<https://www.springboard.com/blog/data-science/machine-learning-gpt-3-open-ai/>

#### Title

\* Journal:

\* Type: [mention research]

\* Link:

\* Keywords / topics: [[NLP ]]

``` abstract

[here you copy-paste abstract]

```

#### key ideas

-

-

-

#### setup

- did they collect data?

- did they run experiments?

- what type of outcome measure? Which models? etc

#### caveats

- mention ideas that you have that might have been done better.

``` conclusions

[here you copy-paste some of the important conclusions]

```

#### Gilardi 2023 ChatGPT Outperforms Crowd-Workers for Text-Annotation Tasks∗

\* Journal: arXiv:2303.15056v1

\* Type: experiment

\* Link: https://arxiv.org/abs/2303.15056

\* Keywords / topics: [[NLP ]]

``` abstract

Many NLP applications require manual data annotations for a variety of tasks,

notably to train classifiers or evaluate the performance of unsupervised models.

Depending on the size and degree of complexity, the tasks may be conducted by

crowd-workers on platforms such as MTurk as well as trained annotators, such as

research assistants. Using a sample of 2,382 tweets, we demonstrate that ChatGPT

outperforms crowd-workers for several annotation tasks, including relevance, stance,

topics, and frames detection. Specifically, the zero-shot accuracy of ChatGPT

exceeds that of crowd-workers for four out of five tasks, while ChatGPT’s intercoder

agreement exceeds that of both crowd-workers and trained annotators for all tasks.

Moreover, the per-annotation cost of ChatGPT is less than $0.003—about twenty

times cheaper than MTurk. These results show the potential of large language

models to drastically increase the efficiency of text classification.

```

#### key ideas

- zero-shot ChatGPT classification (without training) better than MTurk annotation

- temperature par (controls creativity) at .2 leads to inter-coder reliability of >.95

- if temperature at default of 1, it becomes 83%

- After testing several variations, we decided to feed tweets one by one to ChatGPT using the following prompt: “Here’s the tweet I picked, please label it as [Task Specific Instruction (e.g. ‘one of the topics in the instruction’)].”

- coding manual provided with instructions

#### setup

- 2382 tweets (Alizadeh 2022) annotated on:

relevance,

stance regarding section 230 (keep/repeal/neutral);,

topics (six classes);

two kinds of frame detection (content moderation and 14 classes)

- submit tasks to ChatGPT, tested accuracy and intercoder agreement (% overlap)

- of all coded tweets, they selected only those who the human annotators agreed on, and presented these to MTurk and ChatGPT

#### caveats / future work

* as GPT will generate similar texts, it is not surprising that it will always come up with similar annotation. why would it differ? basically, they show that the temperature parameter is critical for high inter-coder reliability (or similarity between different tasks presented to GPT

- from paper: The following questions and steps seem particularly promising: (i) performance of ChatGPT across multiple languages; (ii) performance of ChatGPT across multiple types of text (social media, news media, legislation, speeches, etc.); (iii) implementation of few-shot learning on ChatGPT, compared with fine-tuned models such as BERT and RoBERTa; (iv) of semi-automated data labeling systems in which a model first learns by observing human annotations, and is then used to recommend or even automate the labeling (Desmond et al., 2021)

``` conclusions

```

#### LaFleur 2019 Art is long, life is short: An SDG Classification System for DESA Publications

\* Journal: DESA Working Paper No. 159

\* Type: [mention research]

\* Link: https://www.un.org/esa/desa/papers/2019/wp159\_2019.pdf

\* Keywords / topics: [[NLP ]]

``` abstract

Between the many resolutions, speeches, reports and other documents that are produced each year, the United Nations is awash in text. It is an ongoing challenge to create a coherent and useful picture of this corpus. In particular, there is an interest in measuring how the work of the United Nations system aligns with the Sustainable Development Goals (SDGs). There is a need for a scalable, objective, and consistent way to measure how similar any given publication is to each of the 17 SDGs. This paper explains a proof-of-concept process for building such a system using machine learning algorithms. By creating a model of the 17 SDGs it is possible to measure how similar the contents of individual publications are to each of the goals — their SDG Score. This paper also shows how this system can be used in practice by computing the SDG Scores for a limited selection of DESA publications and providing some analytics.

