مجله‌ی بررسی‌های آمار رسمی ایران
سال ۳۰، شماره‌ی ۲، پاییز و زمستان ۱۳۹۸، صص ۵۱۹- ۵۳۰

کاربرد داده‌های آزاد در آمار رسمی (مطالعه موردی: داده‌های شبکه اجتماعی اینستاگرام)

آرش فاضلی* و سعید فیاض

مرکز آمار ایران

چکیده. مفهوم داده‌های آزاد (Open Data) مبتنی بر این ایده است که برخی از داده‌ها باید به‌صورت رایگان در اختیار همه قرار گیرد تا بتوانند آن را آن‌گونه که می‌خواهند استفاده، آزاد استفاده و منتشر کنند، بدون آنکه با محدودیت حق نشر، حق اختراع و یا سایر محدودیت‌ها مواجه شوند. با گسترش فناوری اطلاعات و ارتباطات هر روز داده های بیشتری تولید می شود و این موضوع فرصت مناسبی را برای مراکز آماری که متولی تولید آمارهای رسمی هستند فراهم می آورد تا با سرعت بیشتر و هزینه کمتر آمارهای رسمی را تولید نمایند. تغییر رویکرد مراکز آماری در نظام آماری مدرن به استفاده از منابع داده ایی جدید باعث شده است تا داده های آزاد مورد توجه قرار گیرد. آمارهای رسمی نیروی کار یکی از مهمترین آمارهای رسمی است که بصورت فصلی انجام می شود و نتایج آن همواره مورد توجه بوده است. از سوی دیگر با گسترش فناوری اطلاعات و ارتباطات و شبکه های اجتماعی شاهد رشد مشاغل بسیاری در این حوزه بوده ایم. در این مطالعه از داده های شبکه اجتماعی اینستاگرام برای برآورد و ابزاری جهت اعتبار سنجی داده های بخش اشتغال استفاده شده است. آموزن های آماری و نتایج بدست آمده نشان داد که ارتباط معنی داری بین آمارهای نیروی کار و مشاغل ایجاد شده در شبکه اجتماعی ایستاگرام وجود دارد و در آمارهای رسمی بخش نیروی کار، داده های شبکه های اجتماعی می تواند معیار مناسبی را برای صحت سنجی آمارهای فراهم آورد.

---

* نویسنده‌ی عهده‌دار مکاتبات
دریافت: ۱۳۹۹/۸/۱۷، پذیرش: ۱۳۹۹/۱۲/۱۳.

واژگان کلیدی: آمارهای رسمی، داده های آزاد، آمارهای نیروی کار، اشتغال پنهان، شبکه اجتماعی اینستاگرام.

۱- مقدمه: داده آزاد چیست؟

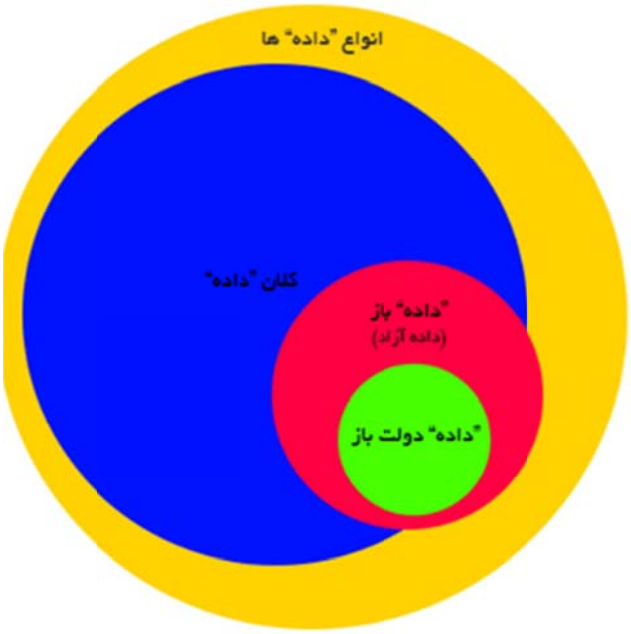
مفهوم داده‌های باز مبتنی بر این ایده است که برخی از داده‌ها باید به‌صورت رایگان در اختیار همه قرار گیرد تا بتوانند آن را آن‌گونه که می‌خواهند استفاده، باز استفاده و منتشر کنند، بدون آنکه با محدودیت حق نشر (Copyright)، حق اختراع(Patent) و یا سایر محدودیت‌ها مواجه شوند و مفاهیم مشابهی با دیگر جنبش های آزاد ، نظیر نرم افزارهای متن باز (Open Source) درست مشابه آنچه در سیستم عامل اندروید شاهد آن هستیم- یا محتوای آزاد(Open Content) دارد. در تعریف بالا سه جنبه مهم پررنگ شده است:

