CENTER FOR DEMOCRACY & TECHNOLOGY

Automation in Content Moderation: Capabilities and Limitations

Centre for European Policy Studies
5 September 2017

About CDT

At the Center of Democracy and Technology, we believe in the power of the internet. Whether it's facilitating entrepreneurial endeavors, providing access to new markets and opportunities, or creating a platform for free speech, the internet empowers, emboldens, and equalizes people around the world.

CENTER FOR DEMOCRACY & TECHNOLOGY

Increasing pressure on content moderation systems

- Content moderation online is foremost an issue of scale
- Most moderation systems operate on flagging-and-review
- Challenges of problematic, even illegal, content increase pressure for moderation systems to be faster and more comprehensive
- Pressure from governments, advertisers, public/users

Leads to increasing focus on automating content moderation processes

Some forms of automated content filtering in use

- Keyword filters
- Hash-matching algorithms (PhotoDNA, ContentID)
- Spam detection tools

These typically rely on

- identifying patterns of behavior (like high rates of spam messages)
- matching previously identified unwanted content (keywords, URLs, image hashes) to new uploads or web browsing activity

And have well-known weaknesses

- Overbroad/unsophisticated
- Easily circumvented

New frontier: Using machine-learning algorithms to determine meaning of text

Definitions:

- Learning algorithm
- Natural language processing
- Sentiment analysis

Development of a sentiment analysis algorithm

CDT's *Digital Decisions* tool: https://cdt.info/ddtool/

Highlights for sentiment analysis tools:

- Acquire training data
- Clean and code training data
- Select data process (e.g. semi-supervised machine learning algorithm)
- Review training output and refine
- Apply to test data and determine accuracy; refine
- Apply to novel data

A few examples

Wikipedia tool/classifier for identifying “toxic” comments on Wikipedia edit pages

Ex Machina: Personal Attacks Seen at Scale, Wulczyn, Thain, and Dixon (2017), https://arxiv.org/pdf/1610.08914.pdf

Sentiment analysis tools developed for use on Twitter

Benchmarking Twitter Sentiment Analysis Tools, Abbasi, Hussan, Dhar (2015), https://pdfs.semanticscholar.org/d0a5/21c8cc0508f1003f3e1d1fbf49780d9062f7.pdf

6 Issues to Consider with Automated Content Analysis Tools

6 Issues to Consider with Automated Content Analysis Tools

1. Bias or lack of representation in training data used to develop the tool.

6 Issues to Consider with Automated Content Analysis Tools

1. Bias or lack of representation in training data used to develop the tool.
2. Complexity of communication and its constant evolution.

6 Issues to Consider with Automated Content Analysis Tools

1. Bias or lack of representation in training data used to develop the tool.
2. Complexity of communication and its constant evolution.
3. Tools are not necessarily transferable across contexts/platforms.

6 Issues to Consider with Automated Content Analysis Tools

1. Bias or lack of representation in training data used to develop the tool.
2. Complexity of communication and its constant evolution.
3. Tools are not necessarily transferable across contexts/platforms.
4. “Accuracy” of tools may be low or not relevant for application.

Multiple ways to think about “accuracy” of a tool

“Accuracy” or “error” can be a measure of:

- Training data vs. test data from the same data set
- Tool vs. novel data & source
- Tool vs. “ground truth”

6 Issues to Consider with Automated Content Analysis Tools

1. Bias or lack of representation in training data used to develop the tool.
2. Complexity of communication and its constant evolution.
3. Tools are not necessarily transferable across contexts/platforms.
4. “Accuracy” of tools may be low or not relevant for application.
5. Human involvement in algorithmic decisions key to mitigate bias/error.

