آزمایشگاه داده و حکمرانی
Data 4 Governance Lab

گزارش

تشریح پالایش هوشمند محتوای نامناسب در رسانه‌های صوت و تصویر فراگیر

امیرحسین سلامی راد، سید احمد حسینی

استودیو فناوری‌های تنظیم‌یار

زمستان ۱۳۹۹

بسم الله الرحمن الرحیم

نوع سند	**گزارش پروژه**
حوزه تخصصی	پژوهشکده‌ی سیاست‌گذاری دانشگاه شریف
برنامه	
عنوان	**تشریح پالایش هوشمند محتوای نامناسب در رسانه‌های صوت و تصویر فراگیر**
نگارنده/نگارندگان	**امیرحسین سلامی‌راد، سید احمد حسینی**
ناظر کیفی	عمادالدین پاینده
تاریخ انتشار	زمستان ۱۳۹۹
مخاطب	سازمان تنظیم مقررات صوت و تصویر فراگیر

جملات کلیدی

کلمات کلیدی

یادگیری ماشینی، انواع محتوای نامناسب، پالایش محتوا

فهرست عناوین

فهرست اشکال

فهرست جداول

۱. مقدمه

بنا بر نظر راقم این سطور، از دیرباز تا کنون، در پارادایم رسانه (به معنای عام آن) دو رهنمون، موسوم به رهنمون‌های ایجابی و سلبی مطرح بوده است که اگر چنانچه این دو در توازن و تعادل قرار گیرند، رسانه به حد ایده‌آل خود نزدیک خواهد شد؛ این دو عبارتند از:

۱. رهنمون ایجابی: نشر محتوای متنوّع و متعدد که به مذاق اقشار مختلف جامعه ــ با فرهنگ‌ها و ارزش‌های متکثر، سلایق و علایق متنوّع، جهت‌گیری‌های سیاسی و عقیدتی گوناگون و در رده‌بندی سنی مختلف ــ خوش آید. هدف از این رهنمون را می‌توان از یک سو، اعطای حق برابری به همه‌ی اقشار مختلف جامعه در نشر و مشاهده‌ی محتوا و از دیگر سو، جذابیت‌بخشی به رسانه و زدودن رسانه از جهت‌گیری‌های سیاسی و حاکمیتی دانست.

۲. رهنمون سلبی: مراقبت و دیده‌بانی از فضای رسانه که مبادا در آن محتوایی نشر یابد که در تضاد و تعارض با انگاره‌های فرهنگی، سیاسی و اجتماعیِ اقشار مختلف جامعه قرار گیرد و در این راستا بایستی از جانب نهادهای ذی‌ربط، قوانین سخت‌گیرانه‌ای وضع گردد.

رسانه‌ها در راستای نشر محتوا بایستی محتاطانه عمل کنند و میان دو بند فوق تعادل و توازنی برقرار نمایند تا نه به دلیل نشر حداکثری محتوا در ورطه‌ی پلورالیسم سقوط کنند و نه به دلیل حذف حداکثری محتوا به خاطر قوانین فوق سختگیرانه، دچار خمودی و کمبود محتوا گردند. لذا نهادهای مختلف در مقوله‌ی «حکمرانی رسانه»[1] تلاش‌هایی را مبذول داشته‌اند که در این راستا می‌توان به مقوله‌ی تعدیل محتوا[2] ــ که به تسامح به آن پالایش محتوا نیز گفته می‌شود ــ اشاره کرد.

همچنین قبل از ورود به رئوس مطالب، لازم است مفهوم «حکمرانی رسانه» که در بالا به آن اشاره شد نیز به درستی تنقیح گردد. حکمرانی رسانه، مفهومی است که توجه زیادی را در میان متخصصان حوزه‌ی ارتباطات به خود جلب کرده است. هرچند در مورد مفهوم حکمرانی، تعریف دقیق و متقنی وجود ندارد اما در یک تعریف جامع، حکمرانی رسانه عبارت است از مجموعه‌ی قواعد و مقرّراتی که هدف آن‌ها ساماندهی نظام‌های رسانه‌ای است. هنگامی که فحول و متخصصین حوزه‌ی ارتباطات از این مفهوم استفاده

1 Media Governance
2 Content moderation

می‌کنند، در واقع آنان به سیاست‌ها و خط مشی‌هایِ جدید در بخش رسانه، نظیر مفاهیمی چون نقش نهادهای تنظیم‌گر[۱]، خود تنظیم‌گیری[۲]، تنظیم‌گری مشارکتی[۳]، تقویت مشارکت شهروندان و توسعه‌ی فرآیندهای مرتبط با تصمیم‌گیری‌های غیررسمی در شبکه‌ها اشاره دارند [1].

امروزه عرصه‌ی حکمرانی رسانه تحت تأثیر تحولات فنّاوری دیجیتال و به موازات رشد و گسترش بازیگران نوظهور رسانه‌ای ــ نظیر پلتفرم‌های اشتراک‌گذاری محتوای آنلاین[٤]، سرویس‌های پخش ویدئو و موسیقی[٥] و شبکه‌های اجتماعی[٦] ــ کسوتی نو به تن کرده و نمودهای جدیدی به خود گرفته است. به گونه‌ای که امروزه شاهد تغییرات مهمی در نظام حکمرانی رسانه و بخش‌های مختلف آن هستیم. افزایش حجم محتوای تولیدی از جانب رسانه‌ها، کاهش هزینه‌های تولید محتوا، ظهور پلتفرم‌های کاربرپدید[۷] (UGC) و افزایش فرآیند سرعت نشر محتوا، از جمله مهمترین عواملی هستند که سبب تغییر در نظام حکمرانی رسانه در مقایسه با حکمرانی در عرصه‌ی رسانه‌های سنتی نظیر رادیو، تلویزیون و نشریات چاپی شده‌اند.

این تغییرات با افزایش همگرایی در فنّاوری و ساختار همراه بوده است به نحوی که مرزهای سیاستی و مقرراتی پیشین را با چالشی جدّی مواجه ساخته است. اگر پیش از این، تلویزیون ملّی به صورت انحصاری، وظیفه‌ی اطلاع‌رسانی، آگهی‌بخشی، تقویت نظام سیاسیِ حاکم، تبلیغات سیاسی و تجاری، تولید محتوای سرگرمی و آموزشی را بر عهده داشت، اینک بی‌شمار بازیگر مختلف در عرصه‌ی رسانه که به لحاظ زیرساخت و عملکرد، کارکرد متفاوتی دارند این بار را بر دوش خواهند کشید؛ لذا نهادهای حکمرانی رسانه نیز بایستی پابه‌پای این تغییرات، تمهیدات و تدبیراتی نو بیاندیشند و علی‌رغم تنظیم‌گری به واسطه‌ی وضع قوانین سلبی ـ ایجابی (Reg)، لاجرم با فنّاوری (Tech) (که در اینجا مراد، فنّاوری‌های مبتنی بر

---

۱ Regulators
۲ Self-regulatory
۳ Co-regulatory
٤ Online Content-Sharing Platforms
٥ Video and Audio streaming services
٦ Social Networks
۷ User generated content

یادگیری ماشینی است) نیز بیش از پیش آشنا شوند و آن را به کار بندند. برای نیل به این هدف، سازمان‌هایِ متولّیِ فنّاوری‌های موسوم به RegTech یا تنظیم‌یار ظهور و بروز می‌یابند که در حالت ایده‌آل، حلقه‌ی وصلی میان نهادهای تنظیم‌گر از یک سو و رسانه‌ها از سویی دیگر می‌باشند و بار تکنولوژیکیِ تنظیم‌گری ـ که مصداق بارز آن توسعه‌ی ماشین‌های کشف محتوای نامناسب است ـ را بر عهده می‌گیرند.

شکل ۱، فنّاوری RegTech به عنوان حلقه‌ی وصلی میان نهادهای تنظیم‌گر و رسانه‌ها

همانطور که پیش از این نیز متذکر شدیم، نهادهای تنظیم‌گرِ رسانه، متکفّلِ دیده‌بانیِ فضای رسانه‌اند و در این راستا قوانینی را وضع می‌کنند. چنانچه یک پلتفرم، محتوایی را نشر دهد که ناقضِ قوانین وضع شده و مغایر با انگاره‌های فرهنگی ـ سیاسی ـ اجتماعی باشد، از سوی نهادهای تنظیم‌گر دستوری مبنی بر حذف آن محتوا در پلتفرم‌های میزبان صادر میگردد و در برخی مواقع جریمه‌هایی نیز برای آن در نظرگرفته می‌شود. برای مثال در آذر ماه ۱۳۹۹، پیرو انتشار قسمت جدید سریال «شام ایرانی» که در آن به یکی از اقوام اصیل ایرانی توهین شده بود، ساترا (سازمان تنظیم مقررات صوت و تصویر فراگیر) به محض دریافت گزارش مردمی از سوی کاربران نسبت به بررسی مورد مذکور اقدام کرد. با بررسی صورت‌گرفته، مشخص شد مورد ذکرشده در تضاد آشکار با فصل «احترام به فرهنگ و اقوام کشور» در ضوابط محتوایی است؛ به همین جهت با دستور ساترا، این قسمت از همه رسانه‌های رسمی منتشرکننده حذف شد [2]. در سال ۲۰۱۸ نیز «نهاد صنعت و تجارت شهرداری پکن»[1] در چین، سوپر اپلیکیشن تاوتیائو[2] را به جهت نشر تبلیغات پزشکی (تبلیغ سه دارو به نام‌های Tongrentang Gegen Yam Capsule Tongrentang Anxin Capsule و Qidong Yangxue Capsule) بدون اخذ گواهیِ بررسی و صحت‌سنجی و نقض «قانون تبلیغات»[3] به میزان ۳ میلیون یوان جریمه کرد. بر اساس «قانون تبلیغات» ــ که به عنوان یک بند

1 Beijing Municipal Administration for Industry and Commerce
2 Toutiao
3 Advertising Law

قانونی در کشور چین شناخته می‌شود ــ هرنوع تبلیغ بر بستر پلتفرم‌ها بایستی پیش از نشر از سوی نهادهای ذی‌ربط، ارزیابی و صحت‌سنجی گردد [3].

همچنین کشور آلمان قانونی موسوم به NetzDG مبنی بر جریمه‌ی پلتفرم‌های صوت و تصویر فراگیر تا سقف ۵۰ میلیون یورو به دلیل عدم حذف محتوای مربوط به نفرت پراکنی وضع نموده است [4]. دولت فرانسه نیز «قوانین محتوای جدید پلتفرم»[1] را با اقتباس از تجربیات مشابه در کشور آلمان وضع کرده است که در آن لزوم پالایش محتوا مورد تأکید قرار گرفته است. کشور استرالیا نیز درصورت نقض قوانین مربوط به انتشار محتوای نامناسب، جریمه‌ای بالغ بر ۱۰ درصدِ درآمد سالانه و در مواردی تا ۳ سال زندان برای مدیران ارشد سازمان را در نظر می‌گیرد. برای برون‌رفت از این تکلفات و جرایم، رسانه‌های صوتی و تصویری موظف هستند تا ــ با استفاده از روش های جدید ــ مبتنی بر یادگیری ماشینی و عوامل انسان محور بر روی محتوای خود دقت نظر داشته باشند و آن‌ها را ارزیابی و پالایش نمایند [4].

به‌عنوان نمونه‌ای دیگر از اقدامات نهادهای قانون‌گذار در راستای پالایش محتوا، می‌توان به الزام کنگره آمریکا مبنی بر یافتن راهکارهای نوین برای حذف محتوای مستهجن جنسی اشاره کرد که با بهره‌گیری از آن و استفاده از ابزارهای مدرن مبتنی بر یادگیری ماشینی، میزبانی چنین محتوایی از ٤٣ درصد در سال ۲۰۱۷، به ۲۵ درصد در سال ۲۰۱۹ رسیده است [5].

پیرامون قوانین وضع شده و موجود در کشورهای توسعه‌یافته در راستای تنظیم‌گری و پالایش محتوا، در ادامه و به عنوان ذکر یک نمونه، تجربه‌ی اتحادیه‌ی اروپا را به تفصیل از نظر می‌گذرانیم.

۲. چارچوب پالایش محتوای نامناسب در اتحادیه اروپا

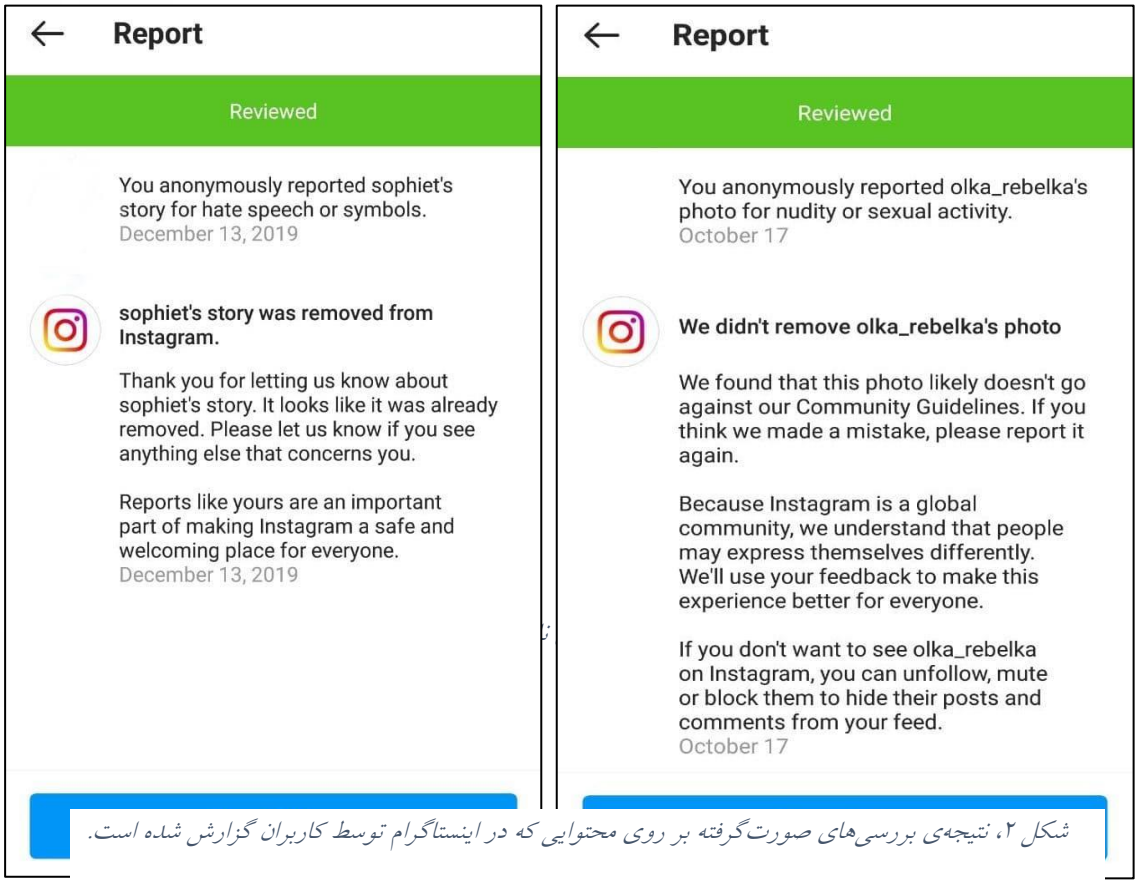
رویکرد پالایش و اصلاح محتوای نامناسب در اتحادیه‌ی اروپا ریسک ـ پایه است. بدین معنا که به دلیل حجم بالای محتوای منتشره و عدم توانایی در نظارت تمام محتوا، فرآیندهای تنظیم‌گریِ کلان به صورت پیشینی و پسینی وجود ندارد و بر مبنای ریسک‌های شناسایی شده، پالایش محتوا صورت می‌گیرد. مثلاً دستورالعمل سرویس‌های رسانه‌ای صوتی و تصویری اروپا (۲۰۱۸) الزامات اساسی را بر یک نوع خاص از پلتفرم‌های آنلاین موسوم به پلتفرم‌های اشتراک‌گذاری ویدئو قرار داده است. طبق این دستورالعمل

1 New Platform Content Rules

این نوع از پلتفرم‌ها موظف هستند اقدامات مقتضی و مناسب به منظور حفظ عموم از محتوای غیرقانونی را به عمل آورند و همچنین اقداماتی برای حفظ کودکان و نوجوانان از محتوای نامناسب انجام دهند. این اقدامات باید متناسب با نوع محتوا، گروهی که بایستی از آن‌ها محافظت شود، منافع مشروط در خطر، اندازه پلتفرم و ماهیت خدمت ارائه شده، باشد [6].