دسترسی: داده باید با هزینه معقول و به‌سادگی در دسترسی باشد؛ در بهترین حالت از طریق بارگیری کردن از اینترنت بتوان آن را به دست آورد. همچنین داده باید به‌گونه‌ای ارائه شود که بتوان در آن تغییراتی ایجاد کرد.

باز استفاده و بازنشر : داده باید تحت شرایطی منتشر شود که به افراد اجازه دهد آن را به شکل دلخواه استفاده و یا بازنشر کنند و یا بتوانند با سایر پایگاه‌های داده ترکیب کنند.

مشارکت همگانی : داده‌ها باید برای همه گروه‌ها و هر تلاشی در هر حوزه‌ای آزاد باشند. برای مثال، نباید حق دسترسی تنها محدود به فعالیت‌های غیرتجاری(Non-commercial) باشد؛ به این معنی که استفاده تجاری از داده‌ها محدود شود. و یا نباید استفاده از داده‌ها تنها برای اهداف خاصی مانند اهداف آموزشی آزاد باشد.

داده‌های باز اغلب از جنس بیگ دیتا (کلان داده Big data) هستند، با این حال مجموعه داده‌های کوچک نیز می توانند بصورت باز منتشر شوند. به همین دلیل این دو تعریف از هم متمایز هستند.



شکل۱- جایگاه داده های آزاد

۲- داده‌های آزاد در ایران

طبق قانون انتشار و دسترسی آزاد به اطلاعات، هر شخص ایرانی حق دسترسی به اطلاعات عمومی را دارد؛ مگر آنکه قانون، این دسترسی را منع کرده باشد و مؤسسات عمومی مکلف‌اند اطلاعات موضوع این قانون را در حداقل زمان ممکن و بدون تبعیض در دسترسی مردم قرار دهد. ابتکارهای عملی مانند سامانه ملی کاتالوگ و مجموعه داده‌های آزاد و کاربردی[۱] و سامانه داده‌نمای تهران[۲] نمونه‌ای از ابتکار عمل در این حوزه است. از مزیت‌های عمده این سامانه دسترسی به داده‌های خام، قابلیت خوانش داده‌ها توسط ماشین(Machine Readability) و دادن دسترسی به توسعه‌دهندگان برای سوار کردن برنامه کامپیوتری بر روی داده‌هاست(Application Programming Interface (API)) [۱].

۳- استفاده از داده های آزاد شبکه اجتماعی

استفاده از داده های آزاد در آمار رسمی توسط بسیاری از سازمان های آماری مورد توجه قرار گرفته است. در این خصوص می توان به استفاده از داده های موبایل برای تولید آمار رسمی اشاره نمود[۴] این موضوع با توجه به اهمیت آن در بخش آماری سازمان ملل منجر به ایجاد یک گروه کاری شده تا بر چگونگی استفاده از این منابع داده ایی برای تولید آمار رسمی بررسی های ویژه ایی انجام گیرد[۳]. منابع داده های آزاد کاربرد بسیاری زیادی می توانند در تولید آمارهای رسمی داشته باشند. این منابع داده می توانند به طوز کامل بعنوان جایگزین داده های جمع آوری شده در طرح های آمارگیری مورد استفاده قرار گیرند و یا بطور منابع کمکی یا جانبی مورد استفاده قرار گیرند. بعنوان مثال می توان به استفاده از داده های فروشگاه های اینترنتی بعنوان منابع داده ایی جایگزین برای تولید شاخص قیمت مصرف کننده ((Customer Price Index (CPI) استفاده نمود و از داده های شبکه های اجتماعی مانند اینستاگرام بعنوان منبع کمکی برای صحت سنجی یا برآورد اطلاعات بخش خاص استفاده نمود که در این مطالعه از این داده به منظور برآوردی از اشتغال پنهان و منبع کمکی برای آمارهای بخش اشتغال استفاده شده است.

