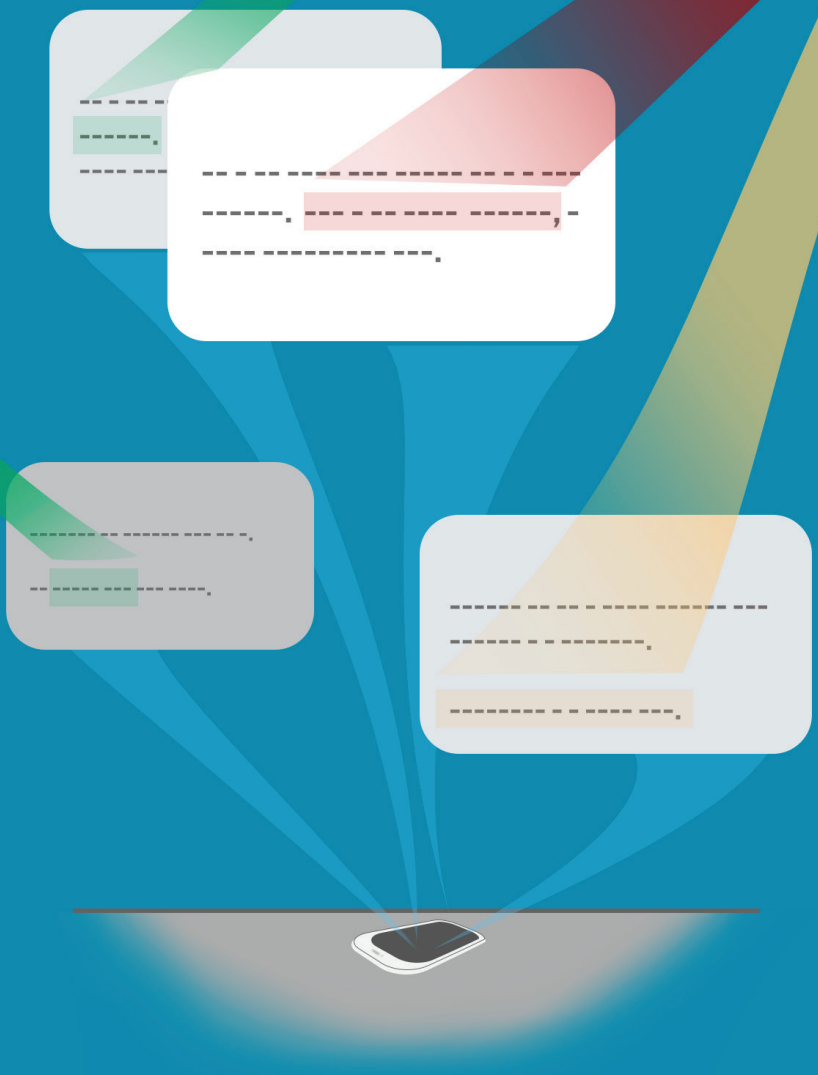
# MIXED MESSAGES?

## THE LIMITS OF AUTOMATED SOCIAL MEDIA CONTENT ANALYSIS



*November 2017*

# Mixed Messages?

*The Limits of Automated Social Media Content Analysis*

*Contributors:*
**Natasha Duarte**, *Policy Analyst*
**Emma Llanso**, *Director, Free Expression Project*
**Anna Loup**, *COMPASS Fellow, University of Southern California*

***Center for Democracy & Technology***

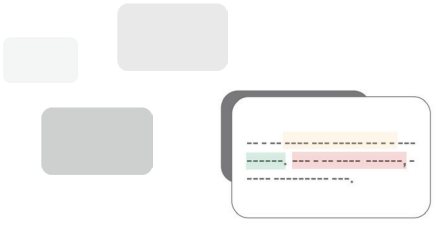
*Designed by Timothy Hoagland*

# EXECUTIVE SUMMARY

Governments and companies are turning to automated tools to make sense of what people post on social media, for everything ranging from hate speech detection to law enforcement investigations. 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. 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 accomplish on a small scale.

Today's tools for automating social media content analysis have limited ability to parse the nuanced meaning of human communication, or to detect the intent or motivation of the speaker. Policymakers must understand these limitations before endorsing or adopting automated content analysis tools. Without proper safeguards, these tools can facilitate overbroad censorship and biased enforcement of laws and of platforms' terms of service.

This paper explains the capabilities and limitations of tools for analyzing the text of social media posts and other online content. It is intended to help policymakers understand and evaluate available tools and the potential consequences of using them to carry out government policies. This paper focuses specifically on the use of natural language processing (NLP) tools for analyzing the text of social media posts. We explain five limitations of these tools that caution against relying on them to decide who gets to speak, who gets admitted into the country, and other critical determinations. This paper concludes with recommendations for policymakers and developers, including a set of questions to guide policymakers' evaluation of available tools.



