Mixed Messages? The Limits of Automated Social Media Content Analysis[1]

Presented at the 2018 Conference on Fairness, Accountability, and Transparency

Natasha Duarte
Emma Llansó
Center for Democracy & Technology
Anna Loup
University of Southern California

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. Today's 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 NLP 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) Natural language processing tools perform best when they are trained and applied in specific domains, and cannot necessarily be applied with the same reliability across different contexts; (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 and intercoder reliability challenges documented in NLP studies warn against widespread application of the tools for consequential decision-making; and (5) Text filters remain easy to evade and fall far short of humans' ability to parse meaning from text. Human

[1] The authors would like to thank Miranda Bogen, Robyn Caplan, Nick Feamster, John Grant, Joseph Lorenzo Hall, Marti Hearst, Brendan O'Connor, Aaron Rieke, and Katherine Stasaski, for their expertise and thoughtful feedback on earlier drafts of this paper.

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.

Introduction

For purposes ranging from hate speech detection to law enforcement investigations, governments and companies want to make sense of what people post on social media.[2] Policymakers routinely call for social media companies to identify and take down hate speech, terrorist propaganda, harassment, "fake news" or disinformation, and other forms of problematic speech.[3] In September 2017 UK Prime Minister Theresa May urged companies to detect and remove all "extremist" content within two hours of it being posted.[4] Other policy proposals have focused on mining social media to inform law enforcement and immigration decisions. The U.S. Department of Homeland Security (DHS) is seeking a contract to build an "extreme vetting" system that would analyze social media posts to predict whether individuals will become "positively contributing member[s] of society" and whether a person "intends to commit criminal or terrorist acts after entering the United States."[5]

These policy proposals can be critiqued on a number of grounds. They often burden fundamental rights and rely on untested assumptions about the efficacy of taking down content or reviewing a person's social media posts. These proposals are also technically infeasible. They wrongly assume that automated technology can accomplish on a large scale the kind of nuanced analysis that humans can accomplish on a small scale. Research and machine learning have helped automated text analysis evolve beyond clunky keyword filters over the past few decades. However, natural language processing (NLP) tools still have major limitations when it comes to parsing the nuanced meaning of human communication, much less detecting the intent or motivation of the speaker. Policymakers must understand these limitations before endorsing or adopting automated NLP tools.

A knowledge gap exists between the data scientists studying NLP and the policy makers advocating for wide adoption of automated content moderation. This gap must be bridged lest governments place unwarranted faith and investment into large-scale automated filtering or social media surveillance that results in overbroad censorship, chilling of speech and association, and disparate impacts for minority communities and non-English speakers. To that end, this paper draws on existing NLP research to explain the capabilities and limitations of these tools for analyzing and classifying text posted on social media platforms. We identify five limitations of NLP that policymakers must understand. We conclude with

---

[2] See, e.g., Heather Stewart, May Calls on Internet Firms to Remove Extremist Content Within Two Hours, The Guardian (Sept. 19, 2017), https://www.theguardian.com/uk-news/2017/sep/19/theresa may-will-tell-internet-firms-to-tackle-extremist-content; Kenneth P. Vogel & Cecilia Kang, Senators Demand Online Ad Disclosures as Tech Lobby Mobilizes, N.Y. Times (Oct. 19, 2017), https://www. nytimes.com/2017/10/19/us/politics/facebook-google-russia-meddling-disclosure.html.

[3] *See* sources cited *supra* note 2.

[4] *See* Stewart, *supra* note 2.

[5] See Immigration & Customs Enforcement Homeland Security Investigations, ICE-HSI Data Analysis Service: Solicitation Number HSCEMD-17-R-0010, FedBizOpps.Gov, June 12, 2017; ICE-HSI, Extreme Vetting Initiative: STATEMENT OF OBJECTIVES (SOO), June 12, 2017, FedBizOpps.Gov; ICE-HSI, Background, June 12, 2017, FedBizOpps.Gov.

recommendations for technical experts to help ensure that policymakers understand these limitations and do not blindly rely on NLP as a solution to complex and nuanced social problems.

I. NLP classifiers for automated social media content analysis

Automated content filtering is not new. Many tools have been developed over the years to identify and filter content, including keyword filters, spam detection tools, and hash matching algorithms.[6] These tools filter web traffic and content based on the existence of certain pre-established keywords, metadata, or patterns. For example, hash-matching algorithms have been used to detect images associated with copyrighted material or child pornography.[7] They identify images by a unique code – a sort of "fingerprint" for a given image – called a hash, and compare them against the hash of known copyrighted or child pornography images.[8] Images identical to a known copyrighted or illegal image can then be automatically flagged or filtered out. Early spam filtering methods used the appearance and frequency of certain words in known spam emails to predict the likelihood that an email was spam.[9] These methods rely on previously seen patterns, files, or keywords to identify unwanted content. These relatively simple tools can be effective at identifying content that contains a known keyword or image or matches a known hash or metadata pattern. But they are not capable of parsing the meaning or context of text, such as whether it contains hate speech or terrorist propaganda, is a lawful use of a copyrighted work, or reveals criminal intent.[10] For these tasks, researchers and industry have begun to turn to machine-learning natural language processing (NLP) tools.

