جایگاه کلان‌داده‌ها در روش پژوهش علوم اجتماعی با معرفی نرم‌افزارهای کاربردی

بهنام لطفی خاچکی[1]

چکیده

افزایش نفوذ اینترنت در بین اقشار مختلف، در کنار گسترش شبکه‌های اجتماعی، انواع پلتفرم‌های مالی، پیام‌رسان‌ها، اپلیکیشن‌های خدماتی و ... به تدریج باعث فزونییافتن مقدار داده‌های تولید شده توسط مردم شده است. حجم فراتر از حد تصور این داده‌ها که در بسترهای گوناگون و با اهداف متفاوت تولید شده‌اند، به تدریج آنها را تبدیل به منبعی غنی برای تحقیقات اجتماعی کرده است. علی‌رغم مباحث زیادی که حول کلان‌داده‌ها و کاربرد هوش مصنوعی در شناسایی و استخراج ایده‌های پژوهشی از دل آنها در حداقل برخی محافل علمی علوم اجتماعی ایران در جریان است، استفاده کاربردی و عملیاتی از این داده‌ها در عمل با دشواری‌های روبرو است. این مقاله سعی دارد ضمن تشریح مسائل پیش‌روی تحلیل داده‌های بزرگ در فضای تحقیقات علوم اجتماعی ایران امروز، به معرفی چند نرم‌افزار و ابزار کاربردی در این حوزه بپردازد. از جمله موارد قابل ذکر می‌توان به نرم‌افزار Gephi (نرم‌افزاری متن‌باز و مبتنی بر شبکه برای تجسم و تحلیل داده‌های پیچیده شبکه‌ای) و نرم‌افزار Pajek (نرم‌افزاری شبکه‌محور و دارای ابزارهایی چون الگوریتم مرتب‌سازی، کاوش و خوشه‌بندی)، نرم‌افزار R Studio، ابزار Ngram، ابزار Looker Studio و فناوری Hadoop اشاره کرد که هریک کاربردهای خاص خود را دارد. در این مقاله سعیشده است با معرفی دقیق‌ترموارد قابل کاربرد در تحلیل کلان‌داده‌ها، مسیری برای خلق ایده‌های پژوهشی ترسیم گردد.

واژگان کلیدی: کلان‌داده، روش پژوهش، نرم‌افزار کاربردی، گفی، مرکز داده

[1] دکتری جامعه‌شناسی، عضو هیأت علمی گروه آموزش علوم اجتماعی دانشگاه فرهنگیان، blotfi66@cfu.ac.ir

مقدمه

پیشرفت فناوری در دهه‌های اخیر جنبه‌های گوناگون زندگی انسان را تحت تأثیر قرار داده است. یکی از مهم‌ترین اتفاقات این حوزه تحولات ارتباطی و انقلاب اطلاعاتی و در پی آن ظهور اینترنت است.

بالا رفتن ضریب نفوذ اینترنت در بین اقشار مختلف، تولید داده در فضای مجازی را افزایش داده است. در دهه اخیر گسترش استفاده از اینترنت از یک طرف و تحولات اجتماعی از طرف دیگر، روند رشد داده‌ها و اطلاعات را به طور چشمگیری افزایش داده است(Tulasi, 2014). با گسترش شبکه‌های اجتماعی، انواع پلتفرم‌های مالی، پیام‌رسان‌ها، اپلیکیشن‌های خدماتی و ... به تدریج مقدار داده‌های تولید شده توسط مردم‌فزونییافته است. به عنوان نمونه‌ای از کلان‌داده‌ها می‌توان به انجام بیش از یک میلیون تراکنش مشتریان در هر ساعت در یک فروشگاه زنجیره‌ای، بیش از ۵۰ میلیارد عکس که روزانه در فیس‌بوک منتشر می‌شود و شش پتابایت داده‌های ماهانه نجومی اشاره کرد که دائماً جمع‌آوری می‌شوند. همینطور در سال ۲۰۲۲ در هر روز به طور میانگین ۱۸,۹ میلیارد جستجو در گوگل صورت گرفته است و تعداد توییت‌های روزانه توئیتر به حدود ۲,۴ میلیاردمورد می‌رسد. حجم فراتر از حد تصور این داده‌ها که در بسترهای گوناگون و با اهداف متفاوت تولید شده‌اند، به تدریج آنها را تبدیل به منبعی غنی برای تحقیقات اجتماعی کرده است؛ منبعی که به آن کلان‌داده[۲]می‌گویند.

کلان‌دادهایدارایی‌های بسیار انبوه، پرشتاب و یا گوناگونی هستند که محصول پیشرفتچشمگیر فناوری اطلاعات بوده و روز به روز مشاغل و حوزه‌های بیشتری را درگیر خود می‌سازد. این اندازه داده‌هانیست که آن را مهم و حیاتی می‌کند، بلکه کارهاییاست که سازمان‌ها و افراد می‌توانند با استفاده ازاین داده‌ها انجام دهند (Beyer & Laney, 2012: ۹۸).کلان‌داده‌ها، پدیده نسبتاً جدیدی بوده و محصول محیط فناورانه‌ای است که در آن محیط، همه چیز می‌تواند به صورت دیجیتال ثبت شده، مورد اندازه‌گیری قرار گیرد و در نتیجه، به داده تبدیل شود. این موضوع اشاره به مفهوم داده‌ای شدن دارد که می‌تواند باعث شود هزاران رخداد که شامل تعدادی عدد، متن، تصویر، صدا و ویدیو است و نیازمند ظرفیت حافظه پتابایتی (معادل هزار ترابایت) است، به صورت همزمان ردیابی و به صورت بلادرنگ اجرا شوند (گنجی، ۱۳۹۹).

دوگ لانی[۳]در سال ۲۰۰۱ در مؤسسه گارتنر برای اولین بار اصطلاح کلان‌داده را برای اشاره به داده‌هایی که نظر حجم و سرعت و تنوع در حال افزایش هستند، به کاربرد. طبق تعریف وی کلان‌داده شامل اطلاعاتی با حجم زیاد است که با روش‌های نوین پردازش و ذخیره‌سازی برای درک بهتر از دنیا و روند تصمیم‌گیری دقیق‌تر مورد استفاده قرار می‌گیرد (کوهزادی و همکاران، ۱۴۰۱: ۳۵). کلان‌داده مفهومی گسترده برای تعریف مجموعه‌های پیچیده و بزرگ از داده‌هاست که به وسیله ابزارهای سنتی، قابل پردازش نیستند و مسائلی چون تحلیل، دریافت، گزینش، اشتراک، انتقال و امنیت داده‌ها با افزایش حجم و تنوع آنها برای کاربران حوزه‌های گوناگون چالش‌برانگیز شده است (روحانی و همکاران، ۱۳۹۸: ۱۲۲). اندازهٔ ایندادهها دائمادر حال رشد است و پردازش آنها جهت مدیریت،

---

۲ Big data
۳ Doug Loney

ذخیره‌سازی، به‌اشتراک‌گذاری،تجزیه و تحلیل، انتقال و جست‌وجو در یک زمان قابلتحمل توسط نرم‌افزارهایی که به‌طور معمول استفاده می‌شوند، دشوار است (Brown; 2015 Choi and Shin; 2011).

کلان‌داده ویژگی‌هایی دارد که آن را از دیگر انواع داده‌ها متفاوت می‌سازد. یکی از اصلی‌ترین ویژگی آن حجم[۴]است که به طور پیوسته در حال رشد می‌باشد. گوناگونی[۵]ویژگی دیگر کلان‌داده است که باعث تنوع زیاد داده‌های دیجیتالی همچون صفحات وب، پایگاه‌های خبری، فایل‌های صوتی، متن‌های شخصی، تصاویر، ویدئوها و... شده است.شتاب[۶]نیزنشانگر سرعت تولید و پردازش داده‌ها در این حوزه است. این سه ویژگی اصلی از دهه گذشته تاکنون مورد توجه بوده است که اصطلاحاً به آن مدل ۳ گفته می‌شود (Chen et al, 2014; Chen & Zhang, 2014). اما با گسترش ورود این مفهوم به حوزه‌های گوناگون، ویژگی‌های جدیدی همچون درستی[۷] تغییرپذیری[۸]استمرار و پیچیدگی[۹]نیز به آن اضافه شده است (Sun, 2018).

همچنین برخی اندیشمندان از زاویه‌ای دیگر ده ویژگی برای کلان‌داده‌ها برشمرده‌اند که بدین شرح است. «حساس» یعنی اگر اطلاعات شخصی افراد عمومی بشود، باعث زیان عاطفی و اقتصادی به آنها می‌گردد. «کثیف» یعنی غالباً داده‌ها با اطلاعاتی همراه هستند که به درد پژوهش نمی‌خورد. «تحت طلسم الگوریتم» یعنی رفتار کاربران در سیستم‌های کلان داده طبیعی نیست و تحت تأثیر اهداف مهندسی سیستم به وقوع می‌پیوندد.«شناور» یعنی بسیاری از سیستم‌های کلان‌داده خصوصاً نوع تجاری آن همواره در حال تغییر هستند. این سیستم‌ها به سه طریق شناوری جمعیتی (تغییر در کاربران سیستم)، شناوری رفتاری (تغییر در نحوه استفاده کاربران از سیستم) و شناوری سیستمی (تغییر در خود سیستم) منجر به تغییراتی در داده‌ها می‌شوند. «نامعرف» یعنی بسیاری از منابع کلان‌داده نمونه‌های معرفی از جمعیت کل نیستند و عمدتاًبه درد سنجش‌های درون‌نمونه‌ای می‌خورند. «خارج از دسترس» یعنی دسترسی به داده‌هایِ در اختیار شرکت‌ها و دولت‌ها، برای پژوهشگران بسیار دشوار است. «ناقص» بودن یعنی بیشتر منابع کلان‌داده اطلاعاتی را که پژوهشگر برای تحقیق بدان نیاز دارند، در خود ندارند. «بدون واکنش» یعنی فرایند اندازه‌گیری و تحقیق در کلان‌داده، بعید است باعث تغییر رفتار کسی بشود. بنابراین مشکل تغییر رفتار افراد مورد مشاهده هنگام تحقیق در اینجا وجود ندارد. «همواره روشن» یعنی سیستم‌های کلان‌داده پیوسته در حال جمع‌آوری داده هستند. «حجیم» یعنی داده‌های با مقدار بسیار زیاد تولید می‌شوند (سالگانیک، ۱۴۰۰).