# Five Key Limitations
## of Automated Social Media Content Analysis Tools

### *I. 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.*

Language use can vary considerably across and within social media platforms, demographic groups, and topics of conversation. The language people use in captions when sharing images of their pets on Instagram has very different characteristics from the language used to discuss major geopolitical events on Facebook. A tool trained to recognize the former cannot be reliably applied to analyze the latter. NLP tools must be trained to recognize the particular type (or "domain") of speech they will be used to analyze; otherwise their performance will suffer. Tools marketed for use "off-the-shelf" on any text, without domain-specific training, should garner skepticism, and policies should not rely on the existence of a one-size-fits-all tool for analyzing social media content.

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

Natural language processing tools can amplify social bias reflected in language. Machine-learning algorithms learn about the world from their training data. Any bias in the text they learn from can be reflected in their outputs if not corrected. Several studies have found, for example, that machine learning models reflect or amplify gender bias in the text used to train them. This type of bias could lead to content moderation decisions that disproportionately censor or misinterpret the speech of certain groups, such as marginalized groups or those with minority views.

Many documented and commercially available natural language processing tools are only effective for English-language text. Reliance on these tools is likely to create disproportionately harmful outcomes for non-English speakers. Non-English text is more likely to be misinterpreted by these tools, possibly creating more unwarranted censorship or suspicion of speakers of languages other than English.

English-language tools may have disparate accuracy levels for minority populations. NLP tools often have trouble with variations in dialect and language use across different demographic and cultural groups of English speakers. Demographic factors such as gender, age, race, ethnicity, and location are associated with different language use patterns. For example, researchers have found that popular NLP tools tend to misidentify African American Vernacular English as non-English (one system identified examples of AAVE as Danish with 99.9% confidence).

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 marginalized groups raises obvious civil and human rights concerns.

### ***III.*** *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.*

To train an automated tool to reliably recognize problematic content, researchers or engineers must have a clear definition of the targeted content. However, the tools described in this paper are often targeted at content that is hard to define. For example, 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. The definitions of targeted content (e.g. hate speech, extremism) used by researchers vary considerably among studies. Policy efforts that rely on precise and comprehensive detection of poorly defined categories of speech are not likely to be successful.

### ***IV.*** *The relatively low accuracy and intercoder reliability achieved in natural language processing studies warn strongly against widespread application of the tools to social media content moderation.*

Among studies using NLP to judge the meaning of text (including hate speech detection and sentiment analysis), the highest accuracy rates reported hover around 80%, with most of the high-performing tools achieving 70 to 75% accuracy. These accuracy rates may represent impressive advancement in NLP research, but they should also serve as a strong caution to anyone considering the use of such 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.

Accuracy itself can be a subjective concept in machine learning. Researchers may define and calculate accuracy in different ways depending on their objectives. In NLP studies, accuracy often refers to how closely a tool came to agreeing with humans' determinations about the content. The use of automated content analysis tools in complex decisionmaking likely warrants different (and more robust) validation methods than the standard measure of "accuracy".

Machine learning tools used to make subjective predictions, such as whether someone will positively contribute to society or is at risk of becoming radicalized, may be impossible to validate. Policymakers adopting these tools would likely be forced to rely on proxies — such as whether human coders would have judged a target's speech as negative toward America — that have limited predictive power.

***V.*** *Even state-of-the-art NLP tools 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 policy makers' expectations. The meaning of language is highly dependent on contextual elements such as tone, speaker, audience, and forum. Because they rely on previously seen features in text, NLP filtering tools are 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. 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.

## Recommendations:

*Use of automated content analysis tools to detect or remove illegal content should never be mandated in law.*

*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 carry substantial risks of overbroad censorship that disproportionately affects already marginalized speakers.*

*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.*

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

## *Questions to Guide Policymakers' Evaluation of Automated Text Analysis Tools*

1. Is this tool intended to be used "out of the box" without customization, or does it allow (or require) you to train or tailor it for a specific use?

2. What are the accuracy rates of the tool in ideal conditions? How does this compare to the accuracy of other tools or methods of identifying the target content?

3. How did the tool developers define the target content? What did the developer do to ensure consistent application of this definition during training and testing?

4. What is the tool's rate of false positive (over-inclusive) and false negative (under-inclusive) errors? What consequence would each type of error have when the tool is used as intended?

5. What data was the tool trained on and where did it come from?

6. How do the speakers represented in the training data compare to the target population in terms of demographics, language use, dialect, subject matter, context, or platform?