Natural language processing (NLP) refers to a set of techniques for using computers to parse text. For the NLP tools described in this paper, the goal of this parsing is usually to predict something about the meaning of the text, such as whether it expresses a positive or negative opinion. Businesses and government entities can purchase off-the-shelf NLP tools designed for a range of purposes, such as determining how consumers feel about a product or brand, translating text, filtering offensive content, and improving spam detection. For example, one company's NLP tool for employers promises to "uncover not only what employees are saying, but also how they feel about topics such as work environment and leadership," and to provide "real-time actionable insights and analysis for improving employee satisfaction and retention."[11]

For this paper, we reviewed existing research documenting the creation and testing of NLP classifiers for tasks such as identifying hate speech, performing sentiment analysis (classifying comments as positive, negative, or neutral), and identifying speech associated with "radicalization" or "extremism."

---

[6] *See, e.g.*, Evan Engstrom & Nick Feamster, The Limits of Filtering: A Look at the Functionality and Shortcomings of Content Detection Tools, Engine (March 2017), http://www.engine.is/the-limits-of-filtering/; Microsoft PhotoDNA, https://news.microsoft.com/download/presskits/photodna/docs/photoDNAFS.pdf.
[7] *See* sources cited *supra* note 6.
[8] *See* Engstron & Feamster, *supra* note 6.
[9] Pieter Arntz, *Explained: Bayesian Spam Filtering*, Malwarebytes Labs (Feb. 17, 2017), https://blog.malwarebytes.com/security-world/2017/02/explained-bayesian-spam-filtering/.
[10] *See, e.g.*, Julia Reda, When Filters Fail, https://juliareda.eu/2017/09/when-filters-fail/.
[11] Ultimate Software, Ultipro HCM Features, https://www.ultimatesoftware.com/UltiPro-Solution-Features-Employee-Surveys.

Most of today's NLP classifiers are trained using examples of text labelled by humans as either belonging to or not belonging to a targeted category of content (e.g. hate speech vs. not hate speech). Each example of text (e.g. each tweet, Facebook post, or email) is called a document, and a collection of documents used to train a classifier is called a corpus (plural: corpora). When building a spam detection tool, for example, one would use a corpus containing both spam and non-spam messages. The spam messages would be annotated as such, so that the model could begin to learn the features and patterns associated with spam and the features that distinguish spam from non-spam. Training corpora are often annotated by human coders, sometimes using crowdsourcing services such as CrowdFlower or Amazon Mechanical Turk. Researchers or developers typically provide definitions for the targeted content (e.g. hate speech, spam, "toxic" comments)[12] or other instructions for annotating the text.

Training corpora are pre-processed to numerically represent features in the text. Features can range from the simple, such as the presence or absence of an individual word, to the complex, such as word embeddings that represent the context in which words appear in a document. Machine-learning tools such as Google's Word2Vec are often used to create word embeddings.[13] The labelled and pre-processed text is used to train machine-learning models to classify new documents. The classifier is tested on a set-aside portion of the training data to determine how closely its classifications matched the human coders.[14]

The advantage and appeal of NLP tools is their ability to process and classify text on a much larger scale than humans. Their scalability makes them attractive solutions to problems like filtering objectionable content from social media platforms. However, when it comes to discerning the subjective meaning and intent behind text, computers still cannot approach humans' language sensitivity and understanding.[15] Even humans struggle with text analysis—for example, with discerning the line between political activism and calls for violence—and NLP is far behind humans in this respect. The next part of this paper discusses the limitations of NLP that make it a problematic solution for automated social media content analysis.

II. Five limitations of NLP tools for social media analysis

1. Natural language processing tools perform best when they are trained and applied in specific domains, and cannot necessarily be applied with the same reliability across different contexts.

[12] *See* Jigsaw, Perspective, https://jigsaw.google.com/projects/#perspective.
[13] *See* Google Code, word2vec, https://code.google.com/archive/p/word2vec/ (Word2vec is a Google open source project); TensorFlow, Vector Representations of Words, https://www.tensorflow.org/tutorials/word2vec.
[14] See, e.g., Microsoft TechNet, Training and Testing Data Sets (2012), https://technet.microsoft.com/en-us/library/bb895173(v=sql.110).aspx.
[15] See Will Knight, *AI's Language Problem*, MIT Tech. Rev. (2016), https://www.technologyreview.com/s/602094/ais-language-problem/.

Language use can vary considerably across social media platforms,[16] demographic groups,[17] and topics of conversation.[18] The NLP tools described in this paper are most effective when they are trained on examples from the same "domain" or context (e.g. posts on the same platform, in the same forum, after a particular event, about a common subject) as the text that the tool will ultimately analyze.[19] Stand-alone tools may be appealing to government agencies and smaller companies that do not have the resources to build or train their own domain-specific tools. However, this appeal is based on the premise that one tool can be reliably applied to multiple domains, and that premise is contradicted by the NLP research.

Ahmed Abbasi, Ammar Hassan, and Milan Dhar demonstrated the importance of domain specificity when they compared fifteen "stand-alone" sentiment analysis tools to five "workbench" tools.[20] Stand-alone tools are those that can be purchased and applied "out of the box" to any data. Workbench tools must be trained using a labeled corpus. In testing tools on publicly available posts on Twitter, Abbasi et al. found that the workbench tools provided higher average accuracy rates: between 67% and 71%, compared to a 56% median average accuracy for the standalone tools. Because the workbench tools were trained on data sets similar to the text they were evaluating, they were able to incorporate domain-specific knowledge.[21]

In the hate speech detection literature, many of the documented tools are trained on samples of text that represent a particular "subtype" of hate speech. Because hate speech is relatively rare as compared to the total volume of social media posts, a random sample of social media posts must be very large to include enough examples of hate speech to train a model.[22] Creating large enough random samples is difficult and expensive.[23] Researchers have avoided this problem by first filtering social media posts with search terms or hashtags thought to be associated with above-average levels of hate speech (e.g., "Islam terror,"

---

[16] *See* Bermingham et al., Combining Social Network Analysis and Sentiment Analysis to Explore the Potential for Online Radicalisation at 3, Proceedings of the International Conference on Advances in Social Network Analysis and Mining (2009) ("Often, when a YouTube user expresses an opinion they simply state it rather than qualifying it with 'I think…' or 'I feel…'. This behaviour is not seen in the blog corpus where authors are keen to distinguish opinion from fact in their posts.)."