اهمیت و ارزش یافتن روزافزون کلان‌داده‌ها برای محققان، سازمان‌ها و حتی کشورها، باعث خلق نوآوری‌هایی در مدیریت و پردازش این اطلاعات حجیم و ابداع روش‌های نوآورنه و خلاقانه برای بهره‌گیری هرچه بیشتر از آنها شده است (رضایی و همکاران، ۱۴۰۰: ۸۳؛ غفاری و همکاران، ۱۳۹۲). تحولات اخیر درعرصه فناوری‌های نوین ارتباطی، وشتاب در تغییر و در قابلیت‌های رو به رشد این فناوری‌ها، چه به لحاظسهولت بهره‌برداری و جاذبه‌های فرمی، و چه به

---

[۴] Volume
[۵] Variety
[۶] Velocity
[۷] Veracity
[۸] Variability
[۹] Complexity

لحاظ محتوایی، شیوه‌های ارزیابی و روش‌شناسی خاص پژوهش را به همراه آورده است. در واقع ظهور کلان‌داده‌ها در حوزه روش‌شناسی باعث ظهور کلان‌روش‌های تحقیقبا تکنیک‌های مقتضی آن نیز شده است (نقیب‌السادات، ۱۴۰۱). در واقع یکی از جنبه‌های تغییرات فناوریو پیدایش داده‌های بزرگ، اثرگذاری بر روش‌های تحقیق در علوم مختلف از جمله علوم اجتماعی است. از آنجایی که علوم اجتماعی و به طور خاص جامعه‌شناسی برای انجام تحقیق با داده‌ها و اطلاعاتی سر و کار دارد که انسان در تولید آن نقش بازی می‌کند و با توجه رشد حجم تولید داده توسط انسان در دهه‌های اخیر، می‌توان شاهد تحول در حوزه روش تحقیق در علوم اجتماعی به واسطه کلان‌داده‌ها بود.

از جمله بسترهای مهم تولید کلان‌داده که برای محققان علوم اجتماعی از اهمیت و جذابیت بالایی برخوردار است، شبکه‌های اجتماعی هستند.با گسترش استفاده از اینترنت بین اقشار مختلف، استفاده کاربران از شبکه‌های اجتماعی روزبه‌روز در حال افزایش است. بر اساس آمارهای موجود حدود ۶۰ درصد جمعیت جهان کاربر اینترنت بوده از این بین حدود ۷۰ درصدکاربران فعال شبکه‌های اجتماعی هستند (حیدری، ۱۳۹۹: ۱۱۰). از جمله علل محبوبیت استفاده از این رسانه‌ها این است که به کاربران فرصت دریافت یا ایجاد و به اشتراک گذاشتن پیام‌های عمومی را با هزینه‌های کم و در همه‌جا را می‌دهد. این ابزارها امکان بهره گرفتن از انواع فرمت‌های داده، از جمله داده‌های متنی، تصاویر، فیلم‌ها، صداها و موقعیت‌های جغرافیایی را فراهم ساخته است.رشد بسیار زیاد استفاده از این قبیل شبکه‌ها خود منجر به انباشت داده‌های عظیم شده است (Stieglitz & et al, 2018). بنابراین تحقیق پیرامون داده‌های مرتبط با شبکه‌های اجتماعی از جمله حوزه‌های تحقیقی کار با کلان‌داده‌ها با رویکرد اجتماعی است.

هم‌اکنون بررسی وضعیت تولید پژوهش با کمک کلان‌داده‌ها نشان می‌دهد عمده تحقیقات انجام شده در کشور حول رضایت مشتریان، بازاریابی، بیمه، کتابداری و بانکداری می‌چرخد. اما این روند به تدریج جای خود را در حوزه علوم اجتماعی نیز باز خواهد کرد.

از سوی دیگر بررسی‌ها نشان می‌دهد کلان‌داده‌ها در حوزه‌ها و اکوسیستم‌های مختلفی تولید شده و مورد استفاده قرار می‌گیرد. این اکوسیستم‌ها که کاربرد هریک در شکل ۱ نشان داده شده است، عبارت‌اند از: سلامت، محیط زیست، بانکداری، فناوری اطلاعات و ارتباطات، انرژی، خرده‌فروشی، رسانه، حمل‌ونقل، تفریحی و ورزشی و سرگرمی، آموزش و تحقیقات (صاحب و فرزین ۱۳۹۶). محقق علوم اجتماعی می‌تواند در هریک از اکوسیستم‌های فوق دست به پژوهش بزند. در واقع منابع کلان‌داده‌ها از رسانه‌های اجتماعی تا داده‌های ماشین، حسگرها، تراکنش‌های اینترنتی و ... می‌توانند در تحقیقات اجتماعی بین‌رشته‌ای مورد استفاده قرار بگیرند.

شکل۱- کاربرد کلان‌داده‌ها در اکوسیستم‌های مختلف (صاحب و فرزین، ۱۳۹۶)

نتیجه یک تحقیق میدانی پیرامون اکوسیستم‌های فعال در حوزه کلان‌دادهدر ایران نیز نشان می‌دهد استفاده از کلان‌داده‌ها بیش از همه در بخش فناوری اطلاعات (۴۷,۱ درصد)، فروش(۲۶,۵ درصد)، بانکداری (۲۳,۵ درصد) و نیز در بخش آموزش و تحقیقات (۲۰,۶ درصد) دیده می‌شود. حوزه‌هایی که در رتبه بعدی قرار می‌گیرند نیز عبارتند ازسلامت، رسانه، انرژی و حمل‌ونقل (دسترنج و همکاران، ۱۳۹۸). این آمار نشانگر اهمیت روزافزون کاربرد کلان‌داده‌ها در بخش پژوهش و تحقیقات است.

در صورتی که پژوهش را بخشی از نظام آموزشی و به ویژه نظام آموزش عالی در نظر بگیریم، باید از تأثیر کلان‌داده‌ها بر این نظام نیز سخن گفت. مطالعهدقیق مفاهیم مرتبط با کلان‌داده‌هـا که تقریباً در دهه اخیر آغاز شده است نشان می‌دهداین مفهوم ابتدا در حوزه تجارت، فناوری اطـلاعات و ارتـباطات و هـم‌اکنون در حوزه آموزش و آموزش عالی مورد توجه جدی قرار گرفته است. سرعت بالای تولید داده‌ها در آینـده‌ای نه چندان دور نه تنها حوزه آموزش، بلکه حوزه پژوهشرا نیز دستخوش تغییر قرار خواهد داد. پیدایش و گسترش کلان‌داده همچون یک نظام دانـش تـوانسته آموزش عالی را به لحاظ تغییر اهداف دانش و نظریه‌های اجتماعی تحت تأثیر قرار دهد؛به طوری که کلان‌داده‌ها پژوهش‌های نوظهور و رویکردهای تحلیلی جدیدی را شکل داده است (Wagner & Ice, 2012؛ Boyd & Crawford, 2012).

با توجه به اهمیت کلان‌داده‌ها برای علوم اجتماعی تحقیق حاضر قصد دارد ضمن تشریح مسائل پیش‌روی تحلیل داده‌های بزرگ در فضای تحقیقات علوم اجتماعی ایران امروز، به معرفی چند نرم‌افزار و مرکز داده/بستر (دیتابیس) کاربردی در این حوزه بپردازد.

روش تحقیق

در تحقیق حاضر با توجه به ماهیت ساختار پژوهش، از روش توصیفی-تـحلیلیاستفاده شده است. همچنین برای گردآوری داده‌ها در این تحقیق از روش اسنادی استفاده شده و محقق با استفاده از مرور مباحث نظری و مطالعه منابع مختلف به دنبال بررسی جایگاه کلان‌داده‌ها در روش تحقیق بوده است.

پژوهش اسنادی[1] استفاده از منابع و اسناد بیرونی برای پشتیبانی از نظریهیا دیدگاهی در مطالعات دانشگاهی است.روش پژوهش اسنادی هم به منزله روشی تام و هم تکنیکی برای تقویت سایر روش‌هادر پژوهش‌های علوم اجتماعی مورد توجه بوده است. در این روش، پژوهشگر داده‌های پژوهشی خود را درباره کنشگران، وقایع و پدیده‌های اجتماعی، از بین منابع و اسناد گردآوری می‌کند. بخش قابل توجهی از پژوهش‌های نظری در علوم اجتماعی،به انحایمختلف از روش اسنادی بهره می‌برند.