این دستورالعمل اقداماتی را برای مقابله با محتوای نامناسب پیشنهاد کرده است [6]:

۱. طراحی سازوکارهای کاربرپسند و شفاف برای گزارش[۱] و علامت‌گذاری[۲] محتوا،
۲. توسعه‌ی سیستمی که از طریق آن، پلتفرم‌های اشتراک‌گذاری ویدئو و تصویر به کاربران خود در ارتباط با تأثیر گزارش و نشانه‌گذاری‌های صورت گرفته، توضیحاتی ارائه می‌دهند. برای مثال اینستاگرام پس از بررسی محتوایی که توسط کاربر گزارش شده است، پیغامی مبنی بر نتیجه‌ی بررسی‌های صورت گرفته بر روی آن محتوا را برای کاربر ارسال می‌کند:



شکل ۲، نتیجه‌ی بررسی‌های صورت‌گرفته بر روی محتوایی که در اینستاگرام توسط کاربران گزارش شده است.

۱ Report
۲ Flagging

۳. توسعه‌ی سیستمی مناسب برای رده‌بندی سنی محتوا توسط کاربران.

٤. طراحی رویه‌های مؤثر، ساده و شفاف برای حل و فصل شکایت‌های کاربران در رابطه با محتوا.

لازم به ذکر است که در این دستورالعمل، صراحتاً به استفاده از یادگیری ماشینی اشاره‌ای نشده است، ولی در برخی موارد خاص مانند مبارزه با تروریسم که متعاقباً به آن خواهیم پرداخت به لزوم استفاده از ابزارهای هوشمند در پالایش محتوای تروریستی اشاراتی شده است.

چارچوب تنظیم‌گری ملّی[1] اتحادیه‌ی اروپا که مرجع تنظیم‌گری رسانه است به مقوله‌ی ارزیابی رسانه‌های صوت و تصویر فراگیر تأکید بسزایی داشته است. در این چارچوب، پنج نوع محتوای غیرقانونی با قواعد و مقررات مشخص تعریف شده و الزامات و قواعد میزبانی از چنین محتوایی نیز تبیین گردیده است [6]:

۱- دستورالعمل مبارزه با تروریسم

در این دستورالعمل بیان شده است که کشورها بایستی در جهت حذف محتوای تروریستی از بستر وب‌سایت‌ها اقدامات لازم را انجام دهند. در ضمن اتحادیه‌ی اروپا سعی دارد تا پا را فراتر نهاده و ارائه‌دهندگان خدمات اینترنتی را نیز موظف به حذف محتوای تروریستی نماید. این قانون در کشورهای مختلف اروپایی از نظر محدودیت زمانی برای حذف محتوا و جرائم و عواقب عدم تبعیت از قانون، متفاوت است. همچنین این قانون به دادستان‌ها و دادگاه‌ها اجازه می‌دهد تا به شرکت‌ها، دستور حذف یا مسدودسازی محتوا ــ در بازه‌ی زمانی بیست‌وچهار ساعته تا چهل‌وهشت ساعته در برخی شرایط خاص ــ را بدهند. در ضمن، توصیه‌های بیشتری نظیر آنچه در ادامه خواهد آمد، برای جلوگیری از انتشار محتوای تروریستی بر بسترهای آنلاین ارائه شده است [6]:

۱. حذف محتوا در ظرف یک ساعت در صورت دستور صادره از نهادهای دارای صلاحیت ملّی، نه فقط نهادهای قضایی.

۲. محتواهای ارجاع داده شده از سوی نهادهای دارای صلاحیت ملّی بایستی در اولویت قرار داده شوند و با سرعت و دقت بیشتری مورد بررسی قرار گیرند.

1 National Regulation Framework

۳. اقدامات پیشگیرانه در راستای پالایش محتوای تروریستی از بستر سرویس‌ها، با به‌کارگیری ابزارهای مبتنی بر یادگیری ماشینی صورت گیرد. طبق گزارش‌ها در سال ۲۰۱۸ از هفتادوهفت هزار محتوای گزارش شده تروریستی، ۸٤ درصد آن‌ها به کمک ابزارهای مبتنی بر یادگیری ماشین از پلتفرم‌های آنلاین کشف شده است.

۲- دستورالعمل سوءاستفاده جنسی از کودکان

در این دستورالعمل، کشورهای عضو اتحادیه، متعهد به حذف و محدودسازی محتوای مرتبط با سوء استفاده‌ی جنسی از کودکان شدند. در سال ۲۰۱۷، اتحادیه‌ی حفاظت از خردسالان در فضای آنلاین، به منظور رسیدگی به خطرات نوظهور خردسالان در فضای آنلاین، پایه ریزی شد. هدف از این انجمن ارائه‌ی ابزارهای قوی و در دسترس برای ارسال بازخورد سریع و آسان، ترویج طبقه‌بندی محتوا و تقویت همکاری میان اعضای اتحادیه و دیگر گروه‌ها است [6].

۳- دستورالعمل مبارزه با نژادپرستی

طبق این دستورالعمل، دولت‌ها موظف‌اند که نژادپرستی، نفرت‌پراکنی و بیگانه‌هراسی را مجازات کنند. برای مبارزه با محتوای نژادپرستی و بیگانه‌هراسی، پلتفرم‌های آنلاین به تدوین قاعده‌ای رفتاری میان خود پرداخته‌اند. بر مبنای این قاعده، پلتفرم‌ها موظف شده‌اند تا [6]:

۱. رویه‌های مشخصی جهت شفاف‌سازی ارائه دهند و به طور مشخص بیان کنند که کاربران از ترویج خشونت و نفت‌پراکنی خودداری نمایند.

۲. فرآیندهای مؤثر و شفاف به منظور بازنگری گزارش‌های محتوای غیرقانونی ایجاد کنند و این بازنگری بر اساس دستورالعمل‌ها و استانداردهای سرویس و همچنین قوانین ملّی صورت گیرد. برای کارمندانِ پلتفرم‌های آنلاین آموزش‌های منظم در ارتباط با تحولات و تغییرات اجتماعی صورت گیرد.

۳. تشویق و ترویج گزارش محتوای غیرقانونیِ نفرت‌پراکن توسط متخصصان، نهادهای مدنی و دیگر افراد صورت پذیرد.

٤- دستورالعمل پالایش اطلاعات گمراه کننده[1] و جعلی[2]

اطلاعات گمراه‌کننده و جعلی همواره از اقسام محتوای غیرقانونی محسوب نمی‌شوند و در مواقع خاص اهمیت دوچندان دارند و بسیار در تقابل با ارزش‌های اروپایی می‌باشند. در این قاعده‌ی خودتنظیم‌گرایانه، پلتفرم‌های آنلاین تعهداتی نظیر مسدودسازی حساب‌های کاربران جعلی، سرمایه‌گذاری بر روی فنّاوری به منظور آگاهی‌بخشی به کاربر در زمان دریافت اطلاعات، اولویت‌بخشی به اطلاعات تأیید شده و معتبر، تسهیل‌گری در دریافت محتوای جایگزین مرتبط با منافع عمومی، شفافیت در تبلیغات سیاسی، افشای هویت حامی مالی تبلیغات سیاسی، تفکیک محتوایی، تقویت جامعه‌ی تحقیقاتی و صحت‌سنجان را پذیرفته‌اند [6].

۳. چارچوب پالایش محتوای نامناسب در چین

پلتفرم‌های چینی، در کنار سایر ارائه دهندگان خدمات اطلاع‌رسانی آنلاین[3]، توسط نهادهای نظارتی دولتی تنظیم می‌شوند و نسبت به محتوایی که میزبان آن هستند، مسئولیت دارند [7]. در سال‌های اخیر، اداره فضای مجازی کشور چین[4] (CAC) ـ بخش تنظیم کننده‌ی محتوای آنلاین چین ـ هدف نظارت اینترنتی خود را از وب‌سایت‌ها و کاربران خاص، به پلتفرم‌های اصلی تغییر داده است؛ لذا تقریباً همه‌ی پلتفرم‌های بزرگ، به دلیل میزبانی محتوای نامناسب مرتباً احضار و مجازات می‌شوند [7].

اداره فضای مجازی کشور چین (CAC) پلتفرم‌های چینی را به اجرای «مسئولیت‌های اصولی»[5] در حاکمیت محتوای آنلاین ملزم می‌کند و مجموعه‌ای از الزامات خاص را برای انجام این کار تعیین کرده

1 Misinformation
2 Disinformation
۳ Online information service providers
٤ The Cyberspace Administration of China
٥ Primary responsibilities

است، از جمله اجرای سیاست ثبت نام[۱]، تأیید نام واقعی[۲]، نظارت بلادرنگ بر محتوا[۳] و ایجاد سازوکار لیست سیاه[٤] کاربران.

ردیف	عنوان	عنوان لاتین
۱	ثبت نام بر مبنای نام واقعی	The real-name registration
۲	تأیید نام واقعی	The real-name verification
۳	نظارت بلادرنگ بر محتوا	Real-time content monitoring
۴	ایجاد سازوکار لیست سیاه کاربران	Blacklisting mechanism

جدول شماره ۱ مجموعه سیاست‌ها و الزامات خاص معیّن شده برای پلتفرم‌های دیجیتالی در راستای اجرای «مسئولیت‌های اصولی» در حاکمیت محتوای آنلاین در کشور چین را نشان می‌دهد [7].

جدول ۱، مجموعه سیاست‌ها و الزامات خاص معین شده برای پلتفرم‌ها

از آن‌جایی که حاکمیت محتوای آنلاین چین عمدتاً بر محتوای حساس سیاسی[۵]، محتوای مبتذل[٦] و شایعات[۷] (یا اطلاعات نادرست) متمرکز است در پلتفرم‌ها نیز چنین محتواهایی مرتباً سرکوب می‌شوند، زیرا حزب کمونیست چین[۸] (CCP) برای مشروعیت بخشیدن به کنترل اینترنت به گفتمان‌های اخلاق‌پایه[۹] تکیه می‌کند.

تقریباً تمامی پلتفرم‌های اصلی چینی، متعلق به شرکت‌های اینترنتی داخلی[۱۰] هستند. چین یک سیستم مجوز متمایز را برای ارائه دهندگان خدمات خبری خصوصی[۱۱] و رسمی[۱۲] ایجاد کرده است. دولت چین

---

۱ The real-name registration
۲ Real-name verification
۳ Real-time content monitoring
٤ Blacklisting mechanism
٥ Politically sensitive content
٦ Vulgar content
۷ Rumours
۸ Chinese Communist Party (CCP)
۹ Morality-based
۱۰ Domestic Internet companies
۱۱ Private
۱۲ Official

برای تقویت کنترل تحریریه[۱] (کنترل انتشار مطالب) در پلتفرم‌های خصوصی، ابتکار موسوم به «سهم ویژه مدیریت»[۲] را آزمایش کرده است.

در عصر پلتفرم‌های باز[۳]، که در آن پلتفرم‌های رسانه‌های دیجیتالی به طور فزاینده‌ای در محوریّت تولید، توزیع و مصرف محتوا قرار دارند، دولت حزب چین[٤] با یک اکوسیستم رسانه‌ای تغییر شکل یافته مواجه شده و بر این اساس، در حال تنظیم چارچوب تنظیم‌گری محتوای آنلاین[٥] مختصّ خود است. تقریباً از سال ۲۰۱٤، نهاد تنظیم‌گر محتوای آنلاین چین، اداره فضای مجازی چین (CAC)[٦] و حکمرانی اینترنت[۷] خود را بر مبنای تمرکز بر پلتفرم‌های دیجیتال سامان‌دهی کرده است [8]. همان‌گونه که در گزارش سالانه سال ۲۰۱۵ در مورد افکار عمومی آنلاین در کشور چین مشاهده می‌کنیم، به وضوح به این موضوع اشاره شده است: «مدیریت دولت بر افکار عمومی در فضای آنلاین[۸] از تمرکز بر فیلتر کردن کلید واژه‌های حساس و برخی افراد مشخّص، به سمت هدف قرار دادن پلتفرم‌های آنلاین، به ویژه پورتال‌های تجمیع اخبار[۹] و شبکه اجتماعی مانند WeChat و Weibo تغییر کرده است» [8].

در سال‌های اخیر، CAC به طور مداوم مقررات خود را در مورد پلتفرم‌ها گسترش داده و به روز می‌کند. در میان یک سری مقررات جدید مربوط به پلتفرم‌ها، به طور کلی برخی از پلتفرم‌های هدف، مانند مقررات مربوط به حساب‌های رسمی[۱۰] در پلتفرم‌ها، نظرات کاربران آنلاین، گروه‌های آنلاین در پلتفرم‌ها و ارزیابی امنیت در فناوری‌ها و برنامه‌های جدید. دیگران بر روی انواع خاصی از پلتفرم‌ها مانند مقررات مربوط به سرویس‌های پیام رسانی فوری[۱۱] مانند WeChat، برنامه‌های جمع آوری اخبار تلفن همراه، سرویس‌های

---

۱ Editorial
۲ Special management share
۳ Open platforms
٤ Chinese party-state
٥ Online content regulatory framework
٦ The Cyberspace Administration of China (CAC)
۷ Internet governance
۸ Online public opinion
۹ Newsaggregating portals
۱۰ Official accounts
۱۱ Instant messaging

پخش زنده، تالارهای گفتگوی آنلاین و خدمات میکروبلاگینگ[۱] متمرکز هستند. این واقعیت که در چنین مدت کوتاهی بیش از ده‌ها مجموعه از مقررات مربوط به پلتفرم‌ها صادر شده است، نه تنها نشان دهنده سیاست‌های متغیّر چین در مورد چارچوب نظارتی اینترنت بوده است، بلکه نشانگر نگرانی‌های دولت حزب چین در مورد تأثیر فزاینده پلتفرم‌ها بر روی کنترل اطلاعات خود است [9].

یکی از الزامات اصلی مقررات فوق این است که پلتفرم‌ها بایستی «مسئولیت اصولی» را در اداره محتوای آنلاین بر عهده داشته باشند. این بدان معناست که از پلتفرم‌ها نه تنها انتظار می‌رود که با سانسور محتوای مدّ نظر دولت مطابقت داشته باشند، بلکه آن‌ها باید در ایجاد قوانین مربوطه و اعمال تعدیل محتوا نقش فعالی داشته باشند. CAC در دستورالعمل‌های خود به پلتفرم‌ها برای انجام مسئولیت‌های حکمرانی خود، در درجه اول شماری از الزامات خاص را به آن‌ها تحمیل کرده است [10]:

- **نظارت بی درنگ بر محتوا**[۲] و ذخیره اطلاعات کاربران برای مدت زمان کمتر از شش ماه.
- اجرای سیاست ثبت‌نام و تأیید نام واقعی[۳]. کاربران پلتفرم‌های نرم افزاری قبل از ارسال هرگونه محتوا باید اطلاعاتی در مورد هویت واقعی خود ارائه دهند و پلتفرم‌ها ملزم به شناسایی هویت کاربران بر اساس شناسه‌های (کدهای) نهادی[٤] منحصر به فرد (برای کاربران نهادی) و شناسنامه-های دولتی (برای افراد حقیقی) هستند.
- ایجاد مکانیزم‌های گزارش دهی کاربر[٥]، رفع شایعات[٦] و لیست سیاه[٧] به پلتفرم‌ها اجازه می‌دهد تا کاربرانی را که به طور جدی قوانین و مقررات مربوطه را نقض کرده‌اند به «لیست سیاه» اضافه کرده و مجازات مناسب را اعمال کنند (به عنوان مثال بستن حساب‌های آن‌ها، ممنوعیت ثبت نام مجدد آن‌ها).

