Synthetically generated text for supervised text analysis

Andrew Halterman*

14 September 2022

Abstract

Supervised text models are often the best tool for categorizing documents into known classes or for extracting information from within documents. However, supervised models are often difficult to employ because of the expense involved in hand-labeling documents, the difficulty of retrieving relevant documents for rare class annotation, and copyright and privacy concerns involved in sharing annotated documents. This paper proposes a partial solution to these three issues, in the form of controlled generation of synthetic text. Recent advances in text generation make it possible to create synthetic documents with desired class labels and in a form that can be broadly shared without copyright or licensing concerns. I demonstrate the usefulness of text generation techniques with three applications: using an off-the-shelf language model prompted with article headlines to generate synthetic news articles describing specified political events for training an event detection system, using a fine-tuned language model to generate synthetic tweets describing the fighting in Ukraine for named entity recognition labeling, and using a task description approach to generate a multilingual corpus of populist manifesto statements for training a sentence-level populism classifier. The article includes a discussion of the ethical concerns inherent in this work along with proposed guidelines for researchers.

Introduction

Supervised learning techniques are the appropriate tools for analyzing text when researchers want to categorize documents into known categories or extract specific pieces of information from within a document. Supervised text analysis carries major costs, however: it depends on collecting potentially expensive human annotations on text, identifying rare event classes for annotation can be difficult, and copyright,

*Michigan State University. ahalterman0@gmail.com

licensing, and privacy concerns can make it difficult or impossible to share annotated texts with other researchers, limiting the field's ability to build on or evaluate previous supervised models. Lowering these costs would make it easier for political scientists to use supervised models when they are the appropriate methodological tool for their research.

This paper proposes addressing these barriers by generating synthetic text with inferred labels using large language models, a set of computational models from natural language processing. Large language models, such as GPT-2, , GPT-3, and many others, are trained on large corpus of text to predict the next word in a sequence of training data. By using efficient transformer-based architectures, models with a very large number of parameters (in the hundreds of millions to billions), and large, diverse sets of training data, language models can generate a sequence of tokens that is likely to follow an input set of tokens. These large language models can already generate news text or political text that is already indistinguishable from human-produced text (Zellers et al. 2019; Kreps, McCain, and Brundage 2022). A set of prompting and fine-tuning techniques allow users to generate text with desired content or style.

The rest of the paper describes how researchers in political science can use large language models to lower the costs of supervised text analysis, including the decisions that researchers face in generating and using synthetic text and the serious ethical pitfalls inherent in using synthetic text. It presents three short applications from political science, illustrating how synthetic text can be used in creating training data for event data detection models, synthetic tweets describing the war in Ukraine training named entity recognition model, and a "zero shot" approach to building a sentence-level populism classifier for studying political manifestos. It demonstrates that synthetic data can be used to create tweets that are difficult to distinguish from real tweets and that synthetic documents can be used to train "zero shot" text classification models with no human annotation at all. However, a marginal labeled synthetic document is generally not as informative as a marginal real document in training models, presenting a tradeoff for researchers between accuracy and the benefits produced by using synthetic text.

The costs of supervised text models

Supervised learning is generally a more difficult approach to text analysis than unsupervised text analysis. The primary cost in developing a supervised model is *labeling*. Human annotations on documents, such as document labels for classification or spans for information extraction tasks, are expensive to collect. Researchers need to define their concept of interest, create a codebook, train annotators, pay them, and conduct quality assurance on the labels they provide. Many supervised learning projects may not occur because labeling costs were prohibitive.

A second obstacle is *retrieval*, namely that annotators need to be provided a set of text that has sufficient positive classes. Because many of the concepts of interest in political science are rare classes (Miller, Linder, and Mebane 2020), a simple random

sample will not retrieve enough positive documents. Currently, researchers will oversample relevant documents using keywords (e.g. Mueller and Rauh 2017), using active learning techniques (Miller, Linder, and Mebane 2020), or by exhaustively annotating an entire corpus (Halterman et al. 2021). Each of these techniques carries drawback in annotation cost, low recall, or dependence on a model to suggest documents to label.

Finally, researchers face *copyright* or other restrictions on their ability to share annotated documents. Once a researcher collects annotations on a document, they are often unable to share the original document because of copyright, licensing restrictions, terms of service, or privacy concerns.

If researchers can generate synthetic text in a controllable way, that is, with the ability to direct the content of the text, that would partially address all three of these issues by producing text with reliable class labels for text classification, with noisy "pseudo" labels for retrieval, and in any case in a form that is free of restrictions on sharing.

Proposed use

Researchers generating synthetic text for supervised learning applications face two sets of questions: how to generate the text and what to do with the generated text afterward.[1]

Applied researcher have three primary options for guiding the content of the synthetic text. These techniques are general, working on current transformer-based neural networks, but also on older technologies such as recurrent neural networks like LSTMs.

First, researchers can use an *off-the-shelf model with simple prompting* strategy. For certain applications, researchers can download a generic pre-trained model (e.g. GPT-2) and write a prompt to generate relevant text. Large language models are trained to generate text that follows previous text, so if a researcher can provide the beginning of a document, a large language model can generate a plausible continuation of the document. For example, news articles can easily be prompted to contain desired content using a headline that cues the desired content. This approach is simple, but will fail if either the domain is not covered well in the language model's training data (e.g. legal statues) or if the desired text is difficult to prompt (e.g. tweets). This approach is illustrated in the first application to generate news stories about military assaults by providing headlines.