[17] *See* Su Lin Blodgett & Brendan O'Connor, Racial Disparity in Natural Language Processing: A Case Study of Social Media African-American English at 1-2, Proceedings of the Fairness, Accountability, and Transparency in Machine Learning Conference (2017), https://arxiv.org/pdf/1707.00061.pdf.

[18] *See* Ahmed Abbasi, Ammar Hassan & Milan Dhar, Benchmarking Twitter Sentiment Analysis Tools, Proceedings of the 9th Language Resources and Evaluation Conference (2014) (finding that "workbench" sentiment analysis tools trained on tweets about specific categories of products and services (e.g., telecommunications) performed better for classifying new tweets about those products or services than uncustomized "stand-alone" tools).

[19] *Id.*

[20] *Id.*

[21] *Id. See also* Leon Derczynski et al., Twitter Part-of-Speech Tagging for All: Overcoming Sparse and Noisy Data, Proceedings of Recent Advances in Natural Language Processing 198 (finding that a part-of-speech tagging system trained on Twitter-specific data performed 26.8% better on token tagging and 12.2% better on sentence tagging than POS tagging systems trained on non-Twitter data).

[22] Anna Schmidt & Michael Wiegand, A Survey on Hate Speech Detection Using Natural Language Processing at 7, Proceedings of the 5th International Workshop on Natural Language Processing for Social Media (2017) ("there are much fewer hateful than benign comments present in randomly sampled data, and therefore a large number of comments have to be annotated to find a considerable number of hate speech instances. This skewed distribution makes it generally difficult and costly to build a corpus that is balanced with respect to hateful and harmless comments.").

[23] *Id.*

"feminazi"), and then drawing their sample from these search results. However, this method tends to disproportionately surface particular subtypes of hate speech. The language used in anti-Muslim hate speech is different from the language used in hate speech against women, black Americans, or LGBTQ people.[24] A classifier trained on a corpus that over-represents a particular subtype of hate speech may be less effective at detecting other subtypes of hate speech.

For example, Pete Burnap and Matthew Williams trained a classifier to detect hate speech using Twitter posts from a two-week period following the murder of Fusilier Lee Rigby, a British Army soldier killed in a terrorist attack that sparked anti-Muslim sentiment.[25] The tweets were compiled by searching the hashtag associated with the attack. The objective of the study was to train a classifier that could help law enforcement find hate speech in the aftermath of an event, which could signal the potential for violence. Burnap and Williams warned that "variance in the way people respond to such [terrorist] events" may affect their tool's ability to identify hate speech in other contexts.[26] Because the model was trained on tweets about a terrorist attack by "Islamic extremists," it is likely that anti-Islamic hate speech was overrepresented in the corpus compared to hate speech against other groups. This might explain some of the study's results: Burnap and Williams found that "hateful terms alone" were almost as predictive of hate speech as hateful terms combined with n-grams.[27] The same hateful terms, or slurs, are likely to reappear frequently within hate speech directed at the same group.

2. Decisions based on automated social media content analysis risk further marginalizing and disproportionately censoring groups that already face discrimination.

a. Natural language processing tools can amplify social bias reflected in language.

As with other machine learning applications, bias in a training corpus will be reflected, and likely amplified, by the resulting classifier. For example, Bolukbasi et al. found that, when trained on articles from Google News, word2vec's word embeddings "exhibit[ed] female/male gender stereotypes to a disturbing extent."[28] When asked, "man is to doctor as woman is to ____," word2vec predicted "nurse." And when asked, "man is to computer programmer as woman is to ___," it predicted "homemaker." The researchers were able to manually correct for these gender biases, but they warned that "the blind application of machine learning runs the risk of amplifying biases present in data."[29] Bias in training data can actually be amplified by the resulting model. For example, Zhao et al.'s study using machine learning to label images found that, while the activity of cooking was about 33% more likely to be associated with

---

[24] *See* William Warner & Julia Hirschberg, Detecting Hate Speech on the World Wide Web, Proceedings of the Second Workshop on Language in Social Media (LSM) 19 (2012), https://dl.acm.org/citation.cfm?id=2390377.
[25] *See* Pete Burnap & Matthew L. Williams, Cyber Hate Speech on Twitter: An Application of Machine Classification and Statistical Modeling for Policy and Decision Making, 7 Policy & Internet 223, 225–26 (2015).
[26] *Id.* at 236.
[27] *Id.*
[28] Tolga Bolukbasi et al., *Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings*, Proceedings of the 30th Conference on Neural Information Processing Systems (NIPS) (2016), https://papers.nips.cc/paper/6228-man-is-to-computer programmer-as-woman-is-to-homemaker-debiasing-word-embeddings.pdf.
[29] *Id.*

females than males in the training corpus, the resulting model associated cooking with females 68% of the time.[30]

