم ریختگی نرمال‌شده با استفاده از روش پیشنهادی را نشان می‌دهد. این ماتریس از طریق طبقه‌بندی نظرات کاربران در مورد محل با استفاده از الگوریتم لجستیک رگرسیون محاسبه شده است. صحت کلی الگوریتم پیش‌بینی ۹۵/۸٪ و زمان اجرای الگوریتم حدود ۱۰۸ ثانیه می‌باشد.




**نگاره۷: ماتریس درهم ریختگی نرمال‌شده با استفاده از روش پیشنهادی در استخراج عملکرد محل**

همانطور که در نگاره قابل مشاهده است، برای محل‌های مرتبط با سرو خوراکی، پیش‌بینی‌ها بسیار به واقعیت نزدیک بوده و درصد خطا بسیار ناچیز است. به‌گونه‌ای که در ۹۸٪

1- Test Dataset

موارد الگوریتم به درستی توانسته محل‌های مرتبط با سرو خوراکی را از روی نظرات کاربران شناسایی و پیش‌بینی کند. همچنین در حدود ۰/۸٪ از این نوع محل‌ها، به عنوان محل‌های دیدنی درنظر گرفته شده‌اند. این امر می‌تواند ناشی از آن باشد که گاهی اوقات افراد از رستوران‌ها و کافی‌شاپ‌ها به عنوان محلی برای سرگرمی و تفریح استفاده می‌کنند. درمورد هتل‌ها نیز صحت پیش‌بینی‌ها ۹۷٪ می‌باشد. هرچند از آنجایی که غالباً در هتل‌ها مواد خوراکی و یا وعده‌های غذایی سرو می‌شود، در حدود ۱/۸٪ از پیش‌بینی مربوط به هتل‌ها به اشتباه به عنوان محل‌های سرو خوراکی دسته‌بندی شده است. حدود ۰/۵٪ نیز به عنوان محل‌های دیدنی درنظر گرفته شده است که می‌تواند گواه این باشد که برخی هتل‌ها جذابیت توریستی دارند. این مسئله درمورد محل‌های دیدنی که با صحت ۹۳٪ درست پیش‌بینی شده‌اند نیز صادق است و در حدود ۳/۸٪ از نظرات مرتبط با محل‌های دیدنی به اشتباه به عنوان محل‌های سرو خوراکی درنظر گرفته شده است. همانطور که اشاره شد، محل‌های سرو خوراکی می‌توانند به عنوان محلی برای تفریح و وقت‌گذرانی افراد تلقی شود. همچنین در برخی از محل‌های دیدنی ممکن است محلی برای سرو نوشیدنی، غذا و غیره وجود داشته باشد.

در خصوص مراکز خرید صحت به نسبت سایر عملکردها کمتر و در حدود ۷٤٪ است، زیرا اولاً تعداد نظرات مرتبط با این نوع عملکرد کمتر است، هرچند این مسئله از طریق وزن‌دار کردن نمونه‌ها تا حدودی حل شده است. ثانیاً در بسیاری از موارد افراد از پاساژها و مراکز خرید به منظور سرگرمی و تفریح بازدید می‌کنند و نه صرف خرید کردن، این مسئله موجب شده حدود ۱۵٪ مراکز خرید به عنوان محل‌های دیدنی طبقه‌بندی شوند. همچنین حدود ۱۱٪ این مراکز خرید به عنوان محل‌های سرو خوراکی درنظر گرفته شده است. از مهم‌ترین دلایل این امر می‌توان به عملکرد خرید مواد خوراکی در این محل‌ها اشاره کرد که خود نوعی خرید به حساب می‌آید. به علاوه در برخی از مراکز خرید محلی برای سرو نوشیدنی، غذا و غیره وجود دارد. از آنجایی که نظرات مربوط به اقامتگاه‌ها نسبت به سایر عملکردها کمتر است، کمترین صحت مربوط به اقامتگاه‌ها با صحت حدود ٦٥٪ می‌باشد. این درحالی است که در ۲۵٪ موارد، اقامتگاه‌ها به عنوان هتل‌ها دسته‌بندی شده‌اند. این نتیجه خیلی دور از انتظار نیست، زیرا اقامتگاه‌ها و هتل‌ها از نظر عملکردی شباهت بسیار زیادی دارند و غالباً به منظور اسکان مسافران و گردشگران استفاده می‌شوند. همچنین به ترتیب ٤/٨٪ و ٤/٦٪ از اقامتگاه‌ها به عنوان محل‌های دیدنی و محل‌های سرو خوارکی طبقه‌بندی شده‌اند. از دلایل عمده آن می‌توان به بوم‌گردی‌ها و آشپزی و سرو غذا در اقامتگاه‌ها اشاره کرد. نتایج ارزیابی روش پیشنهادی به‌طور خلاصه در جدول(۲) نشان داده شده است.

