Like a Good Nearest Neighbor: Practical Content Moderation with Sentence Transformers

Anonymous ACL submission

Abstract

Text classification systems have impressive capabilities but are infeasible to deploy and use reliably due to their dependence on prompting and billion-parameter language models. SetFit (Tunstall et al., 2022) is a recent, practical approach that fine-tunes a Sentence Transformer under a contrastive learning paradigm and achieves similar results to more unwieldy systems. Text classification is important for addressing the problem of domain drift in detecting harmful content, which plagues social media platforms. Here, we propose Like a Good Nearest Neighbor (LaGoNN), a modification to SetFit that requires no additional parameters or hyperparameters but alters input text with information from its nearest neighbor, for example, the label and text, in the training data, making novel data appear similar to an instance on which the model was optimized. LaGoNN is effective at identifying harmful content and generally improves SetFit's performance. To demonstrate LaGoNN's value, we conduct a thorough study of text classification systems in the context of content moderation under four label distributions.[1]

1 Introduction

Text classification is the most important tool for NLP practitioners, and there has been substantial progress in advancing the state-of-the-art, especially with the advent of large, pretrained language models (PLM) (Devlin et al., 2019). Modern research focuses on in-context learning (Brown et al., 2020), pattern exploiting training (Schick and Schütze, 2021a,b, 2022), adapter-based fine-tuning with learned label embeddings (Karimi Mahabadi et al., 2022), and parameter efficient fine-tuning (Liu et al., 2022a). These methods have achieved impressive results on the SuperGLUE (Wang et al., 2019) and RAFT (Alex et al., 2021) few-shot benchmarks, but most are difficult to use because of their reliance on billion-parameter PLMs, pay-to-use APIs, and/or prompting. Constructing prompts is not trivial and may require domain expertise.

[1]Code and data: https://github.com/[REDACTED]

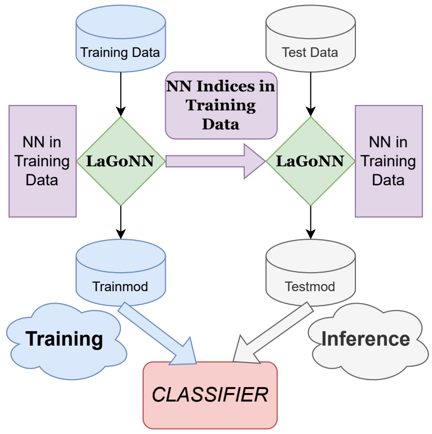

Figure 1: We embed training data, retrieve the text, gold label, and distance for each instance from its nearest neighbor and modify the original text with this information. Then we embed the modified training data and train a classifier. During inference, the NN from the training data is selected, the original text is modified with the text, gold label, and distance from this NN, and the classifier is called.

One exception to these cumbersome systems is SetFit. SetFit does not rely on prompting or billion-parameter PLMs, and instead fine-tunes a pretrained Sentence Transformer (ST) (Reimers and Gurevych, 2019) under a contrastive learning paradigm. SetFit has comparable performance to more unwieldy systems while being one to two orders of magnitude faster to train and run inference.

An important application of text classification is aiding or automating content moderation, which is the task of determining the appropriateness of user-generated content on the Internet (Roberts, 2017). From fake news to toxic comments to hate speech, it is difficult to browse social media without being exposed to potentially dangerous posts that may have an effect on our ability to reason (Ecker et al., 2022). Misinformation spreads at alarming rates (Vosoughi et al., 2018), and an ML system should be able to quickly aid human moderators. While there is work in NLP with this goal (Markov et al., 2022; Shido et al., 2022; Ye et al., 2023), a general, practical, and open-sourced method that is effective across multiple domains remains an open challenge. Novel fake news topics or racial slurs emerge and change constantly. Retraining of ML-based systems is required to adapt this concept drift, but this is expensive, not only in terms of computation, but also in terms of the human effort needed to collect and label data.

SetFit's performance, speed, and low cost would make it ideal for effective content moderation, however, this type of text classification proves difficult for even state-of-the-art approaches. For example, detecting hate speech on Twitter (Basile et al., 2019), a subtask on the RAFT few-shot benchmark, appears to be the most difficult dataset; at time of writing, it is the only task where the human baseline has not been surpassed, yet SetFit is among the top ten most performant systems.[2]

Here, we propose a modification to SetFit, called Like a Good Nearest Neighbor (LAGONN). LAGONN introduces no parameters or hyperparameters and instead modifies input text by retrieving information about the nearest neighbor (NN) seen during optimization (see Figure 1). Specifically, we append the label, distance, and text of the NN in the training data to a new instance and encode this modified version with an ST. By making input data appear more similar to instances seen during training, we inexpensively exploit the ST's pretrained or fine-tuned knowledge when considering a novel example. Our method can also be applied to the linear probing of an ST, requiring no expensive fine-tuning of the large embedding model. Finally, we propose a simple alteration to the SetFit training procedure, where we fine-tune the ST on a subset of the training data. This results in a more efficient and performant text classifier that can be used with LAGONN. We summarize our contributions as follows:

1. We propose LAGONN, an inexpensive modification to SetFit- or ST-based text classification.

2. We suggest an alternative training procedure to the standard fine-tuning of SetFit, that can be used with or without LAGONN, and results in a cheaper system with similar performance to the more expensive SetFit.

3. We perform an extensive study of LAGONN, SetFit, and standard transformer fine-tuning in the context of content moderation under different label distributions.

[2]https://huggingface.co/spaces/ought/raft-leaderboard (see "Tweet Eval Hate").

2 Related Work

There is little work on using sentence embeddings as features for classification despite the pioneering work being five years old (Perone et al., 2018). STs are pretrained with the objective of maximizing the distance between semantically distinct text and minimizing the distance between text that is semantically similar in feature space. They are composed of a Siamese and triplet architecture that encodes text into dense vectors which can be used as features for ML. STs were first used to embed text for classification by Piao (2021), however, only pretrained representations were examined.

SetFit uses a contrastive learning paradigm (Koch et al., 2015) to optimize the ST embedding model. The ST is fine-tuned with a distance-based loss function, like cosine similarity, such that examples with different labels are separated in feature space. Input text is then encoded with the fine-tuned ST and a classifier, such as logistic regression, is trained. This approach creates a strong, few-shot text classification system, transforming the ST from a sentence encoder to a topic encoder.

Work done by Xu et al. (2021) showed that retrieving and concatenating text from training data and external sources, such as ConceptNet (Speer et al., 2017) and the Wiktionary[3] definition, can be viewed as a type of external attention that does not alter the architecture of the Transformer in question answering. Liu et al. (2022b) used PLMs and k-NN lookup to prepend examples that are similar to a GPT-3 query, aiding in prompt engineering for in-context learning. Wang et al. (2022) demonstrated that prepending and appending training data helps PLMs in summarization, language modelling, machine translation, and question answering, using BM25 as their retrieval model (Manning et al., 2008; Robertson and Zaragoza, 2009).

We alter the SetFit training procedure by using fewer examples to adapt the embedding model for

[3]https://www.wiktionary.org/

Training Data	Test Data
"I love this." [positive 0.0] (0)	"So good!" [?] (?)
"This is great!" [positive 0.5] (0)	"Just terrible!" [?] (?)
"I hate this." [negative 0.7] (1)	"Never again." [?] (?)
"This is awful!" [negative 1.2] (1)	"This rocks!" [?] (?)

