طراحی الگوی مفهومی تعدیل‌گری هوشمند محتوا در پلتفرم‌ها (مطالعه تطبیقی یوتیوب، ایکس، آپارات و دیوار)

سمیه لبافی[1*]
دانشیار، پژوهشکدهٔ جامعه و اطلاعات، گروه مطالعات اجتماعی اطلاعات، پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک)، تهران، ایران.

ناهید اکرمی
کارشناسی ارشد مدیریت رسانه، دانشگاه تهران، تهران، ایران.

داتیس خواجه‌ییان
دانشیار، گروه مدیریت رسانه و ارتباطات کسب و کار، دانشکدگان مدیریت، دانشگاه تهران، تهران، ایران.

مدیریت اطلاعات
دوره ۱۱، شماره ۲
پاییز و زمستان ۱۴۰۴

چکیده: پژوهش حاضر با هدف طراحی الگویی مفهومی برای تعدیل‌گری هوشمند محتوای تولیدشده توسط کاربران در پلتفرم‌های دیجیتال انجام شده است. مسئله اصلی، تبیین نحوه پیوند تشخیص خودکار، داوری انسانی، مشارکت کاربران و قواعد پلتفرمی در شناسایی و مدیریت محتوای ناقض سیاست‌هاست. اطلاعات نادرست و اخبار جعلی در این پژوهش یکی از مصادیق محتوای مسئله‌دار به شمار می‌روند و با همه انواع محتوای ممنوع یا آسیب‌رسان یکسان تلقی نمی‌شوند. پژوهش با رویکرد کیفی و در قالب مطالعه تطبیقی چهار مورد انجام شد. داده‌های اصلی از اسناد و سیاست‌های منتشرشده یوتیوب، ایکس، آپارات و دیوار و داده‌های مکمل از شش مصاحبه نیمه‌ساختارمند با متخصصان مرتبط با پلتفرم‌های داخلی و حوزه رسانه‌های دیجیتال به دست آمد. داده‌ها در سه سطح کدگذاری اولیه، تجمیع مفهومی و مقوله‌بندی سازمان یافتند و سپس سازوکارهای تعدیل‌گری در ابعاد مشترک و متناسب با ویژگی هر پلتفرم مقایسه شدند. در مجموع ۵۵۹ کد اولیه، ۱۱۸ کد مفهومی و ۴۱ مقوله در سطح چهار مورد مطالعاتی گزارش شد. تحلیل داده‌ها چهار حوزه اصلی سیاست‌گذاری محتوا، شناسایی و ارجاع موارد مشکوک، تصمیم‌گیری و مداخله، و پاسخگویی و بازبینی را نشان داد. در پلتفرم‌های رسانه‌ای، مدیریت اطلاعات نادرست در کنار سایر تخلفات محتوایی قرار می‌گیرد؛ در بازارگاه آگهی، اصالت اطلاعات آگهی، هویت و امنیت تعاملات اهمیت بیشتری می‌یابد. بر پایه این اشتراک‌ها و تفاوت‌ها، الگویی ترکیبی با نقش‌های متمایز برای فناوری، نیروی انسانی، کاربران و مدیریت پلتفرم ارائه شد. نتیجه پژوهش نشان می‌دهد که طراحی تعدیل‌گری هوشمند محتوا مستلزم تفکیک تشخیص از تصمیم، تناسب مداخله با نوع تخلف و پیش‌بینی سازوکار بازبینی است. الگوی پیشنهادی، چارچوبی برای سازمان‌دهی این وظایف در پلتفرم‌های متفاوت ارائه می‌کند.

کلیدواژه‌ها: تعدیل‌گری محتوا، حکمرانی پلتفرمی، محتوای تولیدشده توسط کاربر، اطلاعات نادرست، تعدیل‌گری هوشمند

---

[1] نویسنده مسئول: labafi@irandoc.ac.ir

مقدمه

رسانه‌های اجتماعی در دهه اخیر از ابزارهای ارتباطی صرف فراتر رفته و به زیرساخت‌هایی تعیین‌کننده در تولید، توزیع و مصرف اطلاعات تبدیل شده‌اند. این پلتفرم‌ها نه‌تنها شیوه تعاملات اجتماعی و سیاسی را دگرگون کرده‌اند، بلکه در شکل‌دهی به افکار عمومی، برجسته‌سازی موضوعات اجتماعی و تعیین دستور کار رسانه‌ای نیز نقش روزافزونی یافته‌اند (Van Dijck et al., 2023). ویژگی اصلی این محیط‌ها آن است که کاربران را از موقعیت مخاطب منفعل خارج کرده و به تولیدکنندگان، توزیع‌کنندگان و بازنشرکنندگان فعال محتوا تبدیل می‌کنند (Bateman et al., 2021). این تحول، از یک سو امکان مشارکت گسترده‌تر شهروندان در فضای عمومی را فراهم ساخته و از سوی دیگر، مرزهای سنتی میان تولیدکننده و مصرف‌کننده اطلاعات را کم‌رنگ کرده است (Van Dijck, 2013).

با گسترش انتشار محتوای تولیدشده توسط کاربران، پلتفرم‌ها با طیفی از مسائل مانند اطلاعات نادرست و گمراه‌کننده، جعل هویت، نقض حریم خصوصی، آزار، نقض مالکیت فکری و تخلفات مرتبط با آگهی‌ها، مواجه شده‌اند. این مصادیق از نظر موضوع، پیامد و معیار تشخیص یکسان نیستند. «اطلاعات نادرست» به محتوای خلاف واقع، فارغ از قصد تولیدکننده، اشاره دارد؛ «اطلاعات گمراه‌کننده عمدی» بر قصد فریب تأکید می‌کند؛ و «اخبار جعلی» به محتوایی با ظاهر خبر اطلاق می‌شود که اطلاعات ساختگی یا تحریف‌شده را عرضه می‌کند. در این مقاله، این مفاهیم زیرمجموعه مسئله گسترده‌تر تعدیل‌گری محتوا هستند و به جای یکدیگر به کار نمی‌روند (Aïmeur et al., 2023).

مسئله هنگامی پیچیده‌تر می‌شود که پلتفرم‌ها در مقام واسطه‌های اطلاعاتی، نه کاملاً ناشر سنتی‌اند و نه صرفاً زیرساخت‌های فنی بی‌طرف. آن‌ها با طراحی الگوریتم‌ها، تعیین قواعد انتشار، حذف یا کاهش نمایش محتوا و تعریف سازوکارهای گزارش‌دهی، در عمل بر دیده‌شدن یا نادیده‌ماندن محتوا تأثیر می‌گذارند. بنابراین، پلتفرم‌ها در برابر محتوای منتشرشده توسط کاربران، با نوعی مسئولیت اجتماعی، حقوقی و اخلاقی مواجه‌اند؛ مسئولیتی که باید میان صیانت از آزادی بیان، کاهش آسیب‌های اطلاعاتی و حفظ اعتماد کاربران توازن برقرار کند. از این منظر، تعدیل‌گری محتوا دیگر یک فعالیت فنی حاشیه‌ای نیست، بلکه به یکی از ارکان حکمرانی پلتفرمی تبدیل شده است.

پلتفرم‌های بزرگ برای مواجهه با این چالش، به‌تدریج از ترکیب نظام‌های انسانی و ماشینی استفاده کرده‌اند. هوش مصنوعی، یادگیری ماشین، پردازش زبان طبیعی و تحلیل الگوهای انتشار می‌توانند محتوای مشکوک را در مقیاس وسیع شناسایی، رتبه‌بندی یا نشانه‌گذاری کنند (Shu et al., 2020). با این حال، تشخیص اخبار جعلی همواره یک مسئله صرفاً فنی نیست. محتوای گمراه‌کننده ممکن است در قالب طنز، کنایه، نقل‌قول، بازنشر انتقادی، روایت چندمعنایی یا متن وابسته به زمینه فرهنگی عرضه شود؛ مواردی که تشخیص آن‌ها برای الگوریتم‌ها دشوار است (Gerbina, 2021). از این رو، هرگونه اتکای مطلق به هوش مصنوعی می‌تواند به حذف نادرست محتوا، باقی‌ماندن محتوای آسیب‌زا یا بازتولید سوگیری‌های الگوریتمی منجر شود.

