DCPAPERS

Proc. Int'l Conf. on Dublin Core and Metadata Applications 2018

Categorization Ethics: Questions about Truth, Privacy and Big Data

Presentation

Joseph A Busch
Taxonomy Strategies, USA
jbusch@taxonomystrategies.com

Keywords: bias; automated categorization; auto-classification; privacy; GDPR

Abstract

This abstract discusses issues related to the inherent bias of automated categorization caused by content collections used to build machine learning models and the impact of the General Data Protection Regulation (GDPR).

1. Introduction

Categorization is a common human behavior and it has many social implications. While categorization helps make sense of the world around us, it also affects how we perceive the world, what we like and dislike, who we feel comfortable with and who we fear. Categorization is affected by our family, culture and education. This can easily lead to classification bias where we create categories and apply them in ways that reflect bias rather than trust. (Mai) Statistical bias is caused by sampling or measurement errors. This plays out in many different contexts such as epidemiology (selection bias), the media (source omission), and machine learning (unsupervised analysis).

2. Inherent bias of automated categorization

In the October 19, 2016 ProPublica video "How Machines Learn to Be Racist," part of a series on machine bias, Julia Angwin mentions a study where researchers analyzed 3 million words from Google news stories. The closest word associated with the phrase "black male" was "assaulted." While the closest phrase associated with "white male" was "entitled to." This is an illustration of the problem with an "unsupervised" analysis to identify closely associated words and phrases. It is very common to use news feeds such as Google news stories or Wikipedia as the content collection to "train" automated categorization algorithms.

How does automated categorization work? All automated categorization is based on analyzing a collection of content to identify patterns. Those patterns are transformed into examples that become "templates" for categories. There are many different scenarios that can be used to identify examples. For images, imagine a collection of examples of "cats" and "chairs." Given enough examples, a pattern emerges that can usually determine whether an image is of a cat or a chair or not of a cat or not of a chair. FIG. 1 illustrates these image recognition rules as Boolean queries.

https://doi.org/10.23106/dcmi.952139150

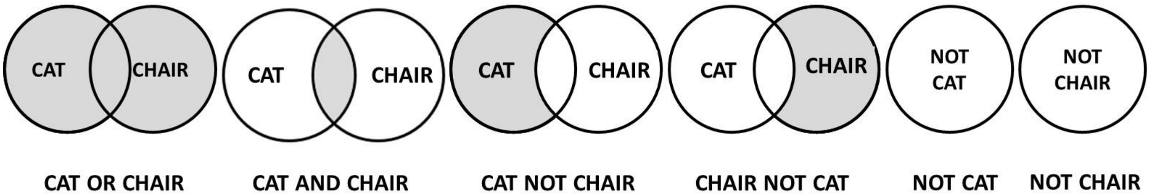
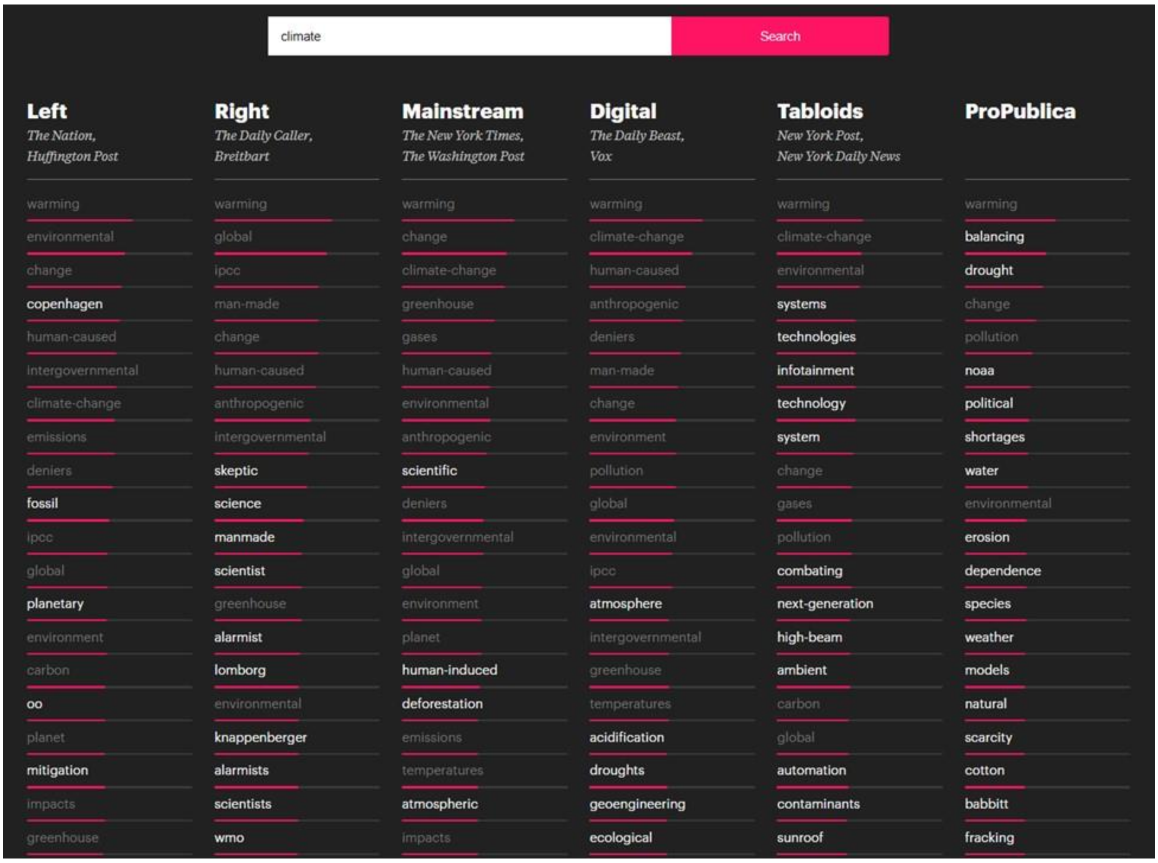