Second, they can use a *fine-tuning* strategy. Pretrained language models are trained on a large set of text to predict the next token in a sequence. Models can be further *fine-tuned* to improve their performance on a downstream task, including to better predict the next token for a specific text domain. Doing so eliminates the need to

[1]Given a set of tokens $X = \{x_1 ... x_n\}$, the probability of the sequence can be decomposed into the probability of each token given the previous: $p(X) = \prod_{i=1}^{n} p(x_i | x_1 ... x_{i-1})$. A large language model learns to approximate the conditional probability of each token and thus can be used to sample plausible following tokens given a sequence of text.

provide a prompt, given that the language model has been fine tuned to generate text from the target domain. This approach is used in the second application to generate synthetic tweets reporting battlefield updates from the war in Ukraine.

Third, researchers can use a *task explanation* approach. Recent very large language models (e.g. GPT-3) are much larger than previous models and trained on much more text and can be prompted with a description of the desired text. For instance, GPT-3 can be provided with a task description like "write a press release in the style of a Republican house member" and obtain a plausible output without the need for a specific prompt or fine tuning on a corpus of press releases. This approach is used in the third application to generate populist party manifestos in 25 European languages.

Researchers in natural language processing are developing more sophisticated techniques for controlled text generation to allow covariates to guide text generation or ensure the factuality of the generated text (e.g. Dathathri et al. 2020; Prabhumoye, Black, and Salakhutdinov 2020; Yogatama, Masson d'Autume, and Kong 2021). Once they mature, these methods will offer additional options for applied researchers to generate controlled text.

Following these steps, a researcher should have synthetic text that closely matches their domain or corpus of interest. They then have two options for how to use it. First, they can treat the text as unlabeled and collect annotations on it in the same way they would with real text, including document-level labels for classification or span-level annotations for information extraction applications. By using controlled synthetic text, they have addressed the copyright or usage restrictions of real text and can share their annotated text freely, and have hopefully addressed the retrieval problem as well. This approach is used in the first and second applications below.

A second option exists when researcher are conducting document classification and believe that their prompting strategy reliably generates documents with the desired class label. In this situation, researcher can train their document classifier directly on the synthetic text, producing a "zero shot" classifier that requires no hand-labeled documents at all.

Ethics

The use of synthetic text presents serious ethical concerns. Synthetic text can include statements that are factually incorrect, articles that are potential conspiracy theories, or offensive statements. As a practical matter, to avoid any possibility of synthetic stories being mistaken for real news, researchers working with synthetic news stories should always attach a disclaimer directly to any synthetic text any time it is saved or stored.[2] The disclaimer should only be removed in-memory as a final step before fitting a model to avoid the possibility of synthetic data being mistake for real text. Annotators should be briefed on the use of synthetic text and the annotation interface

[2]For example, `<!-- SYNTHETIC TEXT! Do not trust the factual content of this text -->`

should clearly state that they are working with synthetic text, which likely contains factual errors. Any synthetic text reported in published work should be clearly marked (e.g. [SYNTH]) and the accompanying text clearly explain its use and potential bias.

Second, while synthetic data is useful for training some kinds of machine learning models, it should never be used to draw any substantive conclusions. While synthetic text may be difficult to distinguish from real text in its style or writing and thus useful for training a model to recognize certain linguistic features, its factual content will be imaginary and thus completely unsuitable for answering substantive questions on its own.

Third, it is well known that models that are pretrained on large amounts of text will learn the biases present in the pretraining data. For example, given the known biased association between racial groups and sentiment, religious groups and political violence, or gender and occupation (Caliskan, Bryson, and Narayanan 2017), researchers should validate that the models that they train on synthetic text are not relying on group stereotypes when making predictions. Curating the text using to fine-tune the model or using prompts that break the association between groups and stereotyped traits offers a partial solution, but greater research into the prevalence and mitigation of these harms is needed.[3]

Finally, if researchers are using synthetic text as a privacy-preserving strategy, as some researchers have suggested (Ororbia II, Linder, and Snoke 2018; Li et al. 2021), they should opt to label their text rather than using zero shot techniques to ensure that they have checked that their synthetic text does not include any mentions of real people or other sensitive information that may have been learned by the fine tuned language model.

Application 1: Generating rare documents for human labeling–training an event data classifier

Event data is a major source of quantitative information for security and international relations scholars (Beieler et al. 2016). Despite decades of investment, no corpus of news text with comprehensive political event labels exists.[4]. The lack of a gold-standard set of labeled news articles prevents makes it difficult to evaluate event classification models or to develop new classifiers as machine learning techniques improve.

There are two major reasons why no such corpus exists: manually labeling stories for event types is a slow and costly process, especially for rare events, and source

[3]Could add to the appendix here. It's pretty easy to e.g. break the Islam–political violence association by manually specifying different religious groups in the prompting headline.

[4]Some datasets with limited coverage exist, e.g. annotation on actions taken by police in India (Halterman et al. 2021). Some event datasets in computer science/computational linguistics also exist, e.g. ACE (Doddington et al. 2004) but have event definitions that are largely not relevant for political science and are only available with a several thousand dollar fee

text generally cannot be shared because of copyright or licensing agreements . The first application shows how both of these challenges can be overcome with the most basic use of a text generator, a simple off-the-shelf language model. A simple way to generate news articles for training data is write a headline that reflects the event type or concept we would like to have a story about. For example, we can write the headline "Foreign 'information operation' spreading disinformation uncovered". We can give this headline to a large language model, in this case GPT-2 (large) and generate a synthetic news story prompted by the title (Radford et al. 2019). We can increase the diversity of the training text by modifying the byline in the prompt to refer to different news sources and cities.