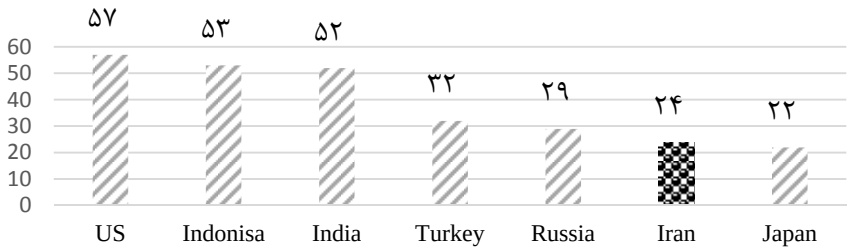
از جمله مهمترین منبع داده های آزاد، داده هایی هستند که روزانه در شبکه اجتماعی و در بستر اینترنت تولید می شوند. با گسترش فناوری اطلاعات و ارتباطات و زیر ساخت های دسترسی به اینترنت سیار و گسترش موبایل های هوشمند هر روز به تعداد کاربران شبکه های اجتماعی افزوده می شود . مهمترین عوامل این موضوع دسترسی آسان، بدون هزینه ، عدم وابستگی به مکان و سرعت بالای انتقال اطلاعات است. در چنین فضایی بسیاری از کاربران استفاده خود را تنها بعنوان یک شبکه ارتباطی محدود نکرده و بسیاری از فعالیت های کسب و کاری خود را در این شبکه ها توسعه داده اند. علاوه بر ویژگی های ارتباطی که این شبکه ها برای کاربران خود فراهم می نمایند در حوزه های کسب و کار امکانات دیگری نیز برای کاربران ایجاد نموده اند که از بسیاری را تشویق به ایجاد کسب و کار مجازی نموده اند.

که از مهترین این عوامل می توان به موارد زیر اشاره نمود.

۱) ایجاد کسب و کار (صفحه اینستاگرام، کانال تلگرام، صفحه فیسبوک و ...) برای افراد حقیقی /حقوقی نیاز به مجوز های قانونی ندارد

۲) تبلیغات در این بستر رایگان بوده و سرعت انتشار اطلاعات در این بستر بسیار بیشتر از روش های سنتی است
۳) رشد کاربران و مشتریان بصورت نمایی است
۴) کسب و کارهای ایجاد شده نیاز به فضای فیزیکی نداشته و می تواند هر از مکانی و هر زمانی شروع شود و به راحتی هم بسته شود.

شبکه اجتماعی اینستاگرام در این مطالعه مورد نظر است، بر طبق آمارهای رسمی منتشر شده در سایت های معتبر[۴] در سال ۲۰۱۹، ایران در رتبه ۷ ام از منظر تعداد صفحات شبکه اجتماعی اینستاگرام قرار گرفته است و همچنین با ۲۴ میلیون صفحه فعال در رتبه دوم پر کاربرترین شبکه اجتماعی بعد از تلگرام با ٤٠ میلیون کاربر قرار گرفته است.