```

#### key ideas

- correspondence between estimated topics and sdg representative texts high (> 60%)

- A second step to validate the classifier is to examine the most frequent words in each group formed by the LDA algorithm.

#### setup

* classifier trained (LDA?) on representative texts (text for UN webpage that describes SDG and section of annual report UN for 2016, 2017, 2018) and then applied to target data (mostly working papers DESA (for each of 267 documents, they know which UN goal is associated. )
* The first step is to collect 17 unique and balanced texts that each represent one of the SDGs. The LDA algorithm is used to estimate an 18 topic model. The extra topic acts as a filter and should capture the common words among all the 17 representative texts.
* bag of words, 18 topic model trained using Mallet. 1 general topic to gather common terms

#### caveats

- not sure how the classifier was trained as LDA deals with unlabeled data?

``` conclusions

expanded by: technical improvement, making the system fully supervised (LDA?)

number of types publication expanded and classified

visual presentation can be improved, has large impact

```

#### Bowman 2023, Eight Things to Know about Large Language Models

\* Journal: Has not been published

\* Type: Descriptive Study

\* Link: https://arxiv.org/abs/2304.00612

\* Keywords / topics: [[NLP ]]

``` abstract

[The widespread public deployment of large language models (LLMs) in recent months has prompted a wave of new attention and engagement from advocates, policymakers, and scholars from many fields. This attention is a timely response to the many urgent questions that this technology raises, but it can sometimes miss important considerations. This paper surveys the evidence for eight potentially surprising points.

]

```

#### key ideas

1. LLMs predictably get more capable with increasing investment, even without targeted innovation.

2. Many important LLM behaviors emerge unpredictably as a byproduct of increasing investment.

3. LLMs often appear to learn and use representations of the outside world.

4. There are no reliable techniques for steering the behavior of LLMs.

5. Experts are not yet able to interpret the inner workings of LLMs.

6. Human performance on a task isn’t an upper bound on LLM performance.

7. LLMs need not express the values of their creators nor the values encoded in web text.

8. Brief interactions with LLMs are often mis-leading.

Some interesting findings:

* the scaling of these LLM is huge, and consists mainly of using more data and parameters instead of increasing complexity of infrastructure
* techniques to train newer LLM models are no longer disclosed, but it seems there are only slight deviations from original models
* even if model fails consistently at a task, a new model being trained much more often (5to10 scale) will do well at that task (scaling up works!)
* Big Bench (benchmark for LLM, [[srivastava et a]] show different trends making scaling law style predictions unreliable (LLM performs better on scale but not consistently on different tasks)
* After model trained, it is discovered that GPT3 can:
  + few-shot learning: ability to learn new task from examples in single interaction
  + chain-of-thought reasoning: write out reasoning on hard tasks ability
* LLMs seem to develop internal representations of the world, which allow them to reason at level of abstraction that is not sensitive to precise language
  + internal representations of color words represents human color perception
  + make inferences about what author believes of knows and use these to predict how document continues
  + internal representations of objects described in stories, which evolve when more info is revealed
  + give instructions on how to draw new objects
  + learn representation of board game at the state of board at each turn
  + distinguish misconceptions from true facts, how likely a claim is true
  + pass tests of common sense reasoning
  + LLM is not just statistical next-word prediction, as these rich representations of the world develop as they are trained. even more, LLM are augmented by interactive training methods, integration with image processing or other software tools
  + often given additional training using reinforcement learning which trains them to give responses that humans find useful without requiring humans to show this behavior/language (same as techniques to outperform humans on games)