در فضای فارسی‌زبان، تنوع بیان زبانی، کنایه، نوشتار غیررسمی و فینگلیش‌نویسی می‌تواند تحلیل خودکار محتوا را دشوار کند. از سوی دیگر، سیاست‌های تعدیل‌گری در پلتفرم‌های داخلی و بین‌المللی در بسترهای حقوقی و سازمانی متفاوتی تدوین می‌شوند. بنابراین، مقایسه سازوکارهای تعدیل‌گری باید هم به عناصر مشترک فرایند و هم به تفاوت مأموریت و نوع محتوای هر پلتفرم توجه کند. پژوهش حاضر با بررسی یوتیوب، ایکس، آپارات و دیوار به این پرسش پاسخ می‌دهد که «مؤلفه‌ها و روابط اصلی یک الگوی مفهومی تعدیل‌گری هوشمند محتوا، بر پایه مقایسه سیاست‌ها و سازوکارهای اعلام‌شده در پلتفرم‌های منتخب، چیست؟» پرسش‌های فرعی عبارت‌اند از: ۱) سازوکارهای شناسایی، ارزیابی و مداخله در هر پلتفرم چگونه است؟ ۲) چه عناصر مشترک و چه تفاوت‌های وابسته به ماهیت پلتفرم وجود دارد؟ ۳) نقش فناوری، نیروی انسانی، کاربران و قواعد پلتفرم در الگوی تلفیقی چگونه تعریف می‌شود؟

مبانی نظری و پیشینه پژوهش

- تنظیم‌گری پلتفرم‌ها

تنظیم‌گری در رسانه‌ها به مجموعه‌ای از تصمیم‌ها، قواعد، سیاست‌ها و سازوکارهای اجرایی اطلاق می‌شود که با هدف کنترل، هدایت و نظارت بر عملکرد رسانه‌ها، حمایت از حقوق کاربران و تأمین منافع عمومی طراحی می‌شود (Puppis, 2010). با گسترش ارتباطات دیجیتال، موضوع تنظیم‌گری از حوزه رسانه‌های سنتی فراتر رفته و پلتفرم‌های آنلاین را نیز در بر گرفته است؛ پلتفرم‌هایی که ماهیت فراملی، سرعت تغییر بالا و ساختارهای فنی پیچیده آن‌ها، تدوین و اجرای قواعد پایدار را دشوار می‌سازد (DeNardis & Hackl, 2015; Gillespie, 2015).

پلتفرم‌ها از یک منظر واسطه‌هایی میان کاربران هستند و بخش عمده محتوای آن‌ها توسط کاربران تولید می‌شود. همین وضعیت، مسئله مسئولیت پلتفرم را به یکی از مباحث اصلی حکمرانی دیجیتال تبدیل کرده است. برخی رویکردهای حقوقی، پلتفرم‌ها را صرفاً واسطه می‌دانند و مسئولیت مستقیم محتوای کاربران را متوجه آن‌ها نمی‌کنند؛ در مقابل، دیدگاه‌های جدیدتر بر این نکته تأکید دارند که پلتفرم‌ها از طریق طراحی الگوریتم‌ها، سیاست‌های انتشار، حذف محتوا و مدل‌های اقتصادی توجه، عملاً در شکل‌دهی به فضای ارتباطی نقش فعال دارند (Balkin, 2004; Godwin, 2003; Gillespie, 2018).

در ایالات متحده، ایده «بندر امن» (Safe Harbor) در چارچوب قوانین مربوط به مسئولیت واسطه‌ها مطرح شده است. بر مبنای این رویکرد، پلتفرم‌ها در مقام واسطه، الزاماً ناشر حقوقی محتوای کاربران تلقی نمی‌شوند؛ اما می‌توانند محتوای خشونت‌آمیز، مستهجن یا آسیب‌رسان را حذف یا محدود کنند (Mueller, 2015). با این حال، این چارچوب که در دوره‌ای پیش از قدرت‌گیری شبکه‌های اجتماعی تدوین شده بود، امروز با پرسش‌هایی درباره مسئولیت واقعی پلتفرم‌ها، قدرت تصمیم‌گیری آن‌ها و امکان پاسخگویی در سطح جهانی مواجه است (Urban et al., 2016).

در اتحادیه اروپا، رویکرد تنظیم‌گری به سمت مسئولیت‌پذیری بیشتر پلتفرم‌ها حرکت کرده است. قانون تنظیم‌گری عمومی محافظت از اطلاعات (GDPR) از سال ۲۰۱۸، پلتفرم‌ها را به رعایت اصول شفافیت،

رضایت آگاهانه و امکان کنترل داده‌ها توسط کاربران ملزم کرده است (Martínez, 2023). قانون بازارهای دیجیتال (DMA) نیز با هدف مقابله با انحصار و تقویت رقابت در بازارهای دیجیتال تصویب شده است (Monti, 2021). قانون خدمات دیجیتال (DSA) نیز پلتفرم‌ها را به شفافیت در تبلیغات، الگوریتم‌ها و سیاست‌های محتوا ملزم می‌کند و برای تخلفات مهم، جریمه‌های سنگین در نظر می‌گیرد (Ponce, 2020).

در ایران نیز با گسترش پلتفرم‌های داخلی و افزایش مصرف محتوای آنلاین، موضوع تنظیم‌گری اهمیت بیشتری یافته است. نهادهایی مانند مرکز پژوهش‌های مجلس، شورای عالی فضای مجازی و سازمان تنظیم مقررات صوت و تصویر فراگیر در فضای مجازی (ساترا) از جمله بازیگران این حوزه محسوب می‌شوند. ساترا با هدف صدور مجوز، نظارت بر عملکرد پلتفرم‌ها و تنظیم مقررات رسانه‌های صوت و تصویر فراگیر ایجاد شده و تلاش کرده است در حوزه کیفیت محتوا، حقوق مخاطبان و نظم رسانه‌ای نقش‌آفرینی کند (آزادی و همکاران، ۱۳۹۸؛ مظاهری و تهرانی، ۱۴۰۱). با وجود این، چالش‌هایی مانند ابهام حقوقی، تنوع خدمات دیجیتال، مقاومت برخی بازیگران و سرعت بالای تغییر فناوری همچنان پابرجاست (آزادی و همکاران، ۱۳۹۸؛ باقری و آصفی، ۱۴۰۰).

- تعدیل‌گری محتوا و تحول آن

تعدیل‌گری محتوا مجموعه اقداماتی است که پلتفرم برای شناسایی، بررسی و مدیریت محتوای تولیدشده توسط کاربران بر اساس قواعد اعلام‌شده و الزامات قانونی به کار می‌گیرد (Common, 2020; Gillespie, 2020). موضوع این فرایند از اطلاعات نادرست فراتر می‌رود و مواردی چون جعل هویت، آزار، نقض حریم خصوصی و تخلف در آگهی‌ها را نیز دربر می‌گیرد. در مقابل، تشخیص اخبار جعلی صرفاً یکی از مسائل تخصصی این حوزه است. از این رو، در پژوهش حاضر «تعدیل‌گری هوشمند محتوا» مفهوم محوری و «اخبار جعلی و اطلاعات گمراه‌کننده» یکی از مصادیق آن است.

تعدیل‌گری محتوا همواره با یک تنش بنیادین همراه است: از یک سو پلتفرم‌ها باید از آزادی بیان، تنوع دیدگاه‌ها و مشارکت کاربران حمایت کنند؛ از سوی دیگر، موظف‌اند از انتشار محتوای مضر، نفرت‌پراکن، گمراه‌کننده، خشونت‌آمیز، ناقض حریم خصوصی یا آسیب‌زننده به کودکان جلوگیری کنند (Balkin, 2017; Flew et al., 2019). این تنش، تعدیل‌گری را از یک عملیات صرفاً فنی به مسئله‌ای سیاسی، اخلاقی و اجتماعی تبدیل کرده است.

