محققان یاهو دقیق ترین الگوریتم تشخیص محتوای توهین آمیز در جملات را توسعه داده اند - دیجیاتو

امین بیگ‌زاده | شنبه، ۰۹ مرداد ۱۳۹۵

تیمی از محققان Yahoo Labs به تازگی با بررسی حجم عظیمی از نظرات ثبت شده در وبسایت این کمپانی، موفق به توسعه الگوریتمی شده اند که می تواند در تشخیص و جلوگیری از نظرات آزاردهنده کاربرد داشته باشد. به استناد وبسایت Technology Review، الگوریتم برنامه نویسی یاهو بهترین ابزار خودکاریست که تا کنون برای فیلتر کردن نظرات آزاردهنده و توهین آمیز طراحی شده است.

بسیاری از روش های فیلترینگ نظراتی که این روزها در سرویس های آنلاین استفاده می شوند، به ترکیبی از واژه های ممنوعه، اصطلاحات رایج و ساختار جملات برای تشخیص محتوای نفرت پراکنی تکیه می کنند. اما محققان یاهو یک گام فراتر گذاشته و یادگیری ماشینی را برای انجام این کار به خدمت گرفته اند.

با استفاده از تکنیک word embedding، که واژه ها را به عنوان یک کمیت برداری و نه صرفا مثبت یا منفی بودن بررسی می کند، سیستم جدید یاهو می تواند جملات دارای محتوای توهین آمیز را تشخیص بدهد، حتی اگر کلمات جمله به تنهایی معنای بدی نداشته باشند.

به ادعای یاهو سیستم مورد بحث در آزمایشات صورت گرفته، 90 درصد مواقع در تشحیص ادبیات توهین آمیز در جملات موفق بوده است. اگرچه عملکرد این الگوریتم بسیار تحسین برانگیز است، اما نباید فراموش کرد که نفرت پراکنی کلامی موضوعی است که دائما در حال تغییر و تحول بوده و شاید حتی یک انسان هم نتواند به طور صد در صد توهین آمیز بودن یک جمله را تشخیص بدهد.

به گفته الکس کرازودومسکی-جونز، محقق فعال در حوزه سوء استفاده های اینترنتی، از هر ده توئیتی که به گروهی از انسان ها برای تشخیص آزاردهنده بودن یا نبودن داده شد، به ندرت پیش آمد که همگی بر سر تشخیص محتوای یک توئیت به توافق برسند. بنابراین خودتان تصور کنید که انجام این کار برای یک رایانه چقدر مشکل خواهد بود.

دیجیاتو