برای نگارش این مقاله ابتدا با بررسی کلیدواژگانی مانند «کلان‌داده» و «داده‌های بزرگ» و «نرم‌افزار تحلیل داده حجیم» مقالات مرتبط شناسایی و مورد مطالعه قرار گرفت. سپس نرم‌افزارهای آموزشی قابل کاربرد در تحلیل کلان‌داده‌ها شناسایی شد. در قدم بعد هریک از این نرم‌افزارها مورد مطالعه دقیق‌تر قرار گرفت تا کاربرد هر کدام متناسب با موضوع تحقیق مشخص شود.

یافته‌ها

در ابتدای این بخش لازم است حول این پرسش که کلان داده‌ها چه کمکی به پیشبرد پژوهش‌های اجتماعی می‌کنند، نکاتی ارائه شود.

روش‌های میدانی در علوم اجتماعی عموماً هزینه‌بر و زمان‌بر بوده و با دشواری‌هایی همراه هستند؛ زیرادر این روش‌ها باید با افراد و گروه‌های مختلف تعامل برقرار کرد. از این رو برخی محققان ممکن است به دلیل مشکلاتی که روش‌های میدانی به همراه دارندبه سراغ روش‌های اصطلاحاً غیرمخل بروند. سنجش غیرمخل[2] بدون مزاحمت با انتشار کتابی با عنوان «سنجش‌های غیرمخل: تحقیق بدون واکنش در علوم اجتماعی»نوشته اوجین ورت و همکاران توسعه پیدا کرد. این شیوه را عموماً نه به عنوان روشی جایگزین، بلکه مکمل روش‌های دیگر دانسته‌اند که هم به اجتناب از مشکلات اخلاقیتحقیقات میدانی کمک می‌کند و هم تأثیر محقق بر افراد مورد مطالعه را از بین می‌برد(بیکر، ۱۳۹۶). با این نگاه می‌توان پژوهش با کمک کلان‌داده‌ها را در ذیل روش‌های غیرمخل قرار داد. اما نکته مهم اینجاست که کار با کلان‌داده‌ها نیاز به مهارت‌های آماری، کار با برنامه‌های کامپیوتریبرای نظم بخشی به داده‌ها، آشنایی با بسترهای تولید داده و امثال آن دارد. از این جهت به لحاظ سهولت می‌توان کار با کلان‌داده‌ها را حتی دشوارتر از تحقیقات مرسوم میدانی دانست که در ادامه به برخی از این مسائل اشاره می‌شود.

---

[1] Documentary research
[2] Unobtrusive measure

➢ چالش‌های کار با کلان داده

بحثی که در اینجا باید بدان اشاره کرد چالش‌های کار با کلان داده است که توضیحاتی پیرامون آن خواهد آمد.

الف) چالش‌ها در ایران

- بخشی از فضای تولید کلان‌داده‌ها مربوط به شبکه‌های اجتماعی و پیام‌رسان‌های مختلفی است که کاربران روزانه اطلاعات مختلفی را از طریق آنها منتشر می‌کنند. همانطور که شواهد نشان می‌دهد به دلیل مسائل امنیتی در بسیاری از این بسترها امکان حضور افراد بدون استفاده از ابزارهای عبور از تحریم و یا فیلترشکن‌ها وجود ندارد. استفاده از شبکه‌های اجتماعی و بسترهای تولید داده توسط کاربران همچون فیس‌بوک، توئیتر، تلگرام، تیک‌تاک، واتساپ، اینستاگرام و ... به طور قانونی ممنوع است. افرادی که در فضای غیررسمی از این بسترها استفاده می‌کنند نیز با دشواریدسترسیمواجه هستند. از این رو برای محققان اجتماعی نیز مشکلاتی برای دسترسی به این داده‌ها به‌وجود آمده است.
- آمارها نشان می‌دهد کاهش حضور افراد در این بسترها بعد فیلترینگ آنها محسوس است و از این رو می‌توان گفت بهره‌گیری از داده‌های این بخش، محققان را با مشکل داده‌های ناقصمواجه می‌کند. این موضوع می‌تواند بحث تعمیم‌پذیری نتایج تحقیقات را تحت‌الشعاع قرار دهد. در واقع در تحلیل حداقل بخشی از فضای تولید کلان‌داده، نمی‌توانشاهد حضور همه افراد بود.
- یکی از چالش‌های دیگر این است که بخش زیادی از داده‌های کلان تولید شده در حوزه‌های مختلف در اختیار شرکت‌های دولتی و خصوصی قرار دارد. از این رو امکان دسترسی به داده‌ها به دلایل مختلف که بیان آنها خارج از حوصله این مقاله است،به راحتی میسر نیست. در موارد زیادی دستیابی به این داده‌ها تقریباً غیرممکن است.البته با توجه به سطح محرمانگی داده و نوع سازمان و ارگانی که داده‌ها را در اختیار دارد، میزان دسترسی به داده‌ها متفاوت می‌شود. با این حال تجربه نگارنده نشان می‌دهد در این حوزه محققان با مشکلات عدیده‌ای مواجه هستند.
- موضوع دیگر محدودیت‌های استفاده از روش‌های بدیع در نظام دانشگاهی کشور است. همانطور که اشاره شد ظهور و گسترش کلان‌داده‌ها موضوع جدیدی است. به همین انجام تحقیقات اجتماعی در این حوزه نیازمند به‌کارگیری روش‌های خلاقانه و نو است. اما در این بخش دو چالش عمده وجود دارد. از یک سو این روش‌های جدید به راحتی از سوی مراکز علمی دانشگاهی قابل پذیرش نیست. از سوی دیگر و به واسطه خصلت بین‌رشته‌ای کار با کلان‌داده‌ها، محققان علوم اجتماعی نیز تسلط و مهارت کافی برای کار با این داده‌ها را ندارند.

ب) چالش‌های عمومی

در کنار چالش‌های پیشین باید گفت به طور کلی استفاده از کلان‌داده‌ها در تحقیقات اجتماعی با مسائل مختلفی روبرو است که در ادامه به برخی از آنها اشاره می‌شود.

- یکی از این موارد رعایت حریم خصوصی و محرمانگیاطلاعات افراد است. از این منظر تولید داده‌های بزرگ

توسط افراد را می‌توان با سه مفهوم «اثر»، «رد پا»یا«اثرانگشت» از هم تفکیک کرد. «اثر» بدین معنی است که افراد در فضای تولید داده کاملاً ناشناس باقی می‌مانند. هرچند بسیاری از شرکت‌ها و بسترهای تولید کلان‌داده در مقام ادعا بر این اصل تأکید دارند، اما در عمل از اطلاعات افراد به شکل‌های مختلف استفاده می‌شود. در اینجا مفهوم «رد پا» مطرح می‌شود که به معنی شناخت تقریبی افراد تولید کننده داده است. بدین معنی که تا حدی مشخص است که این داده توسط چه افرادی و با چه ویژگی‌هایی تولید شده است. اما مفهوم «اثرانگشت» بدین معنی است که هر داده تولید شده توسط یک فرد کاملاً قابل ردیابی است و به راحتی شرکت‌های بزرگ می‌توانند از این داده‌ها استفاده تجاری و یا سیاسی نمایند. این موضوع در تضاد با مباحث مربوط به رعایت حریم خصوصی و عدم استفاده از داده‌ها بدون اجازه تولیدکننده آن قرار می‌گیرد. هرچند شاید بتوان به لحاظ فنی و نرم‌افزاری به اطلاعات افراد دسترسی کامل و با جزئیات پیدا کرد، اما نحوه استفاده از این داده‌ها، هدف محققان و شرکت‌ها از کار با کلان‌داده‌ها و نیز استفاده ابزاری از داده‌های مذکور است که می‌تواند منجر به نقض حریم خصوصی گردد. نمونه تحقیقاتی نیز وجود دارد که در آن علی‌رغم تلاش محققان برای حفظ گمنامی افراد، یافته‌ها به تفکیک فرد قابل شناسایی شده بود. این مسئله با مبحث امنیتداده‌ها نیز ارتباط پیدا می‌کند.