LaGoNN Configuration	**Train Modified**
LABEL	"I love this. [SEP] [positive]" (0)
DIST	"I love this. [SEP] [0.5]" (0)
LabDist	"I love this. [SEP] [positive 0.5]" (0)
TEXT	"I love this. [SEP] [positive 0.5] This is great!" (0)
ALL	"I love this. [SEP] [positive 0.5] This is great! [SEP] [negative 0.7] I hate this." (0)
	Test Modified
LABEL	"So good! [SEP] [positive]" (?)
DIST	"So good! [SEP] [1.5]" (?)
LabDist	"So good! [SEP] [positive 1.5]
TEXT	"So good! [SEP] [positive 1.5] I love this." (?)
ALL	"So good! [SEP] [positive 1.5] I love this. [SEP] [negative 2.7] This is awful!" (?)

Table 1: Toy training and test data and different LaGoNN configurations considering the first training example. Text is in quotation marks and the integer label is in parenthesis. In brackets are the gold label or distance from the NN or both. Train and Test Modified are altered instances that are input into the final embedding model for training and inference, respectively. The input format is "*original text* [SEP] [(NN gold) (label distance)] NN *training instance text*". See Appendix A.5 for examples of LaGoNN ALL modified text.

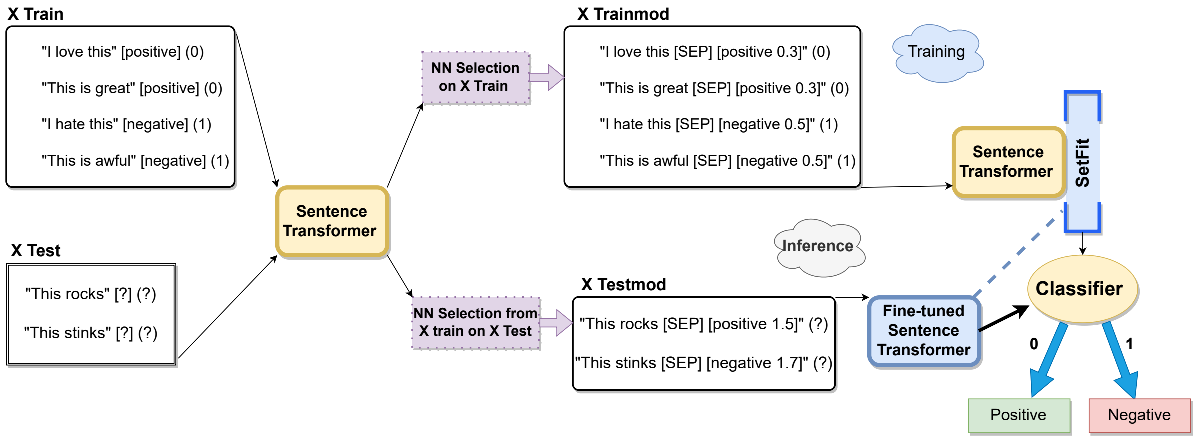

Figure 2: LaGoNN LabDist uses an ST to encode training data, performs NN lookup, appends the NN's gold label and distance, and optionally SetFit to fine-tune the embedding model. We then embed this new instance and train a classifier. During inference, we use the embedding model to modify the test data with its NN's gold label and distance from the training data, compute the final representation, and call the classifier. Input text is in quotation marks, the NN's gold label and distance are in brackets, and the integer label is in parenthesis.

many-shot learning. LaGoNN decorates input text with its NN's gold label, Euclidean distance, and text from the training data to exploit both the ST's distance-based pretraining and SetFit's distance-based fine-tuning objective. Compared to retrieval-based methods, LaGoNN uses the same model for both retrieval and encoding, retrieving only information from the training data for classification.

3 Like a Good Nearest Neighbor

Xu et al. (2021) formulate a type of external attention, where textual information is retrieved from multiple sources and added to text input to give the model stronger reasoning ability without altering the internal architecture. Inspired by this approach, LaGoNN exploits pretrained and fine-tuned knowledge through external attention, but the information we retrieve comes only from data used during optimization. We consider an embedding function, f, that encodes both training and test data, $f(X_{train})$ and $f(X_{test})$. Considering its success on realistic, few-shot data and our goal of practical content moderation, we choose an ST that can be fine-tuned with SetFit as our embedding function.

Encoding and nearest neighbors LaGoNN first uses a pretrained Sentence Transformer to embed training text in feature space, $f(X_{train})$, and NN lookup with scikit-learn (Buitinck et al., 2013) on the resulting embeddings.

Nearest neighbor information We extract text from the nearest neighbor and use it to decorate the original example. We experimented with different text that LaGoNN could use. The first configuration we consider is the gold label of the NN, which we call LABEL. We then consider the Euclidean distance of the NN, which we call DIST, giving the model access to continuous measure of similarity. We then combine these two configurations, appending both the NN's gold label and Euclidean distance, referring to this as LabDist. Next, we consider the gold label, distance, and the text of the NN, which we refer to as TEXT. Finally, we tried the same format as TEXT but for all possible labels, which we call ALL (see Table 1 and Figure 2). Information from the NN is appended to the text following a separator token to indicate this instance is composed of multiple sequences. While the ALL and TEXT configurations are arguably the most interesting, we find LabDist to result in the most performant version of LaGoNN, and this is the version about which we report results. See Appendix A.4.1 for a detailed study of and comparison between all LaGoNN configurations.

Training LaGoNN encodes the modified training data, optionally fine-tunes the embedding model via SetFit, and trains a classifier, $CLF(f(X_{trainmod}))$.

Inference LaGoNN uses information from the nearest neighbor in the training data to modify input text. We compute the embeddings of the test data, $f(X_{test})$, and select and extract information from the NN's training text, decorating the input instance with this information. Finally, we encode the modified data with the embedding model and call the classifier, $CLF(f(X_{testmod}))$.

Intuition The ST's pretraining and SetFit's fine-tuning objective both rely on distance, creating a feature space appropriate for distance-based algorithms, such as our NN-lookup. We hypothesize that LaGoNN's modifications make novel data appear semantically similar to their NNs in the training data, that is, more akin to an instance on which the encoder and classifier were optimized. As LaGoNN utilizes similarity and clear distinctions between classes, we believe it fitting for our use case of content moderation, where it is realistic to have few labels, harmful or neutral, for example.

4 Experiments

4.1 Data and label distributions

We study LaGoNN's performance on four binary and one ternary classification dataset related to the task of content moderation. Each dataset is composed of a training, validation, and test split.

Here, we provide a summary of the five datasets we studied. LIAR was created from Politifact[4] for fake news detection and is composed of the data fields *context*, *speaker*, and *statement*, which are labeled with varying levels of truthfulness (Wang, 2017). We used a collapsed version of this dataset where a statement can only be true or false. We did not use *speaker*, but did use *context* and *statement*, separated by a separator token. Quora Insincere Questions[5] is composed of neutral and toxic questions, where the author is not asking in good faith. Hate Speech Offensive[6] has three labels and is composed of tweets that can contain either neutral text, offensive language, or hate speech (Davidson et al., 2017). Amazon Counterfactual[7] contains sentences from product reviews, and the labels can be "factual" or "counterfactual" (O'Neill et al., 2021). "Counterfactual" indicates that the customer said something that cannot be true. Finally, Toxic Conversations[8] is a dataset of comments where the author wrote with unintended bias[9] (see Table 2).