For example:

> [PROMPT]
>
> *Foreign 'information operation' spreading disinformation uncovered*
>
> *BELGRADE (Reuters)*
>
> [– SYNTHETIC STORY –] In an unusual development this week, Serbian President Tomislav Nikolic called the foreign media a threat to the nation's security.
>
> "Foreign 'information operations' against us and our country (are) the ones spreading the disinformation against us," Nikolic said in […]
>
> [– SYNTHETIC STORY –]

If we provide the same headline but change the byline to "BRUSSELS (local sources)", we can generate a story about misinformation occurring in a completely different context (although note that the story itself is misinformation):

> [PROMPT]
>
> *Foreign 'information operation' spreading disinformation uncovered*
>
> *BRUSSELS (local sources)*
>
> [– SYNTHETIC STORY –]
>
> A report by Belgian daily La Presse has shown that the United States is planning on deploying military 'information operations' in Europe, which could be used for 'hybrid war' scenarios. The report said that the Pentagon is sending US forces to Western Europe to create 'fake news' and create a 'vast digital surveillance network' on political parties, activists, media commentators and'subversives' with […]
>
> [– SYNTHETIC STORY –]

A large-scale annotation effort on synthetically generated news text produces a gold-standard set of copyright-free data that can be employed to train supervised models for event detection.

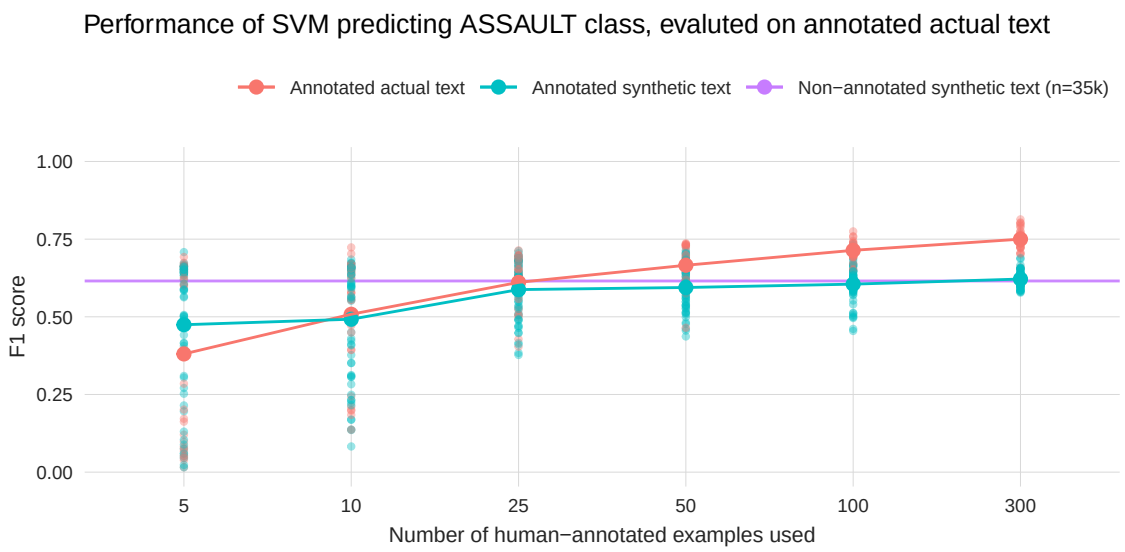

Figure 1: Performance of SVM classifier predicting the ASSAULT class using three sets of training data: annotated real news articles, annotated synthetic articles, and non-annotated synthetic articles (no human labels). Using synthetic documents incurs an accuracy cost, but only after around 50 annotated articles. Using non-annotated synthetic articles (i.e., assuming that every article includes the prompted event type) performs at least as well as labeled synthetic stories, offering a zero-shot classification option. Smaller points indicate 25 random train/test splits, lines show mean performance.

Halterman et al. (2022) describe a new, large-scale annotation effort to produce high quality document-level labels on 12,952 documents for 16 defined event classes. Annotations for the ASSAULT[5] class are used as an evaluation set for this application. Figure 1 compares the performance of classifiers trained on three sets of data: human-annotated actual news stories taken from Factiva, human-annotated synthetic stories, and unlabeled synthetic stories. In the last case, synthetic documents generated with headlines meant to prompt ASSAULT are assumed to do so, and stories generated with non-ASSAULT headlines are assumed to not contain ASSAULT events (see the Appendix for details on the headlines used and the total number of synthetic stories).

Figure 1 shows that a marginal labeled real document improves out-of-sample classifier performance more than an additional labeled synthetic document, indicating that researchers face a tradeoff between accuracy and the retrieval/copyright benefits of synthetic text.[6] The result also show, however, that a model trained exclusively on unannotated synthetic documents performs at least as well as one that uses annotated synthetic documents. Because synthetic documents are essentially free to produce, a large number can be generated and used to train a "zero shot" classifier that performs well for the task.

Application 2: Fine-Tuning Language Models for Copyright-Free Tweets: Identifying Weapons in the Ukraine War

Tweets have become one of the most important sources of text for political scientists but the Twitter terms of use and privacy concerns greatly limit researchers' ability to share or publish their collected tweets. This is an especially significant concern in situations where researchers are collecting expensive annotations on tweets to serve as training data. Researchers generally provide tweets to other researchers in "dehydrated" form, consisting of tweet IDs. Other researchers then "rehydrate" the tweets by querying the Twitter API for the original tweet and merge with the original annotation. If, in the meantime, the original tweet has been deleted, the researcher will not be able to obtain the tweet. While the rate of deletion is generally below 5%, in some politically significant situations, such as tweets related to Brexit, the rate of deletion can be as high as 33% (Bastos 2021). Some research exists on generating synthetic tweets for privacy preservation (Ororbia II, Linder, and Snoke 2018) but not on creating domain-specific synthetic tweets that are difficult to distinguish from real tweets.

