Appendix 1. Method Summaries

Recommendations for Supervised Classifiers

Model Selection

- Language models provide superior classification over bag-of-words classifiers.
- Use a domain adapted model if one is available. Otherwise, models trained for NLI learn faster than base models.

Training Samples

- Ensure manual labelers have sufficient context to accurately label documents.
- Train expert coders if your classification task requires domain specific expertise.

Validation

- Out of sample prediction is the primary test of validity.
- Cross validation can be computationally expensive for language models, so most applications opt to use a test set.

Recommendations for NLI Classifiers

Model Selection

- Use a model pre-trained on multiple NLI data sets.
- DeBERTaV3 Large is currently the only model to achieve performance comparable to supervised classifiers on entailment classification. Use this as a baseline for comparing other models.

Generating Hypotheses

- Match hypotheses to documents based on some relevancy parameter such as keyword matching or topic classification.
- Try classification with a single and multi-hypothesis approach.

Validation

- Calculate the number of samples needed to estimate performance within a confidence interval and label some data.
- Conduct basic sensitivity analysis by classify documents multiple times with synonymous hypotheses.
- Consider using an in-context classifier like GPT-4 as a second point of validating labels.

Recommendations for In-Context Classifiers

Model Selection

- When using an in-context classifier to label training or validation data, prioritize model capability. Currently, GPT-4 and Claude 3 Opus are the most sophisticated models.
- When using an in-context classifier as the the primary classifier, use an open source and reproducible model.
- HuggingFace.co hosts a leaderboard for comparing open source models. The HellaSwag benchmark is useful for assessing a model's inference capabilities.

Prompt Engineering

- Use the system message to prime the model's behavior and the user message to describe the classification task.
- Few-shot classification is inadvisable if zero-shot classification is sufficient due to the potential for over-fitting to prompt examples.
- Chain-of-thought reasoning has inconsistent effects, but significantly increases costs.

Parameters and Decoding

- Turn the temperature parameter to zero.
- Restrict the possible tokens a model can generate to those that represent your classes by biasing the logits.

Validation

- Estimate performance within a margin of error on a small sample of labeled data.
- Conduct sensitivity analysis by classifying the data multiple times with synonymous prompts and showing results are robust across prompts.

Appendix 2. Testing Data and Entailment Phrases

Appendix 2.1 Testing Data

Table 6. Data sources and the number of observations per target

Data set	Source	N	Notes
MTSD	Sobhani, Inkpen, and Zhu 2017	398	Tweets. Each tweet contains multiple politicians. No context provided to labelers.
ConStance	Joseph et al. 2017	98	Tweets. Annotators given a contextual prompt. Assumes anti-Clinton statements are pro-Trump.
KE-MLM	Kawintiranon and Singh 2021	380	Tweets. Annotators given examples of tweets with appropriate labels.
SemEval2016	Mohammad et al. 2016	97	Tweets. No context provided to labelers.
Titling	Berg et al. 2019	161	Tweets. Annotators given the author's photo, location, description, and text the tweet was responding to.
Newsletters	Cormack 2017	1,000	Newsletter Sentences. No context provided to labelers.

Appendix 2.2 Trump Approval Stance Hypotheses

Appendix 2.2.1 Default Hypothesis

The author of this text supports Trump.

Appendix 2.2.2 Alternative Hypotheses

The author of this text is pro Trump.
The author of this text approves of Trump.
The person who wrote this text supports Trump.
The person who wrote this text is pro Trump.
The person who wrote this text approves of Trump.
The author of this document supports Trump.
The author of this document is pro Trump.
The author of this document approves of Trump.