FIG. 1. Image recognition Boolean rules illustrated using Venn diagrams.

It's more complex when the collection is composed of text. In the simplest case, the text is processed using so-called natural language processing or NLP to identify nouns and noun phrases. The nouns and noun phrase occurrences and co-occurrences are counted, and then those counts are weighted based on the length of the analyzed content. Those terms with the highest weighted frequency are then used to characterize the content item. Across the content collection, other content items with similarly weighted high frequency terms are grouped together. New content items are evaluated for similarity to existing ones. Information retrieval services use these automatically generated categorizations to create feeds and make recommendations.

In the story on "How Machines Learn to Be Racist," ProPublica utilized a Google algorithm to identify synonyms (meaning closely associated nouns and noun phrases) by analyzing articles from different categories of news outlets – left, right, mainstream, digital, tabloids, and investigative. This demonstration illustrated in FIG. 2 shows how the point of view of the content collection that is processed affects the resulting list of synonyms which become the rules that define the category.



Left *The Nation, Huffington Post*	Right *The Daily Caller, Breitbart*	Mainstream *The New York Times, The Washington Post*	Digital *The Daily Beast, Vox*	Tabloids *New York Post, New York Daily News*	ProPublica
warming	warming	warming	warming	warming	warming
environmental	global	change	climate-change	climate-change	balancing
change	ipcc	climate-change	human-caused	environmental	drought
copenhagen	man-made	greenhouse	anthropogenic	systems	change
human-caused	change	gases	deniers	technologies	pollution
intergovernmental	human-caused	human-caused	man-made	infotainment	noaa
climate-change	anthropogenic	environmental	change	technology	political
emissions	intergovernmental	anthropogenic	environment	system	shortages
deniers	skeptic	scientific	pollution	change	water
fossil	science	deniers	global	gases	environmental
ipcc	manmade	intergovernmental	environmental	pollution	erosion
global	scientist	global	ipcc	combating	dependence
planetary	greenhouse	environment	atmosphere	next-generation	species
environment	alarmist	planet	intergovernmental	high-beam	weather
carbon	lomborg	human-induced	greenhouse	ambient	models
oo	environmental	deforestation	temperatures	carbon	natural
planet	knappenberger	emissions	acidification	global	scarcity
mitigation	alarmists	temperatures	droughts	automation	cotton
impacts	scientists	atmospheric	geoengineering	contaminants	babbitt
greenhouse	wmo	impacts	ecological	sunroof	fracking

FIG. 2. ProPublic synonym picker illustrates how the point of view of the content affects results.

