# Implications of Large Language Models for Political Conflict Forecasting

Philip A. Schrodt

Parus Analytics LLC
Charlottesville, VA
http://philipschrodt.org
http://eventdata.parusanalytics.com

2024 PaCE Conflict Forecasting Workshop
7-8 March 2024
Trinity College Dublin

PARUS
ANALYTICS

# Outline of presentation

- Considering their novelty and massive resource requirements, these models are surprisingly accessible
- Some lessons-learned from an additional year of work on the PLOVER/POLECAT data set
- Quality and specificity is now more important than quantity
- Dream project for conflict modeling: event-code Wikipedia, then use existing LLM masking models to build a predictive system
- Finishing with wild-ass speculation: the existing hardware will be disrupted because it is demonstrably horribly inefficient!

# A couple core caveats

These technologies are changing astonishingly fast—well, sort of. . .—and any specifics on the major models and resource needs/costs will change within months if not days

That said, after the "Big Bang" with the introduction of ChatGPT and its immediate successors (GPT4 (Open AI/Microsoft), Bing (Google), Llama (Meta)), we've now entered a period of incrementalism

Two useful aphorisms:

- Wayne Gretzky Principle: You skate to where the puck is going to be, not where it is.
- "If an elderly [decidedly] but distinguished [only enough to get invited here] scientist says that something is possible, he is almost certainly right; but if he says that it is impossible, he is very probably wrong." Arthur C. Clark

# These models are phenomenally resource intensive

- This is a really serious disruption to our research expectations since thanks to Moore's Law we've become accustomed to running almost everything on a laptop or occasionally with an inexpensive cloud account
- Estimating core models is a multi-million-dollar proposition just in computing time and energy
- Specialized hardware is also available only to a small number of companies—Open AI is now floating the idea of raising $7-*trillion* for LLM/AGI research—and is in scarce supply
- Required training data for the foundation models involves essentially scraping the entire web

# But it is still remarkably accessible! - 1

- These are all based on a "transformer" paradigm where the massive foundation models are fine-tuned with a much smaller number of targeted training cases.
- In our POLECAT work, a single GPU is sufficient to do work in reasonable amounts of time (but you need a GPU: I burned out an iMac trying to work on just a CPU!)
- HuggingFace, an open-source company valued in August-2023 at a mere $4.5-billion, hosts more than 500,000 open-source models
- Google's "Colaboratory" provides GPU hardware instances for $10/month (!!). Numerous other cloud providers are also available and apparently can get some of the chips

# But it is still remarkably accessible! - 2

- The big companies have clearly adopted a mixed strategy—guess they paid attention in the lectures on zero-sum games!—with respect to models and methods being closed vs open sources
- Technical papers from the major companies are available almost immediately on the web without social-science-style publication delays and paywalls
- Large amounts of post-"Big Bang" work in both industry and academia has focused on making models smaller and more efficient
- Decentralized support via Slack Overflow and the like is quite good in my experience; I have also heard that the new GPT-based programming "assistants" are quite good

# Why massive neural networks could be a game changer for conflict forecasting

- The models have billions of parameters which incorporate extraordinary amounts of information
- The models seem to generalize quite well without a lot of fine tuning
- They are trained by example and are language-neutral
- Like human experts, they are sequence-oriented and are "chunkers" that generalize automatically
- Having been trained on Wikipedia and fiction, they incorporate social "common sense"
- Which leads to the sometimes eerie "zero shot learning": they provide credible answers to questions (and/or languages) they haven't been trained on

# PLOVER/POLECAT event data system

The Ur-documents

- Halterman, Andrew, Philip A. Schrodt, Andreas Beger, Benjamin E. Bagozzi and Grace I. Scarborough. 2023. Creating Custom Event Data Without Dictionaries: A Bag-of-Tricks. Working paper presented at the International Studies Association, March-2023. https://arxiv.org/abs/2304.01331
- Halterman, Andrew, Benjamin E. Bagozzi, Andreas Beger, Philip A. Schrodt, and Grace I. Scarborough. 2023. PLOVER and POLECAT: A New Political Event Ontology and Dataset. Working paper presented at the International Studies Association, March-2023. https://osf.io/preprints/socarxiv/rm5dw

These have almost all the information—around 250 pages —needed to create a comparable system; our sponsor allowed us release virtually everything except a few very specific procedure calls that will change anyway.