شکل۲- کشور ایران از منظر تعداد صفحات فعال در شبکه اجتماعی اینستاگرام

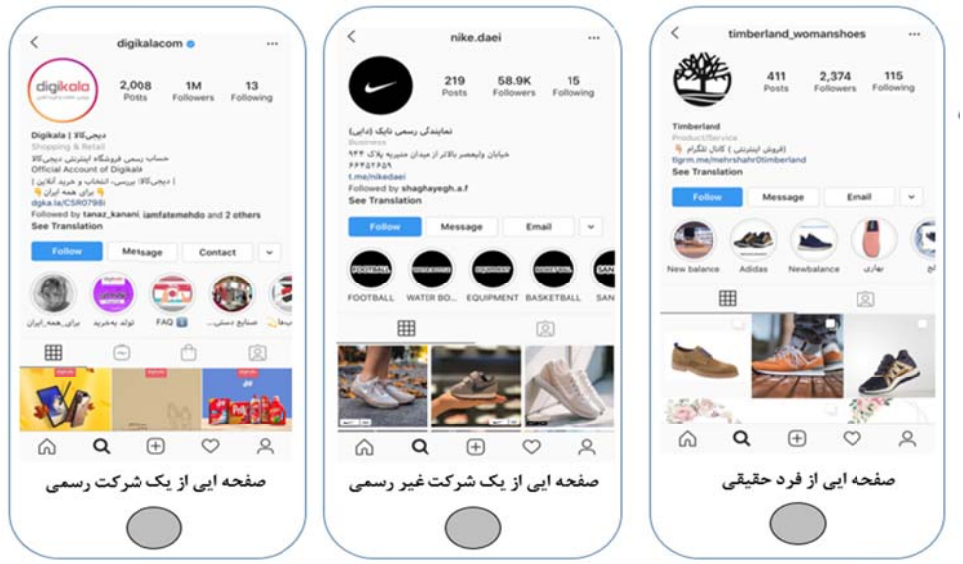
آنچه که در این کسب و کارهای مجازی وجود دارد موضوع اشتغال آن است. موضوع اشتغال همواره بعنوان یکی از دغدغه های دولت است. در این موضوع طرح آمارگیری از نیروی کار ((Labour Force Survey (LFS) که بصورت فصلی توسط مرکز آمار ایران اجرا شده و نتایج آن منتشر می شود، بعنوان مهم ترین آمار رسمی منتشر شده در کشور ر خصوی آمارهای بازار کار است که در آن ماموران آمارگیر در ٤ نوبت در سال از خانوارهای نمونه اطلاعات مربوط به اعضای خانوار و وضعیت اشتغال آنان (شاغل، بیکار و غیرفعال) اطلاعاتی را جمع آوری می نمایند. در این پرسشنامه سوالاتی در مورد کسب و کارهای مجازی پرسیده نمی شود لذا از این مدل کسب و کارها بعنوان اشتغال پنهان می توان نام برد.

سوال اصلی که در این مطالعه به بررسی آن پرداخته شده است این است که چگونه می توان از منبع داده ایی آزاد شبکه اینستاگرام برآورد برای کسب و کار و اشتغال پنهان در راستای طرح آمارگیری نیروی کار ارایه نمود.

بر اساس نتایج طرح آمارگیری برخورداری خانوارها و استفاده افراد از اینترنت در سال ۱۳۹۷[۴]، کشور ایران دارای جمعیت ۶ ساله و بیشتر ۷۲۳۱۸۳۵۰ نفر که از این تعداد ۴۶۳۱۵۵۴۶ نفر کاربر اینترنت برآورد شده اند. ضریب نفوذ اینترنت در کل کشور ۶۴% و ۳۷۷۶۲۸۰۲ نفر دارای تلفن هوشمند بوده اند. در پایان سال ۱۳۹۶ تعداد ۱۳۳۵۵۸۷۵ نفر کاربر شبکه اجتماعی اینستاگرام برآورد شده است.

شکل۳- کسب و کارهای اینترنتی و ویژگی های اشتغال در آنها

صفحات کسب و کار در فضای مجازی دارای سه نوع می باشد (با توجه به اینکه در این مطلاعه شبکه اجتماعی اینستاگرام مورد نظر است لذا مثال های ارایه شده از این شبکه اجتماعی است در صورتی که می توان در مورد سایر شبکه های اجتماعی نیز مصادیق مشابهی را ذکر نمود.