As political scientists begin to collect more data about the ongoing war in Ukraine (Zhukov 2022), supervised models trained on tweets will provide an important source

[5]According to the PLOVER manual, "ASSAULT events are deliberate actions which can potentially result in substantial physical harm' ', including military assaults, attacks, kidnappings, terrorist attacks, ethnic cleansing, torture, beatings, etc.

[6]I use an SVM without tuning the hyperparameters for simplicity and computational ease. Better tuning or more sophisticated models would outperform the SVM, but the interest here is in the relative, not absolute performance.

of information. Transparency and reproducibility will both be greatly helped by releasing the raw training data used to train supervised learning models on tweets. This application shows that synthetic tweets can be generated that are difficult to distinguish from real tweets. A named entity recognition system trained on synthetic tweets reaches the same accuracy as one trained on real tweets, but requires approximately 50% more annotations to reach the same performance.

I first collect a set of around 20,000 real tweets from four Twitter accounts that report detailed information on the fighting in Ukraine.[7] Because the synthetic tweets should closely match the actual tweets and because tweets are more difficult to prompt than news articles, which have a convenient headline–body structure, I opt for a fine-tuning approach to text generation. I fine-tune a large language model, specifically GPT-2, on this set of tweets to generate a model that is tailored to generating tweets about the conflict.[8] By fine-tuning, I can both ensure that the generated text is similar to real tweets about the war in Ukraine, as well as eliminating the need to provide a specific prompt to generate text.

Several hyperparameters control GPT-2's text generation, including the "temperature", "top K", and "top P", which control whether to select a high-probability next token (leading to simple, repetitive tweets) or low probability (leading to more creative but potentially nonsensical tweets).[9]

I introduce a simple adversarial procedure for selecting the best hyperparameters for generating tweets, drawing on the intuition that the harder it is for a classifier to distinguish between real and synthetic tweets, the higher the quality of synthetic tweets. For each of the 56 combinations of generation hyperparameters I consider, I generate 1,000 synthetic tweets. For each batch of synthetic tweets, I sample an additional 1,000 real tweets and split the corpus into a training set (75%) and an evaluation set (25%). I train an SVM classifier to discriminate between real and synthetic tweets by attempting to predict if a tweet is real or synthetic.[10] The classification accuracy varies from 0.98 to a low of 0.64, indicating a large effect of hyperparameters of tweet generation. See the Appendix for details.

Table 1 reports a random sample of synthetic tweets using the best hyperparameters (that is, the ones producing the lowest accuracy for the discriminator model).

I then annotate 1,000 tweets with span-level labels on the specific weapons systems described in the tweets: 200 real tweets and 600 synthetic tweets to serve as training data and an additional 200 real tweets as evaluation data. I train a named entity

[7]Specifically, @uaweapons, @osinttechnical, @oryxspioenkop, and @markito0171.

[8]I use Huggingface's `transformers` library to fine-tune GPT-2 (Wolf et al. 2020). I estimate 20 hours of GPU usage were used for this paper. Using standard New England electricity carbon intensity, this would produce around 2kg of CO_2 emissions, or 8km driven by an average US passenger car. https://mlco2.github.io/impact/#compute.

[9]The full details of these parameters are beyond the scope of this paper. For an overview see Platen (2020).

[10]I use scikit-learn's SVM implementation (Pedregosa et al. 2011). Alternatively, a researcher could assess the synthetic tweets' quality by fitting a structural topic model with the tweet's real/synthetic label as a covariate and examining the difference in topics.

4. [SYNTH] The system is relatively good at engaging low/medium armored targets, like BTRs, MT-LBs, APCs and SPGs
5. [SYNTH] I think people got the wrong impression from today's press conference, where Lukashenko said "I do not fear Western military threats but Russia is prepared to pay a heavy price for any military action.
7. [SYNTH] This is mostly because air defence is weak, and even non TB2s could get shot down. Only a very few aircraft flew today, with the majority of them from the western part of Ukraine. In the north of Ukraine a lack of TB2s has caused large losses. The Ukrainians are probably using the drones to spot artillery strikes.
8. [SYNTH] Tanks on the other side of the Irpin River

Table 1: Selected synthetically generated tweets from a random sample of 10 generated from a GPT-2 model fine-tuned on 20,000 tweets reporting open source intelligence on the war in Ukraine. See the appendix for the full list of 10 randomly selected tweets. Due to Twitter's restrictions on including actual tweets in published work, no comparison is provided for real tweets. GPT-2 generation parameters: top_p= 0.90, top_k= 50, temperature= 1.5, epochs= 3

recognition model to identify mentions of specific weapons in the text.[11] Figure 2 reports the accuracy (span-based F1 score) for the model trained on actual and synthetic tweets at different training set sizes.[12]

Although the face validity of the generated tweets is good, a hand-annotated synthetic tweet only 2/3 as valuable as a labeled actual tweet when training an NER system to recognize weapons in tweets.

Application 3: Synthetic data for zero-shot classifiers–training a sentence-level populist classifier

As attention to populist parties has grown, so too has the methodological work on identifying populism in text (Rooduijn and Pauwels 2011; Dai and Kustov 2022). Recent work has proposed training sentence-level supervised classifiers to recover manifesto-level populist labels, given that no dataset exists that labels populist speech at the sentence level (Di Cocco and Monechi 2021). This approach has been criticized, for, among other things, for relying on document-level labels to train a sentence-level classifier when most sentences in a populist party's manifesto will not be recognizably populist (Jankowski and Huber 2022). This application shows how sentence-level populist text across 25 languages can be generated using a *task description* approach.