#### important citations

More capable models can better recognize the specific circumstances under which they are trained. Because of this, they are more likely to learn to act as expected in precisely those

circumstances while behaving competently but unexpectedly in others: **sycophancy**, where a model answers subjective questions in a way that flatters their user’s stated beliefs, and **sandbagging**, where models are more likely to endorse common misconceptions when their user appears to be less educated [[perez 2022]]

experts say that it is likely that inability to control LLM will cause human extinction [[stein Perlman, 2020]] or catastroph like nuclear war [[Michael 2022]]. moratorium on LLM until governance in place [[bengio 2023]]

ad-hoc techniques that at first seem to provide insight into the behavior of an LLM are

later found to be severely misleading (Feng et al., 2018; Jain & Wallace, 2019; Bolukbasi et al., 2021; Wang et al., 2022).

a model will fail to complete a task when asked, but will then perform the task correctly once

the request is reworded or reframed slightly, leading to the emerging craft of prompt engineering

many of the most cited research papers dealing with LLMs, including many papers that introduce new methods or theories, are not published in peer-reviewed venues. The recent trend toward limiting access to LLMs and treating the details of LLM training as proprietary information is also an obstacle to scientific study.

#### setup

This is a descriptive study hence they do not have any experiment or data collection process. They provide a literature review and pinpoint several aspects of LLMs. They provide example papers for each point they defend.

#### caveats

``` conclusions

[9.1. We should expect some of the prominent flaws of current LLMs to improve significantly.

9.2. There will be incentives to deploy LLMs as agents that flexibly pursue goals.

9.3. LLM developers have limited influence over what is developed.

9.4. LLMs are likely to produce a rapidly growing array of risks.

9.5. Negative results with LLMs can be difficult to interpret but point to areas of real weakness.

9.6. The science and scholarship around LLMs is especially immature

]

```

#### Hsu, LaFleur and Özerbek 2022, Improving SDG Classification Precision Using Combinatorial Fusion

\* Journal: Sensors 2022

\* Type: [Comparison of different methods]

\* Link: https://ieeexplore.ieee.org/document/9730598/

\* Keywords / topics: [[NLP ]]

``` abstract

[Combinatorial fusion algorithm (CFA) is a machine learning and artificial intelligence

(ML/AI) framework for combining multiple scoring systems using the rank-score characteristic (RSC) function and cognitive diversity (CD). When measuring the relevance of a publication or document with respect to the 17 Sustainable Development Goals (SDGs) of the United Nations, a classification scheme is used. However, this classification process is a challenging task due to the overlapping goals and contextual differences of those diverse SDGs. In this paper, we use CFA to combine a topic model classifier (Model A) and a semantic link classifier (Model B) to improve the precision of the classification process. We characterize and analyze each of the individual models using the RSC function and CD between Models A and B. We evaluate the classification results from combining the models using a score combination and a rank combination, when compared to the results obtained from human experts. In summary, we demonstrate that the combination of Models A and B can improve classification precision only if these individual models perform well and are diverse.]

```

#### key ideas

-Having multiple classification systems available necessarily raises the question of

how well they perform relative to some ground truth. Beyond this question, this paper

is concerned with the additional performance that can be gained by combining diverse

classification methodologies in such a way as to improve the performance beyond any

individual method. The combinatorial fusion algorithm is shown to improve on the

classification precision of both models by combining their results.

#### setup

-Model A uses a machine learning clustering algorithm applied to a carefully selected representative sample of documents to generate a classifier (a Latent Dirichlet allocation (LDA)). Model B uses an ontology of terms and the semantic connections between those terms and the SDGs. This model relies on the Semantic Web. In short, a predetermined ontology of SDG terms formalizes the basic schema of the SDG goal-target-indicator-series hierarchy. This ontology allows the creation of a set of Internationalized Resource Identifiers (IRIs) for the SDGs, targets, and indicators.

-They show an innovative way to merge these two models via CFA and get better results.

-Their evaluate the performance of each model using precision @k, k = 1, 3, 5, and 8. For

each document (or publication in general), a human expert gives a scoring system H. A subset of k elements from a set of data items, denoted as Re(H) consisting of those SDGs which are ranked at top k, is considered as a relevant set of SDGs for the document. For Model A, the precision of A at rank k (pre@k) is the number of elements in the intersection of Re(A) and Re(H) divided by k.

