Mixed Messages? The Limits of Automated Social Media Content Analysis [Extended Abstract]*

Natasha Duarte NATASHA@CDT.ORG
Emma Llanso EMMA@CDT.ORG
Anna Loup ALOUP@USC.EDU
Center for Democracy & Technology

Editors: Sorelle A. Friedler and Christo Wilson

Abstract

Governments and companies are turning to automated tools to make sense of what people post on social media. Policymakers routinely call for social media companies to identify and take down hate speech, terrorist propaganda, harassment, fake news or disinformation. Other policy proposals have focused on mining social media to inform law enforcement and immigration decisions. But these proposals wrongly assume that automated technology can accomplish on a large scale the kind of nuanced analysis that humans can do on a small scale. Todays tools for analyzing social media text have limited ability to parse the meaning of human communication or detect the intent of the speaker.

A knowledge gap exists between data scientists studying natural language processing (NLP) and policymakers advocating for wide adoption of automated social media analysis and moderation. Policymakers must understand the capabilities and limits of NLP before endorsing or adopting automated content analysis tools, particularly for making decisions that affect fundamental rights or access to government benefits. Without proper safeguards, these tools can facilitate overbroad censorship and biased enforcement of laws or terms of service.

This paper draws on existing research to explain the capabilities and limitations of text classifiers for social media posts and other online content. It is aimed at helping researchers and technical experts address the gaps in policymakers knowledge about what is possible with automated text analysis. We provide an overview of how NLP classifiers work and identify five key limitations of these tools that must be communicated to policymakers:

1. NLP classifiers require domain-specific training and cannot be applied with the same reliability across different domains. Policies should not rely on the use of a one-size-fits-all tool.
2. Decisions based on automated social media content analysis risk further marginalizing and disproportionately censoring groups that already face discrimination. NLP tools can amplify social bias reflected in language and are likely to have lower accuracy for minority groups who are underrepresented in training data.
3. Accurate text classification requires clear, consistent definitions of the type of speech to be identified. Policy debates around content moderation and social media mining tend to lack such precise definitions.
4. The accuracy achieved in NLP studies does not warrant widespread application of these tools to social media content analysis and moderation.
5. Text filters remain easy to evade and fall far short of humans ability to parse meaning from text. Human review of flagged content remains essential for avoiding over-censorship.

The paper concludes with recommendations for NLP researchers to bridge the knowledge gap between technical experts and policymakers, including (1) Clearly describe the domain limitations of NLP tools; (2) Increase development of non-English training resources; (3) Provide more detail and context for accuracy measures; and (4) Publish more information about definitions and instructions provided to annotators.

Keywords: NLP, content moderation

* Full paper: https://cdt.org/insight/mixed-messages-for-fat-conference-2018/