We study our system by simulating growing

[4] https://www.politifact.com/
[5] https://www.kaggle.com/c/quora-insincere-questions-classification
[6] https://huggingface.co/datasets/hate_speech_offensive
[7] https://huggingface.co/datasets/SetFit/amazon_counterfactual_en
[8] https://huggingface.co/datasets/SetFit/toxic_conversations
[9] https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification/overview

Dataset (and Detection Task)	Number of Labels
LIAR (Fake News)	2
Insincere Questions (Toxicity)	2
Hate Speech Offensive	3
Amazon Counterfactual (English)	2
Toxic Conversations	2

Table 2: Summary of datasets and number of labels. We provide the type of task in parenthesis in unclear cases.

training data over ten discrete steps sampled under four different label distributions: extreme, imbalanced, moderate, and balanced (see Table 3). On each step we add 100 examples (100 on the first, 200 on the second, etc.) from the training split sampled under one of the four ratios.[10] On each step, we train our method with the sampled data and evaluate on the test split. Considering growing training data has two benefits: 1) We can simulate a streaming data scenario, where new data is labeled and added for training and 2) We can investigate each method's sensitivity to the number of training examples. We sampled over five seeds, reporting the mean and standard deviation.

Regime	Binary	Ternary
Extreme	0: 98% 1: 2%	0: 95%, 1: 2%, 2: 3%
Imbalanced	0: 90% 1: 10%	0: 80%, 1: 5%, 2: 15%
Moderate	0: 75% 1: 25%	0: 65%, 1: 10%, 2: 25%
Balanced	0: 50% 1: 50%	0: 33%, 1: 33%, 2: 33%

Table 3: Label distributions for sampling training data. 0 represents neutral while 1 and 2 represent different types of undesirable text.

4.2 Baselines

We compare LaGoNN against a number of strong baselines, detailed below. We used default hyperparameters in all cases unless stated otherwise.

RoBERTa RoBERTa-base is a pretrained language model (Liu et al., 2019) that we fine-tuned with the transformers library (Wolf et al., 2020). We select two versions of RoBERTa-base: an expensive version, where we perform standard fine-tuning on each step (RoBERTa$_{full}$) and a cheaper version, where we freeze the model body after step one and update the classification head on subsequent steps (RoBERTa$_{freeze}$). We set the learning rate to $1e^{-5}$, train for a maximum of 70 epochs, and use early stopping, selecting the best model after training. We consider RoBERTa$_{full}$ an upper bound as it has the most trainable parameters and requires the most time to train of all our methods.

Linear probe We perform linear probing of a pretrained Sentence Transformer by fitting logistic regression with default hyperparameters on the training embeddings on each step. We choose this baseline because LaGoNN can be applied as a modification in this scenario. We select MPNET (Song et al., 2020) as the ST, for SetFit, and for LaGoNN.[11] We refer to this method as Probe.

SetFit Here, we perform standard fine-tuning with SetFit on the first step, and then on subsequent steps, freeze the embedding model and retrain only the classification head. We choose this baseline as LaGoNN also uses logistic regression as its final classifier and refer to this method as SetFit.

***k*-nearest neighbors** Similar to the above baseline, we fine-tune the embedding model via SetFit, but swap out the classification head for a kNN classifier, where $k = 3$. We select this baseline as LaGoNN also relies on an NN lookup. $k = 3$ was chosen during our development stage as it yielded the strongest performance. We refer to this method as kNN.

SetFit expensive For this baseline we perform standard fine-tuning with SetFit on each step. On the first step, this method is equivalent to SetFit. We refer to this as SetFit$_{exp}$.

LaGoNN cheap This method modifies data via LaGoNN before fitting logistic regression. Even without adapting the embedding model, as the training data grow, modifications made to the test data may change. Only the classification head is fit on each step. We refer to this method as LaGoNN$_{cheap}$ and it is comparable to Probe.

LaGoNN On the first step, we use LaGoNN to modify our data and perform standard fine-tuning with SetFit. On subsequent steps, we freeze the embedding model but continue to use it to modify our data. We only fit logistic regression on later steps, referring to this method as LaGoNN. It is comparable to SetFit.

LaGoNN expensive Here we modify our data and fine-tune the embedding model on each step. We refer to this method as LaGoNN$_{exp}$ and

[10] For Hate Speech Offensive, 0 and 2 denote undesirable text and 1 denotes neither.

[11] https://huggingface.co/sentence-transformers/paraphrase-mpnet-base-v2

it is comparable to SetFit$_{exp}$. On the first step, this method is equivalent to LaGoNN.

5 Results

Table 4 and Figure 3 show our results. In the cases of the extreme and imbalanced regimes, the performance of SetFit$_{exp}$ steadily increases with the number of training examples. As the label distribution shifts to the balanced regime, however, the performance quickly saturates or even degrades as the number of training examples grows. LaGoNN, RoBERTa$_{full}$, and SetFit, other fine-tuned PLM classifiers, do not exhibit this behavior. LaGoNN$_{exp}$, being based on SetFit$_{exp}$, exhibits a similar trend, but the performance degradation is mitigated; on the 10^{th} step of Amazon Counterfactual in Table 4 SetFit$_{exp}$'s performance decreased by 9.7, while LaGoNN$_{exp}$ only fell by 3.7.

LaGoNN and LaGoNN$_{exp}$ generally outperform SetFit and SetFit$_{exp}$, respectively, often resulting in a more stable model, as reflected in the standard deviation. We find that LaGoNN and LaGoNN$_{exp}$ exhibit stronger predictive power with fewer examples than RoBERTa$_{full}$ despite having fewer trainable parameters. For example, on the first step of Insincere Questions under the extreme setting, LaGoNN's performance is more than 10 points higher.

LaGoNN$_{cheap}$ outperforms all other methods on the Insincere Questions dataset for all balance regimes, despite being the third fastest (see Table 5) and having the second fewest trainable parameters. We attribute this result to the fact that this dataset is composed of questions from Quora[12] and our ST backbone was pretrained on similar data. This intuition is supported by Probe, the cheapest method, which despite having the fewest trainable parameters, shows comparable performance.

5.1 SetFit for efficient many-shot learning

Respectively comparing SetFit to SetFit$_{exp}$ and LaGoNN to LaGoNN$_{exp}$ suggests that fine-tuning the ST embedding model on moderate or balanced data hurts model performance as the number of training samples grows. We therefore hypothesize that randomly sampling a subset of training data to fine-tune the encoder, freezing, embedding the remaining data, and training the classifier will result in a stronger model.

[12] https://www.quora.com/

To test our hypothesis, we add two models to our experimental setup: SetFit$_{lite}$ and LaGoNN$_{lite}$. SetFit$_{lite}$ and LaGoNN$_{lite}$ are respectively equivalent to SetFit$_{exp}$ and LaGoNN$_{exp}$, except after the fourth step (400 samples), we freeze the encoder and only retrain the classifier on subsequent steps, similar to SetFit and LaGoNN.

Figure 4 shows our results with these two new models. As expected, in the cases of extreme and imbalanced distributions, LaGoNN$_{exp}$, SetFit$_{exp}$, and RoBERTa$_{full}$, are the strongest performers on Toxic Conversations. We note very different results for both LaGoNN$_{lite}$ and SetFit$_{lite}$ compared to LaGoNN$_{exp}$ and SetFit$_{exp}$ on Toxic Conversations and Amazon Counterfactual under the moderate and balanced label distributions. As their expensive counterparts start to plateau or degrade on the fourth step, the predictive power of these two new models dramatically increases, showing improved or comparable performance to RoBERTa$_{full}$, despite being optimized on less data; for example, LaGoNN$_{lite}$ reaches an average precision of approximately 55 after being optimized on only 500 examples. RoBERTa$_{full}$ does not exhibit similar performance until the tenth step. Finally, we point out that LaGoNN-based methods generally provide a performance boost for SetFit-based classification.