شکل ۴- انواع صفحات کسب و کار در اینستاگرام

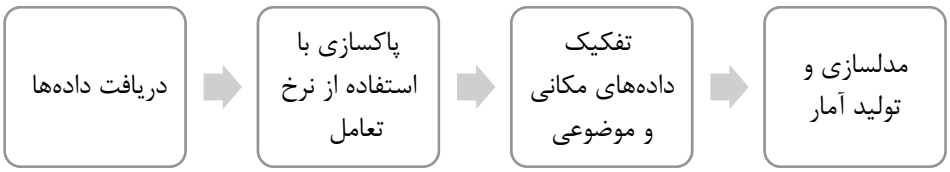
۴- فرایند تجزیه و تحلیل داده ها

اشتغال ایجاد شده در شبکه اجتماعی اینستاگرام با گسترش این شبکه در بین افراد روندی رو به رشد دارد. داده های کاربران این شبکه از سایت اینستاگرام قابل دریافت است هر چند که نحوه دسترسی، سطح دسترسی به ریز داده ها و فراداده ها دارای سیاست های از پیش تعیین شده است، ولی محققان می توانند بسیاری از این داده را دریافت نمایند. داده های شبکه اجتماعی اینستاگرام دارای ویژگی های منحصر به فردی است که عبارتند از:

پشتیبانی از زبان فارسی ، رایگان بودن ایجاد صفحه ، دارای نسخه موبایل اپ، نسخه دکستاپ و سایت، عدم وجود محدودیت در ایجاد تعداد صفحات برای کاربر، عدم نیاز به هر گونه مجوز و زبان برنامه نویسی ، امکان به اشتراک گذاری ویدیو، عکس، گفتگوی مستقیم، داستان ، دریافت عکس العمل بازدید کنندگان توسط لایک و ;کامنت، امکان جستجوی ساده (با علامت #) ، اطلاع رسانی صفحات قلابی یا غیر اخلاقی به سرور با

امکان ریپورت این ویژگی‌ها باعث شد تا نویسندگان این داده ها را دریافت نمایند. پس از دریافت داده های این شبکه مراحل زیر انجام گرفت. نکته مهم این است که که داده های دریافتی دارای ویژگی های مکانی است و می توان آنها را بر اساس ویژگای مکانی بر حسب مناطق استانی طبقه بندی نمود.

شایان ذکر است که در این پاکسازی از نرخ تعامل (عبارتست از نسبت مجموع لایک و کامنت به تعداد بازدید) بعنوان شاخص فعال بودن کسب و کار مجازی استفاده شده است. در صفحاتی که نرخ تعامل بزرگتر مساوی ۱ است این صفحه فعال در نظر گرفته می شود.



شکل ۵- مراحل انجام برآورد اشتغال پنهان با استفاده از داده های اینستاگرام

۵- نتایج و پیشنهادات

در طبقه‌بندهای موضوعی اولیه از داده های دریافتی فراوانی کسب و کار ها برای ۶ ماهه اول سال ۱۳۹۸ بصورت زیر حاصل شده است:

جدول۱- انواع صفحات کسب و کار بر حسب نوع خدمت در اینستا گرام

عنوان حوزه	تعداد	درصد
فروش کالا	۱,۸۵۴,۷۰۵	%۳۶٫۹
تبلیغات	۸۵۹,۷۲۹	%۱۷٫۱
خدمات فنی و سایر موارد	۴۸۹,۴۱۶	%۹٫۷
نرم افزار	۵۲۸,۹۰۸	%۱۰٫۵
آموزش و تولید محتوا	۱,۲۸۹,۵۶۸	%۲۵٫۷
جمع	۵,۰۲۲,۳۲۶	%۱۰۰

بر اساس جدول ۱ بیشترین کسب و کار مجازی برای فروش کالاست که با توجه اینکه ثبت حقوقی و مجوز قانونی وجود ندارد بنابرین امکان استفاده از درگاه پرداخت بانکی در این صفحات وجود نداشته و فروش کالاها و پرداخت وجوه آن از طریق نقل و انتقالات شخصی خواهد بود.