- یکی دیگر از چالش‌های موجود در کلان داده‌ها، عدم تناسب سیستم رایانه با سرعت تولید داده‌ها است.افزایش فزاینده حجم داده‌ها، مشکلات زیادی در جمع‌آوری، ذخیره‌سازی و مدیریت و حفظ امنیت داده‌ها ایجاد می‌کند. همچنین نرخ رشد اطلاعات به صورت نمایی است، درحالی‌که شیوه‌های پردازش اطلاعات به کندی پیشرفت می‌کنند (سنمی علمداری، ۱۳۹۵).امروزه به منظور حل چالش ذخیره‌سازی حجم عظیمی از اطلاعات از رایانش ابری استفاده می‌شود.اگرچه استفاده از رایانش ابری و ذخیره‌سازی اطلاعات در ابر تا حدی مشکل دسترسی به داده‌ها و ذخیره‌سازی اطلاعات را حل کرده است، اما هنوز امنیت اطلاعات در رایانش ابری به عنوان یک چالش مهم تلقی می‌شود. در اینجا برخی مباحث فنی نیز وجود دارد که از ذکر آن صرف نظر می‌شود.
- از جمله مباحث جدی در این حوزه بحث نمونه‌گیری و تعمیم نتایج تحقیقات مبتنی بر کلان‌داده‌ها است. مشخصاً در کار با کلان‌داده‌ها نمی‌توان از شیوه‌های مرسوم نمونه‌گیری استفاده نمود. زیرا از یک طرف همه افراد در تولید کلان‌داده‌ها نقش ندارند و از سوی دیگر برخی کاربران و شرکت‌ها حجم بالاتری از کلان‌داده‌ها را تولید می‌کنند. این موضوع مسائلی چون چارچوب نمونه‌گیری و حجم نمونه و نمایا بودن آن را با دشواری‌هایی مواجه می‌سازد. برخی استدلال می‌کنند در کار با کلان‌داده‌ها هرچند ممکن است نمونه‌گیری دقیق نباشد، اما حجم بالای داده‌ها تا حدی این مشکل را برطرف می‌کند. گروهی نیز به دنبال ابداع روش‌های رایانه‌ای برای رفع نقص داده‌هایکلان هستند؛ از جمله کتاب الما گرمید با این هدف نوشته شده است. به هر حال این چالش همچنان محل بحث است.
- حجم بالای این داده‌ها که می‌توان آن را نقطه قوت دانست، از جهاتی باعث بروز چالش‌هایی می‌شود. این ویژگی به قدری مهم است که در برخی تعاریف مبنا قرار می‌گیرد. به عنوان نمونه موسسه گارتنر کلان‌داده‌ها را دارایی‌های اطلاعاتی با حجم زیاد معرفی می‌کند (Minelli et al, 2012) و شرکت آی.بی.ام آن را داده‌هایی می‌داند

که حجم آنها فراتر از حدی است که بتوان با نرم‌افزارهای رایج به مدیریت، ذخیره و تحلیل آنها پرداخت.گانتز و رینسل نیز آن را داده‌هایی حجیمی اطلاق می‌کنند که ذخیره‌سازی، پردازش و تجزیه و تحلیل آنها از طریق فناوری‌های پایگاه داده سنتی دشوار است (Gantz & Reinsel, 2012). به عنوان شاهد مثال باید اشاره کرد فیس‌بوک روزانه ۶۰ ترابایت اطلاعات جدید را بارگیری می‌کند، یاهو دارای ۱,۴ میلیارد گره وب یا گره‌های گرافیکی است، مایکروسافت در حدود ۱,۱۵ میلیارد جفت پرس و جو داشته است، و گوگل ۲۰ پتابایت در روز پردازش می‌کند (نقیب‌السادات، ۱۴۰۱). با این توضیح باید توجه داشت که این حجم بالا خود چالشی پیش روی محققان برای تحلیل داده‌ها است.

- چالش دیگر نقض هنجارهای رقابتی است. بهره‌گیری از فرایند یادگیری ماشینی و هوش مصنوعی برای پردازش کلان‌داده‌ها، موجب می‌شود تا بنگاه‌ها محصولات و خدمات باکیفیت بهتر و بهای کمتر را در کوتاه‌ترین زمان ممکن ارائه دهند که تداوم این امر به سود فضای کلی رقابت است (یان، ۲۰۱۹).برای مثال، گوگل با استفاده از داده‌های کاربران، مرتباً عملکرد موتور جستجوگر خود را بهبود بخشیده است. شرکت تسلا به کمک داده‌های زیادی که از سیستم ناوبری اتوموبیل‌های خودرانش جمع‌آوری می‌کند، به بهینه‌سازی و اصلاح الگوریتم‌ها و بهبود مداوم عملکرد ماشین‌های خودران خود پرداخته است؛ در حالی که برای رقبایی که به این حجم از داده‌ها دسترسی ندارند، دستیابی به موقعیت رقابتی مشابه، از طرق معمول، دشوار خواهد بود.در حوزه پژوهش‌های اجتماعی نیز انحصار مذکور وجود دارد و این موضوع به گروهی اجازه تحقیقات بیشتر در این حوزه را می‌دهد.

- از سوی دیگر تبانی به کمک کلان‌داده‌ها و هوش مصنوعی با تشکیل کارتل‌های دیجیتال همراه است که در طی آن، بنگاه‌ها به جای استفاده از شیوه‌های معمول، از امکانات مذکور سود می‌جویند. سوءاستفاده از موقعیت مسلط به کمک کلان‌داده‌ها نیز باعث می‌شود برخی شرکت‌ها کاربران خود را به‌طور صریح یا ضمنی درخصوص نحوهٔ گردآوری و استفاده از داده‌های شخصی‌شان ورویه‌های مرتبط با حریم خصوصی فریب دهد و آنها را ترغیب به استفاده از پلتفرم خودکند تا با این تمهید خدعه‌آمیز، به داده‌های بیشتری دست یابد. در حالی که اگر کاربراناز واقعیت امر اطلاع داشته باشند، شاید حاضر به استفاده از آن پلتفرم نشوند. (رهبری، ۱۴۰۱). هرچند این موارد عمدتاً جنبه حقوقی و اقتصادی پیدا می‌کند، اما به جهت مسائل اخلاقی که در پی دارد، می‌توان به عنوان یک مسئله اجتماعی به آن نگریست. از سوی دیگر پژوهش‌های اجتماعی بر روی کلان‌داده‌ها ممکن است در خدمت شکل‌دهی به این رقابت‌های ناسالم، فریب و سوءاستفاده از داده‌ها قرار بگیرد.

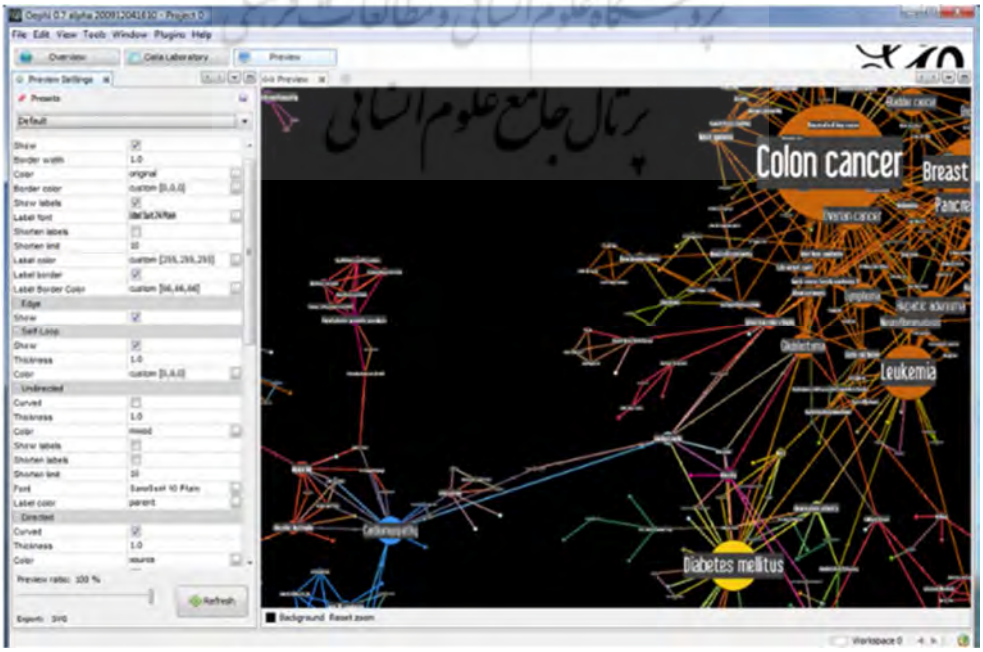
- در نهایت باید به چالش دستکاری افکار عمومی به واسطه تحقیقات اجتماعی بر روی کلان‌داده‌ها اشاره کرد. این دیدگاه مبین الگوریتم‌های نظم‌دهنده تولید داده در فضای شبکه‌های اجتماعی است که تا حدی خودمختاری و کنشگری افراد را زیر سؤال می‌برد. در واقع با پذیرش این نکته که صاحبان بسترهای تولید کلان‌داده با شگردهای اجتماعی، اقتصادی و روانشناختی سعی در جهت‌دهی به رفتار افراد دارند، مسئله صحت داده‌ها مورد سؤال قرار می‌گیرد. این موضوع همچنین مسائل اخلاقی مختلفی را در استفاده از کلان‌داده‌ها به‌وجود می‌آورد.

➢ نرم‌افزارهای کار با کلان داده

قبل از توضیح پیرامون برخی نرم‌افزارها از جمله گفی و پاژک، باید اشاره کرد که برخی از انواع این نرم‌افزار در تحلیل داده‌های شبکه‌های اجتماعی مورد استفاده قرار می‌گیرند. از این رو برخی مفاهیم مشترک دارند. یکی از این مفاهیم گره‌ها (Node) و یال‌ها (Edge) هستند. گره به هر موجودیت دارای ویژگی‌های مورد توجه در تحقیق اشاره دارد؛ مواردی چون فرد، شیء، کالا، کاربر، مکان، رویداد و ... . یال نیز ویژگی‌های مشترک و مشابه و مورد سؤال در تحقیق می‌باشد که پیونددهنده گره‌ها است؛ ویژگی‌هایی چون جنسیت، میزان تخفیف کالا، زمان رخداد، محل وقوع و ... . همچنین تحلیل شبکه هم به شکل آماری انجام می‌شود و هم به شکل بصری و همراه با گراف (شبکه). با این توضیح اهداف تحلیل شبکه‌ها مواردی چون تشخیص نودهای متصل به هم، فاصله بین نودها، میزان اهمیت نودها به واسطه موقعیت‌شان در شبکه، شناخت انجمن‌ها در شبکه، نحوه ایجاد شبکه، شیوه انتشار اطلاعات، شکل‌گیری عقاید در شبکه و مانند آن است.