7. Is the tool trained to interpret communicative elements such as emoticons, emoji, or GIFs?

8. How will the tool adapt to changes in the target population and its language use over time?

9. How does the tool actually perform in real-world scenarios? How will you test this tool and evaluate its accuracy? What effects does it have on the individuals and groups whose speech you are analyzing?

# 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.[1] 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.[2] In September 2017 UK Prime Minister Theresa May urged companies to detect and remove all "extremist" content within two hours of it being posted.[3] 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."[4]

Policy proposals such as these 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, available 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 content analysis tools. Without proper safeguards, these tools can facilitate overbroad censorship and biased enforcement of laws and of platforms' terms of service.

This paper explains the capabilities and limitations of tools for analyzing the text of social media posts and other online content. It is intended to help policymakers understand and evaluate available tools and the potential consequences of using them to carry out government policies. This paper focuses specifically on the use of natural language processing (NLP) tools for analyzing the text of social media posts, though it is important for policymakers to note that social media analysis often involves a combination of methods for processing text, images, and other types of data. These other methods have their own capabilities and limitations.[5]

*In the first part of this paper,* we describe some of the tools and methods available today for inferring meaning from the text of social media posts. *The second part of the paper* discusses five limitations of these tools that policymakers and developers alike must understand when considering the role these tools may play in social media content analysis and moderation. These limitations include:

**(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;

**(3)** Natural language processing tools require clear, consistent definitions of the type of speech to be identified, and policy debates around content moderation and social media mining tend to lack such precise definitions;

**(4)** The relatively low accuracy and intercoder reliability achieved in natural language processing studies warn strongly against widespread application of the tools to social media content moderation; and

**(5)** Even state-of-the-art natural language processing tools remain easy to evade and fall far short of humans' ability to parse meaning from text.

This paper concludes with recommendations for policymakers and developers, including questions to guide policymakers' evaluation of available tools.

## *I. Tools for automated text 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 algorithms.

### *a. Natural language processing models for social media text*

Natural language processing (NLP) *is a discipline of computer science that focuses on 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]

Today's state-of-the-art NLP tools typically use neural networks[12] to analyze a number

of different features in text and *to classify it as either belonging to or not belonging to some targeted category of speech (e.g. hate speech, terrorist content)*. These tools are known as text classifiers. This section walks through a simplified roadmap of how text classifiers are built and how they work.

### *1. Selecting and annotating a training corpus*

Text 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). From these examples, classifiers identify patterns and learn rules for sorting new, unlabeled examples of the targeted content. *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 linguistic 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)[13] or other instructions for annotating the text. They also set the minimum number of people who must annotate each document and the minimum threshold for intercoder agreement (e.g., 75% of coders must agree that a tweet is hate speech for it to be included in the training corpus).

### *2. Representing features in the text*

Training corpora are pre-processed to numerically represent their features, such as the words, phrases, and grammatical structures that appear in the text. Machine-learning models use these features to learn patterns associated with the targeted content. For example, a spam detection model might learn which words occur more frequently in examples labeled as spam than in non-spam examples. Features can range from the simple (individual words) to the more complex (word embeddings, which are described below). Complexity here refers to how much of the information in a document the single feature can represent. Many NLP tools available for purchase today rely on simple features, such as individual words, phrases, and parts of speech. Newer, state-of-the-art methods rely on more complex "word embeddings" that take into account the entire sentence or document. Below, we describe several common feature representations: "bag of words", n-grams, part-of-speech tagging, and word embeddings.

**Bag of words:** The bag-of-words approach to NLP involves representing a corpus as a collection of the individual words contained in the documents. This approach does not consider the order or combinations in which words occur or how the words relate to one another. For example, the text "Bob called Alice" would be represented as the bag of words "Alice, Bob, called." With this method, a model is trained to classify a document based on how similar the words it contains are to words known to appear in targeted content.

Early spam detection tools used the bag-of-words approach to teach classifiers which words appeared more frequently in known spam emails than in other emails. New emails would then be classified as spam or not spam based on how the words

they contained compared to the spam training corpus. This worked relatively well, as long as spam emails contained words the tools had learned to associate with spam.[14] However, spammers could easily learn to evade detection by avoiding these words. The bag-of-words approach does not account for any other features in the text that help elucidate meaning, such as the order of words in a sentence, the context of the sentence in which the word occurs, or the characteristics of the speaker.

