ahalterman0@gmail.com
https://andrewhalterman.com
github.com/ahalterman

Synthetic Text for Supervised Learning

Andy Halterman (MSU Political Science)

MICHIGAN STATE UNIVERSITY

Why is supervised text analysis hard?

1. **LABELING**: labels are expensive to collect
2. **RETRIEVAL**: when a class is rare it's hard to annotate enough positive cases
3. **COPYRIGHT**: most text cannot be freely shared for copyright or privacy reasons, limiting re-use, evaluation, and reproducibility

- **Possible solution**: generate synthetic text with desired content and style.

How do I make good synthetic text?

Language models (e.g. GPT) are trained to predict the next word given a sequence of previous words. The probability of a next token depends on the previous tokens and the model's parameters:

$$\widehat{p}(w_i) = f(w_{i-1}, w_{i-2}, ...w_1, \theta)$$

Thus, to control the generation of synthetic text, *either*:

- **FINE-TUNE**: update θ using domain-specific text, *or*
- **PROMPT**: modify the preceding tokens $w_{i-1}, w_{i-2}...$

When should I FINE-TUNE vs. PROMPT?

- **FINE-TUNING** is useful when prompting is difficult or you have an existing corpus and want to evaluate synthetic data quality.
- **PROMPTING** is useful when you don't have an existing corpus and documents can be prompted.

When **FINE-TUNING**, consider an adversarial technique for tuning the parameters γ that control how to sample from $p(w_i)$. The best γ produces the worst performance for a real vs. synthetic classifier.

App 1: Weapon NER on Ukraine War Tweets

Problem: **COPYRIGHT**.
Setup: **FINE-TUNE** GPT-2 on Ukraine war tweets.

Real tweets are indicated with ✓and synthetic ones with ✗.

✗ Russian equipment losses suffered during the invasion of Ukraine updated with:
1x T-72B3M (abandoned)
1x BMP-2 IFV (destroyed)
1x BTR-80 APC (abandoned)
1x 152mm 2S19 Msta-S SPG (abandoned)
1x 122mm BM-21 MRL (destroyed)
Full list:

✗ A Russian BTR-82A armored personnel carrier was destroyed by the Ukrainian 128th Mountain Assault Brigade in the east of the country.

✗ The Ukrainian SES posted an image of one of the destroyed vehicles, a destroyed Russian T-72B3 variant.

✗ And this is the Ukrainian T-64BV that was struck and destroyed by the Russian forces in Mariupol.

✓ Improvised top attack armor did not save this Russian T-72B3. Reportedly destroyed by a Javelin.

✓ #Ukraine: Two Russian BMP-2 infantry fighting vehicles were recently destroyed by the 93rd Mechanized Brigade of Ukraine on the Eastern front.

✗ A number of destroyed Russian T-72B3s

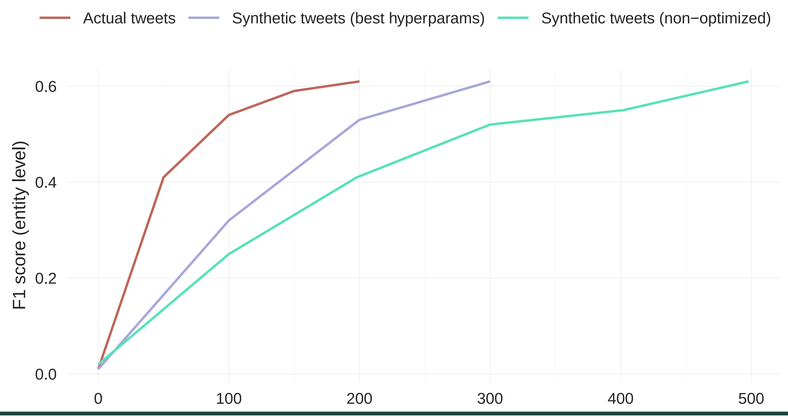
✓ #UkraineWar: Newly added Russian equipment losses:
1x ZALA 421-16E2 reconnaissance UAV (destroyed)
1x T-64BV MBT (destroyed)
1x BTR-80 APC (destroyed)
1x R-166-0.5 signals vehicle (destroyed)
1x Unknown vehicle (destroyed)
Full list:

✗ #Ukraine: Another Ukrainian T-64BV tank was destroyed by the Russian army in #Kyiv Oblast.