تعدیل‌گری انسانی، قدیمی‌ترین و همچنان یکی از مهم‌ترین شیوه‌های مدیریت محتواست. در این روش، تعدیل‌گران آموزش‌دیده محتوای گزارش‌شده یا مشکوک را بررسی می‌کنند و درباره حذف، محدودسازی، برچسب‌گذاری یا حفظ آن تصمیم می‌گیرند. مزیت اصلی این روش، توانایی انسان در فهم زمینه، کنایه، نیت، بافت فرهنگی و حساسیت‌های اجتماعی است. با این حال، مواجهه مداوم با محتوای خشونت‌آمیز، آزاردهنده یا افراطی می‌تواند آثار روانی جدی بر تعدیل‌گران برجای بگذارد (Newton, 2019; Ruckenstein & Turunen, 2020). افزون بر این، حجم عظیم داده‌ها و ضرورت واکنش سریع، اتکای صرف به نیروی انسانی را ناکارآمد می‌کند (Banchik, 2021).

تعدیل‌گری هوشمند، پاسخی به مسئله مقیاس در پلتفرم‌های بزرگ است. در این رویکرد، سامانه‌های هوش مصنوعی، یادگیری ماشین و پردازش زبان طبیعی برای شناسایی اولیه محتوای ناقض قواعد به کار می‌روند (Ballard, 2019). این سامانه‌ها می‌توانند متن، تصویر، صوت و ویدئو را با سرعت بالا پردازش کنند و موارد مشکوک را برای بررسی بیشتر نشانه‌گذاری کنند. با وجود این، محدودیت در درک زمینه، احتمال سوگیری داده‌های آموزشی، دشواری تشخیص طنز و کنایه و خطا در زبان‌های کم‌منبع، موجب می‌شود که تعدیل‌گری هوشمند به‌تنهایی کافی نباشد (Auxier, 2020; Gorwa et al., 2020).

در ادبیات پژوهش، تعدیل‌گری ترکیبی بر تقسیم وظایف میان ابزارهای خودکار، تعدیل‌گران انسانی و سازوکارهای مشارکت کاربر تأکید دارد (Gorwa et al., 2020; Gillespie, 2020). این تقسیم وظایف در همه پلتفرم‌ها به یک شکل اجرا نمی‌شود و بسته به نوع محتوا، ریسک تخلف، امکانات فنی و سیاست‌های سازمانی تفاوت دارد. از همین رو، مقایسه پلتفرم‌ها در این پژوهش بر کارکردهای مشترک تعدیل‌گری استوار است، و فرض یکسان‌بودن رویه‌های اجرایی آن‌ها نیست.

- انواع تعدیل‌گری محتوا

بر اساس زمان مداخله، منبع تشخیص و سطح مشارکت کاربران، تعدیل‌گری محتوا را می‌توان به چند گونه اصلی تقسیم کرد:

- پیش‌تعدیل‌گری (Pre-Moderation): در این شیوه، محتوا پیش از انتشار نهایی بررسی می‌شود. این روش برای جلوگیری از انتشار محتوای آسیب‌زا مفید است، اما ممکن است فرایند انتشار را کُند و پرهزینه کند (Gorwa, 2019).
- پساتعدیل‌گری (Post-Moderation): در این شیوه، محتوا پس از انتشار بررسی می‌شود. این روش انعطاف بیشتری دارد، اما ممکن است محتوای مضر پیش از حذف، به بخشی از مخاطبان برسد (Gillespie, 2018).
- تعدیل‌گری واکنشی (Reactive Moderation): در این روش، کاربران محتوای مشکوک یا آسیب‌زا را گزارش می‌کنند و پلتفرم پس از دریافت گزارش وارد عمل می‌شود. مشارکت کاربران در این روش مهم است، اما پوشش آن کامل نیست (Zeng & Kaye, 2022).
- تعدیل‌گری توزیع‌شده (Distributed Moderation): در این شیوه، هوش مصنوعی، تعدیل‌گران انسانی و کاربران به صورت مکمل عمل می‌کنند. این روش پیچیده‌تر است، اما برای پلتفرم‌های بزرگ و چندزبانه مناسب‌تر به نظر می‌رسد (De Streel et al., 2020).

- چالش‌های تعدیل‌گری هوشمند

تعدیل‌گری هوشمند با وجود ظرفیت‌های مهم خود، با چالش‌هایی بنیادین روبه‌رو است. نخستین چالش، دشواری تشخیص زمینه است. الگوریتم‌ها اغلب نمی‌توانند تفاوت میان نقد، طنز، نقل‌قول، بازنشر آموزشی و محتوای آسیب‌زا را با دقت کافی تشخیص دهند (Bandy & Diakopoulos, 2021; Bacchi, 2021; Young, 2021). چالش دوم، سوگیری الگوریتمی است. داده‌های آموزشی ناقص یا نامتوازن می‌توانند به

حذف ناعادلانه محتوای گروه‌های خاص یا نادیده‌گرفتن محتوای آسیب‌زا در برخی زمینه‌ها منجر شوند (Dowdeswell & Goltz, 2020). چالش سوم، قطبی‌شدن فضای آنلاین و شکل‌گیری حلقه‌های بسته اطلاعاتی است. الگوریتم‌های توصیه‌گر ممکن است کاربران را در معرض محتوایی قرار دهند که باورهای پیشین آنان را تقویت می‌کند و امکان مواجهه با دیدگاه‌های متفاوت را کاهش می‌دهد (Molina & Sundar, 2022). چالش چهارم، سرعت انتشار اخبار فوری است. در موقعیت‌های بحرانی، تأخیر در تعدیل‌گری می‌تواند به انتشار گسترده شایعه منجر شود، اما واکنش شتاب‌زده نیز ممکن است به حذف محتوای مهم یا مستند بینجامد (Dehghan & Glazunova, 2021).

افزون بر این، نبود شفافیت در سیاست‌های تعدیل‌گری و نامشخص بودن معیارهای تصمیم‌گیری، اعتماد کاربران را کاهش می‌دهد (Azzimonti & Fernandes, 2023). حذف بیش‌ازحد محتوا نیز می‌تواند آزادی بیان را محدود کند و به شکل‌گیری احساس سانسور یا تبعیض در میان کاربران منجر شود (Helberger, 2020; Black, 2008). در مقابل، تساهل بیش از حد نیز می‌تواند فضای پلتفرم را به محیطی ناامن و بی‌اعتماد تبدیل کند. بنابراین، چالش اصلی تعدیل‌گری هوشمند، یافتن نقطه تعادل میان مداخله مؤثر، شفافیت، آزادی بیان و مسئولیت اجتماعی است.

روش پژوهش

پژوهش حاضر کیفی و مبتنی بر مطالعه تطبیقی چهار مورد است. واحد تحلیل، «سیاست‌ها و سازوکارهای اعلام‌شده برای تعدیل‌گری محتوای تولیدشده توسط کاربران در هر پلتفرم» تعیین شد. انتخاب موارد، هدفمند و با منطق تنوع ساختاری انجام گرفت: یوتیوب و آپارات برای مقایسه دو پلتفرم ویدئومحور، ایکس برای بررسی محیط تعاملی و متن‌محور، و دیوار برای مطالعه تعدیل‌گری در بازارگاه آگهی، انتخاب شدند. این طراحی، مقایسه کارکردهای مشترک را در کنار شناخت تفاوت‌های ناشی از مأموریت پلتفرم ممکن می‌سازد؛ بنابراین، نتایج به معنای هم‌ارزی محتوای خبری با اطلاعات آگهی یا یکسان‌بودن قواعد چهار مورد نیست. منطق مطالعه چندموردی و مقایسه درون‌موردی و میان‌موردی با مبانی پژوهش موردی سازگار است (Yin, 2018).