This type of bias could lead to content moderation decisions that disproportionately censor certain groups, such as marginalized groups and those with minority views. ProPublica created a tool using word2Vec trained on different "media diets" (left-wing, right-wing, mainstream, digital, tabloid, and ProPublica). When a word is input, each of the six algorithms produces a list of words it estimates are related to that word. ProPublica's tool highlights which of these results are unique to each "media diet".[31] When prompted with the word "abortion", the tool trained on the right-wing media corpus uniquely identified the word "infanticide" as related, while the tool trained on the left-wing media diet uniquely identified the word "anti-choice" as a match.[32] For the term "imma," frequently used in African American Vernacular English (AAVE),[33] only the algorithm trained on a digital media diet recognized the word and produced results. The outputs for "imma" were mostly offensive words that would likely be associated with hate speech or threats, even though "imma" simply means "I'm going to." As the next sections will discuss further, dialects that are underrepresented in mainstream text are more likely to be misinterpreted by algorithms trained on mainstream corpora.

b. Many documented and commercially available natural language processing tools are only effective for English-language text.

Most available NLP tools can only parse English text. As Julia Hirschberg and Christopher D. Manning have pointed out:

> A major limitation of NLP today is the fact that most NLP resources and systems are available only for high-resource languages (HRLs) such as English, French, Spanish, German, and Chinese. In contrast, many low-resource languages (LRLs)—such as Bengali, Indonesian, Punjabi, Cebuano, and Swahili—spoken and written by millions of people have no such resources or systems available.[34]

---

[30] Jieyo Zhao et al., *Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints*, Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP) (2017), https://arxiv.org/pdf/1707.09457.

[31] Jeff Larson, Julia Angwin & Terry Parris Jr., *Breaking the Black Box, How Machines Learn to Be Racist*, Episode 4, Artificial Intelligence, ProPublica (Oct. 19, 2016), https://www.propublica.org/article/breaking-the-black-box-how-machines-learn-to-be-racist?word=Trump.

[32] *Id.*

[33] Other spelling variants include "I'ma" and "I'mma," and "Ima." See, e.g., Jack Sidnell, Outline of African American Vernacular English (AAVE) Grammar (2002), https://cdt.org/files/2017/11/Outline_of_AAVE_grammar___Jack_Sidnell_2002_1_Afr.pdf.

[34] Julia Hirschberg & Christopher D. Manning, *Advances in Natural Language Processing*, 349 Science 261, 261 (July 17, 2015), https://cs224d.stanford.edu/papers/advances.pdf. See also Fredrik Johansson, Lisa Kaati & Magnus Sahlgren, *Detecting Linguistic Markers of Violent Extremism in Online Environments*, in Combating Violent Extremism and Radicalization in the Digital Era, 374–90 (2016), https://www.foi.se/download/18.3bca00611589ae7987820d/1480076542059/ FOI-S--5452--SE.pdf. ("[A]dequate training data and lexical resources are not in abundance for languages other than English."); Schmidt & Wiegand, *supra* note 22 (internal citations omitted) ("With the exception of Dutch and German, we are not aware of any significant research being done on hate speech detection other than on English language data."). The lack of

In fact, it is common for researchers to discard non-English text from a corpus before using it to train a classifier.[35]

Tools that have lower accuracy when parsing non-English text can lead to disproportionately harmful outcomes for non-English speakers. For example, language translation tools using machine learning tend to have lower accuracy for languages that are not well represented on the internet, since the models have fewer examples of those languages to learn from.[36] This becomes problematic when governments rely on machine-learning translations to make decisions affecting people's rights. A Palestinian man was held and questioned by Israeli police relying on an incorrect machine translation of the man's Facebook post.[37] The post, which in fact said "good morning" in Arabic, was translated to "attack them" in Hebrew; police reportedly did not verify the translation with an Arabic speaker before arresting the man.[38] Research into machine-learning translation has made promising strides, but policymakers must understand that these and other NLP tools are not reliable enough to inform high-stakes decisionmaking, especially when the consequences of those decisions are likely to be born disproportionately by groups that are in the minority of online speakers.

c. English-language tools may have disparate accuracy levels for minority populations.

NLP tools also often have trouble with variations in dialect and language usage across demographic and cultural groups of English speakers. Demographic factors such as gender, age, race, ethnicity, and location are associated with different language use patterns.[39] The NLP literature includes several examples of NLP performing less accurately when analyzing the language of female and African-American speakers compared to white male English speakers.[40] Su Lin Blodgett and Brendan

---

resources for languages other than English may be exacerbated online, where a disproportionately high percentage of content is in English and almost all of the content represents only ten languages. *See, e.g.*, Holly Young, *The Digital Language Divide*, Guardian, http://labs.theguardian.com/digital-language-divide/.

[35] Abbasi, Hassan & Dhar, *supra* note 18, at 824; Leandro Silva et al., *Analyzing the Targets of Hate in Online Social Media*, Proceedings of the Tenth International AAAI Conference on Web and Social Media (ICWSM) (2016), https://arxiv.org/pdf/1603.07709.pdf.

[36] *See, e.g.*, An Xiao Mina, *From Digital Divide to Language Divide: Language Inclusion for Asia's Next Billion*, in The Good Life in Asia's Digital 21st Century (2015), https://medium.com/meedan labs/from-digital-divide-to-language-divide-language-inclusion-for-asia-s-next-billion-7792db117844.