---

۱ Microblogging
۲ Real-time content monitoring
۳ Real-name registration and verification
٤ Institutional codes
٥ User-reporting
٦ Rumour-debunking
۷ Blacklisting

تهیه‌ی نظام مدیریت درجه‌بندی (امتیازدهی) و طبقه‌بندی کاربران[۱]. منظور از «درجه بندی[۲]» این است که اگر مشخص شود کاربران برخی قوانین را نقض کرده اند، در این صورت نمرات اعتباری آن‌ها پایین می‌آید و سرویس‌هایی (خدماتی) که می‌توانند استفاده کنند بر این اساس محدود می‌شود. منظور از «طبقه بندی[۳]» این است که برای سرویس‌هایی مانند چت‌های گروهی، پلتفرم‌ها باید آن‌ها را در دسته‌های مختلف با توجه به عواملی از جمله اندازه گروه‌ها و حوزه‌های محتوایی (به عنوان مثال سیاسی، اقتصادی یا سرگرمی) طبقه بندی کنند [10].

برخی از الزامات فوق، از جمله سیاست نام واقعی[٤]، سازوکار لیست سیاه، و سیستم مدیریت درجه‌بندی و طبقه بندی کاربر، برای تقویت ظرفیت آن‌ها در اجرای فرایندهای نظارتی بر کاربران خود، پلتفرم‌های چینی را تحت فشار قرار داده‌اند.

با توجه به اقتدارگرا[٥] بودن دولت چین تعریف محتوای مشکل‌دار[٦] در پلتفرم‌های چین مرزهای خاص خود را دارد. به طور کلی، محتوای مشکل‌دار در پلتفرم‌های چینی را می‌توان به سه دسته طبقه‌بندی کرد [7]: محتوای سیاسی مضر[۷] / و یا حساس؛ محتوای مبتذل[۸]؛ اطلاعات غلط و غیر سیاسی و شایعات[۹]. بخشی از آن به دلیل حجم بالای محتوای بارگذاری شده هر روز، پلتفرم‌های چینی به دلیل میزبانی از انواع مختلف محتوای مشکل‌دار، توسط مقامات نظارتی دولت احضار و مجازات می‌شوند.

تعریف محتوای نامناسب در پلتفرم‌های چینی

---

۱ User-grading-and-classification
۲ Grading
۳ Classification
٤ The realname policy
٥ Authoritarian
٦ Problematic content
۷ Politically harmful
۸ Vulgar
۹ Rumours

با توجه به این‌که چنین محتواهایی ممکن است بر ثبات اجتماعی چین و مشروعیت حاکمیت دولت حزب چین تأثیر بگذارد، مدت مدیدی است که در مقررات آنلاین محتوای نامناسب از نظر سیاسی (به معنای مضر برای دولت ـ حزب) مورد توجه قرار گرفته است [7].

اگرچه در مورد آنچه که «محتوای سیاسی نامناسب» در چین است خط مشخصی وجود ندارد ، اما انواع محتوای سیاسی ممنوع ذکر شده در درجه اول شامل موارد زیر است:

۱. مخالفت با اصول اساسی وضع شده در قانون اساسی چین[۱]، قوانین یا مقررات داخلی کشور چین؛
۲. به خطر انداختن امنیت کشور ، افشای اسرار دولتی،
۳. به خطر انداختن امنیت کشور، افشای اسرار دولتی، براندازی قدرت دولت یا خرابکاری در وحدت کشور؛
٤. آسیب رساندن به عزت ملی و منافع ملی؛
٥. تحریک خصومت قومی یا تبعیض نژادی یا برهم زدن وحدت قومی؛
٦. آسیب رساندن به سیاست‌های ملی مذهبی و تبلیغ خرافات فئودالی[۲]؛
۷. شایعه پراکنی، برهم زدن نظم و ثبات اجتماعی؛
۸. تحریک اجتماعات غیرقانونی، تظاهرات، اعتراض و برهم زدن نظم اجتماعی؛
۹. سازماندهی فعالیت‌ها به نام سازمان‌های غیردولتی غیرقانونی یا سایر مطالب ممنوع شده توسط قوانین و مقررات کشور چین [11].

مجازات پلتفرم‌ها برای تخلف از مقررات

برای مقابله با شیوع محتوای نامناسب در پلتفرم‌های چینی، نهادهای تنظیم‌گر دولتی چین[۳] اغلب به کمپین‌های ضد ابتذال، ضد پورنوگرافی و شایعات که گاه اسم رمزی برای کنترل اطلاعات هستند، متوسل می‌شوند. به عنوان مثال، در نوامبر سال ۲۰۱۸، CAC پلتفرم‌های اصلی را احضار کرده و آن‌ها را ملزم به مقابله با «وضعیّت آشفته»[٤]ی پیرامون خود ـ رسانه‌ها[٥] کرد؛ (خود ـ رسانه‌ها اشاره به حساب‌های رسمی

---

۱ China's Constitution

۲ Feudal superstitions

3 Chinese government regulators
4 Chaotic situation
5 Self-media

در پلتفرم‌ها دارد که توسط افراد اداره می‌شد) [12]. در نتیجه‌ی این اقدام، ۹۸۰۰ حساب خود ـ رسانه‌ای توسط پلتفرم‌ها بسته شد و یا به حالت تعلیق درآمد.

در مثالی دیگر از ژوئن ۲۰۱۷، اداره فضای مجازی پکن (CAB)[1] بزرگترین پلتفرم‌های آنلاین / شرکت های اینترنتی را احضار کرد و به آن‌ها دستور داد تا با پوشش مبتذل و پر سر و صدا رسوایی‌های سلبریتی‌ها و سبک‌های زندگی خودنما و متظاهر[2] مقابله کنند. در عرض چند روز، شصت حساب رسمی سرگرمی محبوب (برخی با میلیون‌ها دنبال‌کننده)[3] در تعدادی از پلتفرم‌ها بسته شد [13].

در آوریل ۲۰۱۸، از سوی نهادهای تنظیم‌گر دولتی دستور داده شد که چهار مورد از محبوب‌ترین پلتفرم‌های جمع‌آوری اخبار در کشور چین، از جمله تاوتیائو به طور موقت از فروشگاه‌های دانلود برنامه‌ی کاربردی خارج شوند. تنها یک روز پس از این دستور، نهاد تنظیم‌گر رادیو و تلویزیون کشور چین (اداره امور رادیویی و تلویزیونی[4] موسوم به SART) به شرکت ByteDance دستور داد برنامه کاربردی تحت عنوان Neihan Duanzi را که یک پلتفرم معروف طنز حاوی محتوای صوتی ـ تصویری است، برای همیشه مسدود کند. «میزبانی محتوای مبتذل» و «جهت گیری اشتباه» از جمله دلایل این اقدام اعلام شده است [14]. وقوع این امر در حالی بود که اگرچه CAC اختیارات کلی در مورد محتوای آنلاین در کشور چین را در اختیار دارد، امّا در خصوص برنامه‌های صوتی و تصویری آنلاین، اختیار قانونی با SART است.

برای درک بهترِ سازوکار موتورهای کشف محتوای نامناسب که قبلاً به آن اشاره کردیم، لازم است در ادامه به تبیین فنّاوری‌های به کار گرفته شده در این موتورها، موسوم به هوش‌مصنوعی و به طور خاص فنّاوری‌های مبتنی بر یادگیری ماشینی بپردازیم.

۴. آشنایی با فنّاوری هوش مصنوعی

رشد و توسعه‌ی روزافزون ابزارهای مبتنی بر هوش‌مصنوعی در دهه‌ی اخیر و افزایش تولید نرم‌افزارهای هوشمند، این شانس را به عموم مردم دنیا داده است که حداقل یکبار در طول زندگی خویش، طعم شیرین استفاده از محصولات فنّاورانه مبتنی بر هوش‌مصنوعی را بچشند و از قدرت و سرعت آنها محظوظ گردند. امروزه استفاده از کاربردهای مختلف هوش‌مصنوعی در زندگی آدمی، امری

[1] The Cyberspace Administration of Beijing (CAB)
[2] Ostentatious lifestyles
[3] Followers
[4] State Administration of Radio and TV

طبیعی و بدوی به‌حساب می‌آید؛ به‌طوری‌که ردّ پای ابزارهای مبتنی بر هوش‌مصنوعی را می‌توان حتی در زندگی روزمره‌ی اهالی کشورهای غیرتوسعه‌یافته نیز مشاهده کرد. به‌عنوان مثالی فراگیر، می‌توان از موتور جستجوی قدرتمند گوگل یاد کرد که شیوه‌ی جستجو نمودن شما را یاد می‌گیرد و متناسب با آنچه که به دنبال آن می‌گردید، نتایج را سفارشی‌سازی می‌کند. به عنوان مثال دیگری از کاربردهای فراگیر هوش مصنوعی در زندگی مردم، می‌توان به قابلیت باز نمودن قفل تلفن‌های هوشمند با تصویر چهره‌ی از پیش تعیین‌شده اشاره کرد که به عنوان یک راهکار هوشمند برای تسهیل فعالسازی تلفن‌های نسل جدید ارائه گردیده است. با این حال، متأسفانه کاربست‌پذیری فوق العاده‌ی ابزارهای هوشمند در زندگی بشر، منجر به ادراک عمومی نسبت به چیستی و چگونگی عملکرد ماشین‌های هوش‌مصنوعی نگردیده و عموم استفاده‌کنندگان از ابزارهای هوشمند، هیچ ایده‌ای در رابطه با تکنولوژی‌های مرتبط با آن ندارند. این مسئله، بر لزوم معرّفی دقیق و دسته‌بندی شاخه‌های مختلف هوش‌مصنوعی (نمودار ۱) صحّه می‌گذارد. هوش‌مصنوعی، شاخه‌ای از علوم کامپیوتر است که در آن به ساخت ماشین‌هایی هوشمند پرداخته می‌شود که مانند انسان عمل کرده و واکنش نشان می‌دهند. برخی از محقّقان، دوران جنگ جهانی دوّم را به عنوان شروع عصر هوش‌مصنوعی در نظر می‌گیرند و معتقدند که تلاش‌های دانشمند انگلیسی آلن تورینگ[1] برای رمزنگاری و کشف پیام‌های نیروهای آلمانی که توسط ماشین انیگما[2] تولید می شدند، سنگ بنایی برای تحقیقات آتی در زمینه هوش‌مصنوعی بوده است [15]. تورینگ ماشینی را هوشمند می‌دانست که بدون القای حس مکالمه با ماشین، به انسان اجازه برقراری ارتباط با آن را بدهد. این مسئله، یعنی ساخت ماشینی که همانند انسان یاد بگیرد و بر پایه شناخت محیط اطراف خود تصمیم‌گیری کرده و عملیاتی را انجام دهد، پایه اصلی علم هوش‌مصنوعی می‌باشد. اما عده‌ای دیگر از محققان حوزه هوش‌مصنوعی، یک کارگاه آموزشی در سال ۱۹۵۶ در کالج دارتموث[3] را به عنوان زادگاه هوش‌مصنوعی معرفی نموده‌اند [16]؛ جایی که دانشمندان برجسته‌ای چون جان مک‌کارتی[4]، ماروین مینسکی[5] و اقتصاددان برجسته

1 Alan Turing
2 Enigma
3 Dartmouth college
4 John McCarthy
5 Marvin Minsky

هربرت الکساندر سایمون[1] به‌همراه جمع کثیری از دانشجویان شان گردهم آمدند و برنامه‌هایی تو سعه دادند که قادر به سخن‌گفتن به زبان انگلیسی، حل سؤالات جبری و اثبات قضایای منطقی بودند.

همانند اختلاف نظرهای موجود در مورد مبدأ و زادگاه هوش‌مصنوعی، پژوهشگران در تعریف هوش مصنوعی نیز با یکدیگر اختلاف نظر دارند و علیرغم تشابه‌های موجود، هنوز تعریف واحدی برای آن ارائه نشده است. این مسئله نشئت گرفته از ارائه‌ی تعاریفی برای صفت «هوشمند» است که سخت می توان محدوده مشخصی برای تعریف آن ارائه کرد. بر اساس یافته‌های ما، خصوصیات زیر جزو قابلیت‌های ضروری برای اطلاق واژه هوشمند به یک فنّاوری است که مورد تأیید عموم محققان نیز می‌باشد:

- پاسخ به موقعیت‌های از قبل تعریف نشده با انعطاف بسیار بالا و بر اساس بانک دانش
- معنا دادن به پیام‌های نادرست یا مبهم
- درک تمایزها و شباهت‌ها
- تجزیه و تحلیل اطلاعات و نتیجه‌گیری
- توانمندی آموختن و یادگرفتن
- برقراری ارتباط دوطرفه

به فرض پذیرش تمامی تعاریف بالا از مؤلفه‌های یک فنّاوری هوشمند، موارد زیر فهرستی از وظایفی است که انتظار می رود یک فنّاوری هوشمند قادر به انجام آنها باشد:

- تولید گفتار
- تشخیص و درک گفتار (پردازش زبان طبیعی انسان)
- دستورپذیری و قابلیت انجام اعمال فیزیکی در محیط طبیعی و مجازی
- استنتاج و استدلال
- تشخیص الگو و بازشناسی الگو برای پاسخ‌گویی به مسائل بر اساس دانش قبلی
- شمایلی گرافیکی یا فیزیکی جهت ابراز احساسات و عکس‌العمل‌های ظریف
- سرعت عکس‌العمل بالا