It needs to be assumed that there is an inherent bias in any collection of content that reflects discourse in a culture at a particular time, or steps need to be taken to obtain representative content—but representative of what? Bias results from models being trained on data that is historically biased. Rebecca Njeri in a 2017 blog post claims that "it is possible to intervene and

https://doi.org/10.23106/dcmi.952139150

address the historical biases contained in the data such that the model remains aware of gender, age and race ***without*** discriminating against or penalizing any protected classes" – (author's emphasis).

2. Impact of GDPR on automated categorization

The General Data Protection Regulation (GDPR) provides rules for protecting personally identifying information (PII), for example, the so-called "right to be forgotten." GDPR applies to processing of personal data, but not to processing of content collections in the public or published domain such as news stories or Wikipedia articles. GDPR restricts the nature of collections used for machine learning excluding anything that includes PII such as social media, customer service records, medical records, etc. Restrictions and work-arounds are already used to aggregate information in a way that obscures the PII. GDPR permits PII to be collected for specified, explicit and legitimate purposes, but does not permit further processing beyond those purposes except "for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes." (Art. 5 GDPR) Thus GDPR provides important restrictions on commercial uses of PII, even aggregated personal information, that has not been explicitly collected for a particular and personally approved purpose.

2.1. Does GDPR have an impact on classification bias?

GDPR requires that personal identifying information be accurate, and that if requested by an individual, that PII be corrected or deleted. GDPR could have an unintended impact on selection bias by allowing deletion of PII leading to incomplete or inadequate representation of a selection class.

3. Conclusions

Individuals can take responsibility for their own perceptions, misperceptions can be pointed out and sometimes changed. But categorization is often imposed on individuals from outside. For information aggregators and information analyzers, the guidelines for appropriate behavior are not always clear, nor is the responsibility for outcomes as a result of errors, bias and worse. GDPR provides some guidelines for aggregation of personal identifying information, but not on categorization bias itself. When errors and bias are commonly held, this can be reflected in the information ecology. The tipping point need not be a majority, truth or based on ethics. It's easy enough to identify cases of mis-categorization, but when should something be done about it?

References

"Art. 5 GDPR Principles relating to processing of personal data." Retrieved on August 20, 2018. https://gdpr-info.eu/art-5-gdpr/.

Dixon, Lucas and others. "Measuring and Mitigating Unintended Bias in Text Classification." Presented at: AAAI/ACM Conference on AI, Ethics, and Society (2018) Retrieved August 19, 2018. https://storage.googleapis.com/pub-tools-public-publication-data/pdf/ab50a4205513d19233233dbdbb4d1035d7c8c6c2.pdf.

Larson, Jeff, Julia Angwin and Terry Parris Jr. "How Machines Learn to Be Racist." (October 19, 2016) Retrieved August 18, 2018. https://www.propublica.org/article/breaking-the-black-box-how-machines-learn-to-be-racist?

Mai, Jens-Erik. "Classification in a social world: bias and trust." 66 *Journal of Documentation* 5: 627-642 (2010) Retrieved August 19, 2018. http://jenserikmai.info/Papers/2010_Classificationinasocialworld.pdf.

Njeri, Rebecca. "How Do Machine Learning Algorithms Learn Bias?" Towards Data Science [blog] (Aug 20, 2017) Retrieved August 19, 2018. https://towardsdatascience.com/how-do-machine-learning-algorithms-learn-bias-555809a1decb.

Wikipedia. "Selection bias." Retrieved August 20, 2018. https://en.wikipedia.org/wiki/Selection_bias.

https://doi.org/10.23106/dcmi.952139150