Automation in Content Moderation

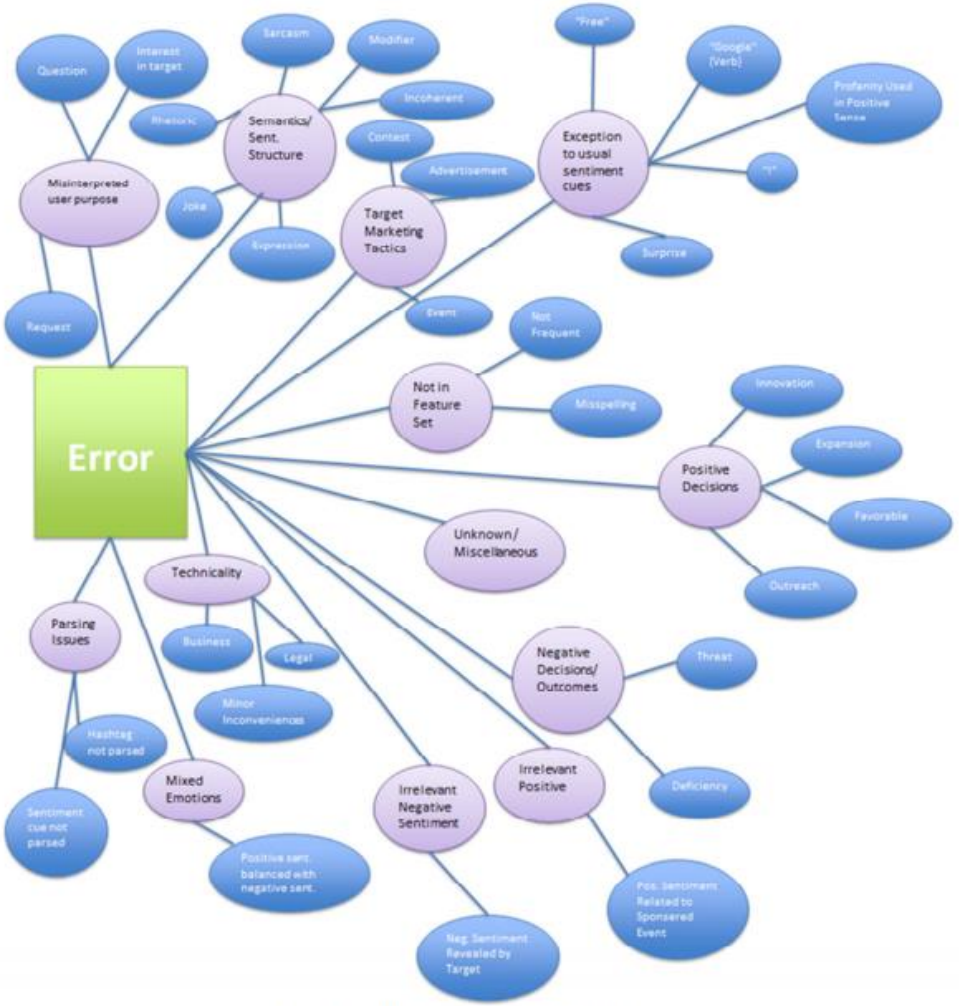

Figure 1: Twitter Sentiment Error Analysis Taxonomy

CENTER FOR DEMOCRACY & TECHNOLOGY

6 Issues to Consider with Automated Content Analysis Tools

1. Bias or lack of representation in training data used to develop the tool.
2. Complexity of communication and its constant evolution.
3. Tools are not necessarily transferable across contexts/platforms.
4. “Accuracy” of tools may be low or not relevant for application.
5. Human involvement in algorithmic decisions key to mitigate bias/error.
6. Transparency in development & application of tools variable, but necessary.

Final Thoughts

- These tools are not a silver bullet for the challenges of content moderation at scale.
- More data is needed to evaluate the state of the technology.
- Human involvement in content moderation processes continues to be vital.
- Content hosts should improve their remedy & appeals processes.
- Policymakers must be savvy consumers of these tools.

THANK YOU | CDT.ORG

Emma Llansó
ellanso@cdt.org

CENTER FOR DEMOCRACY & TECHNOLOGY