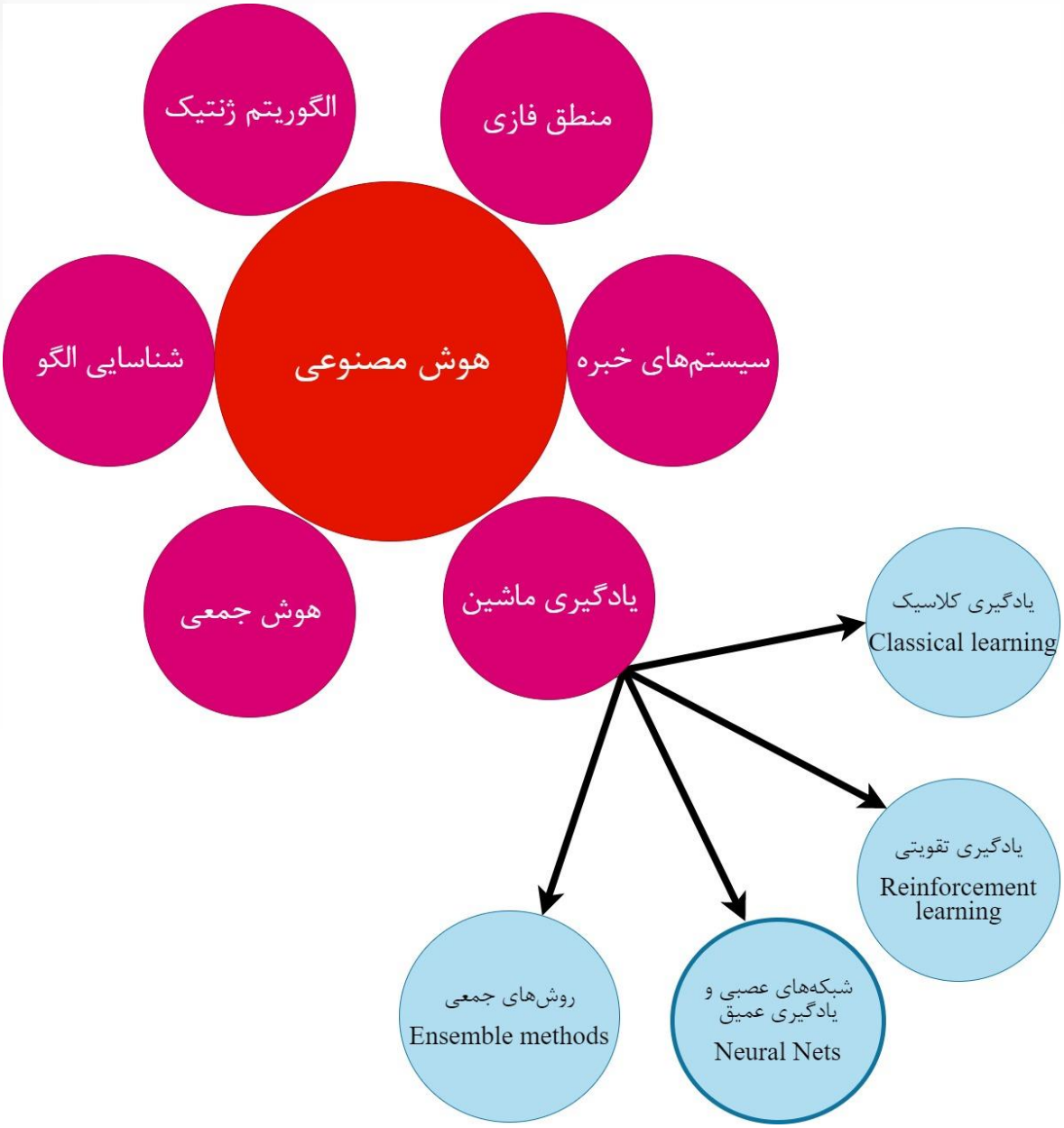
1 Herbert A. Simon

شباهت‌های معنایی میان مؤلفه‌های مذکور برای ابزارهای هوشمند منجر به پیچیده شدن تقسیم‌بندی هوش‌مصنوعی به شاخه‌های مختلف شده است. در واقع، با توجه به هم‌پوشانی مؤلفه‌های هوشمندی، ترسیم درخت فناوری‌های هوش‌مصنوعی با شاخه‌های کاملاً متفاوت، امری محال است؛ با این وجود، زیرمجموعه‌های هوش‌مصنوعی که مورد تأیید عموم محققان می‌باشد، به شرح نمودار ۱ می‌باشد.

همانطور که در نمودار ۱ بدان تأکید شده، یادگیری ماشینی مهم‌ترین شاخه‌ای از هوش مصنوعی است که بیشترین کارکرد را در تنظیم‌گری فضای صوت و تصویر بر عهده دارد. تمامی الگوریتم‌ها و روش‌های یادگیری، زیر مجموعه یادگیری ماشینی (شامل روشهای جمعی[۱]، شبکه‌های عصبی عمیق[۲]، یادگیری تقویتی[۳] و یادگیری کلاسیک[۴]) قابلیت استفاده در کارکردهای مختلفی همچون پردازش سیگنال (صوتی و تصویری) و یا پردازش متن (که از جمله کارکردهای مهم در حوزه تنظیم‌گری فضای صوت و تصویر هستند) را دارا هستند؛ لکن، کارایی و بهره‌وریِ بهترِ روش‌های مبتنی بر شبکه‌های عصبی عمیق یا همان یادگیری عمیق[۵] نسبت به سایر الگوریتم‌های زیر مجموعه هوش‌مصنوعی، در کارکردهایی که پیشتر به آنها اشاره شد، به تأیید همگان رسیده است. بنابراین در این نوشتار، به طور خاص به کاربست‌پذیری شبکه‌های

1 Ensemble methods
2 Deep Neural Network
3 Reinforcement learning
4 Classical learning
5 Deep Learning

عصبی عمیق در سه حوزه کارکردی پردازش صوت، متن و تصویر که در راستای حفظ منافع عمومی بکار گرفته می‌شوند، می‌پردازیم و بدین منظور در ابتدا به شرح اجمالی تعاریف مهم خواهیم پرداخت.



نمودار ۱، درخت فنّاوری هوش مصنوعی (منبع: تشریح شده توسط نویسندگان)

۴-۱ یادگیری ماشینی

عنوان یادگیری ماشینی، به فرآیندی اطلاق می‌گردد که در آن الگوریتم‌ها بدون اتّکا به برنامه‌نویسیِ قانون ـ محور[1] و با استفاده از داده، آموزش داده می‌شوند تا بتوانند در پیش‌بینی نتایج، دقیق‌تر شوند و داده‌های مختلف را بدون برنامه‌ریزیِ مخصوص، پردازش کنند. یادگیری ماشینی به چهار شاخه الگوریتم‌های یادگیری کلاسیک (مبتنی بر مدلهای یادگیری با نظارت[2] و یادگیری بدون نظارت[3])، یادگیری عمیق (مبتنی بر شبکه‌های عصبی عمیق، پیچشی و بازگشتی)، مدلهای جمعی و یادگیری تقویتی تقسیم‌بندی می‌شود. همچنین مهمترین تحول و پیشرفت در حوزه‌ی یادگیری ماشین در سال‌های اخیر، توسعه‌ی شبکه‌های عصبی عمیق برای یادگیری عمیق است. شبکه‌های عصبی، الگوریتم‌هایی هستند که از ساختار بیولوژیک مغز تقلید می‌کنند. بخش عمده از شبکه‌های عصبی «پیشخور»[4] هستند به این معنا که داده‌ها از مجرای هر لایه‌ای از شبکه گذرانده می‌شود به گونه‌ای که خروجی یک لایه، ورودی لایه‌ی بعدی خواهد بود [17]. در این میان، آنچه که در این پژوهش بیشتر مورد توجه است، حوزه یادگیری عمیق است که در سه کارکرد مختلف پردازش تصویر، صوت و متن کاربرد دارد. در ادامه به معرفی هر یک از این سه حوزه پردازشی خواهیم پرداخت.

۴-۱-۱ پردازش تصویر

پردازش تصویر دیجیتال یا به عبارت کوتاه‌تر پردازش تصویر، به شاخه‌ای از پردازش سیگنال گفته می‌شود که با پردازش سیگنال دیجیتال که نماینده تصاویر برداشته شده با دوربین دیجیتال یا اسکن‌شده توسط اسکنر هستند سروکار دارد. پردازش تصویر را گاهاً به اشتباه معادل بینایی ماشین[5] در نظر می‌گیرند؛ حال آنکه بینایی ماشین در پی درک معنا و محتوای تصاویر و برنامه‌ریزی اقدامات متناظر است [18] و در این مسیر از پردازش تصویر جهت درک محتوای تصاویر بهره می‌جوید.

1 Rule-based programming
2 Supervised Learning
3 Unsupervised Learning
4 Feed Forward
5 Machine Vision

ابزارهای پردازش تصویر را بر اساس نوع کارایی آنها به دو بخش آنالیز چهره[1] و آنالیز تصویر[2] تقسیم‌بندی می‌کنند. آنالیز چهره شامل تمامی ابزارهایی است که جهت شناسایی افراد[3]، تشخیص حالات عاطفی[4] و مقایسه دو چهره[5] به‌کار می‌آیند. از مهم‌ترین ابزارهای آنالیز تصویر نیز می‌توان به ابزارهای شناسایی اشیاء[6] در تصویر، تشخیص و دسته‌بندی منظره[7]، برچسب‌گذاری تصویر[8] و نویسه خوان نوری[9] اشاره کرد.

۴-۱-۲ پردازش گفتار

فنّاوری پردازش و تشخیص گفتار نرم‌افزاری است که می‌تواند صوت ورودی را از زبان‌های مختلف دریافت کرده، آن را به‌طور کامل و با دقت به متن تبدیل کرده و محتوای متن را درک کند. برای مثال، به کمک این فنّاوری، رایانه‌ای که توانایی دریافت صدا را دارد (یک کامپیوتر مجهز به میکروفن) این قابلیت را پیدا می‌کند که دستورات صوتی کاربر را متوجه شود و اقدامات متناظر را برنامه‌ریزی کند [19]. سابق بر این از مدلهای مارکوفی پنهان[10] برای پردازش گفتار استفاده می‌شد که با شروع هزاره سوم، این حوزه نیز دچار تغییرات بزرگی شده و از آن پس، برای پردازش گفتار از شبکه‌های عصبی عمیق (یادگیری عمیق) استفاده گردید [20]. اگرچه که در صورت بررسی عمیق عبارات و عناوین در حوزه هوش مصنوعی، مرز مشخصی بین برخی از حوزه‌ها یافت نخواهد شد، اما شاید بتوان فرآیند پردازش گفتار را متشکل از سه سطح با پیچیدگی‌های فنّی مختلف دانست که به شرح زیر می‌باشند:

۱. بازشناسی خودکار گفتار (Automatic Speech Recognition - ASR)
۲. پردازش زبان طبیعی (Natural Language Processing - NLP)
۳. کشف محتوای ضمنی گفتار (Electromyography Recognition - EMG)

---

1 Face Analysis
2 Image Analysis
3 Face Identification
4 Emotion Analysis
5 Face Verification
6 Object Detection
7 Scene Recognition and Classification
8 Image Tagging
9 Optimal Character Recognition (OCR)
10 Hidden Markov Models

بازشناسی گفتار (ASR)، به فرآیند تبدیل گفتار به متن متناظر آن گفته می‌شود که نخستین مرحله از پردازش گفتار است و دقت انجام آن تأثیر بسزایی در کیفیت خروجی تحلیل گفتار دارد.

بخش دوم که به پردازش زبان‌های طبیعی مشهور است، با تحلیل اجزای متن بوجود آمده در مرحله نخست، به بررسی زبانی کلام می‌پردازد؛ در واقع، تحلیل محتوا محور[1] کلام که شامل درک معنای مستقیم متون مستخرج از صوت می‌باشد، توسط NLP صورت می‌پذیرد. برای مثال، کاربری را در نظر بگیرید که به دستیار صوتی هوشمند موجود در تلفن همراه خود دستور تماس با فرد خاصی را می‌دهد؛ تحلیل چنین درخواستی، توسط NLP صورت می‌پذیرد.

اما گاهی اوقات، تحلیل گفتار گوینده بایستی بصورت زمینه محور[2] صورت پذیرد تا مفهوم صحیح و کامل کلام از آن استنباط شود. برای مثال، کنایه‌ها، ضرب‌المثل‌ها، توهین‌های در لفافه و ... از جمله مواردی هستند که NLP قادر به درک و استخراج محتوای آن‌ها نیست؛ در نتیجه، محققان برای کشف محتوای ضمنی گفتار، از بخش سوم پردازش گفتار که همان EMG است، مدد می‌جویند.

۴-۱-۳ پردازش متن

پردازش متن و به بیان دقیق‌تر متن‌کاوی، حوزه‌ای است که با بسیاری از زمینه‌های دیگر هوش‌مصنوعی مانند پردازش زبان‌های طبیعی، داده‌کاوی، کلان‌داده، شبکه‌های عصبی و یادگیری عمیق مرز مشترک دارد که البته هر کدام از این حوزه‌ها به تنهایی با چالش‌ها و گستردگی‌های خاص خود مواجه هستند [21]. منظور از پردازش متن، استخراج اطلاعات با کیفیت از متن است. چنین اطلاعاتی معمولاً از طریق فهم الگوها، گرایش‌ها و معانی و به وسیله یادگیری الگوهای آماری حاصل می‌شود. متن‌کاوی، به فرآیندی اطلاق می‌شود که با ساختاردهی به ورودی‌های متنی (تجزیه اجزای کلام، افزودن برخی ویژگی‌ها و تفاسیر زبانی و حذف موارد اضافی)، شروع شده و با استخراج الگوهای مستتر در داده‌های ساختاریافته ادامه پیدا می‌کند و در نهایت، با ارزیابی و تفسیر خروجی‌ها به پایان می‌رسد. با توجه به آنچه گفته شد، وظایف عمومی پردازش متن را می‌توان شامل دسته‌بندی متون، خوشه‌بندی متون، استخراج معنی و مفهوم، تولید

1 Content-based Analysis
2 Context-based Analysis

رده‌بندی دانه‌ای، تجزیه و تحلیل احساسات، خلاصه کردن اسناد و مدلسازی ارتباط موجودیت‌های کلام دانست [22].

آنچه در بالا شرح دادیم گزیده‌ای از فنّاوری‌های مبتنی بر هوش‌مصنوعی و به ویژه یادگیری ماشینی بود که نقش بسیار مؤثری در زمینه‌ی پالایش محتوای نامناسب ایفا می‌کنند.

در ادامه به بررسی کاربرد پردازش صوت، متن و تصویر در صیانت از منافع عمومی در فضای مجازی پرداخته و همچنین با بررسی نحوه‌ی استفاده از فنّاوری مبتنی بر یادگیری ماشینی در راستای صیانت از منافع عمومی در رسانه‌های صوت و تصویر فراگیر معتبر، سعی در تبیین راهکارهای هوشمند ارائه شده برای حل مسئله‌ی پالایش محتوا در فضای مجازی خواهیم داشت.

نخستین و مهمترین انگیزه رسانه‌ها در ارزیابی محتوای منتشر شده بر بستر آن‌ها، عدم تخطی از قوانین و استانداردهای محتوایی تنظیم‌گران و قانون‌گذاران داخلی و بین المللی است. برای مثال، رسانه‌های چینی بر اساس قانون [23] موظف شده‌اند تا محتوای منتشرشده در پلتفرم‌هایشان را کنترل کنند و در صورت تخطی از این مهم، جرایم سنگینی انتظارشان را خواهد کشید. همچنین، پلتفرم‌های اشتراک محتوا در اتحادیه اروپا از سوی این اتحادیه موظف شده‌اند که محتوای مضر و غیرقانونی (مطابق تعریف اتحادیه اروپا) را از بستر خود حذف کنند [24]. یکی دیگر از انگیزه‌های رسانه‌ها برای پالایش محتوا، تطابق محتوایی با فرهنگ و درک عمومی نسبت به محتوای مناسب است. برای مثال، پینترست[1] در پاسخ به شکایات عمومی مبنی بر وجود مطالبی در پلتفرم خود که حاوی اطلاعات غلط درباره آثار جانبی و تاثیر واکسیناسیون برای انسان‌ها بود، جستجوهای منتهی به آن محتوا را محدود کرد و آن محتوا را از دسترس خارج نمود [25].

۵. بررسی انواع محتوای نامناسب

تقریباً تمامی رسانه‌های پرمخاطب، در پاسخ به الزام حذف محتوای نامناسب از پلتفرم‌های خود، یا اقدام به توسعه راهکار کشف و حذف محتوای نامناسب نموده‌اند و یا از طریق شرکتهای توسعه‌دهنده‌ی

1 Pinterest

راهکار و فنّاوری، نیاز خود را مرتفع ساخته‌اند. در گام نخست و در راستای بررسی چنین اقداماتی، ضروری است که به تبیین مواردی که در عرف عمومی «نامناسب» درنظر گرفته می‌شوند، بپردازیم.

با بررسی ماشین‌های کشف محتوای نامناسب که به‌منظور پالایش محتوا در رسانه‌ها توسعه یافته‌اند، مجموعه‌ای از مواردی که در رسانه‌های صوتی و تصویری به‌عنوان «محتوای نامناسب» درنظرگرفته شده‌اند به شرح زیر می‌باشد.