**جدول۲: نتایج حاصل از ارزیابی روش پیشنهادی**

| نوع محل | دقت | بازیابی | امتیاز F1 |
|---|---|---|---|
| محل‌های دیدنی | ۰/۹۲ | ۰/۹۳ | ۰/۹۲ |
| محل‌های سرو خوراکی | ۰/۹۷ | ۰/۹۸ | ۰/۹۸ |
| هتل‌ها | ۰/۹۸ | ۰/۹۷ | ۰/۹۸ |
| مراکز خرید | ۰/۸۷ | ۰/۷٤ | ۰/۸۰ |
| اقامتگاه‌ها | ۰/۸٦ | ۰/٦٥ | ۰/۷٤ |
| میانگین کلان | ۰/۹۲ | ۰/۸٦ | ۰/۸۸ |
| میانگین وزن‌دار | ۰/۹٦ | ۰/۹٦ | ۰/۹٦ |

همانطور که از نتایج ارزیابی استنباط می‌شود، بهترین نتایج به ترتیب مربوط به محل‌های سرو خوراکی، هتل‌ها و محل‌های دیدنی با امتیاز F1 برابر ۰/۹۸، ۰/۹۸ و ۰/۹۲ می‌باشد. امتیاز F1 مربوط به مراکز خرید و اقامتگاه‌ها به ترتیب با ۰/۸۰ و ۰/۷٤ نسبت به سایرین کمتر است. این موضوع اگرچه تا حدودی وابسته به حجم داده‌ها در هر دسته است، ولی شباهت برخی محل‌ها از نظر عملکرد را اثبات می‌کند. برای نمونه عمده عملکرد هتل‌ها و اقامتگاه‌ها اسکان افراد می‌باشد که این امر موجب شده این دو گاهی به جای یکدیگر در نتایج پیش‌بینی ظاهر شوند.

## ٤- نتیجه‌گیری

با توجه به افزایش روزافزون استفاده از رسانه‌ها و شبکه‌های اجتماعی، محتواهای کاربرتولید رشد چشمگیری داشته است. از میان انواع محتواهای کاربرتولید، محتواهای متنی عموماً به‌صورت غیرساختاریافته به اشتراک گذاشته می‌شوند. این اطلاعات در توصیف عوارض مکانی، غالباً به‌صورت محل‌مبنا بوده و سرشار از اطلاعات غنی در مورد آن محل است. از جمله ویژگی‌های مهمی که در توصیف محل به‌کار گرفته می‌شود، عملکرد محل است. در این تحقیق سعی شد عملکرد محل با استفاده از تحلیل محتواهای متنی کاربرتولید به اشتراک گذاشته شده در وبگاه TripAdvisor توسط کاربران استخراج شود. به این منظور از روش‌های مختلف NLP به‌منظور آماده‌سازی و پیش‌پردازش داده‌ها استفاده شد. سپس مجموعه واژگان ساخته شده برای هر دیدگاه کاربر، به یک طبقه‌بندی‌کننده لجستیک رگرسیون به منظور آموزش مدل داده شد و با استفاده از آن عملکرد محل بر روی داده‌های آزمایشی پیش‌بینی شد. نتایج بیشترین دقت و امتیاز F1 را برای محل‌های سرو خوراکی نشان می‌دهد، در حالی‌که اقامتگاه‌ها به دلیل شباهت عملکردی به هتل‌ها کمترین دقت و امتیاز F1 را دارند، ولی با این وجود نتایج آن‌ها نیز قابل اطمینان است.