[37] Alex Hern, *Facebook Translates 'Good Morning' into 'Attack Them', Leading to Arrest*, Guardian (Oct. 24, 2017), https://www.theguardian.com/technology/2017/oct/24/facebook palestine-israel-translates-good-morning-attack-them-arrest.

[38] *Id.*

[39] Blodgett & O'Connor, *supra* note 17; Dirk Hovy, *Demographic Factors Improve Classification Performance*, Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics & the 7th International Joint Conference on Natural Language Processing, 752–762 (2015), http://aclweb.org/anthology/P15-1073; Dirk Hovy & L. Shannon Spruit, *The Social Impact of Natural Language Processing*, Proceedings of Association for Computational Linguistics (2016); Rachael Tatman, *Gender and Dialect Bias in YouTube's Automatic Captions*, Proceedings of the First Association for Computational Linguistics Workshop on Ethics in Natural Language Processing, 53–59 (2017), http://www. aclweb.org/anthology/W/W17/W17-1606; Maider Lehr, Kyle Gorman, & Izhak Shafran, *Discriminative Pronunciation Modeling for Dialectal Speech Recognition*, Proc. Interspeech (2014).

[40] *See, e.g.*, Blodgett & O'Connor, *supra* note 17; Tatman; supra note 39.

O'Connor found that popular NLP tools tend to misidentify African-American Vernacular English (AAVE) as non-English (one system identified examples of AAVE as Danish with 99.9% confidence).[41] If socioethnic dialects of English are systematically labeled as non-English, NLP algorithms designed to parse English-language statements may overlook those dialects altogether, furthering a cycle of underrepresentation.

Cultural-linguistic bias may be particularly problematic for hate speech detection, since cultural norms play an important role in both how hate is expressed (i.e. the words and phrases used) and whether people perceive something as hate speech. For example, in internal tests, Instagram's DeepText automated hate speech filter incorrectly identified the following sentence as hate speech: "I didn't buy any alcohol this weekend, and only bought 20 fags. Proud that I still have 40 quid tbh."[42] The tool evidently identified "fags" as a slur that marked the statement as hate speech, although the word is also used to refer to cigarettes in colloquial British English and is clearly being used in that sense in the statement.

When platforms or governments adopt automated content analysis tools, the algorithms behind the tools can become the de facto rules for enforcing a web site's terms of service or a country or region's laws. The disparate enforcement of laws or terms of service by biased algorithms that disproportionately censor people of color, women, and other groups raises obvious civil and human rights concerns.

3. Natural language processing tools require 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.

The NLP tools described in this paper are often targeted at content that is hard to define. For example, The U.S. Department of Homeland Security (DHS) has stated its intent to use automation to "evaluate an applicant [for entry into the United States or immigration benefits]'s probability of becoming a positively contributing member of society."[43] This language comes from an executive order of the president, but neither the White House nor DHS has defined what this standard means or how it might be evaluated based on an individual's social media posts.[44]

Among studies evaluating NLP tools for identifying hate speech, there is little agreement on what actually constitutes hate speech. For the most part, each study we reviewed used a slightly different definition.[45]

---

[41] Blodgett & O'Connor, *supra* note 17.

[42] Nicholas Thompson, *Instagram Unleashes an AI System to Blast Away Nasty Comments*, Wired (June 29, 2017), https://www.wired.com/story/instagram-launches-ai-system-to-blast-nasty-comments/. The example 25 Mixed Messages? comment uses a British slang term for cigarettes; this word is also a derogatory term in American English for gay men.

[43] *See* sources cited *supra* note 5.

[44] See Associated Press, *Federal 'Extreme Vetting' Plan Castigated by Tech Experts*, N.Y. Times (Nov. 16, 2017), https://www.nytimes.com/aponline/2017/11/16/us/ap-us-extreme-vetting-artificial intelligence.html?_r=2&mtrref=www.google.com.

[45] *See, e.g.*, Thomas Davidson et al., *Automated Hate Speech Detection and the Problem of Offensive Language*, Proceedings of the International AAAI Conference on Web and Social Media (ICWSM) (2017), https://arxiv.org/pdf/1703.04009.pdf. A survey on hate speech detection by Anna Schmidt and Michael Wiegand referred to hate speech as "a broad umbrella term for numerous kinds of insulting user-created content . . . ."

As far as international standards around illegal hate speech, the International Covenant on Civil and Political Rights (ICCPR) requires that "Any advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence shall be prohibited by law,"[46] but there is not a uniform interpretation of this standard in national laws.[47] In practice, social media platforms' terms of service set the definitions of objectionable content that are applied across their global user base.

Translating an abstract definition into a clearer and more concrete one can make annotation easier, but doing so comes with its own risks. Tools that rely on narrow definitions will miss some of the targeted speech, may be easier to evade, and may be more likely to disproportionately target one or more subtypes of the targeted speech. Some research on using NLP to identify "extremism" or "radicalism" has tried to translate these abstract concepts into components that can be more readily observed in text. Cohen et al. and Johansson et al. have proposed addressing radicalism by using NLP to detect "warning behaviors",[48] behaviors that are said to precede acts of targeted violence.[49] The theory of warning behaviors already oversimplifies a highly complex and difficult-to-predict phenomenon. But the NLP studies simplify this concept even further by focusing on only three warning behaviors that may be more easily identified in text: leakage, the communication of intent to harm a third party; fixation, the increasing preoccupation with a person or cause; and identification, the association of one's self with military, weapons, attackers, etc.[50]

Evidence of "extremism" or "radicalization" is often difficult even for humans to distinguish from other types of speech, such as political activism and news reporting. Furthermore, "extremism" or "terrorist propaganda" is often a moving and subjective target, as new groups can be added to different countries' terrorist watch lists, extreme views can become more mainstream, and vice versa. Definitions that oversimplify an already messy category of speech only exacerbate the problem of effectively detecting it; policy efforts that rely on precise and comprehensive detection of poorly defined categories of speech are not likely to be successful.

