Regulation of Social Media and the Evolution of content: a cross-platform analysis

by Marina Rizzi | Collegio Carlo Alberto & University of Turin

Is self-regulation of social media effective in moderating online con- tent? Or does it shift abusive posts to darker corners of the internet? This paper addresses these questions analyzing the effect of social media regulation from a cross-platform perspective. I exploit an episode of enlargement of Twitter's regulation against racist hate speech and I investigate whether regulation is effective in curbing this harmful content or whether it shifts abusive content to unregulated platforms. I exploit synthetic difference-in-differences to assess the effects of the new Twitter's policy, and I construct text classifiers using supervised methods from natural language processing to predict whether a post can be considered targeted hate speech. I find that racist content decrease in percentage both in Twitter and in Parler, but that users that have accounts in both platform increase their racist posts on Parler after Twitter's policy. Finally, I exploit word embedding to see if racist language changed after the regulation, and I find that terms related to some minorities became more offensive in Parler after Twitter's policy. Overall, my results show a possible substitution effect between platforms, with more abusive content shifting from regulated to less regulated social media platforms.