۵-۱ محتوای نامناسب تصویری (عکس و فیلم)

جدول ۲، محتوای نامناسب تصویری (عکس و فیلم)

محتوای نامناسب	نوع	زیرنوع	مصادیق
جنسی[1]	پورنوگرافی[2]	فیلم	✓ فیلم و انیمیشنی که انواع رابطه و پوزیشن‌های جنسی در روابط غیر هم‌جنس (زن و مرد) و هم‌جنس (مرد با مرد یا زن با زن) را نشان می‌دهد. ✓ هر محتوایی که انواع روابط جنسی حیوانات را دربر می‌گیرد. ✓ هر محتوایی که روابط جنسی انسان با حیوانات را در بردارد. ✓ هر محتوایی که شامل انواع بوسه‌های عاشقانه است.
		انیمیشن	
	برهنگی[3] و نیمه‌برهنگی[4]	مردان	✓ برهنگی کامل مرد و زن و یا برهنگی قسمت‌هایی از بدن که در عرف ناشایست و تحریک‌آمیز است. ✓ نمایان بودن برجستگی بدن زنان و مردان با وجود پوشیدن لباس (کوتاه و چسبان)، به طوری‌که التذاذآور باشد.
		زنان	

1 Sexual
2 Pornography
3 Nudity
4 Partial nudity

	پوشش نامناسب[1]	لباس‌زیر و لباس‌های ورزشی نامناسب	✓ سوتین، بیکینی، شورت و غیره.
	ابزارهای جنسی	–	✓ اسباب‌بازی‌های جنسی[2] و اندام جنسی مصنوعی ✓ نمایش اندام‌های جنسی مردان و زنان در قالب مجسمه (مجسمه کامل یا نیم‌تنه انسان و مجسمه اندام جنسی)، تابلو، نقاشی و سایر قالب‌های هنری و نمایشی. ✓ تبلیغات ابزارهای جنسی
خشونت[3]	سلاح[4]	سلاح گرم	✓ اسلحه کمری، اسلحه اتوماتیک، بمب و غیره. ✓ چاقو، شمشیر، پنجه بوکس و غیره. ✓ تبلیغات سلاح‌های گرم و سرد.
		سلاح سرد	
	جراحت[5]، اذیت و آزار فیزیکی[6]	خون[7]	✓ هرگونه خشونت، به نحوی که موجب جذابیت و از میان رفتن قبح آن شود، همچون صحنه‌های قتل، جنایت، ضرب و شتم، قطع عضو، خونریزی، کشتار جمعی و فردی. ✓ تصویر خفه‌گی یا سوختگی شدید انسان یا حیوان. ✓ نمایش یا تصویرگری خودکشی. ✓ دعوای خیابانی ✓ تصادف انسان با وسایل نقلیه
		سوختگی	
		تصادف	

1 Inappropriate costumes
2 Sexual toys
3 Violence
4 Weapons
5 Wound
6 Physical abuse/harassment
7 Blood

		کودک آزاری[1]	
		آزار جنسی[2]	
		آزار حیوانات[3]	
		نزاع	
مواد مخدر[4]، مشروبات الکلی[5] و قمار[6]	دخانیات	-	✓ قرص، سیگار، جام و بطری شراب تزریق با سرنگ و غیره. ✓ تبلیغات خرید و فروش مواد مخدر، مشروبات الکلی و غیره.
	الکل		
	سرنگ		
	ابزارآلات قمار		
نفرت‌پراکنی[7]	هتک حرمت[8]	توهین به مقدسات و ارزش‌ها	✓ هتک حرمت به نمادهای فرهنگی و مذهبی از طریق تصویر و کاریکاتورهای موهن.
		توهین به فرهنگ‌ها و اقلیت‌ها	

1 Child abuse
2 Sexual abuse
3 Animal abuse
4 Drugs
5 Alcohol
6 Gambling
7 Hate speech
8 Desecration

	فرقه‌گرایی و ترویج تروریسم	علائم و پرچم‌ها	✓ پرچم داعش ✓ علائم شیطان‌پرستی
	نژادپرستی و جدایی‌طلبی		✓ تصویر پرچم جوامع جدایی‌طلب ✓ پانترکیزم

۵-۲ محتوای نامناسب صوتی و متنی

همانطور که قبلاً نیز متذکر شدیم، اولین گام در تحلیل گفتار (صوت) این است که صوت دریافتی به متن تبدیل می‌شود و سپس تحلیل‌های بعدی بر روی متن انجام می‌گیرد؛ از این رو محتوای نامناسب صوتی و متنی را نیز همزمان در جدول ۲ بیان کرده‌ایم.

جدول ۳، محتوای نامناسب صوتی و متنی

نوع	تعریف
جنسی	متن یا صوت آمیخته به الفاظ جنسی نامناسب
خشونت	توصیف و ترویج جراحت و صدمه
	توصیف و ترویج آزار، اذیت و خودکشی
مواد مخدر، مشروبات الکلی و قمار	توصیف و ترویج استفاده از مواد مخدّر، مشروبات الکلی و قمار
نفرت پراکنی و توهین	نفرت‌پراکنی علیه دین، توهین به ادیان آسمانی، کتب مقدس، انبیاء الهی و معصومین (ع) و مقدسات مسلم اسلامی.
	نفرت‌پراکنی و توهین به مسئولین مملکتی

	توهین به قومیت و فرهنگ
	توهین به نژاد
	الفاظ رکیک
اسپم[1]	محتوای ناخواسته یا نامطلوب شامل ارسال پیام به گروه‌ها
اطلاعات گمراه‌کننده و اخبار جعلی[2]	اطلاعات نادرست و مشکوک که منبع موثق علمی ندارند.

لازم به ذکر است که به علّت ماهیت جهانی اینترنت، جغرافیا و فرهنگ تأثیر بسزایی بر روی محتوا دارند؛ زیرا محتوایی که در یک کشور، قانونی در نظر گرفته می‌شود، ممکن است در کشور دیگر قانونی نباشد. برای مثال تبلیغات در مورد فروش سلاح‌های گرم در آمریکا قانونی است اما در بریتانیا غیرقانونی محسوب می‌شود و یا انکار کشتار جمعی یهودیان در آلمان غیرقانونی است اما در اکثر کشورهای دیگر چنین نیست. این بدین معناست که تفاوت‌های ملّی نیز باید در مقوله‌ی پالایش محتوا در نظر گرفته شوند و به همین دلیل جدول انواع محتوای نامناسب در کشورهای مختلف می‌تواند تفاوت‌های جزئی داشته باشد که ناشی از تفاوت‌های فرهنگی و اجتماعی است [17].

حال، نوبت به بررسی کاربرد تکنیک‌های مختلف یادگیری ماشینی در کشف و پالایش هر یک از موارد بالا می‌رسد که در ادامه به آن‌ها خواهیم پرداخت.

۳-۵ پردازش تصویر در پالایش محتوا

سابق بر این، تنها از تکنیک‌هایی چون دسته‌بندی صحنه و کشف اشیا[3] جهت پالایش محتوا استفاده می‌شد که در کنار مزایای بسیار (برای مثال راحتی تعریف و حل مسئله)، معایب پرشماری نیز برای آنها قابل ذکر است. از جمله این معایب می‌توان به دقت کم این روش‌ها برای کشف محتوای نامناسب اشاره کرد؛ در واقع، برای کشف دقیق محتوای نامناسب (در هر سه نوع گفتار و متن و تصویر)، این دو روش

1 Spam
2 Fake news
3 Object Detection

بایستی با سایر روش‌های ارزیابی ادغام شوند. برای مثال، هیچ‌یک از این دو روش توانایی تشخیص تصاویرِ دارای نوشته‌های اهانت‌بار را ندارند. به‌عنوان مثالی دیگر، می‌توان به ابزار کشف پوست بدن[1] (که یکی از زیر مجموعه‌های کشف اشیا به حساب می‌آید) اشاره کرد که به‌عنوان راه حلی برای کشف برهنگی شناخته می‌شود. حال آنکه بر اساس گزارشات، در تصاویر با کنتراست نور بالا به مشکل برمیخورد. مشکل دیگر این روش، عدم تطابق آن با هدف کلی استفاده از تکنولوژی در پالایش محتوا (**تسهیل فرآیندها، افزایش سرعت و دقت**) است؛ با این توضیح که با استفاده از این روش، الگوریتم هوشمند در هر جا که اثری از پوست انسان پیدا کند، علامت (پرچم)[2] می‌زند. این مسئله باعث ایجاد خطاهای بیشمار (نشان‌دار کردن محتوا با کشف پوست دست، صورت و سایر نقاط غیرمرتبط) در ارزیابی تصاویر و افزایش حجم محتوای پرچم‌گذاری‌شده می‌شود.

با توجه به آنچه که گفته شد، مشخص می‌شود که از میان انواع مصادیق محتوای نامناسب که بایستی در ویدئوکلیپ‌ها، تصاویر اشتراکی، و تصاویر پروفایل کشف شود، پردازش تصویر در کشف عناوین زیر نقش اصلی را بازی می‌کند:

- تصاویر شامل انسان عریان و نیمه عریان، اسلحه، مواد مخدر، الکل و یا تبلیغات کالاهای نامناسب و غیره.
- تصویر انسان‌های منفور و یا افرادی که تصاویرشان نبایستی در دسترس عموم قرار بگیرد، علامات گروه‌های افراطی و شیطان‌پرستی و غیره.
- متون نامناسب و یا اطلاعات نادرست تعبیه‌شده در تصاویر با استفاده از نویسه‌خوان نوری.
- تصویر اطلاعات محرمانه افراد (شناسنامه، قراردادهای تجاری و ...).
- تصاویر نامناسب برای کودکان.

۵-۴ پردازش گفتار در پالایش محتوا

با توجه به توضیحاتی که پیرامون هر یک از سه مرحله پردازش گفتار ذکر شد، واضح است که هر سه مرحله‌یِ بازشناسی گفتار، پردازش زبان‌های طبیعی و بازشناسی الکترومیوگرافی در فرآیند پالایش محتوا

1 Skin Detection
2 Flag

کاربرد دارند. برای مثال فرض کنید که یک فایل صوتی (حاوی گفتار) با استفاده از فنّاوری بازشناسی گفتار (ASR) به متن تبدیل شده است؛ حال برای اینکه بدانیم آیا این گفتار مناسب است یا خیر، لازم است تا از دو نظر گفتار را مورد ارزیابی قرار دهیم:

i. آیا گفتار دارای کلمات رکیک و ممنوعه است؟
ii. اگر گفتار شامل کلمات رکیک نیست، آیا به لحاظ معنایی مناسب است یا خیر؟ برای مثال فرض کنید فردی بدون ذکر الفاظ رکیک، اقدام به افشای اطلاعات محرمانه افراد و یا ارگان‌ها بکند و یا تبلیغی انجام دهد که برای کودکان مناسب نیست.

نوع اول ارزیابی با تحلیل‌های مبتنی بر پردازش زبان‌های طبیعی (NLP) صورت می‌پذیرد و برای درک دقیق محتوای جمله (مورد دوم)، از فناوری کشف محتوای ضمنی گفتار (EMG) استفاده می‌شود که متشکل از مجموعه‌ای از ابزارهای مختلف است که به کمک هم به تعیین ماهیت گفتار (یا هدف گوینده) می‌پردازند.

با توجه به آنچه که گفته شد، می‌توان از پردازش گفتار در کشف عناوین زیر که در فرآیند ارزیابی محتوا دارای اولویت هستند، استفاده کرد:

- گفتار شامل الفاظ رکیک
- تحریک گروه‌های سیاسی، ملی، مذهبی و غیره.
- تبلیغات نامناسب
- اطلاعات نادرست
- نفرت‌پراکنی

۵–۵ پردازش متن در پالایش محتوا

از مهمترین ویژگی‌هایی که باعث اهمیت دوچندان پردازش متن در پالایش محتوا شده است، کاربرد آن در تحلیل محتوای تصویری صوتی، علاوه بر پالایش محتوای متنی است. برای مثال، از پردازش متن در ارزیابی محتوای تصویری (بررسی محتوای تعبیه شده در تصاویر)، درک مفهوم و تشخیص کلمات رکیک در گفتار و نوشتار استفاده می‌شود. شایان ذکر است که هر یک از این تحلیل‌ها، جایگزین‌های غیر متنی نیز دارند که متناسب با سرعت و دقت راه حل مدنظر برای مسئله، می‌توانند

انتخاب شوند. به‌طور خلاصه، مهمترین بخش‌هایی که پردازش متن در تحلیل آن به کمک ارزیابی محتوا می‌آید، به شرح زیر است:

- تحلیل متون تعبیه شده در تصاویر و کشف موارد نامناسب
- کشف الفاظ رکیک در گفتار
- کشف نظرات نامناسب
- تحلیل محتوایی گفتار و کشف موارد ممنوعه (توهین به حاکمیت‌ها، افشای اطلاعات محرمانه یا تحریک قومیت‌ها و سایر سخنان تنفرآمیز و ...).

در ادامه به بررسیِ مقوله‌ی پالایش محتوا در سه رسانه‌یِ مهم، تأثیرگذار و پر مخاطب در جهان؛ یعنی وی‌چت، یوتیوب و فیس‌بوک خواهیم پرداخت. این پالایش محتوا، گاه فنّاورانه و مبتنی بر یادگیری ماشینی و گاه انسان‌محور و با دخالت عامل انسانی صورت می‌پذیرد که در این نوشتار بخش فنّاورانه آن مورد تأکید است.

طبعاً به‌دلیل اینکه معماری، فرآیند و قوانین ارزیابی محتوا جزو اطلاعات مهم رسانه‌ها است، دسترسی دقیق به آن گاهاً میسّر نیست و بایستی با استفاده از سایر سرنخ‌ها، تحلیل‌هایی ارائه کرد که در متن به صراحت به آن اشاره شده است. لذا، پیش از آنکه درمورد نحوه ارزیابی هوشمند محتوا در رسانه‌ها صحبت کنیم، به برخی از مهم‌ترین مواردی که هنگام ارزیابی هوشمند محتوا بایستی مدنظر قرار بگیرد، اشاره می‌کنیم:

۱. پردازش محتوا (خصوصاً محتوای ویدئویی) به لحاظ قدرت پردازشی، بسیار سنگین است و ارزیابی فریم به فریم محتوا روش بهینه‌ای نیست (هر چند گاهی تنها گزینه است).
۲. پردازش محتوای تصویری (که مهمترین بخش پردازشی جهت ارزیابی است)، بایستی با پردازش محتوای صوتی و متنی تجمیع شود و به‌عنوان یک مجموعه‌ی کامل، در تصمیم‌گیری درباره حذف یا نگهداری یک محتوا بکار رود. برای مثال، ممکن است یک محتوای آموزشی، اگر در زمینه[1] دیگری منتشر شود، غیر مجاز باشد؛ ولی وقتی در عنوان آن ذکر می‌شود که ویدئو آموزشی است،

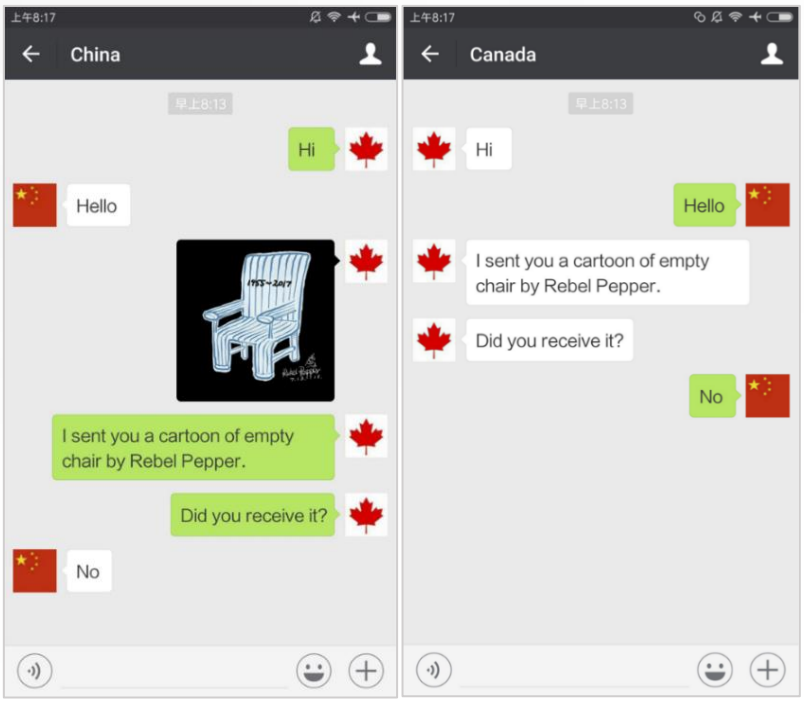
1 Context

نباید حذف گردد. این مهم، از ترکیب تحلیل متنی (عنوان کلیپ) و محتوایی (تصاویر و صوت گوینده) قابل دسترسی است.