-After that,they extend this analysis to an additional 30 publications chosen at random from the corpus of documents, and show that combined models have performance advantages over individual classification models.

-The dataset used in this analysis comprises 267 texts published by the United Nations

between 1995 and 2019. These include major flagship publications, reports by task teams,

reports of the Secretary General, research notes, reports published by ECOSOC, thematic

policy briefs, a full collection of DESA’s working papers, and other texts.

#### caveats

- External validity question: Can we just merge two methods even when their training data are not the same?

``` conclusions

[In summary, we demonstrate that a combination of the two models can improve each

individual model only if these two models are relatively good (in terms of performance

ratio) and they are diverse (in terms of cognitive diversity). In addition to that, model

fusion using combinatorial fusion algorithms was able to improve not only the prediction

but also the data quality with regard to reproducibility by subject experts.

]

```

#### Kim and LaFleur 2020, What does the United Nations “say” about global agenda? An exploration of trends using natural language processing for machine learning

\* Journal: [United Nations Department of Economic and Social Affairs](https://www.un-ilibrary.org/search?value1=United+Nations+Department+of+Economic+and+Social+Affairs&option1=author&noRedirect=true)

\* Type: [mention research]

\* Link: https://www.un-ilibrary.org/content/papers/10.18356/25206656-171

\* Keywords / topics: [[NLP ]]

``` abstract

[How has the focus of the UN General Assembly changed over time and how well is the global agenda expressed in these documents? This paper presents a proof-of-concept classifier to examine the evolution of the global agenda expressed and observed in words of the UN General Assembly resolutions. Using natural language processing to identify four categories of resolutions — Sustainable Development, Justice and Law, Human Rights, and Peace and Security — the analysis of 3,765 UN GA resolutions from 2007 to 2019 reveals the changing areas of focus of the Member States and, as a result, of the UN Secretariat. Sustainable Development is slowly gaining importance in the language in UN resolutions. ]

```

#### key ideas

- In absolute numbers, each of the four categories saw increases in the number of resolutions between the 62nd session and the 73rd session. Interestingly, the number of resolutions on Sustainable Development has remained constant at 57 in the last four sessions of the GA.

\*\*They have a very nice summary of already existing methods so it was quite beneficial to read.

#### setup

-They use text as data to achieve the research objectives in this paper. The data covers a collection of GA resolutions from session 62 (2007-2008) through session 73 (2018-2019), obtained from the UN Dag Hammarskjöld Library.

-More on training data: The annual Secretary-General report gives them well defined examples of language that describes each of the categories they wish to analyze. They also include the contents of the “what we do” section of the UN website, which is separated by categories. In addition to these two sources, they include recent UN reports published for each category. For example, on the issue of terrorism, they include the recent Secretary-General report on the UN’s global counter-terrorism strategy Finally, they include the content of the websites from each UN Department responsible for each of the categories. Only documents and website contents that clearly fit into one or another category were used to build the model.

-They use Mallet and Python.

#### caveats

-In the training model, they only included documents that would fit into at least one of the categories. Is this a problem if we give the model a text that fits into none?

- If the UN increases the number of resolutions we would naturally expect some increase in all four categories anyways. Does their model have to categorize a paper into at least one of the groups? What is the added value of this research?

-”We report here the performance of the selected classifier when compared to a testing dataset of 21 documents drawn randomly from the set of 211 documents used to train the model.” Is this sample size ok to decide in NLP research?

-They also tried an unlabeled classification method following the methodology described in LaFleur (2019) but it failed to distinguish between results. Does that signal low external validity for this paper?

``` conclusions

[In sum, natural language processing reveals that the UN is active in all four categories, namely, Sustainable Development, Justice and Law, Human Rights, and Peace and Security, with a growing number of resolutions over the last eleven sessions. Sustainable Development is an area that is gaining importance, not just claimed by the UN, but by the evidence from the language in the resolutions.]