For the usual IP reasons, the training examples are not available but like any example-based system, these are easily reverse-engineered from open data (which is on Dataverse already) provided a set of texts covering the same period is available.

# PLOVER

Political Language Ontology for Verifiable Event Records

Event, Actor and Data Interchange Specification

Open Event Data Alliance

`http://openeventdata.org/`

DRAFT Version: 1.3

3 August 2023

# Experiments after ISA-23

- Extended refinement of the training cases for the categories
- Experiments with multiple metrics for evaluating quality
  - Metrics within the full training set (N in range [350,500])
  - Metrics on a "canonical" pure out-of-sample expert-annotated evaluation set with a 1:2 positive:negative ratio (precision is quite sensitive to this ratio; recall is not) (N in range [30,300])
  - Metrics on a "balanced" pure out-of-sample expert annotated evaluation set with event frequencies, including stories containing no events, roughly following the frequencies generated by Factiva (N=1024)
- Experiments with various alternative BERT transformer models

# Some illustrative accuracy figures

- Best cases (which are high frequency): e.g. PROTEST, COERCE, ASSAULT
  - All training cases: `Acc: 0.9276 Prec: 0.9801 Recall: 0.9049 F1: 0.9410`
  - Canonical evaluation: `Acc: 0.7662 Prec: 0.7213 Recall: 0.8713 F1: 0.7892`
  - Balanced evaluation: `Acc: 0.6851 Prec: 0.0588 Recall: 0.8333 F1: 0.1099`
- Medium cases : e.g CONSULT, AID, MOBILIZE
  - All training cases: `Acc: 0.9435 Prec: 0.9465 Recall: 0.9388 F1: 0.9426`
  - Canonical evaluation: `Acc: 0.7296 Prec: 0.7471 Recall: 0.6771 F1: 0.7104`
  - Balanced evaluation: `Acc: 0.8931 Prec: 0.4375 Recall: 0.6848 F1: 0.5339`
- Low cases: e.g. AGREE, ACCUSE, REJECT
  - All training cases: `Acc: 0.8439 Prec: 0.8431 Recall: 0.8866 F1: 0.8643`
  - Canonical evaluation: `Acc: 0.6397 Prec: 0.4694 Recall: 0.5000 F1: 0.4842`
  - Balanced evaluation: `Acc: 0.7687 Prec: 0.0167 Recall: 0.5714 F1: 0.0325`

"best/medium/low" determined by F1 on the canonical and balanced

# Overall performance

- The standard metrics—accuracy, precision, recall, and F1—vary quite substantially across estimation runs, primarily due to the random train/test partition (variation is typically 0.05 but as high as 0.20) but also due to round-off variation in the estimation and inference: this seriously increases the time required to evaluate approaches
- Expert curating of training cases typically improved metrics 0.05 to 0.10
- Recall is generally very good; precision much less so.
  - This is the opposite of dictionary-based models—notably ICEWS—which have good precision but very poor recall, at least in our measures
  - Precision could be increased by finding some precision/recall tradeoff, employing some explicit rare events approaches, or adding dictionary-based filters: see supplemental slides

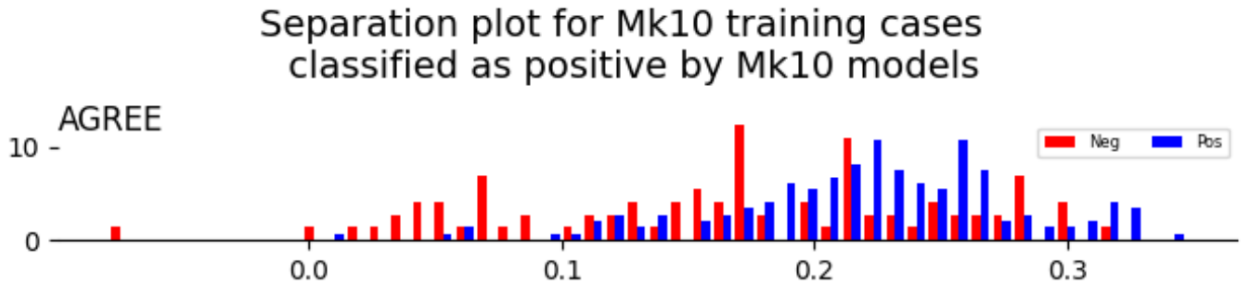
# Some experiments that had limited effects