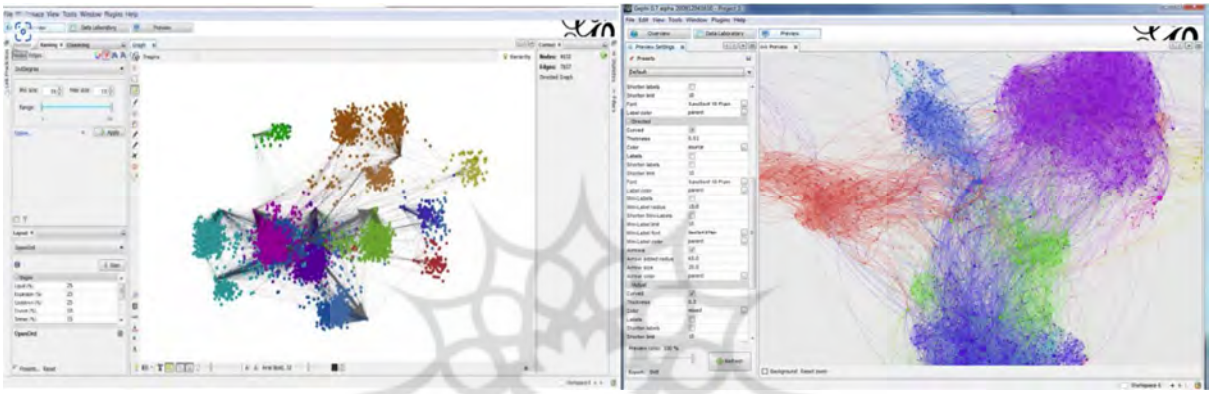
۱) نرم‌افزار Gephi

گفیابزاری برای بصری‌سازی اطلاعات و داده‌های دریافتی از شبکه‌های اجتماعی و سیستم‌های پیچیده است. این نرم‌افزار امکان ترسیم و بهینه‌سازی شکل و رنگ نمودارها و کشف الگوهای پنهان را میسر می‌سازد. در این نرم‌افزار الگوریتم‌هایی برای تشخیص گروه‌ها، آشکارسازی خوشه‌ها و شناخت انجمن‌های مختلف در دل گراف شبکه‌ها وجود دارد. گفی را باید نرم‌افزاری برای اکتشاف بصری و تجزیه‌وتحلیل شبکه‌ها دانست. گفی با مد نظر قراردادن پیوند بین گره‌ها، انواع شبکه‌ها از جمله تیپ‌شناسی اینترنت، شبکه‌های اشتراکی فایل‌ها، شبکه‌های بیولوژیکی، شبکه‌های آنلاین و ارتباطی، شبکه‌های دوستی، شبکه‌های سازمانی، شبکه‌های معنایی و... تحلیل می‌کند. در تصویر زیر نحوه انتشار یک بیماری نشان داده شده است.



تصویر۱: شبکه ترسیم شده از نحوه شیوع یک بیماری

بصری‌سازی در این نرم‌افزار به صورت زمان واقعی صورت می‌گیرد؛ بدین معنی که کاربر می‌تواند این مراحل را مانند یک انیمیشن تماشا کند و از این طریق امکان بررسی رفتار یک شبکه در طول زمان میسر می‌شود. این نرم‌افزار قابلیت ترسیم گراف (شبکه) تا ۱۰۰۰۰۰ گره و ۱۰۰۰۰۰۰ یال را دارد. گفی الگوریتم‌های متعددی با کارکردهای مختلف دارد که قابلیت ویرایش و تغییر گراف‌ها حین اجرا را فراهم می‌کند. به لحاظ ظاهری نیز تنظیمات گوناگونی در نرم‌افزار وجود دارد که می‌توان ضخامت، رنگ، مکان، شکل و نزدیکی یال‌ها و گره‌ها را تغییر داد. رابط کاربری گفی در حال تحلیل یک شبکه وسیع در تصویر زیر آمده است.



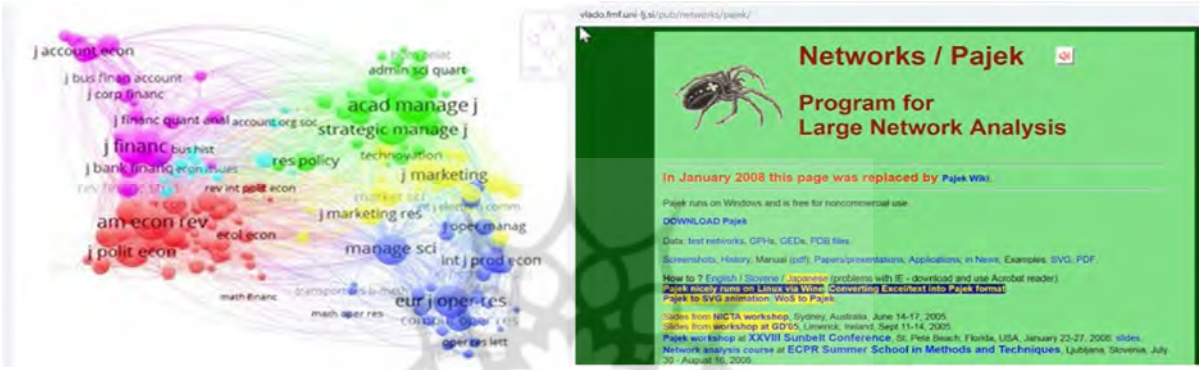
تصویر۲: رابط کاربری گفی با سه بخش مجزا در تحلیل یک شبکه وسیع

۲) نرم‌افزار Pajek

پاژک شباهت‌هایی با نرم‌افزار گفی دارد. این نرم‌افزار امکان ترسیم شبکه و تغییر مشخصات گره‌ها و یال‌ها را دارد. پاژک به کاربر امکان می‌دهد که انجمن‌ها را شناسایی نماید. خوشه‌بندی شبکه و محاسبه مرکزیت آن از دیگر ابزارهای پاژک است. همچنین این نرم‌افزار امکان بررسی نحوه انتشار یک پدیده در شبکه‌ای مشخص را فراهم می‌آورد. رابط کاربری این نرم‌افزار در تصویر ۳ مشاهده می‌شود.

تصویر۳: رابط کاربری پاژک

در پاژک بخش آماری نیز وجود دارد که کاربر می‌توان آن را ویرایش کند و روابط بین گره‌ها را به شکل ماتریس اعداد تحلیل نماید. این نرم‌افزار الگوریتم‌های مختلفی برای نمایش شبکه‌ها دارد که بیشتر جنبه بصری داشته و به کاربر کمک می‌کند از زوایای مختلف، داده‌ها را بررسی نماید. این نرم‌افزار نیز مانند گفی در حالت رخداد در زمان (Time events) تغییرات یک شبکه را در طول کار با داده‌ها مشخص می‌کند. ساخت انواع شبکه‌های تصادفی و دستی نیز از دیگر قابلیت‌های پاژک است. در این نرم‌افزار امکان دانلود نمونه داده از سایت آن برای تمرین وجود دارد. همچنین قابلیت ایجاد شبکه دستی برای کار با بخش‌های مختلف نرم‌افزار در آن ایجاد شده است.



تصویر۴: سایت پاژک و ترسیم یک شبکه علمی به کمک نرم‌افزار

۳) رابط‌های API

یکی از شیوه‌های استخراج الگوهای رفتاری کاربران مورد مطالعه از دل داده‌های کلان موجود، بهره‌گیری از داده‌کاوی و استفاده از نرم‌افزار تولیدشده از سوی پژوهشگران است که عموماً با زبان‌های معمول برنامه‌نویسی از جمله پایتون تولید می‌شوند.

این ابزارها یا به اصطلاح اِی‌پی‌آی‌ها (Application Programming Interface) رابط‌هایی نرم‌افزاری هستند که امکان دسترسی به شبکه‌های اجتماعی را برای نرم‌افزارهای دیگر فراهم می‌کنند و به دیگر برنامه‌ها اجازه می‌دهند که با آن ارتباط داشته باشند (افتاده، ۱۳۹۵). رابط برنامه‌نویسی کاربردی مجموعه‌ای از کدهای برنامه‌نویسی است که برای پرس‌وجوی داده، تجزیه جواب‌ها و ارسال دستورالعمل‌ها در بینیکیا چند پلتفرم نرم‌افزاری مورد استفاده قرار می‌گیرد.از آنجایی که غالباً محققان علوم اجتماعی با زبان‌های برنامه‌نویسی آشنایی کافی ندارند، در این بخش می‌توان از چت‌بات‌های هوش مصنوعی از جمله چت جی‌پی‌تی و یا بینگ برای کدنویسی کمک گرفت.