```

#### Morales, Jagüey and Becerra-Alonso 2022, A Comparison of Multi-Label Text Classification Models in Research Articles Labeled With Sustainable Development Goals

\* Journal: IEEE Access Volume:10

\* Type: [mention research]

\* Link: https://ieeexplore.ieee.org/document/9954368/

\* Keywords / topics: [[NLP ]]

``` abstract

[The classification of scientific articles aligned to Sustainable Development Goals is crucial for research institutions and universities when assessing their influence in these areas. Machine learning enables the implementation of massive text data classification tasks. The objective of this study is to apply Natural Language Processing techniques to articles from peer-reviewed journals to facilitate their classification according to the 17 Sustainable Development Goals of the 2030 Agenda. This article compares the performance of multi-label text classification models based on a proposed framework with datasets of different characteristics. The results show that the combination of Label Powerset (a transformation method) with Support Vector Machine (a classification algorithm) can achieve an accuracy of up to 87% for an imbalanced dataset, 83% for a dataset with the same number of instances per label, and even 91% for a multiclass dataset.]

```

#### key ideas

- This work proposes a comparative study of one multi-class classifier called One-Versus-Rest (OvR) and three multilabel problem transformation methods (Binary Relevance, Label Powerset, and Classifier Chains), applied to four classification algorithms: Naive Bayes, Logistic Regression , Support Vector Machine, and Random Forest, on balanced and imbalanced datasets.

#### setup

- The scientific article datasets were obtained from a free account of Dimensions: a bibliographic database categorized with United Nation’s Sustainable Development Goals (SDG), managed by Digital Science. Dataset creation with 180,852 scientific papers with title

and abstract from organic agriculture 3.0 domain. These scientific articles with SDG multi-label classification range from January 2015 to August 2021.

- They evaluate the models based on their accuracy, F1 score and hamming loss.

#### caveats

- mention ideas that you have that might have been done better.

``` conclusions

[In dataset 2018, independent of the dataset scenario and transformation method, Support Vector Machine is the best classification algorithm with the best overall classification accuracy performance with a difference of 20% (best 91% and worst 71%). The results present LP as the best transformation method, independently of the dataset scenario (balance/imbalanced) and even with any classification algorithm NB, LR, SVM, and RF.

On the other hand, NB combined with any transformation method has the lowest performance with a range of difference between the best (LP-NB) and the worst (BR-NB) result in accuracy of 68 points.

]

```

#### Chan 2023 Harms from Increasingly Agentic Algorithmic Systems

\* Journal: Manuscript submitted to ACM

\* Type: theoretical

\* Link: <https://arxiv.org/abs/2302.10329>

\* Keywords / topics: [[NLP]],[[agentic algorithmic systems]]

``` abstract

Research in Fairness, Accountability, Transparency, and Ethics (FATE) has established many sources and forms of algorithmic harm, in domains as diverse as health care, finance, policing, and recommendations. Much work remains to be done to mitigate the serious harms of these systems, particularly those disproportionately affecting marginalized communities. Despite these ongoing harms, new systems are being developed and deployed which threaten the perpetuation of the same harms and the creation of novel ones. In response, the FATE community has emphasized the importance of anticipating harms. Our work focuses on the anticipation of harms from increasingly agentic systems. Rather than providing a definition of agency as a binary property, we identify 4 key characteristics which, particularly in combination, tend to increase the agency of a given algorithmic system: underspecification, directness of impact, goal-directedness, and long-term planning. We also discuss important harms which arise from increasing agency -- notably, these include systemic and/or long-range impacts, often on marginalized stakeholders. We emphasize that recognizing agency of algorithmic systems does not absolve or shift the human responsibility for algorithmic harms. Rather, we use the term agency to highlight the increasingly evident fact that ML systems are not fully under human control. Our work explores increasingly agentic algorithmic systems in three parts. First, we explain the notion of an increase in agency for algorithmic systems in the context of diverse perspectives on agency across disciplines. Second, we argue for the need to anticipate harms from increasingly agentic systems. Third, we discuss important harms from increasingly agentic systems and ways forward for addressing them. We conclude by reflecting on implications of our work for anticipating algorithmic harms from emerging systems.