منبع اصلی داده‌ها، اسناد عمومی منتشر شدهٔ هر پلتفرم، از جمله شرایط استفاده، راهنماهای جامعه کاربری، سیاست‌های محتوا، سازوکارهای گزارش‌دهی و قواعد مرتبط با حریم خصوصی و رسیدگی به تخلف بوده است. فهرست منابع، نمونه‌هایی از صفحات رسمی بررسی‌شده و تاریخ دسترسی ثبت‌شده به برخی از آن‌ها را نشان می‌دهد. برای تکمیل فهم زمینه اجرایی، شش مصاحبه نیمه‌ساختارمند انجام شده است. سه نفر مرتبط با دیوار، دو نفر مرتبط با آپارات و یک متخصص حوزه پلتفرم‌های رسانه‌ای، بدین منظور انتخاب شدند. بنابراین، مقایسه چهار پلتفرم اساساً بر داده‌های اسنادی استوار است و مصاحبه‌ها در تفسیر تجربه پلتفرم‌های داخلی نقش مکمل دارند؛ داده‌های مصاحبه‌ای به رویه‌های داخلی یوتیوب یا ایکس تعمیم داده نشده‌اند.

تحلیل در سه سطح انجام شد. ابتدا گزاره‌های مرتبط با قواعد محتوا، شیوه شناسایی تخلف، نقش کنشگران، نوع مداخله و رسیدگی پس از تصمیم از داده‌ها استخراج و به کدهای اولیه تبدیل شد. سپس کدهای هم‌معنا در سطح هر پلتفرم تجمیع و در مقوله‌های مفهومی سازمان‌دهی شدند. در گام تطبیقی، مقوله‌ها بر اساس پنج محور «موضوع و قواعد تعدیل‌گری»، «منشأ شناسایی و گزارش»، «بررسی و تصمیم‌گیری»، «نوع مداخله» و «پاسخگویی و بازبینی» کنار یکدیگر قرار گرفتند. این محورها برای سامان‌دهی مقایسه نتایج به کار رفتند. در تفسیر نتایج، میان قواعد منتشرشده پلتفرم، برداشت متخصصان مصاحبه‌شونده و پیشنهادهای هنجاری الگوی نهایی تفکیک شد. تفکیک کدگذاری اولیه از تجمیع و نمایش مقوله‌ها با منطق تحلیل کیفی چندمرحله‌ای انطباق دارد (Miles et al., 2014).

- منطق انتخاب موارد و چارچوب مقایسه

چارچوب مقایسه با تمرکز بر کارکردهای مشترک تعدیل‌گری تنظیم شد. برای هر پلتفرم، موضوع قواعد محتوا، مسیر تشخیص یا گزارش، مرجع بررسی، نوع اقدام و سازوکار بازبینی بررسی شد. مقایسه در دو سطح صورت گرفت. نخست، مقایسه هم‌نوع یوتیوب و آپارات در محیط ویدئومحور؛ دوم، مقایسه میان ایکس و دیوار برای شناسایی اجزایی که به نوع خدمت وابسته‌اند یا قابلیت تعمیم مفهومی دارند.

- جایگاه داده‌های اسنادی و مصاحبه‌ای

در تحلیل، اسناد رسمی برای شناسایی قواعد اعلام‌شده و مسیرهای تعریف‌شده رسیدگی به تخلف مبنا قرار گرفتند. مصاحبه‌ها به روشن‌شدن تجربه اجرایی، ملاحظات حقوقی و محتوایی و محدودیت‌های عملی در پلتفرم‌های داخلی کمک کردند. در گزارش یافته‌ها، گزاره‌های ناظر به اسناد از توضیح‌های حاصل از مصاحبه تفکیک شده‌اند. در مواردی که شواهد صرفاً از سیاست منتشرشده به دست آمده، نتیجه به «سیاست اعلام‌شده» محدود شده است و به عملکرد واقعی پلتفرم تعمیم داده نمی‌شود.

- مسیر تحلیل

جدول یک توالی تحلیل را نشان می‌دهد. اعداد کدها در جدول دو به تفکیک پلتفرم گزارش شده‌اند و چهار خوشه مقایسه‌ای بخش یافته‌ها، بازآرایی مفهومی همین نتایج‌اند. الگوی نهایی از ترکیب کارکردهای مشترک گزارش‌شده با نکات نظری ادبیات تعدیل‌گری استخراج شده است؛ در واقع می توان گفت که اجزای آن در جایگاه «پیشنهاد مفهومی پژوهش» قرار دارند.

جدول ۱. فرایند تحلیل داده‌ها در پژوهش

مرحله تحلیل	منبع داده	تحلیل	خروجی
کدگذاری اولیه	اسناد رسمی و مصاحبه‌ها	استخراج گزاره‌های مرتبط با قواعد، تشخیص، تصمیم و رسیدگی	کدهای اولیه هر مورد
تجمیع مفهومی	کدهای اولیه هر پلتفرم	ادغام مفاهیم هم‌معنا و حفظ تفاوت موضوع تخلف	کدهای مفهومی هر مورد

مرحله تحلیل	منبع داده	تحلیل	خروجی
مقوله‌بندی و مقایسه	کدهای مفهومی و مقوله‌ها	تنظیم مقوله‌های موردی و مقایسه بر پایه کارکردهای مشترک	مقوله‌های موردی و خوشه‌های تطبیقی

یافته‌های پژوهش

یافته‌های اسنادی نشان می‌دهد چهار پلتفرم در تعریف محتوای ناقض قواعد، سازوکار دریافت گزارش و پیش‌بینی اقدام در برابر تخلف اشتراک‌هایی دارند؛ با این حال، موضوعات اصلی نظارت در آن‌ها یکسان نیست. در یوتیوب، ایکس و آپارات، قواعد مربوط به محتوای رسانه‌ای و تعامل کاربران وجود دارد؛ در دیوار، صحت اطلاعات آگهی، هویت، امنیت تعامل و مسئولیت درج‌کننده آگهی اهمیت بیشتری دارد. این تفاوت، حاصل ماهیت خدمات پلتفرم‌هاست. از مصاحبه‌های انجام‌شده درباره دو پلتفرم داخلی، برای توضیح ملاحظات اجرایی و زمینه‌ای این قواعد استفاده شده است.

در گزارش کدگذاری، ۵۵۹ کد اولیه در چهار مورد شناسایی شد؛ ۱۳۹ کد برای یوتیوب، ۱۳۶ کد برای آپارات، ۱۹۶ کد برای دیوار و ۸۸ کد برای ایکس شناسایی شدند. پس از تجمیع، ۱۱۸ کد مفهومی در سطح موارد گزارش شد که به‌ترتیب ۳۰، ۳۲، ۳۱ و ۲۵ کد بودند. این کدها در مجموع در ۴۱ مقوله موردی قرار گرفتند. ارقام یادشده جمع مقوله‌های گزارش‌شده برای هر پلتفرم‌اند و نباید با تعداد مقوله‌های یکتای مشترک میان چهار پلتفرم یکسان تلقی شوند. همچنین تفاوت تعداد کدها به‌تنهایی شاخص پیچیدگی یا کیفیت نظام تعدیل‌گری پلتفرم نیست.

جدول ۲. خلاصه کدگذاری و مقوله‌بندی در پلتفرم‌های مورد مطالعه

پلتفرم	کدهای اولیه	کدهای مفهومی	مقوله‌های نهایی	ویژگی
یوتیوب	۱۳۹	۳۰	۹	ترکیب قواعد جامعه کاربری، برنامه نشانه‌گذاران معتمد و تعدیل‌گری الگوریتمی
آپارات	۱۳۶	۳۲	۹	ترکیب گزارش مردمی، یادگیری ماشین و انطباق با قوانین داخلی
دیوار	۱۹۶	۳۱	۱۳	تأکید بر احراز هویت، امنیت معامله و مسئولیت کاربر در آگهی‌ها
ایکس	۸۸	۲۵	۱۰	تمرکز بر ایمنی، اصالت، حریم خصوصی و برخورد مرحله‌ای با نقض قوانین

- مقوله‌های مشترک تعدیل‌گری

برای مقایسه میان‌موردی، مقوله‌های گزارش‌شده در چهار خوشه محتوایی قرار گرفتند: ۱) حریم خصوصی، هویت و مالکیت؛ ۲) قواعد انتشار و مدیریت محتوا؛ ۳) حقوق کاربران، ایمنی و اخلاق؛ و ۴) اختیار پلتفرم، رسیدگی به تخلف و بازبینی تصمیم. در واقع این خوشه‌ها، سطحی از تجمیع تحلیلی برای

مقایسه‌اند. در هر خوشه، موضوعات اختصاصی پلتفرم‌ها آورده شده است؛ برای مثال، امنیت معامله در دیوار با قواعد انتشار ویدئو در یوتیوب از یک جنس نیست، هرچند هر دو در نظام مدیریت تخلف جای می‌گیرند.