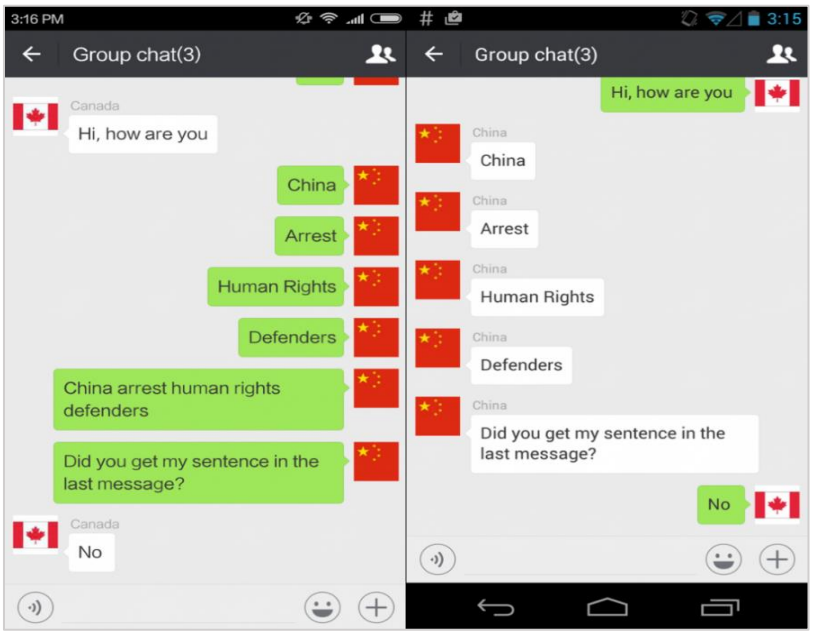
۶. استفاده از یادگیری ماشینی در وی‌چت برای پالایش محتوای تصویری

وی‌چت یک سرویس چند منظوره است که توسط شرکت تنسنت[1] (سرمایه‌گذار اصلی وی‌چت) در چین و در سال ۲۰۱۱ متولد شده است. در حال حاضر حدود ۲ میلیارد کاربر فعال ماهانه از این سرویس استفاده می‌کنند. بر اساس آنالیزی که از پالایش محتوای تصویری در وی چت صورت گرفته‌است مشخص شده است که وی چت از یک سامانه خودکار و بهنگام[2] جهت پالایش محتوا (خصوصاً از نوع تصویر) در محیط‌های گفتگو بهره می‌برد که بنای حذف تصاویر را بر ۱.شباهت بصری آن‌ها با نمایه[3] تصاویر موجود در فهرست محتوای نامناسب موسوم به لیست سیاه محتوا و یا ۲.محتوای نامناسب موجود در تصویر، گذارده است [26]. از این گزاره در می‌یابیم که تمامی پیام‌های تصویری در گفتگوهای وی چت با موتورهای نویسه‌خوان نوری و شباهت‌یاب اثر انگشت (کدهای هش) ارزیابی می‌شوند و در صورت کشف شباهت با نمایه‌های محتوای تصویری ممنوعه، به‌صورت خودکار حذف می‌شوند. برای توضیح بیشتر باید گفت که در روش مبتنی بر نویسه‌خوان نوری، متون تعبیه شده در داخل تصاویر و یا ویدئو توسط نویسه‌خوان استخراج شده و با مجموعه‌ای از کلمات کلیدی حساس مقایسه می‌شود و در صورتی که هر گونه انطباقی مشاهده شود، تصویر اولیه حذف می‌گردد (شکل ۵). لذا، علاوه بر نویسه‌خوان نوری، از تحلیل های متنی نیز در پالایش محتوای تصویری گفتگوهای وی‌چت، استفاده می‌شود.

1 Tencent
2 Realtime
3 Index



شکل ۵، مثال سانسور تصاویر در وی چت



شکل ٤، مثال سانسور محتوای نامناسب در وی‌چت

همچنین در گفتگوهای شخصی و گروهی وی چت، جملات معنی‌دار که مناسب تشخیص داده نمی‌شوند، پس از آنالیز از طریق پردازش متن، حذف می‌گردند (شکل ٤). نکته جالب توجه ارزیابی محتوا در وی چت، بررسی محتوا به دو زبان انگلیسی و چینی است که قابلیت حذف محتوای نامناسب در هر دو را میسر می‌سازد [26].

همچنین شرکت تنسنت سامانه‌ای به نام Instant Messaging را راه اندازی کرده است که بوسیله‌ی آن محتوای متنی نامناسب موجود در ویدئوهای لایو را حذف می‌کند. ادعا شده است که این ابزار محتوای رکیک متنی را تا ۹۰ درصد کاهش داده است. علاوه بر این، تنسنت ابزاری برای بررسی محتوای ویدئویی از لحاظ وجود صحنه‌های مستهجن دارد که در قالب سرویس‌های ابری آن را ارائه می‌کند [27]. این ابزار که Porny Image Recognition Technology نام دارد، به این صورت مورد استفاده قرار می‌گیرد که هر ۱۰ ثانیه (این میزان قابل تغییر است) از ویدئو یک شات می‌گیرد و با بررسی سطح تصاویر مستهجن در آن، تصمیم می‌گیرد که ــ در صورت سلامت محتوا ــ محتوا را تأیید کند و یا آن را برای بررسی نهایی، نشان‌دار کند و ادامه فرآیند ارزیابی را به عامل انسانی بسپارد.

ویدئوهای نشان‌دار شده توسط الگوریتم‌های کشف محتوای نامناسب، در عموم رسانه‌ها برای ارزیابان انسانی ارسال می‌شود تا درباره آن تصمیم‌گیری کنند؛ در واقع بایستی محتوای گزارش شده توسط مردم را نیز به این تعداد اضافه کرد. لکن، با عنایت به حجم بسیار زیاد محتوای بارگذاری شده در بستر رسانه‌ها، ارزیابی سریع این ویدئوها نیازمند استخدام نیروی انسانی بسیار است، که راه حل منطقی به نظر نمی‌رسد.

بطور کلی، گلوگاه ارزیابی محتوا در رسانه، بخش ارزیابی انسانی است که در صورت بهینه‌شدن فرآیندهای آن، بهره‌وری کل فرآیند ارزیابی محتوا در رسانه بطور چشمگیری رشد خواهد کرد. یکی از ایده‌هایی که در این زمینه تا کنون مطرح شده است، ایده خودکارسازی فرآیند ارزیابی و اعتماد کامل به هوش مصنوعی می‌باشد. مطمئناً این روشِ ارزیابی، مخالفان بسیاری دارد و مهم‌ترین دلیل آنها برای مخالفت با این روش، ریسک بسیار زیاد آن در ارزیابی محتوا و اتکای کامل به دقت شبکه‌های هوشمند است. مخالفان این ایده عقیده دارند که در جهانِ پویای کنونی، روندها سریعاً بوجود آمده و دائماً در حال تغییر هستند و از آنجا که آموزش آن به ماشین زمان‌بر است، با استفاده از این روش همواره یک گام از آنچه در جامعه اتفاق می‌افتد، عقب‌تر خواهیم بود. علاوه بر این، از آنجا که ماشین (هر چقدر هم که پیشرفته باشد)، توانایی استنباط مشابه انسان را ندارد، دقت مناسبی در هنگام مواجهه با روندهای ایجاد شده، ندارد. اما

استفاده از این روش مخصوصاً در رسانه‌هایی که محتوای موجود در آنها از سطح حساسیت بالایی برخوردار نیست؛ مانند شرکت‌های فعال در حوزه‌ی بازی آنلاین[1]، مانند Hatch[2] و Sulake[3] (فنلاند)، Roblox[4] و Animal Jam[5] (آمریکا)، Movie Star Planet[6] (دانمارک)، Hyperhippo[7] و Kabam[8] (کانادا) و یا شرکت‌های فعّال در حوزه آموزش مانند Brainly (آمریکا) و حتی شبکه‌های اجتماعی کوچک مانند Yubo[9] (فرانسه) می‌تواند بسیار کارآمد باشد. بدین صورت که یک مدل هوش مصنوعی، بر اساس تصمیمات قبلی ارزیاب‌های انسانی و متناسب با مسئله، یاد می‌گیرد که ۱. گزارشات اشتباه را حذف کند، ۲. اقدامات لازم جهت حل و فصل گزارشاتی که به خطای واضح و آشکار فردی اشاره دارد را انجام دهد و ۳. گزارشاتی که حتما بایستی توسط انسان مورد بازبینی قرار بگیرد را اولویت‌بندی کند.

۷. استفاده از یادگیری ماشینی در یوتیوب برای پالایش محتوا

یوتیوب یک پلتفرم سرویس اشتراک ویدیو با ۲ میلیارد کاربر فعال در ماه می‌باشد که حدود ۷۹ درصد از کاربران اینترنت را جذب خود نموده است [28]. این کاربران از اقصی‌نقاط جهان با سلایق، انگیزه‌ها، ذهنیت‌ها، ادیان و فرهنگ‌های متنوع می‌توانند به این سایت مراجعه کرده و علیرغم مشاهده‌ی ویدیوهای مندرج در آن، پس از ثبت‌نام، ویدیوهایی را بارگذاری نمایند که ممکن است با ارزش‌ها و هنجارهای اجتماعی برخی از جوامع و برای رده‌های سنی بخصوصی ناسازگار و نامناسب باشد؛ لذا از این جهت مقوله‌ی پالایش محتوا در این رسانه بسیار حائز اهمیت است.

پیش از توضیح درباره‌ی سازوکار فرآیند ارزیابی محتوای یوتیوب، بایستی بگوییم که در راستای کاهش بار پردازشی کامپیوتری و فنی مورد نیاز جهت ارزیابی و پالایش محتوا در این پلتفرم، ابزاری در اختیار تولیدکنندگان محتوا قرار گرفته است که با آن می‌توانند به راحتی بخش‌های نامناسب ویدئو خود را تار[10] کنند [29].

---

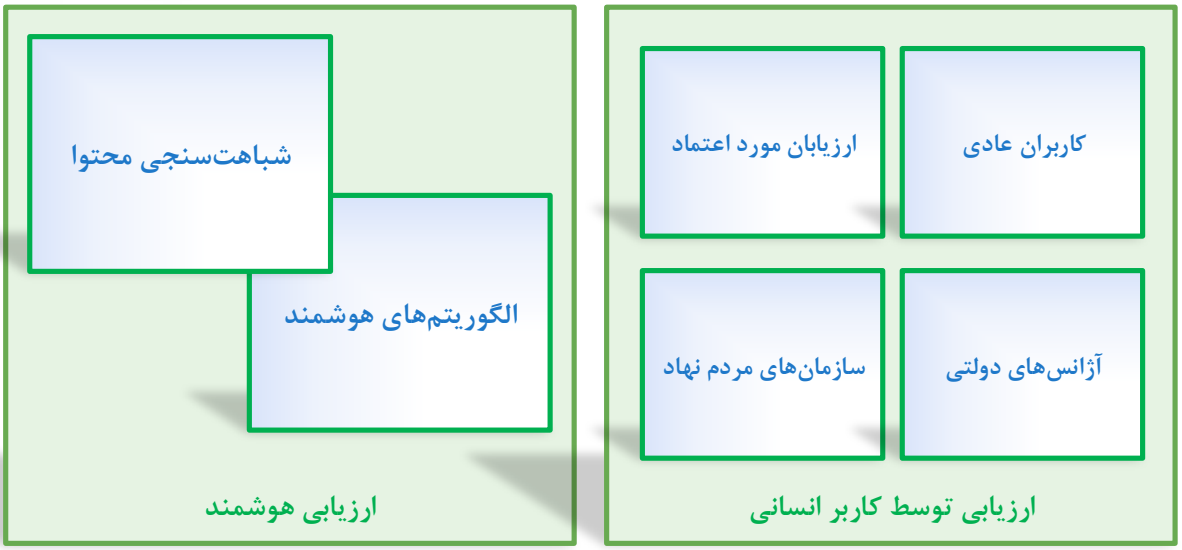
1 gaming
2 https://playhatch.com/
3 https://www.sulake.com/
4 https://www.roblox.com/
5 https://www.animaljam.com/
6 https://corporate.moviestarplanet.com/
7 https://hyperhippo.ca/
8 https://kabam.com/
9 https://yubo.live/en
10 Blur

ارزیابی محتوا در یوتیوب به دو مورد ارزیابی هوشمند و ارزیابی توسط کاربر انسانی (ارزیابی محتوا بر اساس گزارشات/شکایات کاربران[1]) تقسیم می‌شود (شکل ۵). ارزیابی هوشمند خود به دو دسته مجزا قابل تقسیم است که عبارتند از:

۱) ارزیابی محتوا با استفاده از الگوریتم‌های هوشمند پردازشی مبتنی بر یادگیری ماشینی.

۲) ارزیابی محتوا با مکانیزمی مشابه به آنچه که در سامانه Content ID[2] وجود دارد. روش حذف محتوا در این حالت بدین صورت است که موتورهای کشف یوتیوب بر اساس اثر انگشت[3] محتوا، به بررسی تشابه میان ویدئوی ورودی و ویدئوهایی که سابق بر این (با هر روشی) حذف شده است، پرداخته و در صورت وجود تشابه، آن را بدون بررسی بیشتر حذف می‌نماید.

طبق ادعای یوتیوب، در بازه زمانی سه ماهه آوریل تا ژوئن ۲۰۲۰، ۱۱۴۰۱۶۹۶ محتوای ویدئویی از بستر این پلتفرم حذف شده است [30] که سهم هر یک از روشهای ارزیابی به شرح نمودار ۲ است.