5.2 LaGoNN's computational expense

LaGoNN is more computationally expensive than Sentence Transformer- or SetFit-based text classification. LaGoNN introduces additional inference with the encoder, NN-lookup, and string modification. As the computational complexity of transformers increases with sequence length (Vaswani et al., 2017), additional expense is created when LaGoNN appends textual information before inference with the ST. In Table 5, we provide a speed comparison of comparable methods computed on the same hardware.[13] On average, LaGoNN introduced 24.2 additional seconds of computation compared to its relative counterpart.

6 Discussion

Flagging potentially dangerous text presents a challenge even for state-of-the-art approaches. It is imperative that we develop reliable and practical text classifiers for content moderation, such that we can inexpensively re-tune them for novel forms

[13] We used a 40 GB NVIDIA A100 Tensor Core GPU.

Method *Extreme*	**InsincereQs** 1^{st}	5^{th}	10^{th}	Average	**AmazonCF** 1^{st}	5^{th}	10^{th}	Average
RoBERTa$_{full}$	$19.9_{8.4}$	$30.9_{7.9}$	$42.0_{7.4}$	$33.5_{6.7}$	$21.8_{6.6}$	$63.9_{10.2}$	$72.3_{3.0}$	$59.6_{16.8}$
SetFit$_{exp}$	$24.1_{6.3}$	$29.2_{6.7}$	$36.7_{7.3}$	$31.7_{3.4}$	$22.3_{8.8}$	$64.2_{3.3}$	$68.6_{4.6}$	$56.8_{14.9}$
LaGoNN$_{exp}$	$\mathbf{30.7}_{8.9}$	$37.6_{6.1}$	$39.0_{6.1}$	$36.1_{2.3}$	$\mathbf{26.1}_{17.5}$	$\mathbf{68.4}_{4.4}$	$\mathbf{74.9}_{2.9}$	$\mathbf{63.2}_{16.7}$
RoBERTa$_{freeze}$	$19.9_{8.4}$	$34.1_{5.4}$	$37.9_{5.9}$	$32.5_{5.5}$	$21.8_{6.6}$	$41.0_{12.7}$	$51.3_{10.7}$	$40.6_{8.9}$
kNN	$6.8_{0.42}$	$15.9_{3.4}$	$16.9_{4.3}$	$14.4_{3.0}$	$10.3_{0.2}$	$15.3_{4.2}$	$18.4_{3.7}$	$15.6_{2.4}$
SetFit	$24.1_{6.3}$	$31.7_{4.9}$	$36.1_{5.4}$	$31.8_{3.6}$	$22.3_{8.8}$	$32.4_{11.5}$	$42.3_{8.8}$	$34.5_{5.9}$
LaGoNN	$\mathbf{30.7}_{8.9}$	$39.3_{4.9}$	$41.2_{4.7}$	$38.4_{3.0}$	$\mathbf{26.1}_{17.5}$	$31.1_{19.4}$	$33.0_{19.1}$	$30.9_{2.3}$
Probe	$24.3_{8.4}$	$39.8_{5.6}$	$44.8_{4.2}$	$38.3_{6.2}$	$24.2_{9.0}$	$46.3_{4.4}$	$54.6_{2.0}$	$45.1_{10.3}$
LaGoNN$_{cheap}$	$23.6_{7.8}$	$\mathbf{40.7}_{5.9}$	$\mathbf{45.3}_{4.4}$	$\mathbf{38.6}_{6.6}$	$20.1_{6.9}$	$38.3_{4.9}$	$47.8_{3.4}$	$38.2_{9.5}$
Balanced								
RoBERTa$_{full}$	$47.1_{4.2}$	$52.1_{3.6}$	$55.7_{2.6}$	$52.5_{2.9}$	$73.6_{2.1}$	$78.6_{3.9}$	$\mathbf{82.4}_{1.1}$	$78.9_{2.2}$
SetFit$_{exp}$	$43.5_{4.2}$	$47.1_{4.6}$	$48.5_{3.9}$	$48.0_{1.7}$	$73.8_{4.4}$	$69.8_{4.0}$	$64.1_{4.6}$	$69.6_{3.6}$
LaGoNN$_{exp}$	$42.8_{5.3}$	$47.6_{2.9}$	$47.0_{1.7}$	$46.2_{2.0}$	$\mathbf{76.0}_{3.0}$	$73.4_{2.6}$	$72.3_{2.9}$	$72.5_{3.4}$
RoBERTa$_{freeze}$	$47.1_{4.2}$	$52.1_{0.4}$	$53.3_{1.7}$	$51.5_{2.1}$	$73.6_{2.1}$	$76.8_{1.6}$	$77.9_{1.0}$	$76.5_{1.3}$
kNN	$22.3_{2.3}$	$30.2_{2.3}$	$30.9_{1.8}$	$29.5_{2.5}$	$41.7_{3.4}$	$57.9_{3.3}$	$58.3_{3.3}$	$56.8_{5.1}$
SetFit	$43.5_{4.2}$	$53.8_{2.2}$	$55.5_{1.6}$	$52.8_{3.5}$	$73.8_{4.4}$	$79.2_{1.9}$	$80.1_{1.0}$	$78.6_{1.8}$
LaGoNN	$42.8_{5.3}$	$54.1_{2.9}$	$56.3_{1.3}$	$53.4_{3.7}$	$\mathbf{76.0}_{3.0}$	$\mathbf{80.1}_{2.0}$	$81.4_{1.1}$	$\mathbf{79.8}_{1.4}$
Probe	$47.5_{1.6}$	$52.4_{1.7}$	$55.3_{1.1}$	$52.2_{2.5}$	$52.4_{3.4}$	$64.7_{2.5}$	$67.5_{0.4}$	$63.4_{4.4}$
LaGoNN$_{cheap}$	$\mathbf{49.3}_{2.6}$	$\mathbf{54.4}_{1.4}$	$\mathbf{57.6}_{0.7}$	$\mathbf{54.2}_{2.7}$	$48.1_{3.4}$	$62.0_{2.0}$	$65.3_{0.8}$	$60.5_{5.0}$

Table 4: Average performance (average precision $\times$ 100) on Insincere Questions and Amazon Counterfactual. The first, fifth, and tenth step are followed by the average over all ten steps. The average gives insight into the overall strongest performer by aggregating all steps. We group methods with a comparable number of trainable parameters together. The extreme label distribution results are followed by balanced (see Appendix A.2 for additional results).

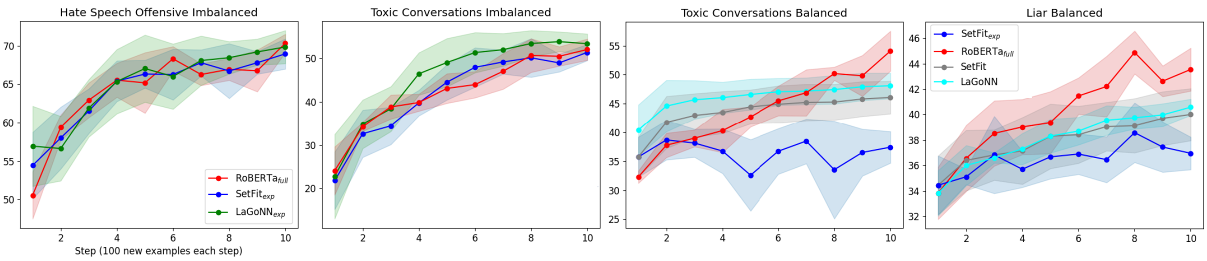

Figure 3: Average performance in the imbalanced and balanced regimes relative to comparable methods. We include RoBERTa$_{full}$ results for reference. The metric is macro-F1 for Hate Speech Offensive, average precision elsewhere.