**N-grams:** N-grams are combinations of words occurring together. "Free speech" is an example of a bigram—a two-word n-gram. The bigram "free speech" has a distinct meaning that is not conveyed by either the word "free" or the word "speech" independently. N-grams can provide a more nuanced analysis of the content of text than can a bag-of-words approach.[15] Some words that are considered offensive out of context can have benign meanings when paired with other words. For example, the term "queer" may be associated with hate speech in some contexts, but the n-gram "openly queer" may indicate a benign or positive use of the word.[16] The ability to recognize n-grams can help classifiers avoid the false positives that would come from relying on keywords alone. However, n-gram analysis still does not account for the context in which an n-gram appears or where it appears in a sentence. It also does not account for words that co-occur in a sentence but are far apart.[17]

**Part-of-speech tagging:** Part-of-speech (POS) tagging involves labeling text according to its grammatical function (e.g., noun, verb, preposition). POS tagging is often combined with a bag-of-words or n-gram approach to provide a more nuanced analysis of the content and structure of text. A POS tag can provide insight into a word's function within a sentence and its relationship to the words around it, which can help elucidate meaning.[18]

**Word embeddings:** State-of-the-art NLP tools rely on more complex features called word embeddings. Tools such as word2vec[19] create word embeddings that represent how words in a corpus are related to one another based on the context in which they appear, including their place and function in a document.[20] Word2vec creates a sort of map, where words that appear in similar contexts are mapped as being close together in meaning.[21] For example, given the sentences "I took my cat to the vet" and "I took my dog to the vet," word2vec would represent "cat" and "dog" as semantically similar (or close together). Word embeddings can help computers achieve more nuanced parsing of language. For example, hate speech filters relying on word embeddings may be harder to evade by simply substituting an offensive or hate-related word with a benign one. A classifier trained on word embeddings might understand that the substituted word was serving a similar function because of the context of the sentence in which it appeared.

### *3. Choosing and training a machine-learning model*

Once features are represented in a corpus, they can be used to train a machine-learning classifier (neural networks are often used for this task). Neural networks process labeled training corpora and create internal rules for how much relative weight to assign to each feature. They use those rules to classify new examples of text as belonging to or not belonging to the targeted category of speech.

### *4. Testing and adjusting the model*

The standard way of testing an NLP classifier is to set aside a portion of the labeled training corpus before training the model and to use

that set-aside text to test the model.[22] During this phase, researchers typically analyze errors to find out where the model fails to correctly identify the targeted content and, of those errors, how many are false positives (benign content classified as the targeted content) or false negatives (targeted content not classified as such). At this point, researchers will often attempt to identify and tweak the feature representations or weights in the model that are contributing to errors or unwanted outcomes.

For some tasks, it may be possible to skip all of these steps and purchase an off-the-shelf text classifier that can be applied to any data. However, as the second half of this paper will address, doing so without customizing the model will likely cause performance of the tool to suffer and will lead to less useful or reliable results.

NLP tools can vary in methodology and sophistication; however, they all measure the objective characteristics of text—its words and grammatical structures—sometimes in an attempt to predict its subjective meaning. In some applications of machine learning, models' ability to learn from large data sets can give them an analytical advantage over humans. However, when it comes to discerning the subjective meaning and intent behind text, computers still cannot approach humans' language sensitivity and understanding.[23] Even humans struggle with text analysis—for example, with discerning the line between political activism and calls for violence—and automated tools are far behind humans.[24] 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 automated social media text analysis tools*

### *a. 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.*

The NLP tools described in this paper are most effective when domain-specific. Domain-specific tools are ones that are trained on examples from a specific context (e.g. posts on the same platform, in the same forum, after a particular event, about a common subject) and used to analyze text within the same context or domain.[25] While researchers have had some success with domain-specific models for classifying text,[26] their results do not suggest that the same tools can be reliably applied in different domains, such as on different social networks or responses to different events. Language use can vary considerably across social media platforms,[27] demographic groups,[28] and topics of conversation.[29]

Researchers 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.[30] 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; typically, this corpus is drawn from the same context to which the tool will be applied. 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 for the standalone tools.[31] Because the workbench tools were trained on data sets similar to the text they were evaluating, they were able to incorporate domain-specific knowledge.[32]

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 research.[I]

In research on automated hate speech detection, most of the documented tools are trained (whether intentionally or not) to identify a specific "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.[33] Creating large enough random samples is difficult and expensive.[34] 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," "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.[35] A classifier trained on a corpus that over-represents a particular subtype of hate speech will likely underperform 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.[36] 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.[37] 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.[II] The same hateful terms, or slurs, are likely to reappear frequently within hate speech directed at the same group. The fact that hateful terms were such a strong predictor of hate speech in this study suggests that it may not perform as well for hate speech directed at other groups.

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

I. Research suggests that there may be cheaper & less time-consuming "bootstrapping" methods for building domain-specific models, by training the model on a large, general set of unlabeled data and then a smaller set of human-annotated examples. However, this research does not negate the necessity of using domain-specific examples to train a classifier. See, e.g., Aliaksei Severyn & Alessandro Moschitti, UNITN: Training deep convolutional neural network for Twitter sentiment classification, Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval) 464 (2015), http://alt.qcri.org/semeval2015/cdrom/pdf/SemEval079.pdf.