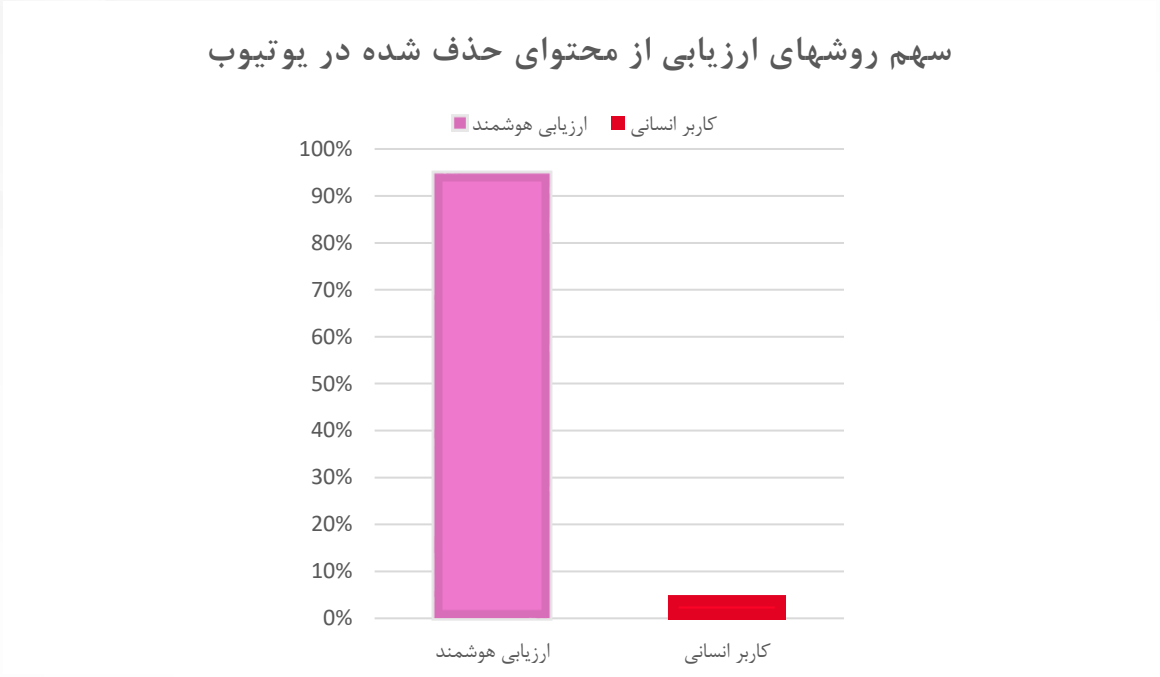


شکل ٦، روش‌های ارزیابی محتوا در یوتیوب

[1] شامل کاربران عادی، ارزیابان مورد اعتماد، NGO ها و یا آژانس‌های دولتی

[2] Content ID ابزار مشابهت‌یاب یوتیوب در راستای حفظ مالکیت فکری افراد بر محتوای تولیدی است [39].

3 Fingerprint



نمودار ۲، سهم روش‌های ارزیابی از محتوای حذف شده در یوتیوب در بازه‌ی زمانی آوریل تا ژوئن ۲۰۲۰

اما در بخش ارزیابی هوشمند محتوا در یوتیوب، هر یک از موارد نامناسبی که سابق بر این معرفی گردید با استفاده از الگوریتم‌های مبتنی بر پردازش صوت، متن و تصویر کشف می‌شوند که در ذیل بدان پرداخته‌ایم.

- انسان عریان: برای کشف هر گونه محتوای شامل برهنگی، ابتدا نقاط مختلف بدن را مشخص نموده و با آموزش ماشین هوشمند کشف برهنگی و نیمه برهنگی، محتوا را ارزیابی و درجه بندی می‌کند [31].
- کودک آزاری: یوتیوب با تعریف مصادیق کودک آزاری مانند تنبیه کودک و ویدئوهای ترسناک و نامناسب برای کودکان، اقدام به کشف این موارد با استفاده از پردازش تصویر می‌نماید.
- کلیپ‌های مروج خشونت و شکار غیرقانونی: این موارد نیز به سادگی با استفاده از الگوریتم‌های پردازش تصویر یوتیوب، قابل کشف است؛ به این صورت که فریم‌های مختلف توسط موتورهای پردازشی ارزیابی شده و امتیازی از میزان خشونت در هر ویدئو به آن تخصیص می‌یابد که مرجع تصمیم‌گیری درباره حذف و یا نگه‌داری آن محتوا خواهد بود [31].

- عبارات توهین آمیز و تحریک‌کننده: اگر این محتوا به‌صورت متن تعبیه شده در تصویر باشد، با استفاده از نویسه خوان نوری (پردازش تصویر) استخراج می‌شود و با تحلیل متن ارزیابی می‌گردد. چنانچه این عبارات در گفتار مستتر باشد، با استفاده از بازشناسی گفتار (ASR) آن را تبدیل به متن کرده و با استفاده از تحلیل متنی سلامت محتوا ارزیابی می‌شود [31].
  - نکته: روشهای دیگری نیز برای ارزیابی عبارات رکیک و توهین آمیز موجود است؛ برای مثال یافتن کلمات کلیدی در گفتار[1] که دیگر نیاز به تحلیل مفهوم کلی گفتار ندارد و فقط لازم است به دنبال کلمات خاصی در صوت بگردد. این روش علیرغم کاربرد بسیار زیاد در محتوای متنی، در حوزه صوت دقت قابل قبولی ندارد و مورد اقبال قرار نمی‌گیرد.
- افشای اطلاعات محرمانه و یا دعاوی کذب هویتی: افشا اطلاعات محرمانه در یوتیوب، با تجمیع الگوریتم‌های پردازش صوت (کشف عبارات فاش کننده اسرار) و متن (عنوان ویدئو که بیانگر افشا اطلاعات است) و تصویر (درک افشا تصویر اسناد محرمانه) قابل کشف است. همچنین، یوتیوب با استفاده از پردازش تصویر، به احراز دعاوی هویتی پرداخته و موارد تخلف را شناسایی می‌کند.

۸. استفاده از یادگیری ماشینی در فیس‌بوک برای پالایش محتوا

فیس‌بوک جزو پرمخاطب‌ترین شبکه‌های اجتماعی برخط با بیش از ۲٫۷ میلیارد کاربر فعال در ماه می‌باشد [32]، این کاربران روزانه مطالب و اطلاعات متنوع و متعددی را در این پلتفرم نشر می‌دهند که برخی از این محتواها مصادیق بارزِ تصاویر نامناسب مانند برهنگی و نیمه‌برهنگی، نفرت‌پراکنی (هر محتوایی که به طور مستقیم به فرد یا گروهی توهین نماید)، اخبار جعلی و غیره هستند که تأثیر بسزایی در فضای حقیقی دارند و در برخی موارد باعث بروز ناامنی و خشونت می‌گردند. به عنوان نمونه در ۳۱ آگوست ۲۰۲۰ یک کهنه سرباز آمریکایی به نام رونی مک نات[2] در یک ویدئو زنده در فیسبوک اقدام به خودکشی با اسلحه کرد. فیس‌بوک قادر به شناسایی و حذف بی‌درنگِ این ویدیوی خشونت‌آمیز نبود و با تأخیر سه ساعته آن را حذف کرد، اما در ظرف همین سه ساعت به دلیل بازدید بالا، این ویدئو در

1 Keyword Extraction in Speech
2 Ronnie McNutt

اینستاگرام، یوتیوب و دیگر پلتفرم‌ها نیز بازنشر شد [33]. همین اتفاق مجدداً توجه‌ها را به سوی مقوله پالایش آنی و بهنگام محتوا و چالشهای آن جلب کرد. واضح است که فیس‌بوک برای امن نگه‌داشتن فضای خود و جلوگیری از نشر چنین فجایعی ناچار است محتوای مندرج در بستر خود را به صورت آنی و بهنگام پالایش و ارزیابی نماید. از آنجا که ارزیابی این حجم از محتوا صرفاً توسط عوامل انسانی ممکن نیست، لذا فیس‌بوک لازم است به روش‌های مبتنی بر یادگیری ماشینی روی آورده و از به روزترین و دقیق‌ترین این ابزارها استفاده نماید.

لازم به ذکر است که پالایش هوشمند محتوا در فیسبوک سابقه‌ی دور و درازی دارد. در سال ۲۰۱٤، با ورود تانتون گیبس[1] به فیس‌بوک ــ که تجربه همکاری با شرکت‌هایی همچون گوگل و مایکروسافت را داشت ــ عصر جدیدی در این سازمان آغاز شد. در آن سال فیس‌بوک با استفاده از فناوری‌ای به نام PhotoDNA که توسط مایکروسافت توسعه یافته بود شروع به شناسایی و حذف تصاویر سوءاستفاده از کودکان، از پلتفرم خود کرد [34]. این فناوری یک کد منحصر به فرد از تصاویر غیرمجاز (که به آن Hash یا Fingerprint می‌گویند) ایجاد می‌کند و از آن برای مقایسه و یافتن دیگر تصاویر غیرمجاز بهره می‌برد [35]. در سال ۲۰۱۸ فیس‌بوک برای اولین‌بار از راهکاری موسوم به PDQ[2] برای پالایش بهنگام تصاویر نامناسب استفاده کرد که از دقت بالایی برخوردار است. الگوریتم PDQ در ابتدا یک کد هش ۲۵٦ بیتی تولید می‌کند. این کد شامل یک زیر عکس ۱٦در۱٦ از عکس اصلی است. سپس هر بیت با توجه به موقعیت خود و بیت‌های اطراف آن به دو حالت ۰ و ۱ تبدیل می‌شود. از این هش بدست آمده برای محاسبه فاصله همینگ[3] بین دو تصویر استفاده می‌شود. به طور متوسط فاصله دو عکس که به صورت رندوم انتخاب شدند حدود ۱۲۸ می‌باشد. در صورتی که این فاصله به کمتر از ۳۰ برسد با دقت خوبی می‌توان اطمینان داشت که دو عکس تطابق دارند.

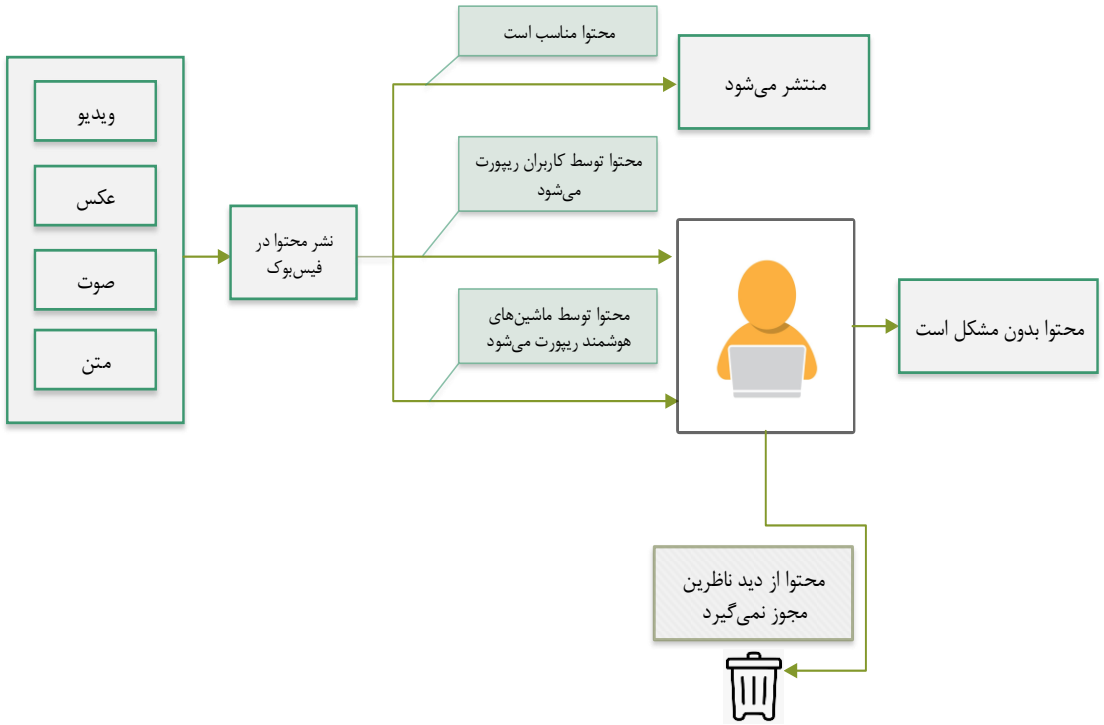
الگوریتم TMK + PDQF[4] نیز از یک نسخه بهینه‌سازی شده از PDQ استفاده می‌کند با این تفاوت که این الگوریتم قادر به محاسبه متغیرهای مربوط به زمان نیز هست، که همین ویژگی باعث می‌شود برای

1 Tanton Gibbs
2 Perceptual algorithm utilising a Discrete Cosine Transform and outputting (amongst others) a Quality metric
3 Hamming distance
4 Temporal Match Kernel

پیدا کردن مشابهت‌های ویدئویی نیز مورد استفاده قرار گیرد. این الگوریتم به صورت کلی شامل ٤ بخش اصلی است:

١. بازسازی ویدئو با نرخ ١٥ فریم در ثانیه
٢. محاسبه و شناخت یک المان توصیف‌گر برای هر فریم
٣. محاسبه میانگین برای بازه‌های مشخص
٤. تولید کد هش بر اساس میانگین‌های محاسبه شده.

رویکرد فیس‌بوک در مورد شناسایی و حذف محتواهای نامناسب را به طور کلی می‌توان به دو بخش گزارش کاربران و نظارت هوشمند که در بالا به اجمال به آن پرداختیم، تقسیم کرد. پس از انتشار یک پست، کاربران می‌توانند موارد ناپسند آن را گزارش دهند تا در روند بررسی قرار گیرد. شکل زیر (شکل ٧) به خوبی روشنگر این موضوع است. در حال حاضر پلتفرم فیس‌بوک بیش از ١٥ هزار ناظر محتوا دارد که مسئول بررسی این محتواها هستند [36]، محتواهای گزارش شده توسط کاربران و یا ماشین‌های هوشمند ارزیابی محتوا ـ که اصطلاحاً به آن‌ها تیکت[1] گفته می‌شود ـ به طور کاملاً اتفاقی و رندوم بین ناظران توزیع شده و ناظران مشغول به بررسی آن‌ها می‌شوند.



شکل ٧، نحوه‌ی ارزیابی محتوا توسط موتورهای هوشمند و ناظرین در فیس‌بوک

1 Ticket

با توجه به حجم بالای گزارش‌ها، هر ناظر می‌بایست روزانه حدود ۱۳۰۰ تیکت را بررسی کند [37] که به دلیل حجم بالای محتوا و خطای انسانی این کار بسیار طاقت‌فرسا است و ممکن است با خطا یا تأخیر در بررسی مواجه شوند. بعلاوه برخی از محتواهایی که توسط الگوریتم‌های پردازشی هوشمند و یا کاربران بعنوان محتوای نامناسب گزارش می‌شوند، بایستی به سرعت توسط نیروی انسانی مورد بازرسی قرار گیرند و سریعاً حذف گردند. این مسئله ناشی از تفاوت حساسیت دسته‌های مختلف محتوای نامناسب می‌باشد؛ برای مثال حساسیت حذف ویدئو فاجعه مسجد شهر کرایست‌چرچ[1] بسیار بیشتر از حذف یک ویدئو شامل برهنگی است. لذا، برای اولویت دهی به محتوای در صف ارزیابی ناظران، فیسبوک از موتورهای هوشمندی بهره می‌برد که محتوا را بر اساس حساسیت، اولویت‌بندی کرده و در اختیار ناظرین قرار می‌دهند. این امر باعث تسریع فرآیند ارزیابی محتوای حساس (مانند آنچه که در پرونده رونی مک‌نات ذکر شد) می‌شود که خود منجر به کاهش پخش ویدئوهای نامناسب در سطح پلتفرم‌ها خواهد شد.

همچنین فیس‌بوک در طی دو سال گذشته، مبارزه با اطلاعات نادرست را در اولویت قرار داده است. یکی از چندین اقدامی که این شرکت برای سنجش اخبار نادرست و کاهش نرخ انتشار آن‌ها انجام می‌دهد، استفاده از صحت‌سنج[2]ها برای بررسی و ارزیابی صحت و درستی مطالب می باشد. این شرکت، با هدف شناسایی سریع‌تر و دقیق‌ترِ طیف وسیعی از اطلاعات نادرست، بررسی و صحت‌سنجی عکس‌ها و فیلم‌ها را به ۲۷ شریک خود در ۱۷ کشور در سرتاسر جهان گسترش داده است.