[11] I use spaCy 3.1.2's small `en_core_web_sm` model as a base and the default training values set by Prodigy (Honnibal and Montani 2017; Montani and Honnibal 2018). Better absolute performance could be achieved with a larger model, but I expect the relative performance to be the same.

[12] I.e., precision = proportion of identified named entities that are correct and recall = proportion of named entities identified by the model.

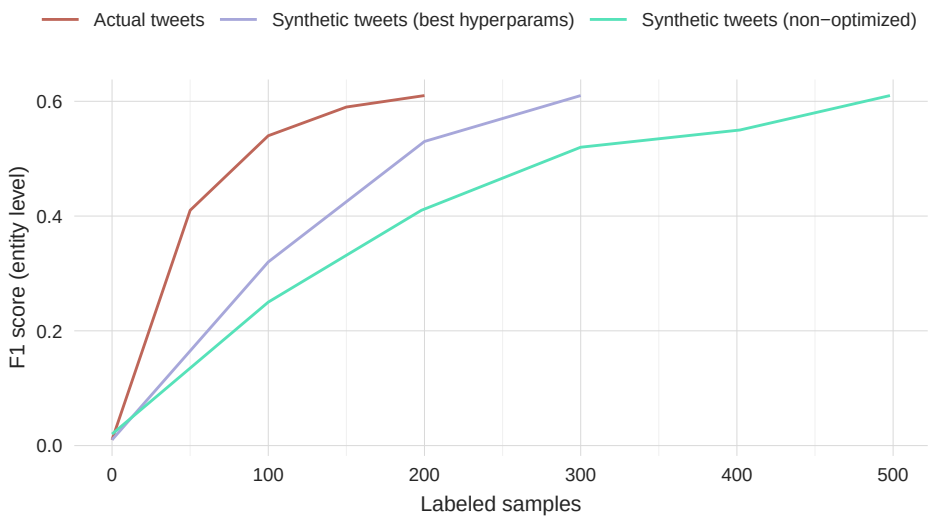

Figure 2: Performance of a named entity recognition model detecting a WEAPON class, trained on annotated actual tweets and annotated synthetic tweets. A model trained on annotated synthetic tweets requires around 500 examples to reach the performance of a model trained on 200 annotated actual tweets.

The synthetic data can be used to train a classifier that identifies populist statements in real manifesto text from European political parties.

Generating synthetic manifesto text is a more challenging task than generating news text or tweets. First, political manifestos are a much rarer form of text than news stories, making it more difficult for the model to accurately generate text. Second, manifestos cannot be easily prompted with headlines in the same way that news stories are.

Rather than relying on relatively simple models like GPT-2, instead we can use GPT-3, a model with around two orders of magnitude more parameters and training data (Brown et al. 2020). In contrast to GPT-2, GPT-3 can use tasks descriptions provided as part of the prompt that describe what the desired text output is. This allows researchers to incorporate a description of political concepts to generate relevant text. For example, given the italicized prompt, GPT-3 generate the synthetic text that follows:

> 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 ten statements that a populist party in Canada might make (in the English language).*
>
> [SYNTH] 1. We need to put Canadian values and interests first, and not let the elites in Ottawa or other countries dictate what's best for us.

[SYNTH] 2. We need to get tough on crime and immigration, and make sure that everyone who comes to Canada is following our laws.

[SYNTH] 3. We need to put the people back in charge of politics.

Because GPT-3 is much larger than previous models and was trained on a multilingual corpus, we can also generate synthetic data in non-English languages, even with an English-language prompt. In this case, we are generating negative examples for our populism classifier, given the italicized prompt:

PROMPT *A political party supports existing political institutions, multiculturalism, globalization, and respect for the existing political process. It supports equality for all people and welcomes immigrants.* `\n\n` *Write ten statements that this party in Finland might make (in Finnish):*

[SYNTH] 1. Uskomme, että markkinatalous tuottaa eniten hyvinvoinnin kaikille, kunhan sen sopeudumiskyky huomioidaan.

[SYNTH, translated] 1. We believe that a market economy produces the most well-being for all, as long as its adaptability is taken into account.

By varying GPT-3's generation hyperparameters and providing prompts for each country and its associated language(s), I collect 1947 synthetic populist and non-populist sentences covering the countries covered by the Comparative Manifesto Project.

I then train a classifier on the synthetic data, specifically a fine-tuned multilingual transformer model,[13] assuming that the sentence pseudo-labels are correct. The model achieves an average F1 score of 0.83 on the synthetic data, ranging from 0.97 for Portugal to 0.4 for Iceland.

I then apply the newly trained classifier to the actual text of manifestos provided by the Manifeso Project (Volkens et al. 2021). Doing so provides some face validity for the method. Given all manifesto sentences from the United Kingdom Independence Party, a populist party, the model identifies the three sentences as having the highest populism scores:

- "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."

The first two sentences reflect a common understanding of populism, while the third is off-topic. The following three sentences are the sentences with the lowest predicted populism scores:

[13] Specifically, XLM-RoBERTa-base using the Huggingface `transfomers` library.

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

None of these low probability sentences are related to populism, providing more support to the claim that our classifier has learned to identify populist rhetoric at the sentence level without any human annotated training data at all. This is known as a zero-shot classifier because no labeled actual text was used to train it.

Evaluating the quality of the sentence-level classifier is difficult given that no hand-annotated data on populism labels exists.

Conclusion

This paper argues that three of the obstacles to supervised text analysis in political science, the problems of labeling, retrieval, and copyright, can be addressed in part by generating synthetic text with the content or style that a researcher desires. Different applications will call for different approaches to generating synthetic text, including using off-the-shelf language models, fine-tuning language models, or using very large models that can be prompted with directions about the desired text. A new adversarial model helps researchers select the optimal hyperparameters to generate synthetic text that is difficult to distinguish from real text. Each of these approaches is illustrated in three short applications from political science, demonstrating that synthetic text can address the retrieval and copyright issues, and sometimes the problem of labeling, albeit with some penalty in accuracy.