II. Hateful terms alone had the same "precision" performance (rate of false positives) as hateful terms combined with n-gram typed dependencies, but had lower "recall" performance (a higher rate of false negatives), meaning that the classifier relying on hateful terms missed some hate speech that did not include the hateful terms.

*1. Natural language processing tools can amplify social bias reflected in language.*

Machine-learning algorithms learn about the world from their training data. Any bias in that text can be reflected in their outputs if not corrected. This is what happened to a tool called word2vec, which learned gender bias from Google News articles. As discussed above in Section I, word2vec represents relationships between words as they are commonly used by comparing the contexts in which words appear. The representations, or word embeddings, that word2vec creates can easily represent societal bias reflected in text.[38] A team of researchers found that, when trained on articles from Google News, word2vec's word embeddings "exhibit[ed] female/male gender stereotypes to a disturbing extent."[39] For example, 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."[40] 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 females than males in the training corpus, the resulting model associated cooking with females 68% of the time.[41]

This type of bias could lead to content moderation decisions that disproportionately censor certain groups, such as marginalized groups or those with minority views. For example, 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".[42] 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.[43] For the term "imma," frequently used in African-American Vernacular English (AAVE),[44] 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.

*2. 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)[III] 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.*[45]

III. High-resource languages, such as English, French, and German, are those for which an abundance of resources, such as annotated training corpora, exist, making it easier to train machine-learning models to recognize those languages.

In fact, it is common for researchers to discard non-English text from a corpus before using it to train a classifier.[46] The majority of internet users are non-English speakers.[47] Tools that cannot parse non-English text at all will be ineffective in identifying all examples of the targeted content, and may miss important dynamics of how words and terms are used in bilingual and multilingual communications. Content hosts facing potential liability might be more likely to preemptively block or take down non-English content that their automated tools cannot process.

Tools that have lower accuracy when parsing non-English text—due to a lack of resources in other languages—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.[48] 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.[49] 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.[50] 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 decision making, especially when the consequences of those decisions are likely to be born disproportionately by groups that are in the minority of online speakers.

*3. 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.[51] The NLP literature includes several examples of NLP performing less accurately when analyzing the language of female and African-American Vernacular English (AAVE) speakers compared to white male English speakers.[52] For example, a 2017 study found that YouTube autocaptioning had a higher error rate for captioning female speakers than for male speakers in videos.[53] Researchers have also found that popular NLP tools tend to misidentify AAVE as non-English (one system identified examples of AAVE as Danish with 99.9% confidence).[54] 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."[55] 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.

### *c. 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]'s probability of becoming a positively contributing member of society."[56] 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.[57]

Among studies evaluating NLP tools for identifying hate speech, there is little agreement on what actually constitutes hate speech.[58] William Warner and Julia Hirschberg defined hate speech as "abusive speech targeting specific group characteristics, such as ethnic origin, religion, gender, or sexual orientation."[59] Other definitions include "offensive language" or slurs.[60] As Ross et al. explained in their study of hate speech annotations,

> *These approaches . . . leave plenty of room for personal interpretation, since there may be differences in what is considered offensive. For instance, while the utterance "the refugees will live off our money" is clearly generalising and maybe unfair, it is unclear if this is already hate speech.*[61]

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,"[62] but there is not a uniform interpretation of this standard in national laws.[63] In practice, social media platforms' terms of service set the definitions of objectionable content that are applied across their global user base. While terms of service may have helped inform some researchers' definitions of targeted content,[64] however, we did not find a study that attempted to detect hate speech based on definitions from a specific platform's terms of service.[65] As we discuss in the next section, vague definitions can lead to different interpretations among coders and make comparability between tools harder.

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. Two studies have proposed addressing extremist content by using NLP to detect "warning behaviors",[66] behaviors that are said to precede acts of targeted violence.[67] This theory already simplifies a highly complex and difficult-to-predict phenomenon. But

the NLP studies oversimplify even further and focus 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.[68]

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.

### *d. The relatively low accuracy and intercoder reliability achieved in natural language processing studies warn strongly against widespread application of the tools to social media content moderation.*

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. For example, if a classifier is designed to predict whether students will do well in school, its accuracy might be tested by comparing its results to those students' test scores, their grades after several years, whether they get into college, or how involved they are in school activities. Each of these metrics could be valid for testing the classifier, but each of them could yield a different result and represent different values (for example, an emphasis on grades versus an emphasis on student engagement). Often, the user's (or developer's) goals shape how they measure accuracy.