فیس‌بوک از یک مدل یادگیری ماشینی که ساخت خود شرکت بوده استفاده می‌کند و از سیگنال‌های تعاملی مختلف ــ از جمله نرخ بازخورد محتوا برای شناسایی محتوای بالقوه نادرست ــ بهره می‌برد و پس از شناسایی تصاویر و فیلم‌های مشکوک، آن‌ها را برای بررسی بیشتر و تایید میزان صحت و درستی، به مراکز صحت‌سنجی می‌فرستد. بسیاری از شرکای صحت‌سنجی، در ارزیابی عکس‌ها و فیلم‌ها تخصص داشته و در فنون تأیید صحت تصاویر، مانند جستجوی عکس معکوس[3] و یا تجزیه و تحلیل فراداده‌های

1 Christchurch Mosque Massacre
2 Fact-checking
[3]جستجوی عکس معکوس: یک فناوری موتور جستجو است که ورودی آن یک فایل تصویری و خروجی آن نتایج مربوط به آن تصویر می باشد. به‌عبارت‌دیگر، مانند جستجوی متن، جستجوی تصویر نیز دارای یک سیستم بازیابی اطلاعات است که برای کمک به یافتن اطلاعات در اینترنت طراحی شده و در واقع این کار با استفاده از فراداده‌ی تصاویر (موقعیت، تاریخ و ساعت عکس، دستگاه گیرنده عکس، فرمت عکس و ...) و همچنین ارتباط تصویر با دیگر تصاویر انجام می‌شود [40].

تصویر، مانند زمان و مکانِ گرفتن آن‌ها، آموزش دیده‌اند. علاوه بر آن، مراکز صحت‌سنجی می‌توانند با ترکیب این مهارت‌ها با سایر روش‌های روزنامه نگاری، مانند استفاده از تحقیقات متخصصان، دانشگاهیان یا سازمان‌های دولتی، صحت یا عدم صحت عکس یا فیلم را مورد ارزیابی قرار دهند.

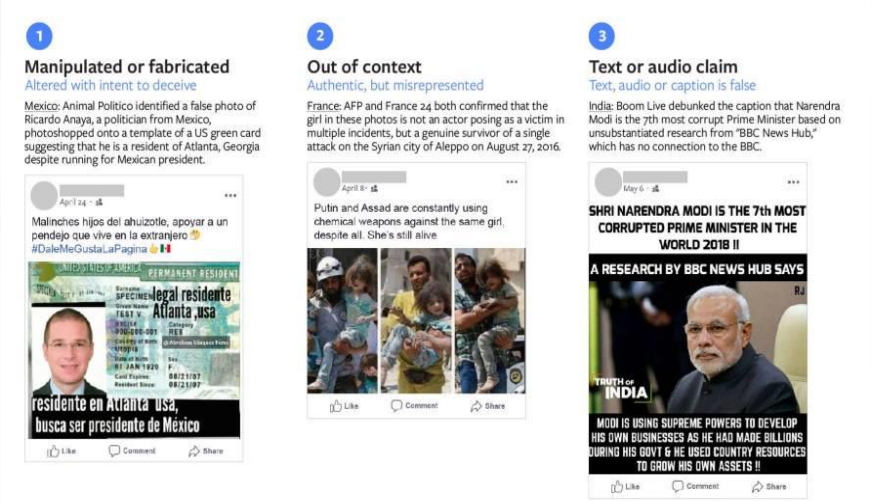
با کسب بازخورد بیشتر از مراکز صحت‌سنجی درمورد عکس‌ها و فیلم‌ها، فیس‌بوک می‌تواند دقت مدل یادگیری ماشین خود را ارتقا دهد. علاوه بر آن، استفاده از سایر فناوری‌های مرتبط با این حوزه نیز می‌تواند در فرآیند شناسایی محتوای کاذب و گمراه کننده مثمر ثمر واقع گردد. به عنوان مثال، بهره‌گیری از فنّاوری نویسه‌خوان نوری (OCR) برای استخراج متن از عکس‌ها و مقایسه آن با توضیحات تصویر، راهکار جدیدی است که برای دستیابی به این هدف مورد استفاده قرار گرفته است.

با این حال شرکت فیس‌بوک تنها به فناوری‌های موجود بسنده نکرده و در حال مطالعه و تحقیق بر روی روش‌های جدیدتری است تا بتواند عکس‌ها و فیلم‌های گمراه کننده بیشتری را شناسایی کرده و برای بررسی بیشتر به صحت‌سنج‌ها ارسال کند.

بر اساس نتایج چندین ماه تحقیق و آزمایش با تعداد زیادی از شرکای شرکت فیس‌بوک، اطلاعات نادرست در عکس‌ها و فیلم‌ها به طور معمول در سه دسته قرار می‌گیرند که عبارتند از [38]:

۱. دستکاری شده یا ساختگی: عکس و فیلم با ابزارهای متفاوت هوش مصنوعی، مانند جعل‌عمیق، ساخته یا دستکاری شده‌اند.
۲. خارج از محتوا[1]: با اینکه اطلاعات مطابق با واقعیت هستند؛ ولی با هدف ضربه زدن به فرد یا سازمان خاصی منتشر شده‌اند.
۳. متن اشتباه: متن خبری، منطبق با عکس بارگذاری شده نیست.

1 Out of context



شکل ۸ اطلاعات نادرست در عکسها و فیلمها در فیس‌بوک در سه دسته مشخص شده قرار می‌گیرد.

۹. اکوسیستم پیشنهادی تنظیم‌گری رسانه‌های صوت و تصویر فراگیر

بنابر نظر راقم این سطور می‌توان اکوسیستمی را در راستای تنظیم‌گری رسانه‌های صوت و تصویر فراگیر پیشنهاد داد. این اکوسیستم (شکل ۹)، دارای چهار المان است که عبارتند از:

۱. نهادهای تنظیم‌گر

- صدور دستورات قهری و اعمال جریمه‌ی رسانه و نهادهای متولّی فنّاوری تنظیم‌یار.

۲. نهادهای متولّی فنّاوری تنظیم‌یار (RegTech)

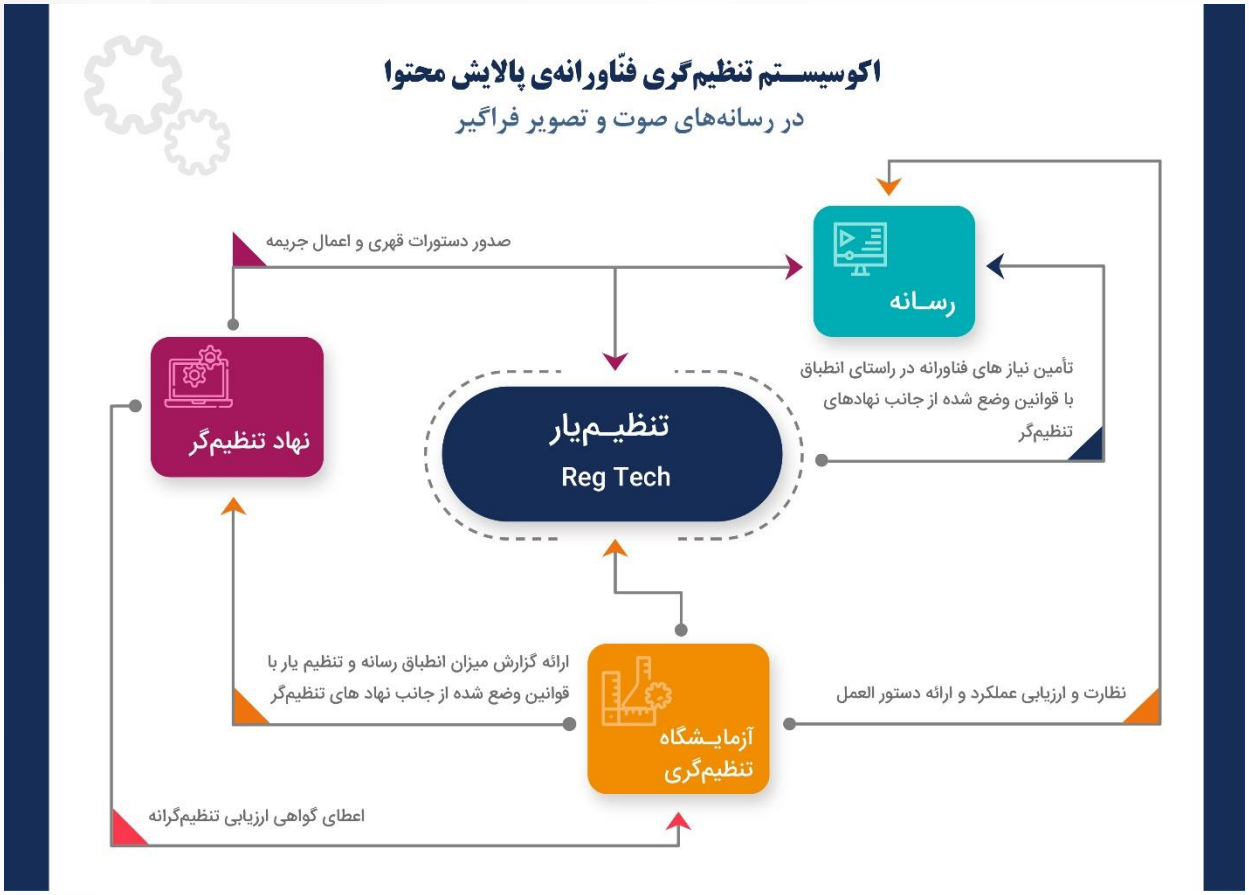
- تأمین نیازهای فنّاورانه در راستای انطباق با قوانین وضع شده از جانب نهادهای تنظیم‌گر برای رسانه‌های صوت و تصویر فراگیر.

۳. آزمایشگاه تنظیم‌گری

- نظارت و ارزیابی عملکرد رسانه و نهادهای متولّی فنّاوری تنظیم‌یار و ارائه‌ی دستورالعمل،
- ارائه‌ی گزارش میزان انطباق رسانه و نهاد متولّی فنّاوری تنظیم‌یار با قوانین وضع شده از جانب نهادهای تنظیم‌گر.

٤. رسانه

نقش هر یک از این المان‌های فوق در شکل ۹ نیز به طور مبسوط مشخص شده است.



شکل ۹، اکوسیستم تنظیم‌گری فنّاورانه‌ی پالایش محتوا در رسانه‌های صوت و تصویر فراگیر

منابع

[۱] M. Puppis" ،Media Governance: A New Concept for the Analysis of Media Policy and Regulation ".*Commiunication, Culture and Critique* ،pp. 134-149 .۲۰۱۰ ،

[۲] Available: https://www.entekhab.ir/fa/news/588741./[متصل].

[۳] Available: https://tech.sina.com.cn/i/2018-11-29/doc-ihpevhcm3575934.shtml.[متصل].

[٤] Available: https://www.bbc.com/news/technology-42510868.[متصل].

[٥] Available: https://bit.ly/2VN0dti, 2019.[متصل].

[٦] A. D. STREEL ،Online Platform's Moderation of Illegal Content Online - Law, Practicies and Option for Reform ،Policy Department for Economics, Scientific and Quality of Life .۲۰۲۰ ،

[۷] J. Wang" ،Regulation of Digital Media Platforms: The case of China ,"The Foundation for Law, Justice and Society, in association with the Centre for Socio-Legal Studies and Wolfson College ،University of Oxford .۲۰۲۰ ،

[۸] Available: http://yuqing.people.com.cn/GB/392071/401685/index.html.[متصل].

[۹] Available: http://www.cac.gov.cn/zcfg/A0909index_1.htm.[متصل].

[۱۰] Available: : http://www.cac.gov.cn/2017-09/07/c_1121624269.htm.[متصل].

[۱۱] Available: http://www.cac.gov.cn/2019-12/20/c_1578375159509309.htm.[متصل].

[۱۲] [متصل]. Available: http://www.ecns.cn/news/2018-11-14/detailifyzrwsr0795942.shtml.

[۱۳] [متصل]. Available: https://www.reuters.com/article/us-chinainternet-censorship-idUSKBN18Z0J3.

[۱٤] [متصل]. Available: : https://www.reuters.com/article/us-china-toutiao/bad-humour-china-watchdog-shuts-toutiao-joke-app-over-vulgar-content-idUSKBN1HI0C4.

[۱٥] "Alan_Turing_and_the_beginning_of_AI [متصل]. "،Available: https://www.britannica.com/technology/artificial-intelligence./

[۱٦] S .& .N. P. Russel"" ،Artificial intelligence: a modern approach .۲۰۰۲ "،"

[

[۱۷] C. Consultants" ،Use of AI in Online Content Moderation.۲۰۱۹ "،

[

[۱۸] [متصل]. Available: https://en.wikipedia.org/wiki/Digital_image_processing.

[

[۱۹] [متصل]. Available: https://en.wikipedia.org/wiki/Speech_processing.

[

[۲۰] [متصل]. Available: https://en.wikipedia.org/wiki/Deep_learning.

[

[۲۱] [متصل]. Available: https://virgool.io/apieco/post2-gjbl07fdwc53.

[

[۲۲] [متصل]. Available: https://www.businessofapps.com/news/youtube-bolsters-efforts-to-remove-inappropriate-content-and-hate-speech./

[۲۳] [متصل]. Available: https://firstmonday.org/article/view/2378/2089. [

[۲٤] [متصل]. Available: https://www.cullen-international.com/events/webinar/2019/06/New-responsibilities-for-content-sharing-platforms-in-the-EU.html. [9

[۲٥] [متصل]. Available: https://www.cigionline.org/articles/why-social-platforms-are-taking-some-responsibility-content. [

[۲٦] [متصل]. Available: https://citizenlab.ca/2019/07/cant-picture-this-2-an-analysis-of-wechats-realtime-image-filtering-in-chats./ [

[۲۷] [متصل]. Available: https://mc.qcloudimg.com/static/qc_doc/a663a282dd120528fd0cb1df406fbeab/doc--Product+Intro.pdf. [

[۲۸] [متصل]. Available: https://www.omnicoreagency.com/youtube-statistics./ [

[۲۹] [متصل]. Available: https://www.diyphotography.net/you-can-now-censor-parts-of-video-directly-in-youtubes-creator-studio./ [

[۳۰] [متصل]. Available: https://transparencyreport.google.com/youtube-policy/removals?hl=en. [

[۳۱] [متصل]. Available: https://www.businessofapps.com/news/youtube-bolsters-efforts-to-remove-inappropriate-content-and-hate-speech./ [

[۳۲] [متصل]. Available: https://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide. [

[۳۳] [متصل]. Available: https://techcrunch.com/2020/09/13/graphic-video-of-suicide-spreads-from-facebook-to-tiktok-to-youtube-as-platforms-fail-moderation-test../ [

[۳٤] [متصل]. Available: https://www.iicsa.org.uk/publications/investigation/internet/part-c-indecent-images-children/c2-detection-images.. [

[۳٥] [متصل]. Available: https://www.wired.com/story/ai-has-started-cleaning-facebook-can-it-finish. [

[۳٦] [متصل]. Available: https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona. [

[۳۷] [متصل]. Available: https://www.wired.com/story/ai-has-started-cleaning-facebook-can-it-finish. [

[۳۸] [متصل]. Available: https://about.fb.com/news/2018/09/expanding-fact-checking. [

[۳۹] [متصل]. Available: https://en.wikipedia.org/wiki/Content_ID_(system). [

[٤۰] [متصل]. Available: https://money.cnn.com/news/newsfeeds/articles/djf500/200807301025DOWJONESDJONLINE000654_FORTUNE5.htm. [