Method	**Time in seconds**
Probe	22.9
LaGoNN$_{cheap}$	44.2
SetFit	42.9
LaGoNN	63.4
SetFit$_{exp}$	207.3
LaGoNN$_{exp}$	238.0
RoBERTa$_{full}$	446.9

Table 5: Speed comparison between LaGoNN and comparable methods. Time includes training on $1,000$ examples and inference on $51,000$ examples.

of hate speech, toxicity, and fake news. LaGoNN exploits semantic similarity and clear boundaries between labels, which we believe is reflected in scenarios with fewer classes, such as quickly filtering out harmful content.

Our results suggest that LaGoNN$_{exp}$ or SetFit$_{exp}$, relatively expensive techniques, can detect harmful content when dealing with imbalanced label distributions, as is common with realistic datasets. This is intuitive from the perspective that less common instances are more difficult to learn and require more effort. An exception would be our examination of Insincere Questions, where

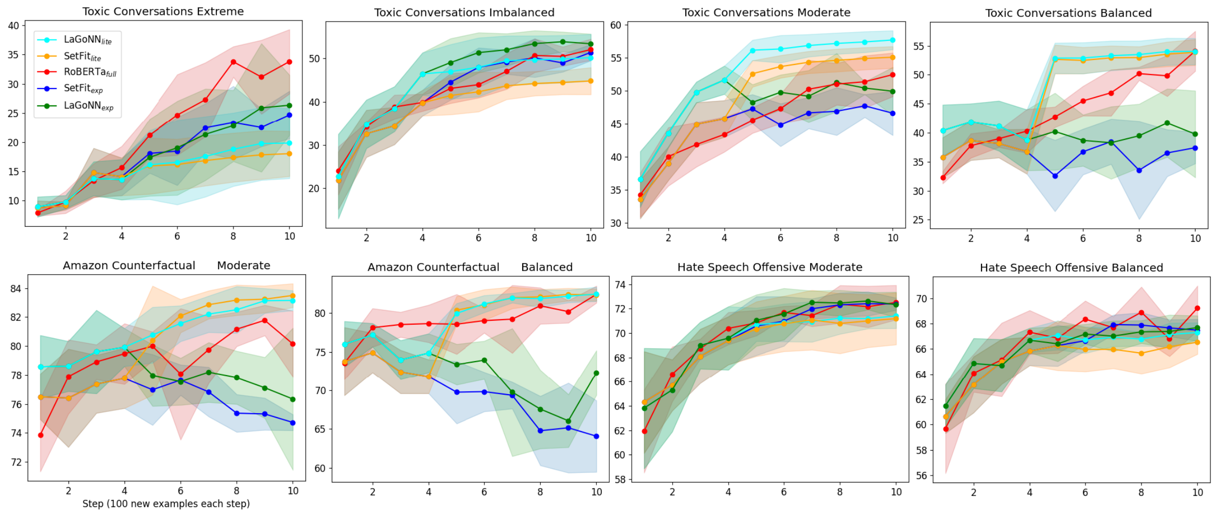

Figure 4: Average performance for all sampling regimes on Toxic Conversations and the moderate and balanced regimes for Amazon Counterfactual and Hate Speech Offensive. More expensive models, such as LAGONN$_{exp}$, SetFit$_{exp}$, and RoBERTa$_{full}$ perform best when the label distribution is imbalanced. As the distribution becomes more balanced, inexpensive models, such as LAGONN$_{lite}$, show similar or improved performance. The metric is macro-F1 for Hate Speech Offensive, average precision elsewhere (see Appendix A.3 for additional results).

LAGONN$_{cheap}$ excelled in the extreme and balanced settings. This highlights the fact that we can inexpensively extract pretrained knowledge if PLMs are chosen with care for related tasks.

Standard fine-tuning with SetFit does not help performance on more balanced datasets that are not few-shot. SetFit was developed for few-shot learning, but we have observed that it should not be applied "out of the box" to balanced, non-few-shot data. This can be detrimental to performance, directly affecting our own approach. However, we have observed that LAGONN can stabilize SetFit's predictions and reduce its performance drop. Figures 3 and 4 show that when the label distribution is moderate or balanced (see Table 3), SetFit$_{exp}$ plateaus, yet cheaper systems, such as LAGONN, continue to learn. We believe this is due to SetFit's fine-tuning objective, which optimizes an ST using cosine similarity loss to separate examples belonging to different labels in feature space, assuming independence between labels. This may be too strong an assumption as we optimize with more examples, which is counter-intuitive for data-hungry transformers; RoBERTa$_{full}$, optimized with cross-entropy loss, generally showed improved performance as we added training data.

When dealing with balanced data, it is sufficient to fine-tune the Sentence Transformer via SetFit with 50 to 100 examples per label, while 150 to 200 instances appear to be sufficient when the training data are moderately balanced. The encoder can then be frozen and all available data embedded to train a classifier. This improves performance and is more efficient than full-model fine-tuning. LAGONN is directly applicable to this case, boosting the performance of SetFit$_{lite}$ without introducing trainable parameters. In this setup, all models fine-tuned on Hate Speech Offensive exhibited similar, upward-trending learning curves, but we note the speed of LAGONN relative to RoBERTa$_{full}$ or SetFit$_{exp}$ (see Figure 4 and Table 5).

7 Conclusion

We have proposed LAGONN, a simple and inexpensive modification to Sentence Transformer- or SetFit-based text classification. LAGONN does not introduce any trainable parameters or new hyperparameters, but typically improves SetFit's performance. To demonstrate the merit of LAGONN, we examined text classification systems in the context of content moderation under four label distributions on five datasets and with growing training data. To our knowledge, this is the first work to examine SetFit in this way. When the training labels are imbalanced, expensive systems, such as LAGONN$_{exp}$ are performant. However, when the distribution is balanced, standard fine-tuning with SetFit can actually hurt model performance. We have therefore proposed an alternative fine-tuning procedure to which LAGONN can be easily utilized, resulting in a powerful, but inexpensive system capable of detecting harmful content.

8 Limitations

In the current work, we have only considered text data, but social media content can of course consist of text, images, and videos. As LaGoNN depends only on an embedding model, an obvious extension to our approach would be examining the modifications we suggest, but on multimodal data. This is an interesting direction that we leave for future research. We have also considered English data, but harmful content can appear in any language. The authors demonstrated that SetFit is performant on multilingual data, the only necessary modification being the underlying pretrained ST. We therefore suspect that LaGoNN would behave similarly on non-English data, but this is not something we have tested ourselves. In order to examine our system's performance under different label-balance distributions, we restricted ourselves to binary and ternary text classification tasks, and LaGoNN therefore remains untested when there are more than three labels. This was an intentional design choice to exploit similar examples in cases with fewer classes and clearer label boundaries. This choice, we believe, is reflective of realistic content moderation settings where fewer labels can be used to filter harmful content. We did not study our method when there are fewer than 100 training examples, and investigating LaGoNN in a few-shot learning setting is fascinating topic for future study. Finally, we note that our system could be misused to detect undesirable content that is not necessarily harmful. For example, a social media website could detect and silence users who complain about the platform. This is not our intended use case, but could result from any classifier, and potential misuse is an unfortunate drawback of all technology.