۴) ابزارNgram و Trends

انگرام (Google Ngram) یکی از خدمات گوگل محسوب می‌شود که امکان جست‌وجو در میان پرطرفدارترین کلمات به کار رفته در کتاب‌های گردآوری شده در کتابخانه مجازی گوگل را فراهم می‌کند. قابلیت دیگر این سرویس امکان جستجو و تعداد تکرار هر عبارت در متن کتاب‌های لاتین ۵۰۰ سال اخیر است. علاوه بر این می‌توان در یک بازه زمانی

مشخص، در رابطه با یک جمله خاص و یا عبارت مشخص در طول سخنرانی‌های افراد مطرح جهان نیز تحقیق کرد. نتایج تمام این جستجوها بر اساس یک نمودارارائه می‌شود. این نرم‌افزارنشان می‌دهد کلمات در گذر زمان چه تغییراتی می‌کنند و به‌کمک آن می‌توان هر کلمه‌ای را در بیش از ۵ میلیون کتابی که از سال ۱۵۰۰ میلادی تا کنون منتشر شده‌اند از نظر نوع استفاده در جملات و تغییراتی که با گذشت زمان در آن‌ها به‌وجود آمده جستجو نمود.ابزار انگرام میزان فراوانی یک کلمه یا عبارت را در کتب منتشر شده نشان می‌دهد. البته این ابزار از زبان فارسی پشتیبانی نمی‌کند، اما می‌تواند برای منابع انگلیسی بسیار مفید باشد. مزیت این ابزار در آن است که می‌توان همزمان چندین ویژگی عبارت را جستجو کرد.

ابزار دیگری که به انگرام شباهت دارد گوگل ترندز (Google Trends) است.با این ابزار می‌توان روند داغ‌ترین جستجوها در سطح جهانی را بررسی نمود و با استفاده از مرور بر اساس تاریخ یا جستجوهای برتر در دسته‌بندی‌های مختلف پرجستجوترین کلمات را حتی بر اساس آی.پی هر کشور مشاهده نمود. علاوه بر این گوگل ترندز به سوالات کاربر در مورد انتخاب کلمات کلیدی در کشورها، مناطق و شهرهای مختلف دنیا پاسخ داده و امکان تحلیل مناسبی در مورد مقایسه کلمات کلیدی ایجاد می‌کند. این ابزار نشان می‌دهد در هر لحظه افراد مختلف دنیا به چه چیزهایی بیشتر اهمیت می‌دهند و به آن توجه می‌کنندو این موارد در طول زمان چه تغییراتی دارند. نمونه کار با این دو ابزار در تصویر ۵ آمده است.

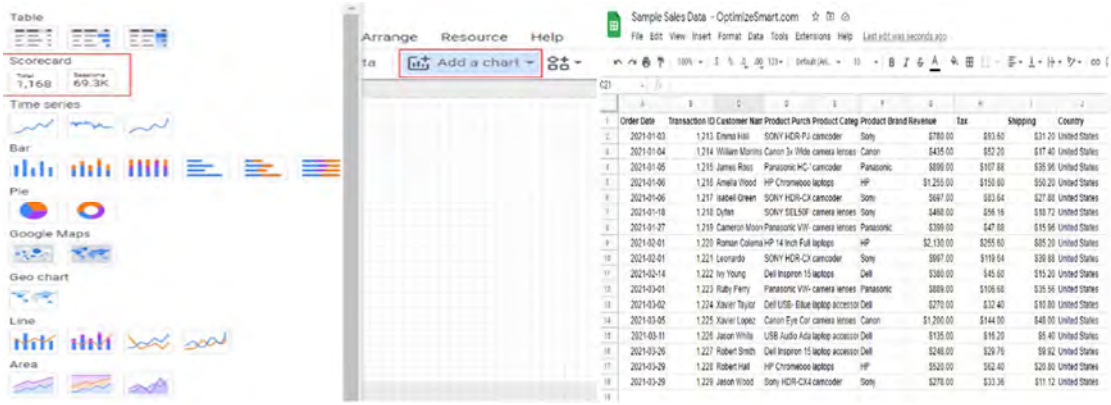
تصویر۵: روند تغییرات واژگان و عبارات در کتب (سمت راست؛ انگرام)و جستجوها (سمت چپ؛ ترندز)

۵) ابزار Looker Studio

لوکر استودیوکه سابقاً با نام گوگل دیتا استودیو (Google Data Studio) شناخته می‌شد، ابزاری برخط برای تبدیل داده‌ها به گزارش‌ها و داشبوردهای اطلاعاتی قابل تنظیم است. لوکر استودیو طیف وسیعی از ویژگی‌های ضروری را برای تجسم و مصور کردن[۱۷]داده‌ها، و ادغام و یکپارچگی آن با منابع داده‌های مختلف ارائه می‌دهد.این ابزار ترسیم نمودارهای گرافیکی و جداول تعاملی کاربردی از هر داده و تلفیق داده‌ها از منابع مختلف را فراهم می‌کند. تجسم داده‌ها

[۱۷] visualization

به تفسیر و حفظ آنها کمک می‌کند. محققان برای اینکه بتوانندیافته‌های مبتنی بر داده خود را معنادارتر و متقاعدکننده‌تر کنند، باید هنر داستان‌گویی را با تجسم داده‌ها بیاموزند. نمونه‌ای از داده‌های عددی و بصری در تصویر ۶ دیده می‌شود.



تصویر۶: محیط کار با لوکر استودیو

این سرویس می‌تواند از طریق بیش از ۶۰۰ کانکتور شریک به حداکثر ۸۰۰ کانال داده مختلف متصل شود و دسترسی فوری به تقریباً هر نوع داده‌ای را بدون نیاز به کد نوشتن یا نرم‌افزار امکان‌پذیر کند.لوکر استودیو داشبوردهای مشترکی را برای شرکت‌های بزرگ با حجم قابل توجهی از داده ارائه می‌دهد. این ابزار به کسب‌وکارها برای پیدا کردن منابع ترافیک سایت و رتبه‌بندی آنها کمک می‌کند. لوکر استودیو با تبدیل اطلاعات مربوط به رفتار کاربران به نمودارهای دقیق این فرصت را به محقق می‌دهد تا فعالیت‌های مختلف را به واضح ترین شکل ممکن مشاهده و حتی رفتار آینده کاربران در فضای اینترنت را پیش‌بینی نماید.

هرچند این ابزار در حوزه بازاریابی، کارآفرینی و کسب‌وکار بیشتر شناخته شده است، اما در تحقیقات اجتماعی مبتنی بر داده‌های بزرگ نیز می‌توان از آن بهره برد.

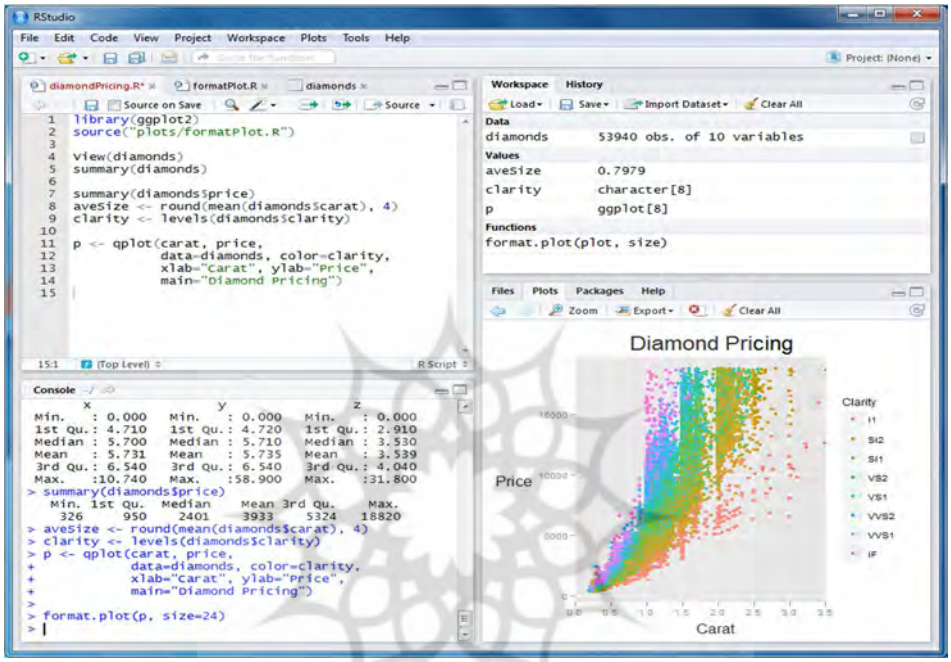
۶) نرم‌افزار R

R یکی از قدرتمندترین و پرطرفدارترین نرم‌افزارهای آماری در سراسر دنیا میباشد که با رشد چشمگیری مورد توجه دانشجویان و محققان در زمینه‌های مختلف قرار گرفته است. یکی از حوزه‌های کاربردی برای استفاده از این نرم‌افزار تحلیل کلان‌داده‌ها است.

این زبان برنامه‌نویسی و محیط نرم‌افزاری آن یعنی R Studio به دلیلِ سادگی و دارا بودنِ پکیج‌های مختلفِ آماری و داده‌ای، کاربردِ فراوانی در بین متخصصینِ علوم‌داده پیدا کرده است. زبانِ R هم مانند پایتون جز پرکاربردترین زبان‌ها در حوزه تحلیل داده، یادگیری‌ماشین و همچنین یادگیری عمیق است. این زبان قابلیت اتصال به پایگاه داده‌های مختلف را داشته و ابزارهای قدرتمندی نیز جهت مصورسازیداده‌ها دارد.نرم‌افزار مذکور برای متن‌کاوی و داده‌کاوی اطلاعاتِ گسترده و وسیع نیز مورد استفاده قرار می‌گیرد. با استفاده از زبان R و پکیج‌های مختلفِ آن، به جای اینکه

الگوریتم‌ها و روش‌های مختلفِ داده‌کاوی را یکی‌یکی پیاده‌سازی کنیم، می‌توانیم به راحتی از نسخه پیاده‌سازی شده و قدرتمند این الگوریتم‌ها استفاده کنیم.