A researcher might wonder whether the step of generating synthetic text is needed at all. If a large language model can reliably generate text with a desired label, it should also be able to apply that label directly to actual text (Ornstein, Blasingame, and Truscott 2022). While models are likely to improve greatly in the near future, the synthetic text approach has several benefits over a zero shot classifier approach. First, many of the best models, such as GPT-3, are hosted by third parties that require payment for their use. While still cheaper than hand-coding a large corpus of text (Ornstein, Blasingame, and Truscott 2022), paying to obtain annotations on a large corpus of text can quickly become expensive. Second, hosted models change rapidly, raising difficulties for future researchers in replicating earlier work. Finally, and most significantly, zero-shot models are often opaque and difficult to evaluate. By using large language models to generate synthetic text and using more well understood models to do the classification step, including bag-of-words models, researchers can evaluate the quality of the generated text and employ classifiers that are faster to run and easier to understand.

This approach to generating synthetic text is applicable to a wide range of tasks. Future work can explore the use of synthetic text to evaluate pre-analysis plans for analyzing free-form text in survey responses (e.g. Wood-Doughty, Shpitser, and Dredze 2021), to allow greater transparency in interviews or field notes while preserving privacy, and in developing improved techniques for guiding the content and quality of the synthetic text.

Acknowledgements

This work was first presented as a poster at PolMeth 2022 at Washington University in St. Louis. Thank you to Adam Lauretig, Erin Rossiter, and Brandon Stewart for helpful comments on the poster. Thanks also to the MIT SuperCloud and Lincoln Laboratory Supercomputing Center for providing the computing resources (Reuther et al. 2018).

This work grew out of two ongoing collaborations. The event classification application uses hand-annotated data from a project with Benjamin Bagozzi, Phil Schrodt, Andy Beger, Grace Scarborough. The populism application draws on ongoing substantive work with Shahryar Minhas, Christian Houle, Nicolas Bichay.

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

References

Bastos, Marco. 2021. "This Account Doesn't Exist: Tweet Decay and the Politics of Deletion in the Brexit Debate." *American Behavioral Scientist* 65 (5): 757–73.

Beieler, John, Patrick T Brandt, Andrew Halterman, Erin Simpson, and Philip A Schrodt. 2016. "Generating Political Event Data in Near Real Time: Opportunities and Challenges." In *Computational Social Science*, edited by R. Michael Alvarez. Cambridge University Press.

Brown, Tom B, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, et al. 2020. "Language Models Are Few-Shot Learners." *Advances in Neural Information Processing Systems* 33 (1877–1901).

Caliskan, Aylin, Joanna J Bryson, and Arvind Narayanan. 2017. "Semantics Derived Automatically from Language Corpora Contain Human-Like Biases." *Science* 356 (6334): 183–86.

Dai, Yaoyao, and Alexander Kustov. 2022. "When Do Politicians Use Populist Rhetoric? Populism as a Campaign Gamble." *Political Communication*, 1–22.

Dathathri, Sumanth, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2020. "Plug and Play Language Models: A Simple Approach to Controlled Text Generation." *ICLR*.

Di Cocco, Jessica, and Bernardo Monechi. 2021. "How Populist Are Parties? Measuring Degrees of Populism in Party Manifestos Using Supervised Machine Learning." *Political Analysis*, 1–17.

Doddington, George R, Alexis Mitchell, Mark A Przybocki, Lance A Ramshaw, Stephanie Strassel, and Ralph M Weischedel. 2004. "The Automatic Content Extraction (ACE) Program-Tasks, Data, and Evaluation." In *LREC*, 2:1.

Halterman, Andrew, Katherine A Keith, Sheikh Muhammad Sarwar, and Brendan O'Connor. 2021. "Corpus-Level Evaluation for Event QA: The IndiaPoliceEvents Corpus Covering the 2002 Gujarat Violence." *Findings of the Association for Computational Linguistics*.

Halterman, Andrew, Philip A Schrodt, Andreas Beger, Benjamin E Bagozzi, and Grace Scarborough. 2022. "PLOVER and POLECAT: A New Political Event Ontology and Dataset." *Working Paper*.

Honnibal, Matthew, and Ines Montani. 2017. "SpaCy 2: Natural Language Understanding with Bloom Embeddings, Convolutional Neural Networks and Incremental Parsing." *To Appear*.

Jankowski, Michael, and Robert A Huber. 2022. "When Correlation Is Not Enough: Validating Populism Scores from Supervised Machine-Learning Models."

Kreps, Sarah, R Miles McCain, and Miles Brundage. 2022. "All the News That's Fit to Fabricate: AI-Generated Text as a Tool of Media Misinformation." *Journal of Experimental Political Science* 9 (1): 104–17.

Li, Jianfu, Yujia Zhou, Xiaoqian Jiang, Karthik Natarajan, Serguei V. S. Pakhomov, Hongfang Liu, and Hua Xu. 2021. "Are Synthetic Clinical Notes Useful for Real Natural Language Processing Tasks: A Case Study on Clinical Entity Recognition." *Journal of the American Medical Informatics Association : JAMIA*.

Miller, Blake, Fridolin Linder, and Walter R Mebane. 2020. "Active Learning Approaches for Labeling Text: Review and Assessment of the Performance of Active Learning Approaches." *Political Analysis*, 1–20.

Montani, Ines, and Matthew Honnibal. 2018. "Prodigy: A New Annotation Tool for Radically Efficient Machine Teaching." *Artificial Intelligence* to appear.