✓ #Ukraine: In the vicinity of #Sumy Oblast, Ukrainian forces destroyed a Tiger-M IMV, a BTR-82A, and 3x supply trucks. [Corrected]

✗ #Ukraine: Another Ukrainian T-64BV tank was destroyed by the Russian army in #Kyiv Oblast.

Named entity recognition model performance with real and synthetic tweets:



App 2: Sentence-level Populism Detection

Problem: **LABELING**. No sentence-level populism annotations exist (Di Cocco and Monechi 2021, Jankowski and Huber 2022), and collecting them, especially across languages, is expensive.
Setup: **PROMPT** GPT-3 with task description, including language prompt.

Prompt: *Populist rhetoric sees politics as a conflict with good, common, or "real" people on one side, and out-of-touch, evil or self-serving elites on the other.*
Write three statements that an Austrian populist party might make (in the German language).

Output:

1. [SYNTH] Europa ist in Gefahr! Wir brauchen eine starkes Österreich, um unsere Werte zu verteidigen!
2. [SYNTH] Die Globalisierung ist ein Angriff auf unser Heimatland.
3. [SYNTH] Die Regierung ist völlig außer Kontrolle und hat keine Ahnung, was die Menschen wirklich wollen.

Next, I trained a populism classifier (XLM-RoBERTa-base) on the synthetic sentences and used this model to generate predicted populism scores for actual manifesto text. For example, the three UKIP sentences with highest *p*(populist):

- This is how utterly ridiculous the Common Fisheries Policy is: it is destroying our fishing industry and we must take back control from Brussels."
- Politics is corrupted by self-interest and big business.
- This is a terrible legacy to leave our children and grandchildren.

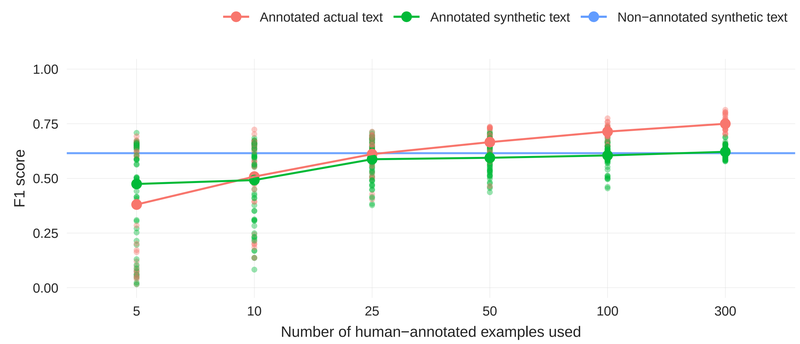
...and the lowest *p*(populist):

- Extend the period during which discharged service personnel are able to access the specialist DMHS scheme from six months to two years.
- CONTENTSINTRODUCTION TO OUR MANIFESTO
- This will be supported by the inclusion of FGM awareness into safeguarding training for teachers, school staff and governors.

App 3: Detecting Political Events

Problems: **RETRIEVAL**, **COPYRIGHT**, **LABELING**.
Setup: **PROMPT** an off-the-shelf GPT-2.
Example prompt: "Bomb detonates in downtown capital BOGOTA (Reuters) –"

Performance of SVM predicting ASSAULT class, evaluted on annotated actual text



Ethics & Acknowledgements

- Synthetic text saved on disk should always be marked and the warning removed only temporarily and in-memory for training models. Consider including researcher name and project description. E.g.:
  `<!-- SYNTHETIC TEXT! May be factually incorrect and offensive. Generated by Andy Halterman for... -->`
- Annotators should always be told that they are working with synthetic text.

Portions of this work work were sponsored by the Political Instability Task Force (PITF). The PITF is funded by the Central Intelligence Agency. The views expressed in this document are the author's alone and do not represent the views of the US Government.