در تحقیقات آتی تلاش می‌شود کارایی سایر روش‌های انتخاب ویژگی و همچنین الگوریتم‌های طبقه‌بندی با استفاده از یادگیری ماشین در استخراج عملکرد محل بررسی و مقایسه شود. همچنین در صورت امکان عملکرد محل به صورت جزئی‌تر و خاص‌تر استخراج شود. برای نمونه بین انواع محل‌های دیدنی تمایز قائل شده و بتوان آن‌ها را از یکدیگر تفکیک نمود. همچنین می‌توان یک سیستم توصیه‌گر به‌منظور پیشنهاد محل به کاربران براساس احساسات و فعالیت‌های به اشتراک گذاشته شده توسط آنان طراحی، توسعه و پیاده‌سازی نمود. به‌علاوه پیشنهاد می‌شود تأثیر روابط مکانی در استخراج عملکرد محل بررسی شود.

## ٥- منابع و مآخذ

1- Adams, B., & Janowicz, K. (2015). Thematic signatures for cleansing and enriching place-related linked data. International Journal of Geographical Information Science, 29(4), 556-579.

2- Adams, B., & McKenzie, G. (2013). Inferring thematic places from spatially referenced natural language descriptions. In Crowdsourcing geographic knowledge (pp. 201-221): Springer.

3- Alazzawi, A. N., Abdelmoty, A. I., & Jones, C. B. (2012). What can I do there? Towards the automatic discovery of place-related services and activities. International Journal of Geographical Information Science, 26(2), 345-364.

4- Alexander, C. (2002). The Phenomenon of Life: BOOK ONE The Nature of Order: An Essay on the Art of Building and The Nature of the Universe.

5- Couclelis, H. (1992). Location, place, region, and space. Geography's inner worlds, 2, 15-233.

6- "Decennial Census P.L. 94-171 Redistricting Data". U.S. Census Bureau. Retrieved August 12, 2020.

7- Fan, K., Zhang, D., Wang, Y., & Zhao, S. (2015). Discovering urban social functional regions using taxi trajectories. Paper presented at the 2015 IEEE 12th Intl Conf on Ubiquitous Intelligence and Computing and 2015 IEEE 12th Intl Conf on Autonomic and Trusted Computing and 2015 IEEE 15th Intl Conf on Scalable Computing and Communications and Its Associated Workshops (UIC-ATC-ScalCom).

8- Gibson, J. J. (1977). The theory of affordances. Hilldale, USA, 1(2).

Goodchild, M. F. (2015). Space, place and health. Annals of GIS, 21(2), 97-100.

9- Han, H., Yu, X., & Long, Y. (2015). Discovering functional zones using bus smart card data and points of interest in beijing. arXiv preprint arXiv:1503.03131.

10- Han, J., & Kamber, M. (2001). Data mining concepts and techniques, Morgan Kaufmann Publishers. San Francisco, CA, 335-391.

11- Hartshorne, R. (1969). Perspective on the Nature of Geography.