NLP models learn to identify content based on examples, which are typically labeled by humans. However, some researchers have struggled to achieve acceptable intercoder reliability, indicating 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] In one study that achieved very low agreement between coders' annotations of text as hate speech, the researchers 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.[70] 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.[71]

If intercoder reliability can be achieved, then accuracy typically measures 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.[72] 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 application 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 may also bias the training data toward majoritarian views of what is "hateful" or "toxic" in a way that fails to recognize the wholly legitimate expression of minority voices. 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 studies using NLP to judge the meaning of text (including hate speech detection and sentiment analysis), the highest accuracy rates reported hover around 80%; in many studies, the highest accuracy rates reported were around 70 to 75%.[73] These accuracy rates may represent impressive advancement in NLP research, but they should also serve as a strong caution to anyone considering the use of such 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.[IV]

Moreover, domain-specific hate speech detection tools that achieve relatively high accuracy rates in studies would likely see a drop in accuracy when applied to the diverse, dynamic speech environment of a social media platform in the wild. This is supported by the gap in accuracy that Abassi et al. found between stand-alone and workbench sentiment analysis tools; tools that achieve high accuracy in one context may suffer when exposed to new contexts and ways of speaking.[74] More research is needed to test these tools in environments similar to active social media platforms, with a wider variety of novel speech and communication patterns that change over time.

Most of the tools described in this paper attempt to detect relatively uncommon types of speech[75] or predict rare events, such as terrorist activity.[76] Rare events are inherently difficult to predict.[77] Imagine you wanted to build a model for predicting a person's likelihood of committing terrorist acts in the future based on their social media posts. Because terrorist acts are so rare, if you were to always guess that a given person would not commit terrorism, you would almost always be correct, meaning you would have a very high accuracy rate—easily higher than a predictive models based on social media posts. Similarly, Warner & Hirschberg found that an anti-semitic speech detection tool was 91% accurate when it always guessed that a text was *not* anti-semitic.[78] In other words, it outperformed the highest-performing tools documented in the NLP literature.

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 be more likely to

IV. Even an extremely high 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. *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. 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.

have false positives for certain speakers or types of speech than for others. In the criminal justice context, some risk assessment algorithms have been shown to have higher false positive rates for black people and higher false negative rates for white people.[79] Given the research showing that NLP tools have trouble identifying socioethnic dialects, accuracy rates are likely to be lower for those communities than for the majority.

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).[80] 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.[81] 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.

As a minimum requirement, technology endorsed or adopted by the government should work well. When machine learning predictions are used to carry out public policy, the government must have ways to validate them. However, machine learning tools used to make subjective predictions, such as whether someone will positively contribute to society or be at risk of becoming radicalized, may be impossible to validate. Policymakers adopting these tools would likely be forced to rely on proxies—such as whether human coders would have judged a target's speech as negative toward America—that have limited predictive power.

### *e. State-of-the-art NLP tools 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. NLP is only concerned with the features of the text itself, which cannot give a full picture of its meaning. Abbasi et al.'s study testing sentiment analysis tools found that the most common errors involved things like jokes, sarcasm, and literary devices.[82] 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."[83]

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.[84] Even a sophisticated neural network could be tripped up by novel uses of punctuation and other characters, which are often stripped during text parsing and classification. White supremacists have also used innocuous terms, including the names of companies ("Google," "Skype," and "Yahoo") as stand-ins for racial and ethnic slurs.[85] Even if dynamic content moderation tools eventually adapted to recognize 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.

Researchers have found that considering information beyond the text, such as demographic information about the speaker, can improve NLP accuracy for hate speech detection.[86] Schmidt and Wiegand theorized that:

> *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.*[87]

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).[88] In other words, instead of annotating randomly selected tweets as offensive or not, Xiang et al. automatically treated all tweets by "offense-pro" users as offensive and all tweets by "law-abiding" users as not offensive.[89] 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.[90] Dadvar et al. (2013) used user-based features, such as users' message history, to detect cyberbullying.[91]

However, using information about the speaker to adjudicate speech raises additional human rights and censorship concerns. Incorporating assumptions about speakers into automated content moderation tools may improve accuracy, but it also means that certain speakers' speech will be more likely to be removed because of who they are. 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 and Question Guide for policymakers, users, developers.*

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 under-inclusive 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.**

Moreover, policymakers must understand that 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.**

Further, 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.**

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

*(next page)*

To assist policymakers and others in evaluating the strengths and weaknesses of these tools, we provide the following question guide:[V]

1. Is this tool intended to be used "out of the box" without customization, or does it allow (or require) you to train or tailor it for a specific use?

2. What are the accuracy rates of the tool in ideal conditions? How does this compare to the accuracy of other tools or methods of identifying the target content?

3. How did the tool developers define the target content? What did the developer do to ensure consistent application of this definition during training and testing?