نرم‌افزار آماری Rتا سال‌های قبل قادر به تجزیه و تحلیل کلان‌داده‌ها نبود،اما امروزه با توجه به شیوه‌های متنوعی که برای تحلیل داده‌های کلان ابداع شده است، یکی از پراستفاده‌ترین نرم‌افزاری تحلیلی محسوب می‌شود (قربانی، ۱۳۹۵). محیط این نرم‌افزار در تصویر زیر قابل مشاهده است.



تصویر۷: محیط کار با نرم‌افزار R

این ابزار بسته‌های آماده متعددی دارد (بالغ بر ۵۰۰۰ پکیج الحاقی) که برای انجام تحقیق روی کلان‌داده‌ها می‌توان از آن استفاده کرد. بنابراین کار با آن الزاماً نیاز به تسلط بر برنامه‌نویسی ندارد.یادگیریR، تحلیلگر داده را تا حد زیادی از یادگیری ابزارهایدیگربی‌نیاز می‌کند و او را در یادگیری مفاهیم و روش‌ها و به کار بردن ایده‌های جدید متمرکز نگه می‌دارد و همیندلیلنقش

به‌سزایی در برانگیختن و رشد و تحلیلگران داده دارد.

۷) فناوری Hadoop

در تحلیل کلان‌داده استفاده از روش‌های سنتی برای ذخیره‌سازی و پردازش، کاری وقت‌گیر و هزینه‌بر است. به همین دلیل فناوری‌هایی مانند هادوپ از طریق برقرار کردن امکان ذخیره‌سازی هر نوع داده در یک محیط توزیع شده و پردازش آن‌ها به صورت موازی به کمک محققان آمده‌اند (جعفری، ۱۴۰۱: ۲۱).هادوپ یک زبان رایانش موازی با تقسیم داده‌های حجیم بین چند کامپیوتر است. هادوپ یک سکوی نرم‌افزار متن‌باز می‌باشد که از طریق آن می‌توان به سادگی حجم بسیار بزرگی از داده‌ها را ذخیره، تحلیل و پردازش کرد. برای مثال هادوپ می‌تواند برای کارهایی نظیر

شاخص‌گذاری صفحات وب، داده‌کاوی، تحلیل فایل‌ها، یادگیری ماشین، تحلیل‌های مالی، شبیه‌سازی‌های علمی و یا تحقیقات در زمینه بیوانفورماتیک مورد استفاده قرار بگیرد (افسری شولی و همکاران، ۱۳۹۵). سکوی هادوپ این امکان را دارد که مجموعه داده‌هایی با حجمی در مقیاس چندین پتابایت را مدیریت و پردازش کرد. پردازش موازی داده‌ها و وجود یک سیستم فایل توزیع شده این امکان را فراهم می‌کند که بتوان با سرعت بیشتری داده‌ها را پردازش کرد(یوسفی کندول، ۱۳۹۷).

چالش‌های مختلف کار با کلان‌داده منجر به ظهور پلتفرم‌های جدیدی مانندHadoop Apache شده است که می‌تواند مجموعه داده‌های بزرگ را به راحتی مدیریت کند. همچنین بسته نرم‌افزاریRHadoopکه برای نرم‌افزار R طراحی شده است، محقق را قادر می‌سازد که با استفاده از تکنیک‌های آماری داده‌های کلان را تحلیل کند.

بحث و نتیجه‌گیری

کلان‌داده پدیده جدیدی است که روزبه‌روز در حوزه‌های مختلف تولید شده و نیاز به پردازش و تحلیل دارد. دلایل مختلفی وجود دارد که کلان‌داده را برای پژوهش اجتماعی بسیار ارزشمند می‌سازد. نخست اینکه پژوهشگران را قادر می‌سازد تا در مورد پیش‌بینی‌ها و یافته‌هایتحقیقاتمختلف، قضاوت کنند. دوم اینکه منابع کلان‌داده، اندازه‌گیری‌های دقیق برای سیاست‌گذاری را فراهم می‌کنند. درنهایت، منابع کلان‌داده به پژوهشگران در برآورد علی پدیده‌ها بدون اجرای آزمایش، کمک می‌کنند. از سوی دیگر منابع کلان‌داده پژوهشگران را قادر می‌سازند تا به نظریه‌پردازی تجربه‌محور بپردازند؛ یعنی پژوهشگران با انباشت محتاطانه واقعیت‌های تجربی، الگوها و مسائل می‌توانند نظریات جدیدی بسازند. این نگاه همان شکل دادن به نظریه زمینه‌ای و داده‌محور است (سالگانیک، ۱۴۰۰: ۸۸).

با این حال استفاده درست از هرکدام از این مزیت‌ها نیازمند چیزهایی بیشتر از داده هستند. در واقع استفاده از کلان‌داده نیازمند پژوهشگرانی استکه بتوانند پرسش‌های جذاب، خلاقانه و مهمی مطرح کنند. این موضوع زمانی میسر می‌شود که محققان با روش‌های عملی کار با کلان‌داده‌ها آشنا باشند. مقاله حاضر سعی نمود ضمن اشاره جایگاه کلان‌داده‌ها در فرایند انجام پژوهش‌های اجتماعی، به چالش‌های پیش‌رو در این حوزه اشاره کند. چالش‌های کنونی در این حوزه از پژوهش، برای ارائه نتایج منطقی و قابل اتکاء باید رفع گردد؛ این امر بر عهده پژوهشگران این عرصه است. در این پژوهش تلاش شد با معرفی تعدادی از بسترها و نرم‌افزارهای کاربردی در حوزه تحلیل کلان‌داده‌ها، برای محققان علوم اجتماعی مسیری گشوده شود. بدیهی است این موارد بخشی از دنیای بزرگ و دائماً روی به تغییر و پیشرفت فناوری است.

دانش مقتضی در این عرصه از طریق فعالین این حوزه‌های علمی باید گسترش یابد. تغییر رویکردهای پژوهشی از تمرکز بر روی خُردروش‌ها به کلان‌روش‌ها باید مورد توجه قرار بگیرد. با روش‌های جدید کار با کلان‌داده‌ها نه تنها می‌توان روندها را شناسایی نمود، بلکه می‌توان تغییرات جدید در عرصه جامعه را نیز مورد ارزیابی قرار داد. در صورت عدم تحلیل دقیق و شناخت محتواهای عظیم تولید شده به زعم دانیل بل سوخت اطلاعات رخ می‌دهد (نقیب‌السادات، ۱۴۰۱)

همانطور که ذکر شد کلان‌داده‌ها ویژگی‌های مختلفی دارند که در پژوهش بدان‌ها اشاره شد. باید توجه داشت که برخی از این ویژگی‌ها برای تحقیقات اجتماعی مطلوب هستند؛ یعنی عموماً و نه همیشهبه نفع پژوهش اجتماعی عمل می‌کنند. «بدون واکنش»بودن، این چالش پژوهش‌های اجتماعی که وقتی افراد مطلع می‌شوند که مورد مشاهده هستند، رفتارشان را تغییر می‌دهند، مرتفع می‌شود؛ چالشی که در روش تحقیق بدان اثر هاثورن نیز گفته می‌شود.«همواره روشن»بودن مزیت‌هایی چون مطالعه پیشامدهای غیرمنتظره و نیز برآوردهای بی‌درنگ و در زمان واقعی را ممکن می‌کند.«حجیم» بودن نیز محقق را قادر به مطالعه پیشامدهای کمیاب، مطالعه ناهمگونی و مطالعه تفاوت‌های کوچک می‌سازد.

از سوی دیگر نباید از نظر دور داشت که برخی ویژگی‌های کلان‌داده غالباً برای پژوهش نامناسب در نظر گرفته می‌شوند؛ در واقع این ویژگی‌هاهم عموماً و نه همیشه، مشکلاتی را برای پژوهش ایجاد می‌کنند. «حساس»بودن می‌تواند مسائل مرتبط با حریم خصوصی را تشدید کند.«کثیف» بودن ممکن است محقق را دچار خطا نماید. «تحت طلسم الگوریتم» بودنباعث القای الگوهای رفتاری خاص به واسطه مهندسی پیشرفته سیستم‌های کلان‌داده که عموماً از دید متخصصان علوم اجتماعیپنهان است، می‌شود. «شناور» بودناین مسئله را ایجاد می‌کند که هر الگوی بدست آمده از کلان‌داده، هم می‌تواند از تغییر مهمی در جهان واقعی حاصل شده باشد و هم می‌تواند حاصل تغییر کاربران یا سیستم باشد. «نامعرف»بودن برای پرسش‌هایی که به تعمیم یافته‌ها به کل جمعیت نیاز دارد، مشکل جدی شمرده می‌شود. «خارج از دسترس»بودن،مشکلات جدی در مسیر تحقیقبرای پژوهشگران به‌وجود می‌آورد.در نهایت «ناقص» بودن می‌تواند منجر به از دست دادن سه نوع اطلاعات مهم برای پژوهشگران شامل اطلاعات جمعیت‌شناختی شرکت‌کنندگان، رفتارشان در پلتفرم‌های دیگر و داده لازم برای عملیاتی‌سازی سازه‌های نظری بشود.