9 Ethics Statement

It is our sincere goal that our work contributes to the social good in multiple ways. We first hope to have furthered research on text classification that can be feasibly applied to combat undesirable content, such as misinformation, on the Internet, which could potentially cause someone harm. To this end, we have tried to describe our approach as accurately as possible and released our code and data, such that our work is transparent and can be easily reproduced and expanded upon. We hope that we have also created a useful but efficient system which reduces the need to expend energy in the form expensive computation. For example, LaGoNN does not rely on billion-parameter language models that demand thousand-dollar GPUs to use. LaGoNN makes use of GPUs no more than SetFit, despite being more computationally expensive. We have additionally proposed a simple method to make SetFit, an already relatively inexpensive method, even more efficient.

References

Neel Alex, Eli Lifland, Lewis Tunstall, Abhishek Thakur, Pegah Maham, C. Jess Riedel, Emmie Hine, Carolyn Ashurst, Paul Sedille, Alexis Carlier, Michael Noetel, and Andreas Stuhlmüller. 2021. RAFT: A real-world few-shot text classification benchmark. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Gaël Varoquaux. 2013. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pages 108–122.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the 11th International AAAI Conference on Web and Social Media*, ICWSM '17, pages 512–515.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Ullrich K. H. Ecker, Stephan Lewandowsky, John Cook, Philipp Schmid, Lisa K. Fazio, Nadia Brashier, Panayiota Kendeou, Emily K. Vraga, and Michelle A. Amazeen. 2022. The psychological drivers of misinformation belief and its resistance to correction. *Nature Reviews Psychology*, 1(1):13–29.

Rabeeh Karimi Mahabadi, Luke Zettlemoyer, James Henderson, Lambert Mathias, Marzieh Saeidi, Veselin Stoyanov, and Majid Yazdani. 2022. Prompt-free and efficient few-shot learning with language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3638–3652, Dublin, Ireland. Association for Computational Linguistics.

Gregory Koch, Richard Zemel, Ruslan Salakhutdinov, et al. 2015. Siamese neural networks for one-shot image recognition. In *ICML Deep Learning Workshop*, volume 2, page 0. Lille.

Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin Raffel. 2022a. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. *arXiv preprint arXiv:2205.05638*.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022b. What makes good in-context examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, Dublin, Ireland and Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, USA.

Todor Markov, Chong Zhang, Sandhini Agarwal, Tyna Eloundou, Teddy Lee, Steven Adler, Angela Jiang, and Lilian Weng. 2022. A holistic approach to undesired content detection. *arXiv preprint arXiv:2208.03274*.

James O'Neill, Polina Rozenshtein, Ryuichi Kiryo, Motoko Kubota, and Danushka Bollegala. 2021. I wish I would have loved this one, but I didn't – a multilingual dataset for counterfactual detection in product review. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7092–7108, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Christian S. Perone, Roberto Pereira Silveira, and Thomas S. Paula. 2018. Evaluation of sentence embeddings in downstream and linguistic probing tasks. *arXiv preprint arXiv:1806.06259*.

Guangyuan Piao. 2021. Scholarly text classification with sentence bert and entity embeddings. In *Trends and Applications in Knowledge Discovery and Data Mining*, pages 79–87, Cham. Springer International Publishing.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Sarah T. Roberts. 2017. *Content Moderation*, pages 1–4. Springer International Publishing, Cham.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389.

Timo Schick and Hinrich Schütze. 2021a. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269, Online. Association for Computational Linguistics.

Timo Schick and Hinrich Schütze. 2021b. It's not just size that matters: Small language models are also few-shot learners. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2339–2352, Online. Association for Computational Linguistics.

Timo Schick and Hinrich Schütze. 2022. True few-shot learning with Prompts—A real-world perspective. *Transactions of the Association for Computational Linguistics*, 10:716–731.

Yusuke Shido, Hsien-Chi Liu, and Keisuke Umezawa. 2022. Textual content moderation in C2C marketplace. In *Proceedings of the Fifth Workshop on e-Commerce and NLP (ECNLP 5)*, pages 58–62, Dublin, Ireland. Association for Computational Linguistics.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. In *Advances in Neural Information Processing Systems*, volume 33, pages 16857–16867. Curran Associates, Inc.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1).

Lewis Tunstall, Nils Reimers, Unso Eun Seo Jo, Luke Bates, Daniel Korat, Moshe Wasserblat, and Oren Pereg. 2022. Efficient few-shot learning without prompts. *arXiv preprint arXiv:2209.11055*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science*, 359(6380):1146–1151.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Shuohang Wang, Yichong Xu, Yuwei Fang, Yang Liu, Siqi Sun, Ruochen Xu, Chenguang Zhu, and Michael Zeng. 2022. Training data is more valuable than you think: A simple and effective method by retrieving from training data. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3170–3179, Dublin, Ireland. Association for Computational Linguistics.

William Yang Wang. 2017. “Liar, liar pants on fire”: A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 422–426, Vancouver, Canada. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Yichong Xu, Chenguang Zhu, Shuohang Wang, Siqi Sun, Hao Cheng, Xiaodong Liu, Jianfeng Gao, Pengcheng He, Michael Zeng, and Xuedong Huang. 2021. Human parity on commonsenseqa: Augmenting self-attention with external attention. *arXiv preprint arXiv:2112.03254*, abs/2112.03254.

Meng Ye, Karan Sikka, Katherine Atwell, Sabit Hassan, Ajay Divakaran, and Malihe Alikhani. 2023. Multilingual content moderation: A case study on reddit. *arXiv preprint arXiv:2302.09618*.

A Appendix

A.1 Observations about LaGoNN

Here, at the suggestion of an anonymous reviewer, we include a little background on LaGoNN. We originally attempted to use Sentence Transformers/SetFit as a retrieval model that would modify input text and then pass this input to a Transformer-based classifier, such as RoBERTa, instead of back into the ST as in LaGoNN. We experimented with different ST retrieval models and Transformer classifiers, but this system was often beaten by baselines, and performant versions were too expensive to justify their use. The failure of this system is what ultimately inspired LaGoNN. We had hoped to construct a system that did not need to be updated after step one and could simply perform inference on subsequent steps, an active learning setup. While the performance of this version of LaGoNN did not degrade, it also did not appear to learn anything and we found it necessary to update parameters on each step. We additionally tried fine-tuning the embedding model via SetFit first before modifying data, however, this hurt performance in all cases. We include this information for transparency and because we find it interesting.

A.2 Additional results for initial experiments

Here we provide additional results from our initial experimental setup that, due to space limitations, could not be included in the main text. We note that a version of LaGoNN outperforms or has the same performance of all methods, including our upper bound RoBERTa$_{full}$, on 54% of all displayed results, and is the best performer relative to Sentence Transformer-based methods on 72%. This excludes LaGoNN$_{cheap}$. This method showed strong performance on the Insincere Questions dataset, but hurts performance in other cases.