4. What is the tool's rate of false positive (over-inclusive) and false negative (under-inclusive) errors? What consequence would each type of error have when the tool is used as intended?

5. What data was the tool trained on and where did they come from?

6. How do the speakers represented in the training data compare to the target population in terms of demographics, language use, dialect, subject matter, context, or platform?

7. Is the tool trained to interpret communicative elements such as emoticons, emoji, or GIFs?

8. How will the tool adapt to changes in the target population and its language use over time?

9. How does the tool actually perform in real-world scenarios? How will you test this tool and evaluate its accuracy? What effects does it have on the individuals and groups whose speech you are analyzing?

V. This is not intended to be a comprehensive list of questions that must be answered before relying on a tool for automated text analysis; rather, we wish to highlight examples of critical questions to ask.

# ENDNOTES

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.
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. Stewart, *supra* note 2.
4. *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.
5. *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/; Cathleen O'Grady, *Tiny, Blurry Pictures Find the Limits of Computer Image Recognition*, Ars Technica (Feb. 20, 2016), https://arstechnica.com/science/2016/02/tiny-blurry-pictures-find-the-limits-of-computer-image-recognition/.
6. *See* Engstrom & Feamster, *supra* note 5.
7. *See, e.g.,* Microsoft PhotoDNA, https://news.microsoft.com/download/presskits/photodna/docs/photoDNAFS.pdf.
8. *See* sources cited *supra* notes 5–7.
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.
12. For an explanation of neural networks, see Ophir Tanz & Cambron Carter, Neural Networks Made Easy, Tech Crunch (Apr. 13, 2017), https://techcrunch.com/2017/04/13/neural-networks-made-easy/.
13. *See* Jigsaw, Perspective, https://jigsaw.google.com/projects/#perspective.
14. *See* Jeremy J. Eberhardt, *Bayesian Spam Detection*, 2 Scholarly Horizons at 4 (2015), http://digitalcommons.morris.umn.edu/cgi/viewcontent.cgi?article=1024&context=horizons.
15. 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).
16. See 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 (2017), https://arxiv.org/pdf/1703.04009.pdf.
17. *See* Burnap & Williams, *supra* note 15.
18. *See, e.g.,* Yorick Wilks & Mark Stevenson, *The Grammar of Sense: Using Part-of-Speech Tags as a First Step in Semantic Disambiguation*, 4 J. Natural Language Engineering 135 (1998).
19. Google Code, word2vec, https://code.google.com/archive/p/word2vec/ (Word2vec is a Google open source project).
20. *See* Jan Bussieck, Demistifying Word2vec, Deep Learning Weekly (Aug. 22, 2017), *https://www.deeplearningweekly.com/blog/demystifying-word2vec*.
21. *See, e.g.,* TensorFlow, Vector Representations of Words, https://www.tensorflow.org/tutorials/word2vec.

22. *See, e.g.,* Microsoft TechNet, Training and Testing Data Sets (2012), https://technet.microsoft.com/en-us/library/bb895173(v=sql.110).aspx.
23. *See* Will Knight, *AI's Language Problem*, MIT Tech. Rev. (2016), https://www.technologyreview.com/s/602094/ais-language-problem/.
24. *See, e.g., infra* notes 25–37 & 82–91.
25. *See, e.g.,* Ahmed Abbasi, Ammar Hassan & Milan Dhar, *Benchmarking Twitter Sentiment Analysis Tools*, Proceedings of the 9th Language Resources and Evaluation Conference (2014) (finding that trained "workbench" tools incorporating domain specific knowledge performed better for analyzing sentiment on Twitter than one-size-fits-all "stand-alone" tools); Dirk Von Grunigen et al., *Potential Limitations of Cross-Domain Sentiment Classficiation*, Proceedings of the 5th International Workshop on Natural Language Processing for Social Media (SocialNLP) (2017) (finding that sentiment analysis tools trained for one domain performed poorly in "foreign domains").
26. *See, e.g.,* Burnap & Williams, *supra* note 15.
27. *See* Bermingham et al., C*ombining 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.)."
28. *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.
29. *See* Abbasi, Hassan & Dhar, *supra*, note 25 (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).
30. *Id.*
31. *Id.*
32. *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).
33. 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.").
34. *Id.*
35. *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.
36. Burnap & Williams, *supra* note 15.
37. *Id.* at 236.
38. *See* 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.
39. *Id.*
40. *Id.*

41. 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.
42. 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.
43. *Id.*
44. 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.
45. 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 33 (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 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*, The Guardian, http://labs.theguardian.com/digital-language-divide/.
46. Abbasi, Hassan & Dhar, *supra* note 25, 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.
47. *See, e.g.,* Young, *supra* note 45.
48. *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.
49. Alex Hern, *Facebook Translates 'Good Morning' into 'Attack Them', Leading to Arrest*, The Guardian (Oct. 24, 2017), https://www.theguardian.com/technology/2017/oct/24/facebook-palestine-israel-translates-good-morning-attack-them-arrest.
50. *Id.*
51. Blodgett & O'Connor, *supra* note 28; 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).
52. *See,e.g.,* Blodgett & O'Connor, *supra* note 28; Tatman; *supra* note 51.
53. Tatman, *supra* note 51.
54. Blodgett & O'Connor, *supra* note 28.
55. Nicholas Thompson, I*nstagram 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

comment uses a British slang term for cigarettes; this word is also a derogatory term in American English for gay men.