4. The accuracy and intercoder reliability challenges documented in NLP studies warn against widespread application of the tools for consequential decision-making.

Schmidt & Wiegand, *supra* note 22. Another study defined hate speech as "a particular form of offensive language that makes use of stereotypes to express an ideology of hate." John T. Nockleby, Hate Speech, in Leonard W. Levy, Kenneth L. Karst & Dennis J. Mahoney, eds., Encyclopedia of the American Constitution, 1277–1279 (2000). Yet another defined it as "any offense motivated, in whole or in a part, by the offender's bias against an aspect of a group of people." Silva et al., *supra* note 35.

[46] International Covenant on Civil and Political Rights ("ICCPR"), G.A. Res. 2200A, Art. 20(2).

[47] Article 19, 'Hate Speech' Explained, A Toolkit (2015), https://www.article19.org/data/files/medialibrary/38231/Hate_speech_report-ID-files--final.pdf. The United States, for example, restricts only speech that is intended to incite imminent violence.

[48] Katie Cohen et al., *Detecting LInguistic Markers for Radical Violence in Social Media*, 26 Terrorism and Political Violence 246–256 (2014), https://www.foi.se/download/18.3bca00611589ae798781dd/1480076522235/FOI-S--4619--SE.pdf; Johansson et al., *supra* note 33.

[49] See J. Reid Meloy, Identifying Warning Behaviors of the Individual Terrorist (April 20, 2016), https://command.columbusstate.edu/readingassignments/auxiliaryreadinglists/FBI-Perspective-Identifying Warning-Behaviors-of-the-Individual-Terrorist.pdf.

[50] Cohen et al., *supra* note 48; Johansson et al., *supra* note 34.

In most studies documenting machine learning classifiers, researchers report their results in terms of "accuracy." However, it is important for policymakers to understand that researchers may define and calculate accuracy in different ways depending on their objectives. In the NLP studies we reviewed, accuracy typically meant how close the classifier came to matching the human coders' results. In other words, a tool that identified hate speech with 80% accuracy would make the same decision as the human coders 80% of the time.[51] This suggests that the goal of NLP is to process speech in the same way that the majority of humans (as represented by the coders) would. This may make sense if the goal is to translate text from one language to another for humans to understand, or to take down content that most social media users would find objectionable. However, for many policy questions or potential applications of NLP tools, the majoritarian view about the likely meaning of a statement is not necessarily the most salient analysis. For example, the fact that a majority of reviewers would consider a particular statement "terrorist propaganda" does not necessarily indicate that the person who uttered the statement has an intent to commit an act of terrorism.

Moreover, human judgment of language can be informed by personal and cultural bias; testing for intercoder reliability may help mitigate this bias in some cases, but it will not necessarily negate the impact of majoritarian views about what is "hateful" or "toxic," which may be shared by the majority of coders. Thus, use of automated content analysis tools in more complex decisionmaking likely warrants different (and more robust) validation methods than the standard measure of "accuracy."

Among the NLP studies we reviewed, the highest accuracy rates reported hovered around the 70% to 80% range.[52] These accuracy rates were typically achieved under ideal conditions: models carefully trained using domain-specific examples labeled by humans and at least reviewed by the researchers. While this level of accuracy often represents impressive advancement in NLP research, it should also serve as a strong caution to anyone considering the use of NLP tools in a decision-making process. An accuracy rate of 80% means that one out of every five people is treated "wrong" in such decision-making; depending on the process, this would have obvious consequences for civil liberties and human rights. Even an accuracy rate of 99% will lead to a high volume of erroneous decisions when applied at scale. For example, Facebook receives approximately 1 million notifications of content that allegedly violates its Community Guidelines every day.[53] A 99% accuracy rate in their content moderation decisions would mean that as many as 10,000 posts or accounts were erroneously taken down (or left online) every day.

---

[51] Some studies use a "gold standard corpus" to calculate accuracy. A gold standard corpus is usually one that is annotated by experts or one for which the researchers verify the labels that the majority of coders assigned to each document. See Warner & Hirschberg, *supra* note 24, at 22.

[52] *See, e.g.*, Mark Clieliebak et al., *A Twitter Corpus and Benchmark Resources for German Sentiment Analysis*, Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media 45–51 (2017); Abbasi, Hassan & Dhar, *supra* note 18; Njagi Dennis Gitari et al., *A Lexicon-based Approach for Hate Speech Detection*, 10 Int'l J. Multimedia & Ubiquitous Engineering 215–30 (2015); Nemanja Djuric et al., *Hate Speech Detection with Comment Embeddings*, Proceedings of the 24th International Conference on World Wide Web 29-30 (2015), http://www.www2015.it/documents/proceedings/companion/p29.pdf; Irene Kwok & Yuzhou Wang, *Locate the Hate: Detecting Tweets Against Blacks*, Proceedings of the 27th AAAI Conference on Artificial Intelligence (2013), https://pdfs.semanticscholar.org/db55/11e90b2f4d650067ebf934294617eff81eca.pdf.