12- Hill, L. L. (2000). Core elements of digital gazetteers: placenames, categories, and footprints. Paper presented

23- Purves, R. S., Winter, S., & Kuhn, W. (2019). Places in information science. Journal of the Association for Information Science and Technology, 70(11), 1173-1182.
24- Relph, E. (1976). Place and placelessness (Vol. 1): Pion.
25- Su, S., Lei, C., Li, A., Pi, J., & Cai, Z. (2017). Coverage inequality and quality of volunteered geographic features in Chinese cities: Analyzing the associated local characteristics using geographically weighted regression. Applied geography, 78, 78-93.
26- Tao, H., Wang, K., Zhuo, L., & Li, X. (2019). Re-examining urban region and inferring regional function based on spatial–temporal interaction. International journal of digital earth, 12(3), 293-310.
27- "Top 8 Cities by GDP: China vs. The U.S." Business Insider, Inc. July 31, 2011. Retrieved July 1, 2018. For instance, Shanghai, the largest Chinese city with the highest economic production, and a fast-growing global financial hub, is far from matching or surpassing New York, the largest city in the U.S. and the economic and financial super center of the world."New York City: The Financial Capital of the World". Pando Logic. October 8, 2015. Retrieved July 1, 2018.
28- Tuan, Y.-F. (1979). Space and place: humanistic perspective. In Philosophy in geography (pp. 387-427): Springer.
29- Vasardani, M., Tomko, M., & Winter, S. (2016). The cognitive aspect of place properties. Paper presented at the International Conference on GIScience Short Paper Proceedings.
30- Winter, S., Baldwin, T., Tomko, M., Renz, J., Kuhn, W., & Vasardani, M. (2021). Spatial concepts in the conversation with a computer. Communications of the ACM, 64(7), 82-88.
31- Zhou, T., Liu, X., Qian, Z., Chen, H., & Tao, F. (2020). Automatic Identification of the Social Functions of Areas of Interest (AOIs) Using the Standard Hour-Day-Spectrum Approach. ISPRS International Journal of Geo-Information, 9(1), 7.
32- Zhou, X., & Zhang, L. (2016). Crowdsourcing functions of the living city from Twitter and Foursquare data. Cartography and Geographic Information Science, 43(5), 393-404.
33- https://doi.org/10.1111/tgis.12999
at the International Conference on Theory and Practice of Digital Libraries.
13- Hobel, H., Fogliaroni, P., & Frank, A. U. (2016). Deriving the geographic footprint of cognitive regions. In Geospatial data in a changing world (pp. 67-84): Springer.
14- Jordan, T., Raubal, M., Gartrell, B., & Egenhofer, M. (1998). An affordance-based model of place in GIS. Paper presented at the 8th Int. Symposium on Spatial Data Handling, SDH.
15- Khoury, R., Karray, F., & Kamel, M. (2006). Extracting and representing actions in text using possibility theory. Paper presented at the Proceedings of the 3rd annual e-learning conference on Intelligent Interactive Learning Object Repositories (i2LOR 2006).
16- Mocnik, F.-B. (2022). Putting geographical information science in place–towards theories of platial information and platial information systems. Progress in Human Geography, 03091325221074023.
17- Mocnik, F.-B., & Westerholt, R. Places Across Cultures.
18- Noulas, A., Scellato, S., Mascolo, C., & Pontil, M. (2011). Exploiting semantic annotations for clustering geographic areas and users in location-based social networks. Paper presented at the Fifth International AAAI Conference on Weblogs and Social Media.
19- Papadakis, E., & Blaschke, T. (2017). Place-based GIS: Functional Space. Paper presented at the AGILE PhD School.
20- Papadakis, E., Gao, S., & Baryannis, G. (2019). Combining Design Patterns and Topic Modeling to Discover Regions That Support Particular Functionality. ISPRS International Journal of Geo-Information, 8(9), 385.
21- Papadakis, E., Resch, B., & Blaschke, T. (2016). A Function-based model of Place. Paper presented at the International Conference on GIScience Short Paper Proceedings.
22- Purves, R. S., Clough, P., Jones, C. B., Arampatzis, A., Bucher, B., Finch, D., . . . Vaid, S. (2007). The design and implementation of SPIRIT: a spatially aware search engine for information retrieval on the Internet. International Journal of Geographical Information Science, 21(7), 717-745.