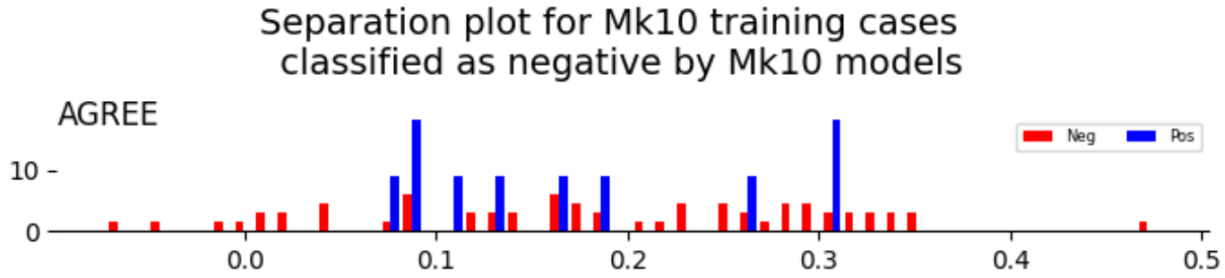
- Choice of models in the BERT family makes little difference
- "Semi-supervised" removal of outlying training cases based on the FP/FN cases as predicted by a model had inconsistent effects
- Getting sufficient out-of-sample evaluation cases for rare categories is problematic: to get 100 positives in a balanced set for a category such as AGREE, SANCTION or REJECT with around 1% incidence, you need 10,000 curated cases, 5,000 of those positive.

  (COOPERATE, CONCEDE, and SUPPORT are even lower, $\approx$0.5%)

# Models don't seem to separate the cases very well

so just changing the thresholds will have only limited effects



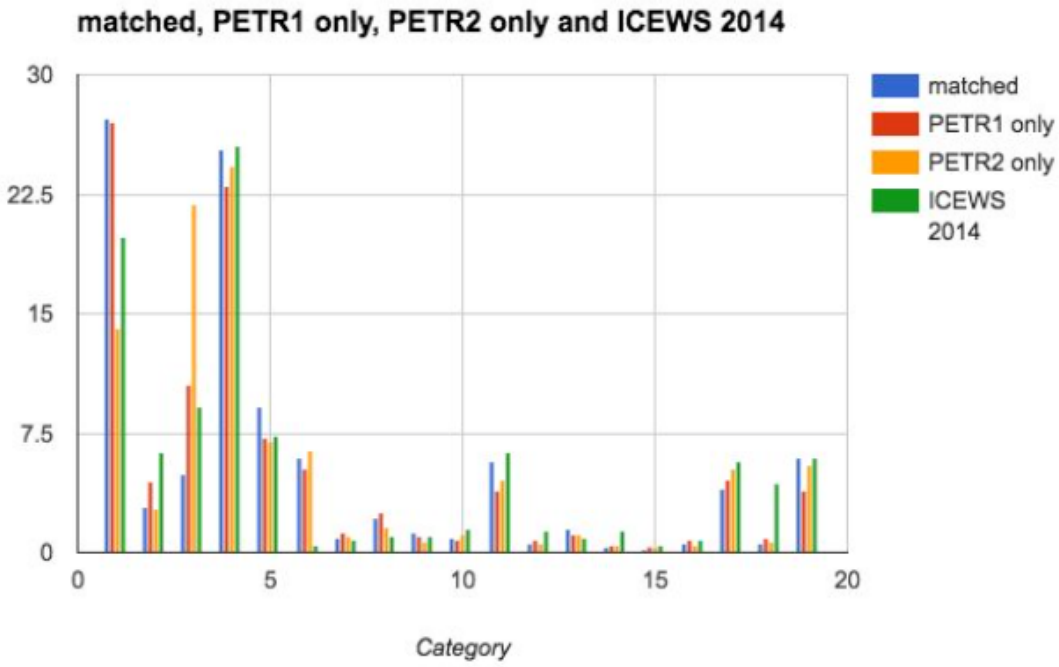




Caveats:

- "Mk10" are the older training cases, before further annotating
- Not all of the categories are this bad

# Accidental discovery

Marginal distribution of events per month in ICEWS is virtually constant over two decades!




This is likely due to filtering by Factiva as part of the contract with the US government, as Factiva provided only a subset of its news stories, and that subset was designed to be focused on political events.