[53] See Sarah Ashley O'Brian, *Facebook gets 1 million user violation reports a day*, CNNTech (March 12, 2016), http://money. cnn.com/2016/03/12/technology/sxsw-2016-facebook-online-harassment/index.html.

Some studies also struggled to achieve acceptable agreement among coders annotating the training corpus, suggesting that people have a hard time agreeing on whether a social media post falls into an objectionable category such as hate speech or extremism.69[54] Ross et al. observed very low agreement between coders' annotations of text as hate speech. Ross et al. concluded that identifying hate speech should not be a binary yes or no question and suggested that people's cultural backgrounds and personal sensibilities play a significant role in whether they perceive content as hate speech.[55] Schmidt and Wiegand have pointed out that there are very few details in the hate speech detection literature about how texts have been annotated, which makes it difficult to evaluate how error or bias may be occurring.[56]

Overall accuracy is not the only important measure for evaluating automated content filtering tools. The ratio and distribution of false positives to false negatives are just as important. A tool may have a high accuracy rate but an unacceptable false positive rate (meaning it too often filters out benign speech). A tool may also have higher error rates for certain groups of speakers—for example, those using socioethnic dialects.

Some NLP studies analyzing social media content assume the general rule that false negatives and false positives should be balanced (the rate of each type of error should be close to equal).[57] However, this assumption ignores the particular stakes of decisions that affect a person's human rights, liberty interests, or access to benefits. For example, when enforcing a limitation on the freedom of expression, the state must demonstrate that the limitation is necessary and achieves a legitimate aim; the presumption or default is against censorship.[58] In any content moderation process, these values would dictate having a higher false negative rate—erring on the side of leaving speech posted—and a lower false positive rate. When fundamental rights such as free expression are at stake, people who develop and use NLP tools cannot default to general rules about distributing error without considering the consequences. In decisions made in the criminal justice or immigration contexts, the question of whether a person is exposed to false-positive or false-negative error could mark the difference between life and death.

---

[54] *See* Björn Ross et al., *Measuring the Reliability of Hate Speech Annotations: The Case of the European Refugee Crisis*, Proceedings of the Third Workshop on Natural Language Processing for Computer Mediated Communication 6, 8 (2016), https://www.linguistics.rub.de/bla/nlp4cmc2016/ross.pdf (finding that overall agreement among coders labeling examples as hate speech or not hate speech was very low. Measured using Krippendorff's alpha (a standard measurement for intercoder reliability), agreement was between = 0.18 and = 0.29, where Krippendorff recommends a minimum of = 0.80, or a minimum of 0.60 "for applications where some uncertainty is un-problematic"); Davidson et al., *supra* note 45 (reporting a 92% intercoder agreement score provided by the online coding platform CrowdFlower, but noting that, of the 5% of tweets that were labelled as hate speech by a majority of coders, only 1.3% were unanimously labelled as hate speech, "demonstrating the imprecision of the Hatebase lexicon." Tweets that did not receive agreement from a majority of coders were not used to train the classifier).

[55] Ross et al., *supra* note 54, at 9.

[56] Schmidt & Wiegand, *supra* note 22. But see Warner & Hirschberg, *supra* note 24, at 21–22.

[57] *See, e.g.*, Burnap & Williams, *supra* note 25, at 235.

[58] *See, e.g.*, ICCPR General Comment No. 34, ¶ 35 (describing the limits on a State's ability to restrict freedom of expression under the ICCPR). *See generally, e.g.,* Near v. Minnesota, 283 U.S. 697 (1931).

5. NLP filters remain easy to evade and fall far short of humans' ability to parse meaning from text.

Today's NLP tools can do more than their predecessor keyword filters, but their ability to parse language falls far short of many policymakers' expectations. The meaning of language is highly dependent on contextual elements such as tone, speaker, audience, and forum. Abbasi et al.'s study testing sentiment analysis tools found that the most common errors involved things like jokes, sarcasm, and literary devices.[59] NLP tools that cannot reliably distinguish jokes and sarcasm from serious statements are particularly ill-suited to the task of filtering social media posts for dangerous content such as threats or terrorist propaganda. Often, context and minor semantic differences separate hate speech from benign speech. For example, the term "slant" is a slur often used to insult the appearance of people of Asian descent, but "The Slants" is an Asian American band whose members chose the name in part "to undercut slurs about Asian Americans that band members heard in childhood."[60]

Because they rely on previously seen features in text, NLP filtering tools are also easy to evade. As social media companies have begun to accelerate their efforts to monitor and take down hate speech, speakers are coming up with new ways to communicate hate against target groups while avoiding detection. For example, triple parentheses have been used on Twitter to indicate in a derogatory way that someone is Jewish.[61] White supremacists have also used innocuous terms, including the names of companies ("Google," "Skype," and "Yahoo") as stand-ins for racial and ethnic slurs.[62] Even if content moderation tools adapt to learn these patterns (a process that would require the accumulation of a significant amount of the novel derogatory uses of the term), users seeking to convey hateful messages could quickly adapt and begin using different novel terms and phrases. Human review of flagged content (whether flagged by users or by automated tools) remains essential for avoiding over-censorship and catching nuances in language use that a classifier might miss.