Mueller, Hannes, and Christopher Rauh. 2017. "Reading Between the Lines: Prediction of Political Violence Using Newspaper Text." *American Political Science Review*, 1–18.

Ornstein, Joseph T., Blasingame Elise N., and Jake S. Truscott. 2022. "How to Train Your Stochastic Parrot: Deep Language Models for Political Texts." *PolMeth Conference Paper*.

Ororbia II, Alexander G, Fridolin Linder, and Joshua Snoke. 2018. "Using Neural Generative Models to Release Synthetic Twitter Corpora with Reduced Stylometric Identifiability of Users." *arXiv Preprint arXiv:1606.01151*.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, et al. 2011. "Scikit-Learn: Machine Learning in Python." *Journal of Machine Learning Research* 12: 2825–30.

Platen, Patrick von. 2020. "How to Generate Text: Using Different Decoding Methods for Language Generation with Transformers." *Hugging Face Blog*.

Prabhumoye, Shrimai, Alan W Black, and Ruslan Salakhutdinov. 2020. "Exploring Controllable Text Generation Techniques." *arXiv Preprint arXiv:2005.01822*.

Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. "Language Models Are Unsupervised Multitask Learners," 9.

Reuther, Albert, Jeremy Kepner, Chansup Byun, Siddharth Samsi, William Arcand, David Bestor, Bill Bergeron, et al. 2018. "Interactive Supercomputing on 40,000 Cores for Machine Learning and Data Analysis." In *2018 IEEE High Performance Extreme Computing Conference (HPEC)*, 1–6. IEEE.

Rooduijn, Matthijs, and Teun Pauwels. 2011. "Measuring Populism: Comparing Two Methods of Content Analysis." *West European Politics* 34 (6): 1272–83.

Volkens, Andrea, Tobias Burst, Werner Krause, Pola Lehmann, Nicolas AND Regel Matthieß Theres AND Merz, Bernhard Weels, and Lisa Zehnter. 2021. "The Manifesto Data Collection. Manifesto Project (MRG/CMP/MARPOR)." *Version 2021a, Berlin: Wissenschaftszentrum Berlin Für Sozialforschung (WZB)*. https://doi.org/https://doi.org/10.25522/manifesto.mpds.2021a.

Wolf, Thomas, Julien Chaumond, Lysandre Debut, Victor Sanh, Clement Delangue, Anthony Moi, Pierric Cistac, et al. 2020. "Transformers: State-of-the-Art Natural Language Processing." In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45.

Wood-Doughty, Zach, Ilya Shpitser, and Mark Dredze. 2021. "Generating Synthetic Text Data to Evaluate Causal Inference Methods." *arXiv Preprint arXiv:2102.05638*.

Yogatama, Dani, Cyprien de Masson d'Autume, and Lingpeng Kong. 2021. "Adaptive Semiparametric Language Models." *Transactions of the Association for Computational Linguistics* 9: 362–73.

Zellers, Rowan, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. "Defending Against Neural Fake News." In *Advances in Neural Information Processing Systems 32*.

Zhukov, Yuri M. 2022. "VIINA: Violent Incident Information from News Articles on the 2022 Russian Invasion of Ukraine." In *Ann Arbor: University of Michigan, Center for Political Studies*. https://github.com/zhukovyuri/VIINA.

Appendix

Headlines for prompting ASSAULT events

Synthetic news stories describing ASSAULT stories were prompted by using around 40 headlines (see below). To increase the geographic and stylistic diversity of the corpus, I followed the following process. For each headline, I sampled five cities from the list below to generate a dateline and varied the news source. For each unique (headline, city, source) combination, I then generated five random stories.

Negative examples were generated from headlines prompting other event types (protest, meeting, etc.), not shown for space reasons.

cities = {"Abuja", "Kabul", "Belgrade", "Zagreb", "Khartoum", "Vienna", "Dhaka", "Brussels", "Minsk", "Kinshasa", "Beijing", "Bogota", "Sao Paulo", "Havana", "Berlin", "Prague", "Moscow", "Washington", "Cairo", "Jerusalem", "Delhi", "Tehran", "Rome", "Amman", "Beirut", "Tokyo", "Nairobi", "New York", "Panama City", "Oslo", "Damascus", "Bangkok", "Istanbul", "London", "Abu Dhabi"}

sources = {"Reuters", "AFP", "(local sources)"}

```
"Activist held for ransom by rebel forces"
"Plane hijacked in suspected terrorist act"
"Police beat dozens of protestors"
"Syrian military tortured prisoners, new report states"
"Evidence of torture uncovered in government prison"
"Two prisoners put to death"
"Execution carried out in Iranian prison"
"Two women raped in capital"
"New reports of sexual violence in ongoing conflict"
"Al Qaeda leader killed in targeted drone strike"
"Iranian scientist assassinated"
"Russian dissident killed in apparent assassination"
"Two civilians hacked to death with machetes in rural area"
"Buildings destroyed in shelling"
"Israeli bulldozers clear houses"
"Four shot in ongoing gun battle"
"Police fire on demonstrating crowd"
"Truck bomb kills three soldiers"
"Explosions rock city"
"Landmine kills three"
"Bomb detonates in downtown capital"
"Dozens killed in suicide bombing"
"Suicide bomber kills three"
"Man detonates explosive vest at checkpoint"
"US tanks and armored vehicles begin assault in Iraqi town"
```

```
"Heavy artillery shelling continues"
"Police disperse protest with water cannons and tear gas"
"Police fire weapons in the air to disperse mob"
"Indian police use lathi charge to break up protest"
"Hundreds expelled from homes in ethnic cleansing"
"Ethnic cleansing ongoing in conflict"
"Serbian forces expel Bosnians from villages in cleansing operation"
"Civilians slaughtered in massacre"
"Syrian air force uses chemical weapons against civilians"
"Four killed in sarin gas attack"
"Anthrax attack infects three"
"Four killed in air strike"
"War planes pummel rebel positions"
"Allied aircraft enforce no-fly-zone, shooting down Iraqi fighter plane"
"Air Force UAV destroys enemy targets"
"Drone strikes increase as conflict intensifies"
"Man hacked to death with machete"
"Angry mob throws rocks and bottles"
"Local opposition leader beaten with baseball bat"
"Terrorist group releases poison gas, killing three"}
```