در اسناد بررسی‌شده، استفاده از ابزارهای خودکار، رسیدگی انسانی و دریافت گزارش کاربران به‌عنوان اجزای قابل شناسایی فرایند تعدیل‌گری مطرح‌اند؛ اما میزان اتکا به هر جزء و ترتیب مداخله در چهار پلتفرم یکسان گزارش نشده است. از این رو، نتیجه تطبیقی پژوهش وجود کارکردهای مکمل در سطح الگوی مفهومی است، و اثبات اجرای یک فرایند واحد یا سنجش میزان کارایی ابزارهای هوشمند در همه موارد نیست.

جدول ۳. تطبیق کارکردهای تعدیل‌گری در چهار مورد

محور مقایسه	یوتیوب و آپارات	ایکس	دیوار
نوع محتوای اصلی	ویدئو و تعاملات مرتبط	متن و گفت‌وگوی تعاملی	آگهی و اطلاعات تجاری
قواعد	انتشار ویدئو و تخلفات محتوایی	ایمنی، اصالت و قواعد تعامل	اصالت آگهی، هویت و امنیت تعامل
مسیر شناسایی	گزارش کاربر و ابزارهای بررسی	گزارش کاربر و قواعد رسیدگی	بررسی آگهی و گزارش تخلف
نوع مداخله	محدودسازی یا حذف محتوای ناقض قواعد	محدودیت محتوا یا حساب	رد، اصلاح یا محدودسازی آگهی
حد مقایسه	مقایسه هم‌نوعِ رسانه ویدئویی	مقایسه میان‌نوعِ محیط تعاملی	مقایسه میان‌نوعِ بازارگاه

- اطلاعات نادرست و سایر مصادیق تخلف محتوایی

در پلتفرم‌های رسانه‌ای، اطلاعات نادرست و محتوای گمراه‌کننده یکی از موضوعات سیاست‌گذاری محتواست و ممکن است با اقداماتی مانند برچسب‌گذاری، محدودسازی نمایش یا حذف مواجه شود. این موضوع با سایر دسته‌های تخلف، از جمله آزار، جعل هویت، نقض حریم خصوصی و نقض مالکیت فکری، همپوشانی کامل ندارد. از این رو، نتایج مربوط به این دسته‌ها در تحلیل به‌عنوان مصادیق متفاوت قواعد تعدیل‌گری حفظ شده‌اند و به‌طور یکسان «اخبار جعلی» نامیده نشده‌اند.

در دیوار، مسئله محوری بیشتر به اصالت اطلاعات آگهی، صحت ادعاهای تجاری و پیشگیری از سوءاستفاده در تعاملات مرتبط است. این موارد از نظر منطق نظارت با مسئله اطلاعات گمراه‌کننده ارتباط دارند، اما ماهیت آن‌ها لزوماً خبری نیست. در آپارات نیز قواعد انتشار ویدئو و رسیدگی به گزارش کاربران در بستر رسانه‌ای تعریف می‌شود. بنابراین، مقایسه این موارد در سطح سازوکارهای تعدیل‌گری انجام شده است و بر پایه فرض یکسان‌بودن مصادیق تخلف نیست.

- نقش انسان و هوش مصنوعی در تعدیل‌گری

داده‌های اسنادی و تفسیرهای حاصل از مصاحبه‌های داخلی، سه نقش متمایز در فرایند تعدیل‌گری دارد. ابزارهای خودکار برای شناسایی یا اولویت‌بندی موارد مشکوک، نیروی انسانی برای رسیدگی به موارد نیازمند

قضاوت زمینه‌ای، و کاربران برای گزارش تخلف یا اعتراض به تصمیم است. این تقسیم‌بندی، کارکردهای اصلی الگوی پیشنهادی را مشخص می‌کند؛ با این حال، از اسناد عمومی نمی‌توان درباره سهم دقیق تصمیم‌های خودکار و انسانی یا عملکرد فنی سامانه‌های هر پلتفرم نتیجه‌گیری کرد. در مدل مفهومی پژوهش، ارزیابی انسانی به مواردی اختصاص می‌یابد که تشخیص تخلف به فهم زمینه، ملاحظات حقوقی یا تفسیر معنای محتوا وابسته است. گزارش کاربر نیز می‌تواند مستقل از تشخیص خودکار، آغازگر بررسی باشد. بر این اساس، انسان و الگوریتم دو مسیر مکمل برای شناسایی و رسیدگی به موارد مشکوک به شمار می‌روند.

- مقایسه پلتفرم‌های داخلی و خارجی

مقایسه میان‌موردی دو دسته تفاوت را آشکار کرد. دسته نخست به نوع خدمت مربوط است. در پلتفرم‌های رسانه‌ای، سیاست‌های انتشار و تعامل محتوایی محور اصلی‌اند؛ در بازارگاه آگهی، اعتبار اطلاعات معامله و هویت کاربران برجسته‌تر است. دسته دوم به بستر حقوقی و سازمانی مربوط می‌شود. اسناد پلتفرم‌های داخلی بر انطباق با الزامات حقوقی ایران تأکید دارند، درحالی‌که پلتفرم‌های بین‌المللی مجموعه‌ای از قواعد عمومی و سازوکارهای رسیدگی در محیط چندملیتی منتشر می‌کنند. این یافته‌ها ناظر به متن سیاست‌های بررسی‌شده‌اند و نه ارزیابی میزان اجرای آن‌ها. در عین تفاوت‌ها، پنج کارکرد در سطح تطبیقی قابل ردیابی است که شامل تعیین قواعد، شناسایی مورد مشکوک، بررسی و تصمیم، اعمال مداخله و رسیدگی پس از تصمیم، است. این اشتراک کارکردی، مبنای تدوین الگوی مفهومی قرار گرفت. بنابراین، مدل نهایی قرار نیست سیاست واحدی را بدون توجه به ماهیت خدمات بر همه پلتفرم‌ها تحمیل کند؛ بلکه چارچوبی برای تنظیم روابط میان کارکردهای مشترک و تصمیم‌های وابسته به زمینه فراهم می‌کند. نتایج مربوط به قواعد محتوا، دسته‌های تخلف، گزارش‌دهی و امکان مداخله عمدتاً از اسناد عمومی پلتفرم‌ها به دست آمده است. داده‌های مصاحبه‌ای، به دلیل تمرکز بر دیوار و آپارات، در تفسیر ملاحظات اجرایی دو مورد داخلی به کار رفته‌اند. چهار خوشه مفهومی گزارش‌شده، جمع‌بندی میان‌موردی پژوهش‌اند؛ در مقابل، توصیه‌هایی مانند معیار ارجاع به انسان، ثبت مبنای تصمیم و چرخه بازنگری نیز، در بخش مدل مفهومی ارائه می‌شوند.

الگوی مفهومی تعدیل‌گری هوشمند محتوا

الگوی مفهومی پیشنهادی با اتکا به خوشه‌های تحلیلی یافته‌ها و مبانی نظری تعدیل‌گری ترکیبی اجرا شد. این الگو هفت رکن دارد: ورود محتوا، تشخیص هوشمند، گزارش‌دهی کاربران، ارزیابی انسانی، مداخله پلتفرم، شفافیت و پاسخگویی، و یادگیری و به‌روزرسانی. پنج رکن نخست وظایف عملیاتی را پوشش می‌دهند و دو رکن پایانی سازوکارهای نظارت بر تصمیم و اصلاح فرایند را توضیح می‌دهند. این ارکان به‌صورت یک زنجیره خطی و یکسان در همه پلتفرم‌ها اجرا نمی‌شوند؛ گزارش کاربران و تشخیص خودکار می‌توانند دو مسیر مستقل آغاز بررسی باشند.