جدول ۲- اطلاعات صفحات کسب و کار در اینستاگرام و طرح نیروی کار در سال ۱۳۹۶ در برخی از استانها [۲]

استان	تعداد صفحات طی ۶ ماه	تعداد صفحات بهار	تعداد صفحات تابستان	شاغل بهار	شاغل تابستان	نرخ بیکاری بهار	نرخ بیکاری تابستان
تهران	۳۶۳،۴۱۶	۲۷۲،۵۶۲	۹۰،۸۵۴	۴،۲۰۱،۸۰۵	۴،۱۹۱،۵۸۹	۱۰٫۲	۱۰٫۸
البرز	۶۰،۲۶۸	۴۷،۶۱۲	۱۲،۶۵۶	۸۳۵،۵۹۳	۸۸۱،۴۰۰	۱۴٫۱	۱۳
اصفهان	۳۲،۰۹۳	۲۱،۵۰۲	۱۰،۵۹۱	۱،۶۷۳،۱۰۸	۱،۶۶۷،۷۸۱	۱۰٫۸	۱۰٫۴
هرمزگان	۲۸،۹۲۹	۹،۲۵۷	۱۹،۶۷۲	۵۲۹،۳۴۴	۵۲۶،۷۷۳	۱۶٫۲	۱۴٫۲
آذربایجان شرقی	۱۰،۸۹۸	۴،۲۵۰	۶،۶۴۸	۱،۲۰۱،۸۹۸	۱،۲۸۷،۰۱۵	۱۰٫۳	۹٫۱
خوزستان	۳۲،۱۸۱	۲۶،۳۸۸	۵،۷۹۳	۱،۲۷۶،۲۲۲	۱،۳۰۵،۲۹۸	۱۶٫۳	۱۴٫۶
خراسان رضوی	۱۷،۴۷۸	۳،۶۷۰	۱۳،۸۰۸	۱،۹۶۸،۷۹۵	۲،۰۷۷،۸۹۶	۹٫۴	۹
گیلان	۱۵،۴۶۹	۹،۴۳۶	۶،۰۳۳	۹۹۶،۶۱۴	۹۲۲،۲۸۷	۸٫۳	۹٫۶
زنجان	۹،۶۴۳	۴،۶۲۹	۵،۰۱۴	۳۵۸،۱۱۸	۳۸۳،۶۱۹	۸٫۴	۵٫۲
قزوین	۹،۲۴۱	۳،۴۱۹	۵،۸۲۲	۴۱۱،۴۵۳	۴۳۱،۱۳۹	۱۰٫۸	۸٫۲
کرمان	۱۴،۷۸۲	۳،۵۴۸	۱۱،۲۳۴	۸۶۵،۰۳۸	۸۹۵،۵۲۷	۱۲٫۲	۱۰٫۸

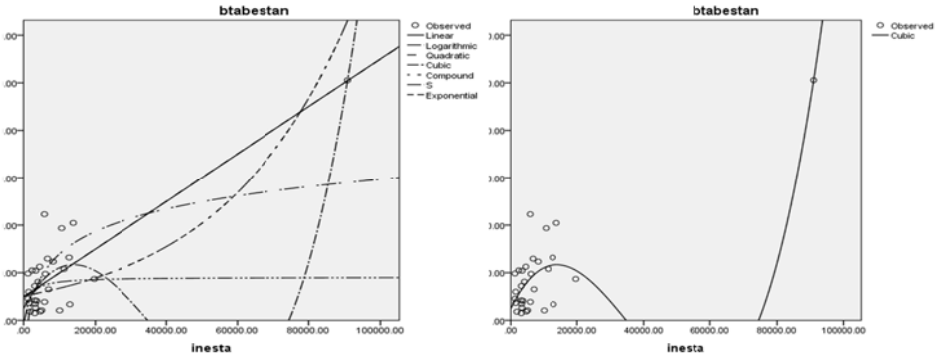
آنچه که بر اساس داده های فوق مورد نظر است آزمون فرض زیر خواهد بود :

$$\begin{cases} H_0: \text{ارتباط معنی دار بین کسب و کارهای مجازی و نرخ اشتغال وجود ندارد} \\ H_1: \text{ارتباط معنی دار بین کسب و کارهای مجازی و نرخ اشتغال وجود دارد} \end{cases}$$

بر اساس داده ها مدل رگرسیون درجه سوم بر روی داده ها برازش داده شده و بر اساس نتایج فرض ادعا مورد پذیرش واقع گردید.