56. *See* sources cited *supra* note 4.
57. *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.
58. 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 . . . ." Schmidt & Wiegand, supra note 33. 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 46.
59. Warner & Hirschberg, *supra* note 35.
60. *See* Nockleby, *supra* note 58; Zeerak Waseem & Dirk Hovy, *Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter*, Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL): Human Language Technologies (HLT), 88–93 (2016).
61. 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, 7 (2016), https://www.linguistics.rub.de/bla/nlp4cmc2016/ross.pdf.
62. International Covenant on Civil and Political Rights ("ICCPR"), G.A. Res. 2200A, Art. 20(2).
63. 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.
64. *See* Ross et al., *supra* note 61, at 7.
65. Platforms' definitions of hate speech and other content prohibited on their platforms may not always be clear or intuitive to users or researchers. *See, e.g.,* Julia Angwin & Hannes Grassegger, *Facebook's Secret Censorship Rules Protect White Men from Hate Speech But Not Black Children*, Pro Publica (June 28, 2017), https://www.propublica.org/article/facebook-hate-speech-censorship-internal-documents-algorithms.
66. 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 45.
67. *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.
68. Cohen et al., *supra* note 66; Johansson et al., *supra* note 45.
69. *See* Ross et al., *supra* note 61, at 8. (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 $\alpha = 0.18$ and $\alpha = 0.29$, where Krippendorff recommends a minimum of $\alpha = 0.80$, or a minimum of 0.60 "for applications where some uncertainty is un-problematic"); Davidson et al., *supra* note 58 (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).

70. Ross et al., *supra* note 61, at 9.
71. Schmidt & Weigand, *supra* note 33. *But see* Warner & Hirschberg, *supra* note 35, at 21–22.
72. 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 35, at 22.
73. *See, e.g.,* Mark Cieliebak 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 25; 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.
74. Abbasi, Hassan & Dhar, *supra* note 25.
75. For example, even in a study designed to increase the prevalence of hate speech in a corpus of tweets, only 5% of the 14,802 tweets were labelled as hate speech by a majority of coders. Davidson et al., supra note 58. See also Warner & Hirschberg, supra note 35 (in a corpus of text relating to Judaism and Israel, 91% of examples were not anti-Semitic); Schmidt & Wiegand, *supra* note 33, at 7 ("[T]here are much fewer hateful than benign comments present in randomly sampled data.").
76. *See, e.g.,* Daniel Bier, *By the Numbers: Europe's Terror Problem*, Foundation for Economic Education (March 31, 2016), https://fee.org/articles/by-the-numbers-europes-terror-problem/.
77. *See , e.g.,* The MITRE Corporation, JASON Program Office, Rare Events (Oct. 2009), https://fas.org/irp/agency/dod/jason/rare.pdf ("There is no credible approach that has been documented to date to accurately anticipate the existence and characterization of WMD-T [weapons of mass destruction-terrorism] threats."); National Research Council of the National Academies of Science, Protecting Individual Privacy in the Struggle Against Terrorists: A Framework for Program Assessment (2008), https://www.nap.edu/catalog/12452/protecting-individual-privacy-in-the-struggle-against-terrorists-a-framework (finding that terrorist identification via data mining or "any other known methodology" was "neither feasible as an objective nor desirable as a goal of technology development efforts").
78. Warner & Hirschberg, *supra* note 35.
79. *See* Julia Angwin et al., *Machine Bias*, ProPublica (May 23, 2016), https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.
80. *See, e.g.,* Burnap & Williams, *supra* note 15, at 235.
81. *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).
82. Abbasi, Hassan & Dhar, *supra* note 25.
83. 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.
84. 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/.
85. *Id.*

86. *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).
87. Schmidt & Wiegand, *supra* note 33.
88. Xiang et al., *supra* note 86, at 1980.
89. Xiang et al. reported a 5.4% improvement in classification using this method, compared to "keyword matching." Xiang et al., *supra* note 86, at 1984.
90. Dadvar et al. (2012), *supra* note 86.
91. Dadvar et al. (2013), *supra* note 86.