جدول ۴. ارکان الگوی مفهومی تعدیل‌گری هوشمند محتوا و کارکردهای آن‌ها

لایه/رکن الگو	کارکرد اصلی	خروجی مورد انتظار
ورودی محتوا	محتوای منتشرشده یا در آستانه انتشار	ثبت ورودی برای بررسی
تشخیص هوشمند	تحلیل اولیه با ابزارهای خودکار در صورت وجود زیرساخت	اولویت‌بندی و ارجاع موارد مشکوک
گزارش‌دهی کاربران	گزارش مستقل از تشخیص خودکار و اعتراض به تصمیم	ثبت گزارش تخلف یا درخواست بازبینی
ارزیابی انسانی	بررسی محتوای نیازمند قضاوت انسانی	تصمیم زمینه‌مند در موارد ارجاعی
مداخله پلتفرم	حفظ، توضیح، محدودسازی یا حذف محتوا طبق قواعد	اعمال اقدام متناسب با نوع تخلف
شفافیت و پاسخگویی	اطلاع‌رسانی، اعتراض و بازنگری تصمیم	ثبت مبنا و فراهم‌کردن امکان بازبینی
یادگیری و به‌روزرسانی	استفاده از نتایج رسیدگی برای اصلاح فرایند	بازنگری در قواعد و ابزارهای تشخیص

در الگوی پیشنهادی، محتوای منتشرشده یا در آستانه انتشار می‌تواند از مسیر تشخیص خودکار، گزارش کاربران یا بررسی معمول پلتفرم وارد فرایند رسیدگی شود. پس از ارزیابی نوع تخلف و میزان ابهام، تصمیم متناسب از میان حفظ محتوا، افزودن توضیح، محدودسازی، حذف یا اعمال محدودیت بر حساب انتخاب می‌شود. ارجاع به ارزیابی انسانی برای موارد وابسته به زمینه یا دارای پیامد مهم پیش‌بینی شده است. مشارکت کاربران دو کارکرد متفاوتِ گزارش محتوای مشکوک و اعتراض به تصمیم اتخاذشده دارد. گزارش‌ها خود به معنای اثبات تخلف نیستند و باید در کنار شواهد و قواعد پلتفرم بررسی شوند. اعتراض نیز ورودی مرحله بازبینی است و نتیجه آن می‌تواند به تأیید یا اصلاح تصمیم قبلی منجر شود. تفکیک این دو کارکرد، امکان طراحی فرایندی پاسخگو و کاهش خطاهای ناشی از گزارش‌های نادرست یا هماهنگ را فراهم می‌کند.

شفافیت و پاسخگویی در این مدل، لایه‌ای پس از تصمیم و در عین حال ناظر بر کل فرایند است. ثبت مبنای تصمیم، اطلاع‌رسانی متناسب به کاربر، امکان اعتراض و تحلیل موارد نقض‌شده، داده لازم برای بازنگری قواعد و بهبود ابزارهای تشخیص را فراهم می‌سازد. از این رو، رابطه ارکان مدل بازگشتی است، نتایج رسیدگی و بازبینی به اصلاح سیاست‌ها و فرایندهای تشخیص بازمی‌گردد. مدل حاضر یک مدل مفهومی مبتنی بر تحلیل تطبیقی است و ارزیابی میدانی کارایی آن در دامنه پژوهش قرار نداشته است.

رکن «ورود محتوا» به دامنه محتوای تولیدشده توسط کاربران مربوط است؛ «تشخیص هوشمند» و «گزارش‌دهی» به مسیرهای شناسایی؛ «ارزیابی انسانی» و «مداخله» به رسیدگی و اعمال قواعد؛ و «شفافیت و پاسخگویی» و «یادگیری و به‌روزرسانی» به بازبینی و اصلاح تصمیم‌ها ارتباط دارند. این نگاشت، پیوند میان ابعاد مقایسه و طراحی مدل را روشن می‌کند. در عین حال، سازوکارهایی چون ثبت بازخورد و اصلاح الگوریتم‌ها، پیشنهادهای طراحی حاصل از تلفیق یافته‌ها با ادبیات نظری‌اند و نباید به‌عنوان رویه اجراشده در همه پلتفرم‌های مورد مطالعه گزارش شوند.

بحث

یافته‌های پژوهش نشان داد که تعدیل‌گری هوشمند محتوا را نمی‌توان به مسئله فنی تشخیص اخبار جعلی فروکاست. در چهار پلتفرم بررسی‌شده، قواعد محتوا، نوع خدمت، شیوه گزارش‌دهی، تصمیم‌های انسانی و امکانات مداخله در کنار یکدیگر معنا پیدا می‌کنند. نتیجه اصلی مقایسه آن است که سازوکارهای مشترک تعدیل‌گری را می‌توان در سطح کارکرد شناسایی کرد، درحالی‌که موضوع تخلف و شیوه اجرای آن وابسته به ماهیت هر پلتفرم باقی می‌ماند.

از منظر نظری، پژوهش حاضر نشان می‌دهد که تعدیل‌گری محتوا را باید بخشی از حکمرانی پلتفرمی دانست. پلتفرم‌ها با تدوین سیاست‌های محتوا، طراحی الگوریتم‌های توصیه‌گر، تعیین سازوکارهای گزارش‌دهی و انتخاب نوع جریمه، در عمل بر ساختار گفت‌وگوی عمومی اثر می‌گذارند. بنابراین، تعدیل‌گری صرفاً عملیاتی برای حذف محتوای نامناسب نیست، بلکه سازوکاری برای توزیع قدرت ارتباطی، مدیریت اعتماد و تعریف مرزهای گفتار مشروع در فضای دیجیتال است.

از منظر کاربردی، نتایج پژوهش بر ضرورت طراحی سامانه‌های چندلایه تأکید دارد. پلتفرم‌هایی که فقط بر گزارش کاربران متکی‌اند، ممکن است در مواجهه با حجم زیاد محتوا کُند عمل کنند یا تحت تأثیر گزارش‌های هماهنگ و مغرضانه قرار گیرند. پلتفرم‌هایی که صرفاً بر هوش مصنوعی تکیه می‌کنند، با خطر خطاهای زمینه‌ای و حذف ناعادلانه محتوا مواجه‌اند. همچنین، نظام‌های صرفاً انسانی نیز از نظر هزینه، سرعت و فشار روانی پایدار نیستند. بنابراین، ترکیب متوازن این عناصر، راهکار واقع‌بینانه‌تری برای مدیریت اخبار جعلی است.

مقایسه پلتفرم‌های داخلی و خارجی نشان می‌دهد که تفاوت محیط حقوقی و نوع خدمت، بر اولویت‌های سیاست‌گذاری محتوا اثر می‌گذارد. یوتیوب و آپارات به دلیل ماهیت ویدئومحور، مبنای مقایسه مستقیم‌تری دارند؛ ایکس ویژگی‌های محیط تعاملی و گفت‌وگویی را نشان می‌دهد؛ و دیوار مرزهای کاربرد الگو در یک بازارگاه آگهی را روشن می‌کند. بنابراین، ارزش تحلیلی انتخاب موارد در مشاهده تنوع زمینه‌ها و تشخیص کارکردهای مشترک است، و رتبه‌بندی کیفیت تعدیل‌گری پلتفرم‌ها، نیست.

از منظر کاربردی، الگوی پیشنهادی برای تدوین رویه‌های روشن‌تر در زمینه تشخیص و ارجاع، مستندسازی مبنای تصمیم، تفکیک نقش انسان و سامانه هوشمند و ایجاد مسیر اعتراض قابل استفاده است. انتقال این چارچوب به پلتفرم‌های فارسی‌زبان مستلزم انطباق آن با نوع محتوا، قواعد حقوقی و ظرفیت اجرایی هر پلتفرم خواهد بود. به همین دلیل، پیشنهادهای این بخش از یافته‌های توصیفی تفکیک می‌شوند و در جایگاه دلالت‌های کاربردی پژوهش قرار می‌گیرند. برای رسیدگی به اطلاعات نادرست و سایر تخلفات محتوایی، حذف تنها گزینه ممکن نیست. متناسب با نوع تخلف و خطر احتمالی، اقداماتی مانند توضیح زمینه، برچسب‌گذاری، کاهش نمایش، محدودسازی بازنشر و حذف قابل طرح‌اند (Gillespie, 2022). انتخاب میان این اقدامات به قواعد پلتفرم و ارزیابی مورد وابسته است. این طیف مداخله، در الگوی مفهومی به‌عنوان مجموعه گزینه‌های تصمیم‌گیری پیش‌بینی شده است.