# Quality is more important than quantity

- Approach prior to LLMs: toss as much data as possible at the model, then use reduction-of-dimensionality methods to get the most important characteristics, leaving most of the remaining noise either in lower dimensions or to be sorted out by later statistical methods maximizing signal-to-noise.
- This appears to fail with models at the scale of LLMs, which have billions of parameters, as they "learn" all of the junk as well. This was the conclusion of a major Microsoft paper, and our experience as well
- Our very best model is based on carefully selected and coded PROTEST events based on an existing full-text database with a single—and therefore relatively consistent—expert coder
- Early training cases generated by a heterogeneous group of often unmotivated annotators resulted in lousy models

# Additional quality issues: category definitions

- Since many of the PLOVER events are very rare, even in a long time series, it has been far more difficult to find representative cases
  - We also encountered this issue in the development of CAMEO ca. 2000-2002, and abandoned some categories when we couldn't find examples for the manual.
  - Microsoft and other LLM researchers assert that high quality *synthetic* data, easily produced in quantity with contemporary LLMs (not possible when we started the work) are superior to low-quality real-world examples
- PLOVER is relatively immature and even with two expert coders we don't have complete consensus on the definitions of some of the rare categories
- While LLM-based models are not as sensitive to vocabulary as dictionary-based models, since they deal with synonyms via word and sentence embeddings, categories still need to have reasonably distinct vocabulary

# An interesting possible project...

- Problem facing conflict forecasting: Getting *current* news stories for a predictive model is very easy using web scrapping, which with contemporary software libraries can be done with a few dozen lines of code and at least when I last experimented, some of the major international sources, notably Reuters and Xinhua, could be easily accessed

- Getting *historical* cases, however, is very difficult since these are largely locked down by two companies, Factiva and Lexis-Nexis, and even getting a couple of decades of those stories is quite expensive.

- We nonetheless have an open and readily-downloadable source of historical cases that systematically covers as far back as we have written records: Wikipedia.

# So. . .

- Download Wikipedia, which is straightforward, and in fact in POLECAT actor resolution is done using a Wikipedia download (see ISA-23 papers)
- Code all of the conflict found in Wikipedia—and wow, amateur military historians *love* contributing to Wikipedia!—into PLOVER events
- Train one (or more) of the newer, resource-efficient LLMs on that *event data* using the same methods applied to natural language text, which has far lower dimensionality, and thus presumably resource requirements, than language.
- Use these for your predictive models
- <**snark**> Write up results, attempt to publish but are thwarted by Reviewer #2 who wants to examine the [2-billion] coefficients and their standard errors, leave academia in disgust and immediately secure $300,000/year job with Microsoft/Google/Meta <**/snark**>

# The promised wild-assed conjecture

- The human brain consumes about 0.3 kWh per day, and let's say can respond to a GPT-like query in a minute, so around 0.0002 kWh/query even if the brain were only working on that query (which it isn't: it's also keeping a lot of physical processes in balance, otherwise you die)
- The latest Nvidia neural-network-oriented GPU, the H100, uses about 700W, and let's say a query response uses 10 H100s for ten seconds, so expenditure of 0.194 kWh/query: four orders of magnitude difference
- This hasn't adjusted for memory, where the brain has $10^{14}$ synapses, vs $10^{10}$ parameters in the largest LLMs, probably giving us a couple more orders of magnitude

Any system whose energy use is several orders of magnitude less efficient than an existing architecture is waiting for disruption.

# And it might go a lot deeper than that. . . neural networks are an architecture optimized for prediction

Lisa Feldman Barrett. *Seven and a Half Lessons About the Brain* 2020:

Pattern-based prediction is the *fundamental* purpose of the cognitive system: the evolutionary advantages are simply too great, both in terms of avoiding predators and being a predator, and in allocating a finite energy budget. This got going during the Cambrian period.