Some studies have suggested that considering information beyond the text, such as demographic information about the speaker, can improve NLP accuracy for hate speech detection.[63] Schmidt and Wiegand theorized that:

---

[59] Abbasi, Hassan & Dhar, *supra* note 18.
[60] Richard Sandomir, *Ruling Could Help Washington Redskins in Trademark Case*, N.Y. Times (Dec. 22, 2015), https://www.nytimes.com/2015/12/23/sports/football/washington-redskins-trademark nickname-offensive-court-ruling.html.
[61] Nikhil Sonnad, *Alt-right Trolls are Using These Code Words for Racial Slurs Online*, Quartz (Oct. 1, 2016), https://qz.com/798305/alt-right-trolls-are-using-googles-yahoos-skittles-and-skypes-as code-words-for-racial-slurs-on-twitter/.
[62] *Id.*
[63] *See, e.g.*, Guang Xiang et al., *Detecting Offensive Tweets Via Topical Feature Discovery Over a Large Scale Twitter Corpus*, Proceedings of the 21st Association for Computing Machinery (ACM) International Conference on Information and Knowledge Management 1980–1984 (2012); Maral Dadvar et al., *Improved Cyberbullying Detection Using Gender Information*, Proceedings of the Twelfth Dutch-Belgian Information Retrieval Workshop (DIR) 22–25 (2012); Maral Dadvar et al., *Improving Cyberbullying Detection with User Context*, Proceedings of the European Conference in Information Retrieval (ECIR) 693–696 (2013).

> Having some background information about the user of a post may be very predictive. A user who is known to write hate speech messages may do so again. A user who is not known to write such messages is unlikely to do so in the future.[64]

Xiang et al. trained an "offensive content" classifier by constructing features from a corpus of "offense-pro twitterers" (users who often used offensive words) and "law abiding twitterers" (users who rarely used offensive words).[65] Instead of annotating randomly selected tweets as offensive or not, Xiang et al. treated all tweets by "offense-pro" users as offensive and all tweets by "law-abiding" users as not offensive.[66] Dadvar et al. (2012) used "gender-specific" features (terms more commonly used by men than women on Myspace) to classify harassing speech by men on Myspace.[67] Dadvar et al. (2013) used user-based features, such as users' message history, to detect cyberbullying.[68]

However, using information about the speaker to adjudicate speech raises additional human rights and censorship concerns. Taking the identity or characteristics of the speaker into account may occasionally make sense, such as when white users direct racial slurs at black users. But incorporating assumptions about certain speakers into automated tools could also result in unfair disparate enforcement of a website's terms of service. Rather than creating tools that reinforce stereotypes in an effort to improve content analysis, policymakers and platform operators should understand the limits of available tools and cabin their use accordingly, such as by maintaining human reviewers as central to the content analysis and moderation process.

III. Recommendations for policymakers

1. As a general matter, natural language processing tools designed to identify hate speech, terrorist propaganda, and other kinds of problematic speech are relatively inaccurate and ineffective. They are prone to both over- and underinclusive results, and their error rates will tend to disproportionately affect already marginalized groups and speakers. **Use of automated content analysis tools to detect or remove illegal content should never be mandated in law.**

2. Regulatory requirements on content intermediaries to comprehensively review user-uploaded content, or to complete reviews of flagged content in short periods of time, are effectively mandates to use automated content analysis tools. As governments, industry, researchers, civil society, and other stakeholders consider policy responses to illegal content online, we must keep in mind that **use of automated content analysis tools carries substantial risks of overbroad censorship that disproportionately affects already marginalized speakers.**

---

[64] Schmidt & Wiegand, *supra* note 22.
[65] Xiang et al., *supra* note 63, at 1980.
[66] Xiang et al. reported a 5.4% improvement in classification using this method, compared to "keyword matching." *Id.* at 1984.
[67] Dadvar et al. (2012), *supra* note 63.
[68] Dadvar et al. (2013), *supra* note 63.

3. As policymakers consider implementing content analysis tools in government processes, it is essential that they keep the significant limitations of these tools in mind. Given the weaknesses of these tools, **government programs must not use automated content analysis tools to make decisions that affect the rights, liberties, or access to benefits of individuals or groups.**

4. **Any use of automated content analysis tools should be accompanied by human review of the output/conclusions of the tool.**

IV. Recommendations for researchers

The NLP research community has great potential to bridge the knowledge gap between technical experts and policy makers so that policy decisions are not based on misunderstandings of or unreasonable expectations for how NLP technology works. The following is a non-exhaustive list actions we recommend researchers take to ensure that they are making responsible representations about NLP tools for social media content

1. Evaluate and clearly acknowledge the domain limitations of documented NLP tools. Where appropriate, test and compare tools' effectiveness across data sets representing different platforms, topics, demographics, and subtypes of speech. This type of evaluation is particularly important for commercial standalone tools that claim to be useful out-of-the box.

2. Describe any training corpora used to develop a classifier, including the sources and sampling methods, any material removed and reasons for removing it, and the individuals and groups whose representation in the data may be disproportionate. Consider any important context or features represented in non-English text, emoticons and other uses of punctuation, and other elements before discarding them.

3. Improve and support the development of NLP tools and resources for non-English languages. Consider retaining text that uses a combination of English and non-English languages to improve classifiers' accuracy for multilingual speakers on social media.

4. Provide more context for accuracy rates and other metrics to help policymakers and advocates evaluate the usefulness of NLP tools and methods.

5. Publish more detailed information about annotation procedures and definitions and/or instructions given to coders. Describe how targeted speech was defined and include any empirical support for definitions.

6. When possible and practical, evaluate tools in settings that represent real world content moderation scenarios. For example, consider attempting to use social media sites' terms of services for definitions and compare results of using a classifier to take down objectionable content with and without human review.