در نهایت، برای پلتفرم‌های فارسی‌زبان، توسعه ابزارهای بومی پردازش زبان طبیعی و داده‌های آموزشی معتبر اهمیت ویژه دارد. بسیاری از خطاهای هوش مصنوعی در تعدیل‌گری ناشی از ضعف در شناخت بافت زبانی و فرهنگی است. بنابراین، سرمایه‌گذاری در داده‌های برچسب‌خورده فارسی، مدل‌های تشخیص زمینه، سامانه‌های تحلیل روایت و ابزارهای ارزیابی چندزبانه، می‌تواند کیفیت تعدیل‌گری هوشمند را در فضای فارسی به‌طور معناداری افزایش دهد.

نتیجه‌گیری

پژوهش حاضر با هدف طراحی الگوی مفهومی تعدیل‌گری هوشمند محتوای تولیدشده توسط کاربران و با مقایسه چهار پلتفرم یوتیوب، ایکس، آپارات و دیوار انجام شد. یافته‌ها نشان داد که باوجود تفاوت نوع خدمت و بستر حقوقی، تعیین قواعد، شناسایی محتوای مشکوک، ارزیابی، مداخله و بازبینی تصمیم‌ها کارکردهای قابل مقایسه‌اند. اخبار جعلی و اطلاعات گمراه‌کننده یکی از مصادیق این فرایند به شمار می‌روند و در همه پلتفرم‌ها مصداق واحدی ندارند. الگوی پیشنهادی در هفت رکن، وظایف فناوری، نیروی انسانی، کاربران و مدیریت پلتفرم را نشان می دهد. در این الگو، تشخیص خودکار و گزارش کاربران می‌توانند به‌صورت موازی مورد مشکوک را شناسایی کنند؛ ارزیابی انسانی در موارد نیازمند تفسیر زمینه‌ای به کار می‌آید؛ و تصمیم پلتفرم با سازوکار اطلاع‌رسانی، اعتراض و بازبینی تکمیل می‌شود. نتیجه بازبینی نیز می‌تواند مبنای اصلاح قواعد و ابزارهای تشخیص قرار گیرد.

سهم علمی پژوهش در صورت‌بندی یک چارچوب تطبیقی مبتنی بر کارکردهای مشترک تعدیل‌گری در چهار نوع بستر و تفکیک آن از مصادیق اختصاصی تخلف است. سهم کاربردی آن، ارائه نقشه‌ای مفهومی برای تنظیم روابط میان تشخیص هوشمند، قضاوت انسانی، گزارش کاربران و پاسخگویی نهادی در پلتفرم‌های فارسی‌زبان است. این نتیجه بر تحلیل سیاست‌های منتشرشده و مصاحبه‌های مکمل استوار است. بر این اساس، تدوین طبقه‌بندی روشن تخلفات، تعریف معیارهای ارجاع به انسان، ثبت مبنای تصمیم، فراهم‌کردن مسیر اعتراض و بازنگری دوره‌ای قواعد پیشنهاد می‌شود. برای پلتفرم‌های فارسی‌زبان، توجه به داده‌های زبانی بومی و آموزش تعدیل‌گران در تشخیص زمینه نیز اهمیت دارد. اجرای این پیشنهادها باید با مأموریت، ظرفیت فنی و الزامات حقوقی هر پلتفرم متناسب شود.

محدودیت اصلی این پژوهش، اتکا به سیاست‌ها و اسناد در دسترس عمومی برای مقایسه چهار پلتفرم و تمرکز مصاحبه‌های مکمل بر تجربه داخلی است. ازاین‌رو، ارزیابی دقت فنی سامانه‌ها، نسبت واقعی تصمیم‌های انسانی و خودکار و آثار تصمیم‌ها بر رفتار کاربران در دامنه این مطالعه قرار نداشت. پژوهش‌های آینده می‌توانند الگوی پیشنهادی را با داده‌های عملیاتی، مطالعه تجربه کاربران و آزمون‌های میدانی اعتبارسنجی کنند.

منابع

آزادی، محمدحسین، ساعی، محمدحسین، و البرزی دعوتی، هادی (۱۳۹۸). تفاوت‌های خبر جعلی و شایعه از منظر اسناد علمی، کارشناسان و نظریه‌پردازان ژورنالیسم، خبر و علوم ارتباطات. *مطالعات میان‌رشته‌ای ارتباطات و رسانه*، ۲(۱)، ۱۱-۳۶.

آپارات (بی‌تا). *قوانین و دستورالعمل‌های جامعه کاربران آپارات*. https://www.aparat.com/community-guideline

باقری، س.، و آصفی، ح. ر. (۱۴۰۰). تحلیل گفتمان اخبار جعلی در رسانه‌های اجتماعی: مطالعه موردی انتخابات ریاست‌جمهوری سال ۱۴۰۰. *مطالعات رسانه*، ۲۷(۲). ۵۹–۸۴.

دیوار(بی تا). *مرکز پشتیبانی دیوار*. https://divar.ir/help

مظاهری، م. م.، و تهرانی، ا. ح. (۱۴۰۱). مطالعه تطبیقی الگوهای حکمرانی فضای مجازی در جهان. *مطالعات بین‌المللی*، ۵۸(۱۵)، ۱۳۱–۱۵۶.

Aïmeur, E., Amri, S., & Brassard, G. (2023). Fake news, disinformation and misinformation in social media: A review. *Social Network Analysis and Mining, 13*(1). https://doi.org/10.1007/s13278-023-01028-5

Auxier, B. (2020, January 27). Most Americans support right to have some personal info removed from online searches. Pew Research Center. https://www.pewresearch.org/short-reads/2020/01/27/most-americans-support-right-to-have-some-personal-info-removed-from-online-searches/

Azzimonti, M., & Fernandes, M. (2023). Social media networks, fake news, and polarization. *European Journal of Political Economy, 76*, 102256. https://doi.org/10.1016/j.ejpoleco.2022.102256

Bacchi, U. (2021, December 20). Online content moderation: Can AI help clean up social media? Thomson Reuters Foundation. https://news.trust.org/item/20211217152644-j0521

Balkin, J. M. (2004). Digital speech and democratic culture: A theory of freedom of expression for information society. *New York University Law Review, 79*(1), 1–55.

Balkin, J. M. (2017). Free speech in the algorithmic society: Big data, private governance, and new school speech regulation [Working paper]. *SSRN*. https://doi.org/10.2139/ssrn.3038939

Ballard, J. (2019, April 29). Most conservatives believe removing content and comments on social media is suppressing free speech. *YouGov*. https://yougov.com/en-us/articles/23164-content-moderation-social-media-free-speech-poll

Banchik, A. V. (2021). Disappearing acts: Content moderation and emergent practices to preserve at-risk human rights–related content. *New Media and Society*, *23*(6). https://doi.org/10.1177/1461444820912724

Bandy, J., & Diakopoulos, N. (2021). Curating quality? How Twitter's timeline algorithm treats different types of news. *Social Media + Society*, *7*(3). https://doi.org/10.1177/20563051211041648

Bateman, G., Abdel Haleem, H., & Majumdar, A. (2021). Is user-generated social media content useful for informing planning and management of emergency events? An investigation of an active shooting event in a U.S. airport. *Case Studies on Transport Policy*, *9*(3), 1015–1025. https://doi.org/10.1016/j.cstp.2021.05.003

Black, J. (2008). Constructing and contesting legitimacy and accountability in polycentric regulatory regimes. *Regulation & Governance*, *2*(2), 137–164. https://doi.org/10.1111/j.1748-5991.2008.00034.x

Common, M. F. (2020). Fear the Reaper: How content moderation rules are enforced on social media. *International Review of Law, Computers and Technology*, *34*(2). https://doi.org/10.1080/13600869.2020.1733762

De Streel, A., Defreyne, E., Jacquemin, H., Ledger, M., Michel, A., Innesti, A., Goubet, M., & Ustowski, D. (2020). *Online platforms' moderation of illegal content online*. European Parliament.