In cases when SetFit-based methods do outperform our system, the performances are comparable, usually within a point, yet they can be quite dramatic when LaGoNN-based methods are the strongest. Below, we report the mean average precision $\times 100$ for all methods over five seeds with the standard deviation, except in the case of Hate Speech Offensive, where the evaluation metric is the macro-F1. Each table shows the results for a given dataset and a given label-balance distribution on the first, fifth, and tenth step followed by the average for all ten steps. In the table caption we provide a summary/interpretation of the results for a given setting. The Liar dataset seems to be the most difficult for all methods. This is expected because it likely does not include enough context to determine the truth of a statement.

Method *Imbalanced*	**Insincere-Questions** 1^{st}	5^{th}	10^{th}	Average
RoBERTa$_{full}$	$39.8_{5.5}$	$\mathbf{53.1}_{4.6}$	$\mathbf{55.7}_{1.2}$	$50.6_{4.4}$
SetFit$_{exp}$	$43.7_{2.7}$	$52.2_{1.9}$	$53.8_{0.9}$	$51.4_{2.9}$
LaGoNN$_{exp}$	$\mathbf{44.5}_{4.5}$	$52.7_{2.4}$	$55.4_{2.0}$	$\mathbf{51.8}_{3.0}$
RoBERTa$_{freeze}$	$39.8_{5.5}$	$44.1_{3.6}$	$46.3_{2.4}$	$44.0_{2.0}$
kNN	$23.9_{2.2}$	$30.3_{3.0}$	$31.6_{2.4}$	$30.0_{2.1}$
SetFit	$43.7_{2.7}$	$47.6_{1.6}$	$50.1_{2.1}$	$47.6_{1.8}$
LaGoNN	$\mathbf{44.5}_{4.5}$	$48.1_{2.2}$	$50.3_{1.7}$	$48.1_{1.9}$
Probe	$40.4_{4.2}$	$49.4_{2.3}$	$52.3_{1.7}$	$49.0_{3.3}$
LaGoNN$_{cheap}$	$40.8_{4.3}$	$51.1_{2.4}$	$54.5_{1.4}$	$50.4_{4.0}$

Table 6: LaGoNN and LaGoNN$_{exp}$ are the strongest performers on the first step, but are overtaken by RoBERTa$_{full}$ on later steps. The average of all steps shows that LaGoNN$_{exp}$ is the overall strongest performer, but we note that LaGoNN$_{cheap}$ shows comparable performance to RoBERTa$_{full}$ despite being much less expensive.

Method *Moderate*	**Insincere Questions** 1^{st}	5^{th}	10^{th}	Average
RoBERTa$_{full}$	$48.1_{2.3}$	$\mathbf{54.7}_{1.9}$	$\mathbf{57.5}_{1.5}$	$53.9_{2.9}$
SetFit$_{exp}$	$48.9_{1.7}$	$53.9_{0.7}$	$54.2_{1.5}$	$52.3_{1.6}$
LaGoNN$_{exp}$	$\mathbf{49.8}_{1.6}$	$52.2_{1.9}$	$53.2_{3.3}$	$52.0_{1.4}$
RoBERTa$_{freeze}$	$48.1_{2.3}$	$50.2_{2.2}$	$52.0_{1.4}$	$50.2_{1.4}$
kNN	$28.0_{2.4}$	$33.9_{2.8}$	$33.6_{2.0}$	$33.5_{1.9}$
SetFit	$48.9_{1.7}$	$53.6_{1.9}$	$55.8_{1.7}$	$53.3_{2.2}$
LaGoNN	$\mathbf{49.8}_{1.6}$	$54.4_{1.3}$	$56.9_{0.5}$	$\mathbf{54.2}_{2.2}$
Probe	$45.7_{2.1}$	$52.3_{1.8}$	$54.4_{1.1}$	$51.4_{2.5}$
LaGoNN$_{cheap}$	$45.7_{2.2}$	$54.4_{1.6}$	$56.4_{0.6}$	$53.2_{3.2}$

Table 7: LaGoNN and LaGoNN$_{exp}$ are the strongest performers on the first step, but are overtaken by RoBERTa$_{full}$ on later steps. The average of all steps shows that LaGoNN is the overall strongest performer, but we note that LaGoNN$_{cheap}$ shows comparable performance to RoBERTa$_{full}$ despite being much less expensive.

Method *Imbalanced*	**Amazon Counterfactual** 1^{st}	5^{th}	10^{th}	Average
RoBERTa$_{full}$	$68.2_{4.5}$	$\mathbf{81.0}_{1.7}$	$\mathbf{82.2}_{1.0}$	$79.2_{3.9}$
SetFit$_{exp}$	$72.0_{2.1}$	$78.4_{2.8}$	$78.8_{1.2}$	$78.0_{2.1}$
LaGoNN$_{exp}$	$\mathbf{74.3}_{3.8}$	$80.1_{1.4}$	$79.0_{1.6}$	$\mathbf{79.5}_{1.9}$
RoBERTa$_{freeze}$	$68.2_{4.5}$	$75.0_{2.2}$	$77.0_{2.4}$	$74.2_{2.6}$
kNN	$51.0_{4.1}$	$60.0_{3.1}$	$61.3_{2.1}$	$59.7_{3.0}$
SetFit	$72.0_{2.1}$	$74.4_{2.3}$	$76.7_{1.8}$	$74.8_{1.4}$
LaGoNN	$\mathbf{74.3}_{3.8}$	$76.1_{3.6}$	$77.3_{3.2}$	$76.1_{1.0}$
Probe	$46.6_{2.8}$	$60.3_{1.4}$	$64.2_{1.2}$	$59.2_{5.2}$
LaGoNN$_{cheap}$	$38.2_{3.2}$	$55.3_{1.8}$	$61.0_{1.2}$	$54.4_{6.7}$

Table 8: LaGoNN and LaGoNN$_{exp}$ are the strongest performers on the first step, but are overtaken by RoBERTa$_{full}$ on later steps. However, the average of all steps shows that LaGoNN$_{exp}$ is the overall strongest performer.

Method *Moderate*	**Amazon Counterfactual** 1^{st}	5^{th}	10^{th}	Average
RoBERTa$_{full}$	$73.9_{2.5}$	$80.0_{1.0}$	$80.1_{2.3}$	$79.1_{2.1}$
SetFit$_{exp}$	$76.5_{1.6}$	$77.0_{2.4}$	$74.7_{0.5}$	$76.5_{1.0}$
LaGoNN$_{exp}$	$\mathbf{78.6}_{2.2}$	$78.0_{2.1}$	$76.3_{4.9}$	$78.2_{1.0}$
RoBERTa$_{freeze}$	$73.9_{2.5}$	$76.6_{1.4}$	$78.5_{0.7}$	$76.4_{1.7}$
kNN	$54.5_{3.1}$	$64.2_{1.9}$	$66.6_{1.3}$	$64.7_{3.5}$
SetFit	$76.5_{1.6}$	$80.6_{0.5}$	$81.2_{0.3}$	$80.0_{1.4}$
LaGoNN	$\mathbf{78.6}_{2.2}$	$\mathbf{81.2}_{1.4}$	$\mathbf{81.6}_{1.1}$	$\mathbf{80.8}_{0.9}$
Probe	$52.3_{2.0}$	$64.1_{1.8}$	$67.2_{1.4}$	$63.1_{4.3}$
LaGoNN$_{cheap}$	$47.3_{3.4}$	$60.7_{1.5}$	$65.2_{1.4}$	$59.5_{5.2}$

Table 9: LaGoNN$_{exp}$ and LaGoNN are the strongest performers on the first step, but LaGoNN is strongest classifier on subsequent steps and is also the overall strongest performer based on the average over all steps.