Appendix 2.3 COVID-19 Hypothesis Set 1

Dimension	Non-compliant	Compliant
Death	The author of this tweet believes covid has not increased deaths.	The author of this tweet believes covid has increased deaths. The author of this tweet believes covid deaths are neutral.
Flu	The author of this tweet believes covid is like the flu.	The author of this tweet believes covid is not like the flu. The author of this tweet believes nothing about covid and the flu.
Masks	The author of this tweet believes masks are bad.	The author of this tweet believes masks are good. The author of this tweet believes masks are neutral.
Lockdowns	The author of this tweet believes lockdowns are bad.	The author of this tweet believes lockdowns are good. The author of this tweet believes lockdowns are neutral.
Vaccines	The author of this tweet believes vaccines are bad.	The author of this tweet believes vaccines are good. The author of this tweet believes vaccines are neutral.
Social Distancing	The author of this tweet believes social distancing is bad.	The author of this tweet believes social distancing is good. The author of this tweet believes social distancing is neutral.
COVID-19 General	The author of this tweet believes the pandemic is not dangerous. The author of this tweet believes covid is not dangerous. The author of this tweet believes coronavirus is not dangerous.	The author of this tweet believes the pandemic is dangerous. The author of this tweet believes the pandemic is neutral. The author of this tweet believes covid is dangerous. The author of this tweet believes covid is neutral. The author of this tweet believes coronavirus is dangerous. The author of this tweet believes coronavirus is neutral.

Appendix 2.4 COVID-19 Hypothesis Set 2

Dimension	Non-compliant	Compliant
Death	The author of this tweet believes COVID death counts are wrong.	The author of this tweet believes many people have died from COVID. The author of this tweet does not express an opinion about COVID deaths.
Flu	The author of this tweet believes COVID is similar to the flu.	The author of this tweet does not believe COVID is similar to the flu. The author of this tweet does not compare COVID and the flu.
Masks	The author of this tweet opposes wearing masks.	The author of this tweet supports wearing masks. The author of this tweet does not express an opinion about wearing masks.
Lockdowns	The author of this tweet opposes lockdowns.	The author of this tweet supports lockdowns. The author of this tweet believes lockdowns are bad for the economy. The author of this tweet believes lockdowns save lives. The author of this tweet does not express an opinion about lockdowns.
Vaccines	The author of this tweet opposes vaccines.	The author of this tweet supports vaccines. The author of this tweet believes vaccines are dangerous. The author of this tweet believes vaccines are safe. The author of this tweet does not express an opinion about vaccines.
Social Distancing	The author of this tweet opposes social distancing.	The author of this tweet supports social distancing. The author of this tweet does not express an opinion about social distancing.
COVID-19 General	The author of this tweet does not believe the pandemic is dangerous. The author of this tweet does not believe COVID is dangerous. The author of this tweet does not believe Coronavirus is dangerous.	The author of this tweet believes the pandemic is dangerous. The author of this tweet believes the pandemic is neutral. The author of this tweet believes COVID is dangerous. The author of this tweet believes COVID is neutral. The author of this tweet believes Coronavirus is dangerous. The author of this tweet believes Coronavirus is neutral. The author of this tweet does not express an opinion about the pandemic. The author of this tweet does not express an opinion about Coronavirus. The author of this tweet does not express an opinion about COVID.

Appendix 3. Glossary

Attention: Also known as "self attention," is a mechanism used by transformer neural networks to estimate the importance of different words within a text sample.

Batch Size: The number of training examples used in one iteration or weight update during training. Higher batch sizes require more memory usage but increase computational efficiency.

BERT (Bidirectional Encoder Representations from Transformers): A language model that processes input sequences (text) in both directions, allowing it to understand the context of a word based on both its preceding and following words. This bidirectional approach overcame the limitations of previous unidirectional models and significantly improved the performance of various natural language processing tasks.

Chain-of-Thought Reasoning: A method of prompting generative language models that requests they explain their reasoning before providing a conclusion. This can sometimes improve model performance on complex tasks.

Context: Information relevant to the stance of interest. Available context consists of information the document contains, and the knowledge base of the classifier. Missing context increases document ambiguity.

CPU: Central Processing Unit. The primary component of a computer responsible for executing instructions from software. The CPU is generally too slow for training and running large language models.

DeBERTa (Decoding-enhanced BERT with Disentangled Attention): Another extension of BERT that introduces two novel techniques: disentangled attention and enhanced mask decoder. Disentangled attention separates information about a tokens content and sentence position, allowing the model to capture richer semantic patterns. The enhanced mask decoder improves the model's ability to reconstruct masked tokens, leading to better understanding of natural language.

DeBERTaV3: DeBERTaV3 uses a more advanced pre-training approach to improve upon DeBERTa. DeBERTaV3 has demonstrates state-of-the-art NLI classification among models similar models such as BERT, RoBERTa, and Electra.