بنابر آنچه ذکر شد هرچند منابع کلان‌داده در همه جا وجود دارند و به شکل‌های مختلف می‌توان به آن دسترسی پیدا کرد، اما بهره گرفتن از آنها در راستای انجام پژوهش‌های اجتماعی به مهارت بالایی نیاز دارد.همانطور که سالگانیک اشاره می‌کند «در کار با کلان‌داده ناهارمجانی نداریم و اگر برای جمع کردن داده‌ها سخت‌کوشی نکنید، برای تحلیل آن باید زمان زیادی اختصاص دهید.»

در پایان باید اشاره کرد یکی از زمینه‌های تقویت این عرصه، آموزش روش‌های جدید پژوهش و تحقیق در کلان‌داده‌ها به محققان علاقه‌مند به این عرصه است. همچنین وارد کردن آموزه‌های جدید به سرفصل‌های درسی در برنامه بازنگری دروس دوره‌های کارشناسی، کارشناسی ارشد و دکتری رشته‌های مرتبط با علوم اجتماعی توسط وزارت علوم، تحقیقات و فناوری پیشنهاد می‌شود. تأمین منابع آموزشی، کتب درسی و راهنماهای اجرای طرح‌های تحقیقاتی با روش‌های جدید در عرصه وب‌کاوی، داده‌کاوی و متن‌کاوی از دیگر پیشنهادهای قابل ارائه است. از سوی دیگر نحوه دسترسی پژوهشگران به داده‌های کلان با لحاظ کردم مسئله امنیت داده و حفظ محرمانگی آن، باید تسهیل گردد.

منابع

- ✓ افتاده، جواد. (۱۳۹۵)، تحلیل شبکه‌های اجتماعی. تهران: انتشارات ثانیه.
- ✓ افسری شولی، فاطمه؛ هارونآبادی، علی و شامحمدی، امین. (۱۳۹۵). تحلیلی از دادههای عظیم در شبکههای اجتماعی. دومین کنفرانس ملی رویکردهای نوین در مهندسی کامپیوتر و برق. رودسر: باشگاه پژوهشگران جوان و نخبگان.
- ✓ بیکر، ترز ال (۱۳۹۶)، نحوه انجام تحقیقات اجتماعی، ترجمه هوشنگ نایبی.تهران: نشر نی.
- ✓ جعفری، سحر (۱۴۰۱)، کلان‌داده و فناوری هدوپ، مطالعات علوم کاربردی در مهندسی، دوره هشتم، شماره ۳، ۲۹-۲۱.
- ✓ حیدری، الهام (۱۳۹۹)، بررسی کلان داده‌ها در شبکه‌های اجتماعی، نشریه تحقیقات جدید در علوم انسانی، دوره جدید، شماره ۲۶، ۱۰۹-۱۲۶.
- ✓ دسترنج، رویا؛ قاضی نوری، سیدسپهر؛ دسترنج، نسرین؛ شایان، علی (۱۳۹۸)، ارزیابی اکوسیستم کلان داده در ایران با استعاره از مدل ارزیابی اکوسیستم هزاره، نشریه پردازش و مدیریت اطلاعات،دوره ۳۴، شماره ۴، ۱۶۴۲-۱۶۱۳.
- ✓ رضایی، زهرا؛ میرحسینی، زهره؛ سپهر، فرشته (۱۴۰۰)، آیندهپژوهی تاثیر کلان داده بر مدیریت و خدمات کتابخانه‌های عمومی کشور و ارائه مدل راهبردی، نشریه علوم و فنون مدیریت اطلاعات، دوره ۷، شماره ۲، ۱۱۰-۸۲.
- ✓ روحانی، شادی؛ رشیدی، زهرا؛ فریدونی، سمیه (۱۳۹۸)، ارائه چارچوبی مفهومی برای به‌کارگیری کلان‌داده‌ها در سیاست‌گذاری آموزش عالی، نامه آموزش عالی، دوره دوازدهم، شماره ۴۵، ۱۴۶-۱۲۱.
- ✓ رهبری، ابراهیم (۱۴۰۱)، تحلیلی بر چالش‌های حقوق رقابتی کلان‌داده‌ها، نشریه تحقیقات حقوقی، شماره ۹۸، ۳۲۰-۲۹۵.
- ✓ سالگانیک، متیو جی (۱۴۰۰)، در آمدی بر علوم اجتماعی محاسباتی،مترجم:عبدالحسین کلانتری، محمدحسین قطبی، ابراهیم دهنوی، انتشارات سروش.
- ✓ سنمی علمداری، یعقوب. (۱۳۹۵). مروری بر کلان داده‌ها .اولین همایش ملی نگرشی نوین در مهندسی برق و کامپیوتر. کرمانشاه: دانشگاه آزاد اسلامی واحد کرمانشاه.
- ✓ صاحب، طاهره؛ فرزین، هادی (۱۳۹۶)، گزارش تحلیل اکوسیستم کسب‌وکارهای مبتنی بر کلان‌داده‌ها. پروژهٔ تدوین نقشهٔ راه کلان‌داده‌ها. تهران: مرکز تحقیقات مخابرات ایران.
- ✓ غفاری قدیر، ج.؛ روشندل اربطانی، ط.؛ ضیایی، م. (۱۳۹۲)، تدوین سناریوهای متصور برای آینده نهاد رسانه‌ای کتابخانه‌های عمومی ایران، تحقیقات اطلاع‌رسانی و کتابخانه‌های عمومی، دوره ۱۹، شماره ۳، ۷۴-۸۹.

- قربانی، حمید (۱۳۹۵)، تجزیه و تحلیل کلان داده با RHadoop ، همایش بین‌المللی کلان‌داده، دانشگاه کاشان.
- کوهزادی، فواد؛ بوداقی خواجه نوبر، حسین؛ علوی متین،یعقوبو قره بیگلو، حسین(۱۴۰۱)،طراحی مدل تجزیه و تحلیل رفتار مشتریان مبتنی بر کلان‌داده با استفاده از روش فراترکیب و دلفی، مطالعات رفتار مصرف‌کننده ، دوره نهم، شماره ۱، ۳۲-۵۴.
- گنجی، کیانوش (۱۳۹۵)، تجزیه و تحلیل داده‌ها؛ تجزیه و تحلیل کلان داده‌ها و کاربرد آن در حسابرسی صورت‌های مالی، نشریه حسابدار رسمی، شماره ۳۴، ۸۲-۸۵.
- نقیب‌السادات، سیدرضا (۱۴۰۱)، تحلیل شبکه‌های اجتماعی با روش داده‌کاوی وب (وبکاوی) داده‌کاوی روابط بین نسلی و ارزش‌های خانوادگی درشبکه‌های اجتماعی، نشریه علوم خبری، دوره ۱۱، شماره ۴، ۱۵۸-۱۷۳.
- یوسفی کندول، وحید. (۱۳۹۷). چهارچوب کاری برای پردازشهای عظیم، کنفرانس بینالمللی برق، کامپیوتر و مکانیک ایران. تهران: دبیرخانه دائمی کنفرانس.

- Beyer, M. & Laney, D. (2012). The Importance of 'Big Data': A Definition, 17. Available at: https://www.gartner.com/en/documents/2057415/the-importance-of-big-data-a-definition.
- Boyd, D. & Crawford, K. (2012). Critical questions for Big Data.Communication & Society, 15 (5), 662–679.
- Brown B. 2011. Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute. Available at: https://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/ big-data-the-next-frontier-for-innovation (accessed July 20, 2018).
- Chen, C. P. , & Zhang, C. Y. (2014). Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. Information Sciences, 275, 314-۳۴۷.
- Chen, M. , Mao, S. , & Liu, Y. (2014). Big data: A survey. Mobile networks and applications, 19(2), 171-209.
- Gantz, J. , & Reinsel, D. (2012). The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east. IDC iView: IDC Analyze the future, ۲۰۰۷(۲۰۱۲), ۱-۱۶.
- Minelli, M. , Chambers, M. , & Dhiraj, A. (2012). Big data, big analytics: emerging business intelligence and analytic trends for today's businesses. New Jersey: John Wiley & Sons.
- Shin D. H., & M. J. Choi. 2015. Ecological views of big data: Perspectives and issue. Telematics and Informatics 32 (2): 311-320.

- ✓ Stieglitz, S. , Mirbabaie, M. , Ross, B. , & Neuberger, C. (2018). Social media analytics–Challenges in topic discovery, data collection, and data preparation. International Journal of Information Management, 39, 156-1۶۸.
- ✓ Sun, Z. (2018). 10 Bigs: Big Data and Its Ten Big Characteristics. Retrieved from: https://www. researchgate. net/profile/Zhaohao_Sun/publication/322592851_10_Bigs_Big_Data_and_It s_Ten_Big_Characteristics.
- Tulasi, B. (2013). Significance of Big Data and analytics in highereducation. International Journal of Computer Applications, 68(14), 23–2۵.
- ✓ Wagner, E. & Ice, P. (2012). Data changes everything: deliveringon the promise of learning analytics in higher education.EDUCAUSE Review, July/August, Pp.33–4۲.
- ✓ Yun, John (2019), The Role of Big Data in Antitrust The GAI Report on the Digital Economy,; Available at: https://gaidigitalreport.com/wp-content/uploads/2020/11/YunThe-Role-of-Big-Data-in-Antitrust.