In most mammals, and certainly primates, this extends to evaluating and positioning oneself in hierarchies and accounts for a great deal of cognitive effort. (Robert M. Sapolsky. *Behave: The Biology of Humans at Our Best and Worst*, 2017)

# Thank you

Email:
schrodt735@gmail.com

Slides:
`http://eventdata.parusanalytics.com/presentations.html`

Blog with lots of extended commentary on event data:
`http://asecondmouse.org`

# Supplementary slides

# PLOVER events

- AGREE
- CONSULT
- SUPPORT
- CONCEDE
- COOPERATE
- AID
- RETREAT
- REQUEST
- ACCUSE
- REJECT
- THREATEN
- PROTEST
- SANCTION
- MOBILIZE
- COERCE
- ASSAULT

# Examples of PLOVER modes

RETREAT

- withdraw (from territory)
- release (captives)
- return (property)
- disarm
- ceasefire
- access (allow third party access)
- resign

COERCE

- seize
- restrict
- ban
- censor
- martial law
- arrest
- deport
- withhold

# PLOVER contexts (partial list)

- military
- diplomatic
- executive
- legal
- intelligence
- legislative
- political institutions
- pro-democracy
- pro-autocracy
- economic
- reparations

# "New generation event coder"

- Use the massive neural network "transformer models" that have been developed by Google, Amazon, Facebook, and Microsoft and are largely open-sourced
- Use training examples (250 to 500 per category) rather than dictionaries
- Observations are 512-token texts rather than single sentences (this is consistent with Google BERT family of models)
- Approaches for components of PLOVER
  - Events: transformer classification models
  - Mode and context: support vector machine models
  - Actors and locations: transformer "question answering" models linking to open databases such as Wikipedia and Geonames
  - Miscellaneous fine tuning (e.g. compound actors): dependency parsing

# Risks in machine learning models

- Over-fitting
- It is not clear that political early warning has a sufficient number of cases to take advantage of methods which require large amounts of data. In rare events analysis, continuity is critical: we need to be able to generate comparable data across many years and different projects
- ML models are generally atheoretical, and the rich parameter spaces mean it is often difficult to impossible to ascertain the relative importance of independent variables
- Some models—notably "deep learning"—are quite new and may have features we don't fully understand
- In many instances, ML models show only marginal improvements over well-understood methods such as logistic regression when applied across a wide set of out-of-sample problems

# The very finite set of widely used ML methods

- Support vector machines
- Clustering, typically using k-means
- Random forests, a relatively recent ensemble variation on the older method of decision trees. To date, these have been the most successful models, probably because they easily accommodate heterogeneous data
- Genetic algorithms
- Logistic regression, which not infrequently is "embarrassingly effective"
- And...

# Recurrent neural networks (LLM/deep learning)

- These appear to be able to extract pretty much all available signal in a set of data
- They are hugely computationally expensive but now benefit from specialized hardware ("GPU"s) originally developed for accurately rendering splattering zombie brains in video games. Apple's M1/M2 chips have neural network hardware on the chip.
- Neural networks have always been good at dealing with missing data—which they treat as information—and non-linear relationships
- Incredible amounts of work is currently being done with these by organizations with vast resources, and much of this is open source
- LLMs notoriously "hallucinate", a very serious risk in policy-oriented research

# Reflections on the precision vs. recall problem -1

We generally don't have these now-standard machine learning metrics for past human-coded data, instead we mostly have coder accuracy against a codebook and/or other coders. This doesn't tell us very much

- Simple "accuracy" is very easy to achieve for rare events, though a lot of the human-coded "accuracies" were probably more like "precision"
- Recall is exceedingly hard to measure since it requires large numbers of negative cases, whereas human-coding projects tend to focus on positive cases
- Metrics are highly dependent on the ratio of cases in the source data

# Reflections on the precision vs. recall problem - 2

When dealing with human analysts, precision is more important than recall because people notice mistakes in the data they are looking at—false positives—but only rarely notice data that are missing.

- Analysis of "fault trees" for nuclear reactors found if experts were presented with a tree containing scores of possible ways a reactor could fail, they typically added about a dozen new contingencies. Remove half of the branches from the original tree, and they still typically added about a dozen new contingencies.

PLOVER is also a relatively new system and some of the categories may not be sufficiently well-defined, either by the experts or as implemented in the training cases.

# Reflections on the precision vs. recall problem - 3

There are also some key differences between ICEWS/CAMEO and POLECAT/PLOVER

- POLECAT codes entire stories whereas ICEWS codes single sentences, so an event category is more likely to occur tangentially or by inference in a story.
- At least three styles of news stories are common in Factiva:
  - Highly focused "breaking news" stories with a single event
  - Summary stories that may have multiple related events, e.g. a protest, police suppression of the protest, and various actors condemning the suppression
  - "Today's headlines..." stories covering a number of unrelated events (we've been trying to filter these out as they aren't part of the training cases)