جدول ۳- نتایج حاصل از برازش مدل رگرسیون بر روی داده ها

	Unstandardized Coefficients		Standardized Coefficients		
Coefficients	B	Std. Error	Beta	t	Sig.
inesta	۱۴٫۴۰۸	۸٫۴۹۴	۲٫۴۱۲	۱٫۶۹۶	۰٫۰۱۰۱
inesta ** 2	-۰٫۰۰۱	۰٫۰۰۱	-۹٫۹۹۵	-۱٫۱۶۱	۰٫۰۲۵۶
inesta ** 3	۵٫۹۶۸E-۹	۰٫۰۰۰	۸٫۴۷۲	۰	۰
(Constant)	۲۴۷۱۷٫۷۴۱	۲۷۴۳۰٫۷۵۷		۰٫۹۰۱	۰٫۰۳۰٫۰۳۷۶۷۶



شکل ۶- مدل رگرسیون: تعداد شاغلان و تعداد صفحات کسب و کاری در اینستاگرام

۶- خلاصه نتایج

- شبکه های اجتماعی با توجه به تعداد رو به رشد و زیاد کاربران در ایران و ضریب نفوذ بالای اینترنت پر سرعت می تواند به عنوان یک منبع مستقیم در تولید آمارهای رسمی باشد و یا منبع ثانویه برای چک کردن صحت آمارها مورد استفاده قرار گیرد.

- با توجه به اینکه تعداد اشتغال تحت تاثیر بسیاری از عوامل آشکار و پنهان است، مدل پیشنهادی می تواند توسعه داده شده و حتی برای برآورد نرخ بیکار و اشتغال مورد استفاده قرار گیرد
- با توجه به روند افزایش اشتغال پنهان در کشور در بویژه در شبکه های اجتماعی، بایداین نوع کسب و کارها در طرح آمارگیری نیروی کار لحاظ شود که در حال حاضر در طرح حاضر سوالی در این خصوص از افراد پرسیده نمی شود و به خاطر اهمیت موضوع اخیرا سازمان بین المللی کار به دنبال تدوین راهنماهایی در این خصوص بوده است
- در بین استان هایی که نرخ بیکاری روند کاهشی داشته است (مانند همدان و قزوین) شاهد رشد زیاد کسب و کار های اینترنتی و در نتیجه اشتغال پنهان بوده ایم. همچنین در استانهایی که نرخ بیکاری افزایش یافته / تغییر محسوسی نداشته است (مانند سیستان) ، تعداد کسب و کار های اینترنتی روند کاهشی / بدون تغییر بوده است
- در این مطالعه تنها شبکه اجتماعی اینستاگرام مورد بررسی قرار گرفته است در حالی که شبکه های دیگری نیز وجود دارد که تجزیه و تحلیل داده های آنها می تواند در بهبود نتایج آمارهای رسمی موجود مفید باشد.
- پارادیتا بعنوان یکی از منابع ثانویه برای تولید آمارهای رسمی هستند. فرایند های کنونی جمع آوری و تولید داده های باید به گونه ایی تغییر یابد تا پارادیتای باکیفیت را تولید نمایند.

توضیحات

1. http://data.gov.ir
2. http://data.tehran.ir
3. Working Group on Open Data
4. www.statcounter.com

مرجع‌ها

[۱] سامانه انتشار و دسترسی آزاد به اطلاعات، https://foia.iran.gov.ir/web/guest/home.

[۲] مرکز آمار ایران (۱۳۹۶). چکیده نتایج طرح آمارگیری از نیروی کار سال ۱۳۹۶، دفتر جمعیت، نیروی کار و سرشماری.

[۳] مرکز آمار ایران (۱۳۹۷). چکیده نتایج طرح آمارگیری برخورداری خانوارها و استفاده افراد از اینترنت در سال ۱۳۹۷، دفتر صنعت، معدن و زیرساخت، گروه آمارهای ICT.

[4] UN (2017). Global Working Group on Big Data for Official Statistics, Handbook on the use of Mobile Phone data for Official Statistics.

آرش فاضلی
فوق لیسانس صنایع
تهران، خیابان فاطمی، خیابان رهی معیری، مرکز آمار ایران.
رایانشانی: arashfazeli61@gmail.com

سعید فیاض
فوق لیسانس صنایع
تهران، خیابان فاطمی، خیابان رهی معیری، مرکز آمار ایران.
رایانشانی: saeed.fayyaz@gmail.com