Decoding: The process used to create output sequences from a generative language model.

Domain Adaptation: Tuning a model trained on one domain to perform well on a different domain.

Electra: An evolution of BERT that improves upon the pre-training methods used. It trains two transformer models simultaneously: a generator that predicts masked tokens, and a discriminator that distinguishes the generated tokens from the original ones.

Encoding: The process used by a language model to convert text into high dimensional vector representations of the text's meaning.

Few-shot Classification: A setting where a language model is provided with a small number of labeled examples for each category and is expected to generalize to classify new instances accurately.

Generative Language Models: Language models that are trained to generate human-like text by predicting the probability distribution of the next word or token based on the previous context. Examples include GPT-4, Claude 3, and Llama 2.

Generative Pretrained Transformer (GPT): A type of transformer-based language model that is pretrained on a large corpus of text to predict the next token in a sequence. This allows it to generate human-like text.

GPU: Graphics Processing Unit. A specialized computer component designed to accelerate the processing of images and parallel computations. GPUs are frequently used to train language models.

Hyperparameter: Settings that control the behavior of a machine learning algorithm or language model. These include such as learning rate, batch size, or training epochs.

Hyperparameter Sweep: Systematically testing different combinations of hyperparameters to find the optimal configuration for a task or dataset.

Hypothesis: the proposition on which an author may take a stance. A stance statement.

In-context Learning: A paradigm where a language model learns to perform a new task by describing the task in a prompt, without explicit fine-tuning.

Learning Rate: A hypterparameter determining how much the model's weights are updated during training iterations. Higher rates allow faster convergence but risk overshooting the optimal solution. Lower rates risk getting trapped in locally optimal solutions and missing globally optimal solutions.

Logit Bias: A technique used to adjust the output distribution of a language model by adding a bias term to the logits (pre-normalized output scores). This allows for control over the model's generation or classification behavior.

Natural Language Inference (NLI): The task of determining the logical relationship between a pair of documents (entailment, contradiction, or neutrality). Models that are effective at NLI can be used as universal classifiers.

Prompt: A plain language input sequence provided to a generative language model to elicit a desired response.

Prompt Engineering: The process of designing and refining prompts to optimize the performance of generative language models on specific tasks.

RoBERTa (Robustly Optimized BERT Pretraining Approach): An improved version of BERT. It uses the same model as BERT but addresses some of its shortcomings by employing different pre-training strategies and more training data.

Sentiment: Positive or negative emotional valence.

Stance: How an individual would answer a proposition.

Stance Detection: Text sample T entails stance S to author A when a human reading T with

context C would infer that T expresses support for S.

Supervised Classification: A machine learning task where the model is trained on labeled data to classify new, unseen instances into predefined categories. Supervised classifiers are task specific.

System Message: A special type of prompt used in conversational AI systems to set the tone, persona, or expected behavior of the language model during the conversation.

Temperature: A hyperparameter that controls the randomness of the output from a generative language model. A temperature of zero leads to the most deterministic outputs.

Textual Entailment: Text sample T entails hypothesis H when a human reading T would infer that H is most likely true.

Token: A token is the basic unit of input and output in natural language processing models. It can be a word, sub-word, character, or any other meaningful unit that the model operates on. Text data is typically tokenized (split into tokens) before being fed into the model.

Training Epochs: One complete pass through the entire training dataset. Multiple epochs are typically required for the model to converge and achieve optimal performance.

Transfer learning: A machine learning technique where a model is pre-trained on a general set of tasks such as next word or next sentence prediction, and then adapted or fine-tuned for a more specific task. Transfer learning can improve the performance and reduce the training necessary for supervised classifiers.

Transformers: A type of neural network architecture that employs self-attention mechanisms to capture long-range dependencies between words in sequences. Transformers consist of an encoder portion that creates semantic representations of documents, and a decoder portion that uses these representations to generate text. Transformers are the the backbone of many modern language models from BERT to GPT-4. BERT type models primarily use the encoder portion, while GPT type models primarily use the decoder.

Zero-shot Classification: The ability of a language model to classify documents into categories it has not been explicitly trained on. NLI and in-context classifiers are capable of zero-shot classification.