Selecting generation hyperparameters for synthetic tweets

I generate synthetic tweets using 56 combinations of parameters:

- epoch $\in \{1, 3\}$
- top_p $\in \{0.8, 0.90, 0.95, 0.99\}$
- temperature $\in \{0.3, 0.5, 0.7, 1, 1.3, 1.5, 1.8\}$
- top_k $\in \{50\}$ (keep fixed)

Although the discriminator approach is useful for this application, it is not applicable to the other two applications. Applications 1 and 2 are trying to generate text with a specific label–either an assault event or populism–while the Twitter application is solely intended to generate text that matches a specific corpus of tweets. If applied to the other applications, any ability to discriminate between real and synthetic could result either from poor quality synthetic text, or because the synthetic text contains a rare class (assault or populism).

Sample of generated tweets

The first ten tweets generated by a fine tuned GPT-2 model using the optimal generation hyperparameters obtained from the process described above: top_p= 0.90, top_k= 50, temperature= 1.5, training epochs=3.

1. [SYNTH]Yes, this is the USS Abraham Lincoln (DDG 71), which was in the Mediterranean a few weeks ago. And yes, you absolutely need to have a GoFund Ukraine account to donate money to charity organizations, including The Red Cross.
2. [SYNTH] An Osint Bunker article detailing the recent Russian buildup near Kyiv
   You can click on the link to be taken to the actual map
3. [SYNTH] Russian military is getting massed in Belarus, most likely to set up new staging areas.
4. [SYNTH] The system is relatively good at engaging low/medium armored targets, like BTRs, MT-LBs, APCs and SPGs
5. [SYNTH] I think people got the wrong impression from today's press conference, where Lukashenko said "I do not fear Western military threats but Russia is prepared to pay a heavy price for any military action.
6. [SYNTH] And as usual people are falling for the bait and trying to equate this to some sort of new high crime, especially when we see examples of this already in the news.
   We already saw it yesterday with @Nrg8000
7. [SYNTH] This is mostly because air defence is weak, and even non TB2s could get shot down. Only a very few aircraft flew today, with the majority of them from the western part of Ukraine. In the north of Ukraine a lack of TB2s has caused large losses. The Ukrainians are probably using the drones to spot artillery strikes.
8. [SYNTH] Tanks on the other side of the Irpin River
9. [SYNTH] Russian forces pushed back from Kharkiv tonight
   Kherson
   Oblast
10. It doesn't even have infrared sensors - only a SINCGARS system. This basically tells you what its main purpose is.

Table 2: Synthetically generated tweets from a GPT-2 model fine-tuned on 20,000 tweets reporting open source intelligence on the war in Ukraine. Due to Twitter's restrictions on including actual tweets in published work, no comparison is provided for real tweets. GPT-2 generation parameters: top_p= 0.90, top_k= 50, temperature= 1.5, epochs=3

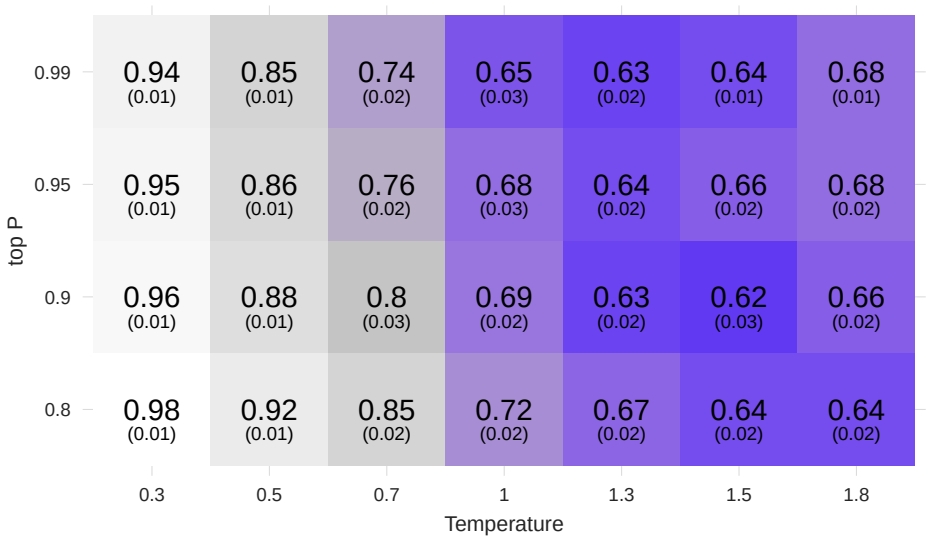

Figure 3: Effect of GPT-2 generation hyperparameters on synthetic tweet quality. Accuracy is the mean out-of-sample accuracy of an SVM classifier trained to discriminate between real and synthetic tweets. Lower accuracy implies better synthetic tweet quality, as the model has a harder time distinguishing real and synthetic tweets. Numbers in parentheses report standard deviation across 10 runs (varying the sampled real tweets and the train/test split).