

Why so little supervised text analysis?

Supervised text analysis is often the best tool for a project but it faces several obstacles:

1. **LABELING**: labels are expensive to collect from human annotators
2. **RETRIEVAL**: when a class is rare, it's hard to find positive cases in a corpus to annotate
3. **COPYRIGHT**: most text cannot be freely shared, limited re-use and evaluation

Possible Solution: Synthetic Text

Can we use large language models (e.g. GPT) help address these problems?

Three approaches to generating useful text:

1. **off-the-shelf + prompt**: use a pre-trained model (e.g. GPT-2) and write a prompt to generate relevant text.
2. **fine-tune**: fine-tune a language model on the domain (e.g. tweets, speeches) and generate similar text
3. **task explanation**: use a very large model (e.g. GPT3) and describe the desired text.

Once we have synthetic text, we have two options for getting labels:

1. **label synthetic text**: collect human annotations on the synthetic text
2. **use "soft"/"distant" labels as-is**: assume the synthetic text has the desired label

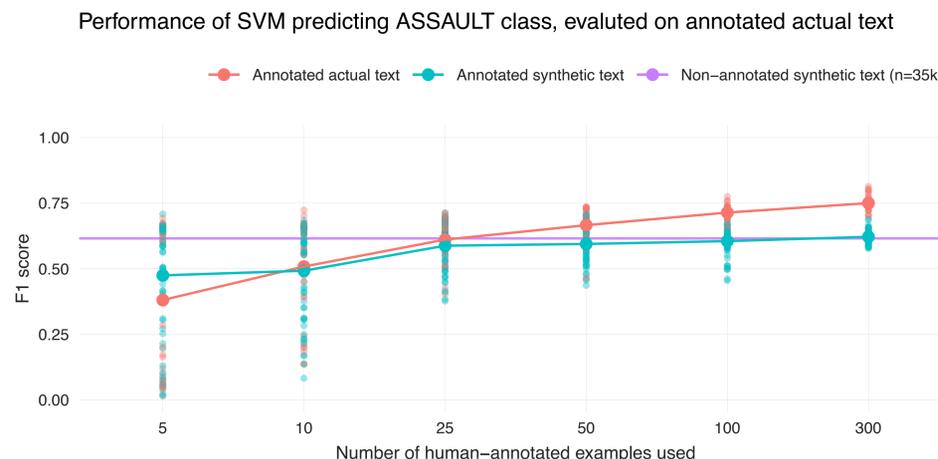
Conclusions

1. Off-the-shelf GPT-2, fine-tuned GPT-2, and "task prompted" GPT-3 can generate text that's almost indistinguishable from original text.
2. However, models trained on synthetic text underperform models trained on original text. Retrieval and copyright benefits may be worth it, though.
3. "Soft labels" from generated text may be promising for zero-shot text classification.

Detecting Political Events

Problems: RETRIEVAL, COPYRIGHT, LABELING.

Setup: off-the-shelf GPT-2 with headlines prompting ASSAULT-type events.



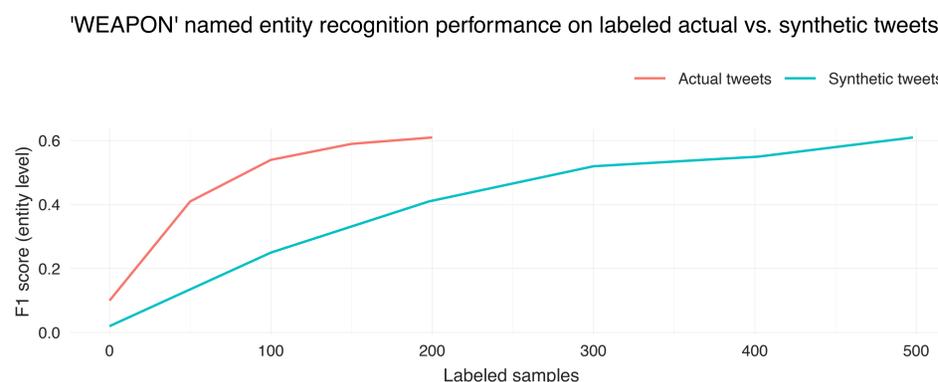
Identifying Weapons in Ukraine War Tweets

Problem: COPYRIGHT.

Setup: fine-tuned GPT-2 on Ukraine war tweets (no prompt)

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

- | | |
|---|---|
| <p>✗ 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.</p> | <p>✓ #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.</p> |
|---|---|



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: 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. Populist parties generally oppose multiculturalism and globalism and want to protect their national way of life.*

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, train a model (XLM-RoBERTa-base) on the synthetic sentences. Use this model to generate predicted populism scores for actual manifesto text. For example, the three UKIP sentences with highest $p(\text{populist})$:

- ▶ 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 $p(\text{populist})$:

- ▶ 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.

Ethics and Acknowledgements

- ▶ Synthetic text saved on disk should **always** be marked and the warning removed only temporarily and in-memory. E.g.:
<!-- SYNTHETIC TEXT! Do not trust the factual content of this text -->. Consider also author name + project.
- ▶ Annotators should **always** be told that they are working with synthetic text.

This poster draws on ongoing joint work:

- ▶ Populism: Shahryar Minhas, Christian Houle, Nicolas Bichay
- ▶ Event classification: Benjamin Bagozzi, Phil Schrod, Andy Beger, Grace Scarborough.