Dehghan, E., & Glazunova, S. (2021). 'Fake news' discourses: An exploration of Russian and Persian tweets. *Journal of Language and Politics*, *20*(5), 741–760. https://doi.org/10.1075/jlp.21032.deh

DeNardis, L., & Hackl, A. M. (2015). Internet governance by social media platforms. *Telecommunications Policy*, *39*(9), 761–770. https://doi.org/10.1016/j.telpol.2015.04.003

Dowdeswell, T., & Goltz, N. (2020). The clash of empires: Regulating technological threats to civil society. *Information & Communications Technology Law*, *29*(2), 194–217. https://doi.org/10.1080/13600834.2020.1735060

Flew, T., Martin, F., & Suzor, N. (2019). Internet regulation as media policy: Rethinking the question of digital communication platform governance. *Journal of Digital Media & Policy*, *10*(1), 33–50. https://doi.org/10.1386/jdmp.10.1.33_1

Gerbina, T. V. (2021). Science disinformation: On the problem of fake news. *Scientific and Technical Information Processing*, *48*(4), 290–298. https://doi.org/10.3103/S0147688221040092

Gillespie, T. (2015). Platforms intervene. *Social Media + Society*, *1*(1). https://doi.org/10.1177/2056305115580479

Gillespie, T. (2018). *Custodians of the internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press. https://doi.org/10.12987/9780300235029

Gillespie, T. (2020). Content moderation, AI, and the question of scale. *Big Data and Society*, *7*(2). https://doi.org/10.1177/2053951720943234

Gillespie, T. (2022). Do Not Recommend? Reduction as a Form of Content Moderation. *Social Media and Society*, *8*(3). https://doi.org/10.1177/20563051221117552

Godwin, M. (2003). *Cyber rights: Defending free speech in the digital age (Rev. and updated ed.)*. MIT Press.

Gorwa, R. (2019). What is platform governance? *Information, Communication & Society*, *22*(6), 854–871. https://doi.org/10.1080/1369118X.2019.1573914

Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data and Society*, *7*(1). https://doi.org/10.1177/2053951719897945

Helberger, N. (2020). The political power of platforms: How current attempts to regulate misinformation amplify opinion power. *Digital Journalism*, *8*(6), 842–854. https://doi.org/10.1080/21670811.2020.1773888

Martínez, M. B. (2023). Platform regulation, content moderation, and AI-based filtering tools: Some reflections from the European Union. *Journal of Intellectual Property, Information Technology and E-Commerce Law*, *14*(1), 211–225. https://www.jipitec.eu/jipitec/article/view/375

Miles, M. B., Huberman, A. M., & Saldaña, J. (2014). *Qualitative Data Analysis: A Methods Sourcebook* (3rd ed.). SAGE Publications.

Molina, M. D., & Sundar, S. S. (2022). When AI moderates online content: Effects of human collaboration and interactive transparency on user trust. *Journal of Computer-Mediated Communication*, *27*(4), Article zmac010. https://doi.org/10.1093/jcmc/zmac010

Monti, G. (2021). The Digital Markets Act: Improving its institutional design. *European Competition and Regulatory Law Review*, *5*(2), 90–101. https://doi.org/10.21552/core/2021/2/4

Mueller, M. L. (2015). Hyper-transparency and social control: Social media as magnets for regulation. *Telecommunications Policy*, *39*(9), 804–810. https://doi.org/10.1016/j.telpol.2015.05.001

Newton, C. (2019, February 25). The trauma floor: The secret lives of Facebook moderators in America. The Verge. https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona

Ponce, A. (2020). The Digital Services Act Package: Reflections on the EU Commission's Policy Options. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3699389

Puppis, M. (2010). Media governance: A new concept for the analysis of media policy and regulation. *Communication, Culture & Critique, 3*(2), 134–149. https://doi.org/10.1111/j.1753-9137.2010.01063.x

Ruckenstein, M., & Turunen, L. L. M. (2020). Re-humanizing the platform: Content moderators and the logic of care. *New Media and Society, 22*(6), 1026–1042. https://doi.org/10.1177/1461444819875990

Shu, K., Mahudeswaran, D., Wang, S., & Liu, H. (2020). Hierarchical propagation networks for fake news detection: Investigation and exploitation. *Proceedings of the International AAAI Conference on Web and Social Media, 14*(1), 626–637. https://doi.org/10.1609/icwsm.v14i1.7329

Urban, J. M., Karaganis, J., & Schofield, B. L. (2016). Notice and takedown in everyday practice. *UC Berkeley Public Law Research Paper*, 2755628.

Van Dijck, J. (2013). *The culture of connectivity: A critical history of social media*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199970773.001.0001

Van Dijck, J., De Winkel, T., & Schäfer, M. T. (2023). Deplatformization and the governance of the platform ecosystem. *New Media & Society, 25*(12). https://doi.org/10.1177/14614448211045662

Yin, R. K. (2018). *Case Study Research and Applications: Design and Methods* (6th ed.). SAGE Publications.

Young, G. K. (2021). How much is too much: The difficulties of social media content moderation. *Information and Communications Technology Law, 31*(1). https://doi.org/10.1080/13600834.2021.1905593

Zeng, J., & Kaye, D. B. V. (2022). From content moderation to visibility moderation: A case study of platform governance on TikTok. *Policy & Internet, 14*(1), 79–95. https://doi.org/10.1002/poi3.287

Conceptual Design of Intelligent Moderation in Media Platforms: A Comparative Study of YouTube, X, Aparat, and Divar

Somayeh Labafi *[1]
Associate Professor, Research Institute for Society and Information, Department of Social Studies of Information, Iranian Research Institute for Information Science and Technology (IranDoc), Tehran, Iran

Nahid Akrami
Master's Degree in Media Management, University of Tehran, Tehran, Iran

Datis Khajeheian
Associate Professor, Department of Media Management and Business Communication, Faculty of Management, University of Tehran, Tehran, Iran.

Abstract
This article aims to propose a conceptual model for the intelligent moderation of fake news on media platforms, capable of explaining the relationship between artificial intelligence, human oversight, and user participation in the processes of identifying, evaluating, and reducing the dissemination of false and misleading content. The significance of this issue lies in the fact that social networks, despite their extensive role in information circulation and public participation, have also become environments for the rapid spread of misinformation, the polarization of public opinion, and the erosion of social trust. This study adopts a qualitative approach based on comparative analysis and documentary research. To this end, the content moderation policies, guidelines, and practices of four platforms—YouTube, X, Aparat, and Divar—were examined and coded. To complement the documentary data, semi-structured interviews were conducted with experts from the social media ecosystem, including product managers, technical, legal, and content specialists, as well as experts in the field of media platforms. The data were analyzed through three stages: open coding, conceptual aggregation, and categorization. The findings indicate that effective moderation of fake news on media platforms does not rely on a single-dimensional mechanism; rather, it operates as a multilayered and hybrid system. Within this system, artificial intelligence algorithms are used for the initial and scalable detection of content; human moderators are responsible for contextual assessment, identifying borderline cases, and making final decisions; users contribute to the monitoring process through reporting and feedback; and platform policies define the legal and operational framework for addressing violations. The comparative analysis also showed that domestic platforms, while drawing on international experiences, have incorporated local legal and cultural requirements into their moderation policies. The study concludes that intelligent moderation is most effective when a balance is established among the speed and scalability of artificial intelligence, the interpretive sensitivity of human oversight, the transparency of platform policies, and responsible user participation. Sole reliance on artificial intelligence increases the risk of contextual errors and algorithmic bias, while complete dependence on human moderation is constrained by limitations related to speed, cost, and psychological pressure. Therefore, the proposed model emphasizes a hybrid, accountable, context-sensitive, and reviewable approach to content moderation.

Keywords: Content Moderation, Platform Regulation, Social Networks, Fake News, Intelligent Moderation

[1] Corresponding Author: labafi@irandoc.ac.ir