```

#### key ideas

- Fairness, Accountability, Transparency, and Ethics (FATE) has already indicated that algorithmic systems play a role in causing harm

- argues that there are many negative externalities of algorithmic systems (ex. perpetuation of power relationships, generation toxic language, informational harms)

- no strong regulatory barriers

- agentic algorithmic systems = counter view that developers have full control over these systems, the interactive and responsive qualities of ML-based systems lead to unforeseen outcomes)

- contribution: 1) characteristics that increase agency of algorithmic systems, 2) need for anticipating harms from increasingly agentic systems, and 3) highlight some harms

#### agentic algorithmic systems

Systems have increasing agency as

* underspecification: when not specified how goal is accomplished
* directness of impact: when system affects world without mediation by human
* goal-directedness: trained to reach an quantifiable objective (reversed?)
* long-term planning: trained to make decisions temporarily dependent on one another, make predictions over long time horizon

Agents behave intentionally

principal-agent theory: principal delegates tasks to agent to achieve their goals; different incentives and information (principal does not tell agent how to complete task - underspecification)

* our focus on agency also shares many commonalities with work from the FATE community on establishing the harms of automated decision-making (ADM). ADM involves the use of algorithms to make decisions or enact policies without human intervention.
* reinforcement learning, has major focus on construction of agents to achieve goal encoded in reward
* more agentic systems are likely to be more effective than less agentic systems
* reinforcement-learning based recommendation systems (RLRS) in today’s social

media platforms warrant additional reason for concern. (have incentives to change or manipulate users’ internal states (e.g. preferences, beliefs, and psychology) for the purposes of increasing the metrics the RLRS systems are optimizing.)

the coding elite concentrates power by controlling the algorithms underlying the modern digital world, using that power to affect politics for their own gains.

reward hacking [Krakovna et al. 2020; Skalse et al. 2022], which is when a system exploits a reward signal to achieve a goal in an unforeseen, perhaps undesirable way.

#### Joshi A Knowledge Organization System for the United Nations Sustainable Development Goals

\* Journal:

\* Type: [mention research]

\* Link:

\* Keywords / topics: [[NLP ]]

``` abstract

This paper presents a formal knowledge organization system (KOS) to represent the United Nations Sustainable Development Goals(SDGs). The SDGs are a set of objectives adopted by all United Nations member states in 2015 to achieve a better and sustainable future. The

developed KOS consists of an ontology that models the core elements of the Global SDG indicator framework, which currently includes 17 Goals, 169 Targets and 231 unique indicators, as well as more than 450 related statistical data series maintained by the global statistical community to monitor progress towards the SDGs, and of a dataset containing these elements. In addition to formalizing and establishing unique identifiers for the components of the SDGs and their indicator framework, the ontology includes mappings of each goal, target, indicator and data series to relevant terms and subjects in the United Nations Bibliographic Information System (UNBIS) and the EuroVoc vocabularies, thus facilitating multilingual semantic search and content linking.

```

#### key ideas

- formal knowledge organization system (KOS) that has been developed for representing the United Nations Sustainable Development Goals (SDGs).

-The global indicator framework lists 247 indicators but has only 231 unique indicators, see for duplicates: https://unstats.un.org/sdgs/indicators/indicators-list

- full taxonomy of sdg goals and indicators: <https://metadata.un.org/sdg/?lang=en>

- link goals, targets and indicators <https://sustainabledevelopment.un.org/LinkedSDGs/about>

-

#### setup

* seems to be an overview of how knowledge graphs can be used for sdg goals. It mainly shows how UN is busy trying to analyze the goals, the hierarchy of goals, targets and indicators.

#### caveats

- it seems that although they try to integrate different frameworks, it is not clear how the reader can make use of these findings.

``` conclusions

The SDG KOS is an attempt to provide stakeholders a means to publish the

data using common terminologies and URIs centred around the SDG concepts,

thus helping break information silos, promote synergies among communities,

and enhance the semantic interoperability of different SDG-related data and

information assets made available by various sectors of society.

```