Method *Extreme*	**Toxic Conversations** 1^{st}	5^{th}	10^{th}	Average
RoBERTa$_{full}$	$7.9_{0.5}$	$21.2_{3.7}$	$\mathbf{33.8}_{5.5}$	$21.9_{9.3}$
SetFit$_{exp}$	$8.8_{1.2}$	$18.1_{3.4}$	$24.7_{4.1}$	$17.6_{5.5}$
LaGoNN$_{exp}$	$8.9_{1.7}$	$17.4_{6.6}$	$26.4_{5.2}$	$17.9_{6.0}$
RoBERTa$_{freeze}$	$7.9_{0.5}$	$12.8_{2.4}$	$19.1_{3.2}$	$13.5_{3.5}$
kNN	$7.9_{0.0}$	$8.7_{0.4}$	$8.7_{0.2}$	$8.5_{0.3}$
SetFit	$8.8_{1.2}$	$13.1_{2.5}$	$16.3_{3.0}$	$13.0_{2.6}$
LaGoNN	$8.9_{1.7}$	$13.8_{3.9}$	$17.1_{4.8}$	$13.4_{2.6}$
Probe	$\mathbf{13.1}_{2.8}$	$\mathbf{24.6}_{2.6}$	$30.1_{2.1}$	$\mathbf{23.9}_{5.6}$
LaGoNN$_{cheap}$	$11.3_{2.2}$	$21.7_{2.7}$	$27.4_{2.3}$	$21.3_{5.3}$

Table 10: Probe is strongest performer on every step, except the 10^{th} where it is overtaken by RoBERTa$_{full}$. If we average over all steps, we see that Probe is the strongest performer. We note, however, that LaGoNN and LaGoNN$_{exp}$ outperform SetFit and SetFit$_{exp}$ on all steps.

Method *Imbalanced*	1^{st}	**Toxic Conversations** 5^{th}	10^{th}	Average
RoBERTa$_{full}$	$\mathbf{24.1}_{5.6}$	$43.1_{3.4}$	$52.1_{2.5}$	$42.4_{8.2}$
SetFit$_{exp}$	$21.8_{6.6}$	$44.5_{4.1}$	$51.4_{1.9}$	$42.1_{9.3}$
LaGoNN$_{exp}$	$22.7_{9.8}$	$\mathbf{49.1}_{5.6}$	$\mathbf{53.4}_{2.3}$	$\mathbf{45.6}_{9.8}$
RoBERTa$_{freeze}$	$\mathbf{24.1}_{5.6}$	$31.2_{4.4}$	$34.0_{4.0}$	$30.5_{3.1}$
kNN	$11.5_{2.5}$	$14.7_{4.0}$	$15.3_{3.2}$	$14.6_{1.1}$
SetFit	$21.8_{6.6}$	$26.7_{5.3}$	$30.2_{4.0}$	$26.6_{2.7}$
LaGoNN	$22.7_{9.8}$	$27.6_{8.9}$	$30.3_{8.7}$	$27.4_{2.4}$
Probe	$23.3_{2.7}$	$33.0_{2.8}$	$37.1_{1.8}$	$32.5_{4.2}$
LaGoNN$_{cheap}$	$20.5_{3.2}$	$31.1_{3.2}$	$35.6_{1.8}$	$30.5_{4.6}$

Table 11: RoBERTa$_{full}$ and RoBERTa$_{freeze}$ are the strongest performers on the first step, but are overtaken by LaGoNN$_{exp}$ for the subsequent steps. The overall strongest performer based on the average over all steps is LaGoNN$_{exp}$.

Method *Moderate*	1^{st}	**Toxic Conversations** 5^{th}	10^{th}	Average
RoBERTa$_{full}$	$34.2_{3.4}$	$45.5_{1.9}$	$52.4_{3.3}$	$45.7_{5.6}$
SetFit$_{exp}$	$33.6_{2.9}$	$47.2_{2.2}$	$46.6_{3.3}$	$44.3_{4.3}$
LaGoNN$_{exp}$	$\mathbf{36.6}_{4.2}$	$\mathbf{48.2}_{2.7}$	$\mathbf{49.9}_{3.7}$	$\mathbf{48.0}_{4.4}$
RoBERTa$_{freeze}$	$34.2_{3.4}$	$38.4_{2.1}$	$39.5_{1.8}$	$38.0_{1.5}$
kNN	$19.4_{1.9}$	$21.5_{3.4}$	$22.4_{2.9}$	$21.6_{0.8}$
SetFit	$33.6_{2.9}$	$39.2_{2.9}$	$41.6_{2.7}$	$38.6_{2.4}$
LaGoNN	$\mathbf{36.6}_{4.2}$	$42.7_{3.7}$	$45.0_{3.5}$	$42.0_{2.5}$
Probe	$29.0_{2.7}$	$36.1_{1.2}$	$39.1_{1.5}$	$35.5_{3.3}$
LaGoNN$_{cheap}$	$26.1_{2.7}$	$34.3_{1.3}$	$37.5_{1.8}$	$33.6_{3.6}$

Table 12: LaGoNN and LaGoNN$_{exp}$ are the strongest performers on the first step and LaGoNN$_{exp}$ remains the strongest for subsequent steps, also being the strongest classifier overall based on the average.

Method *Balanced*	1^{st}	**Toxic Conversations** 5^{th}	10^{th}	Average
RoBERTa$_{full}$	$32.3_{1.1}$	$42.7_{1.8}$	$\mathbf{54.1}_{3.4}$	$43.8_{6.3}$
SetFit$_{exp}$	$35.7_{3.4}$	$32.6_{6.2}$	$37.4_{2.7}$	$36.5_{1.9}$
LaGoNN$_{exp}$	$\mathbf{40.4}_{4.4}$	$40.2_{6.6}$	$39.8_{7.5}$	$40.0_{1.2}$
RoBERTa$_{freeze}$	$32.3_{1.1}$	$39.2_{1.5}$	$41.0_{0.6}$	$38.5_{2.4}$
kNN	$17.4_{0.8}$	$23.7_{2.6}$	$24.3_{2.7}$	$23.1_{2.0}$
SetFit	$35.7_{3.4}$	$44.5_{2.9}$	$46.1_{2.8}$	$43.6_{2.9}$
LaGoNN	$\mathbf{40.4}_{4.4}$	$\mathbf{46.6}_{2.7}$	$48.1_{2.2}$	$\mathbf{46.1}_{2.2}$
Probe	$29.5_{2.4}$	$35.9_{0.9}$	$40.2_{0.9}$	$36.1_{3.5}$
LaGoNN$_{cheap}$	$26.8_{2.7}$	$34.5_{1.3}$	$38.5_{0.8}$	$34.4_{3.7}$

Table 13: LaGoNN and LaGoNN$_{exp}$ are the strongest performers on the first step. LaGoNN remains the strongest until the 10^{th}, where it is overtaken by RoBERTa$_{full}$. Overall, LaGoNN is the strongest classifier based on the average. Note the performance of SetFit$_{exp}$ and LaGoNN$_{exp}$. While both degrade after the first step, LaGoNN$_{exp}$'s performance drop is dramatically mitigated.

Method		**Hate Speech Offensive**		
Extreme	1^{st}	5^{th}	10^{th}	Average
RoBERTa$_{full}$	$30.2_{1.4}$	$43.5_{2.5}$	$\mathbf{51.2}_{2.2}$	$\mathbf{44.3}_{7.4}$
SetFit$_{exp}$	$30.3_{0.8}$	$\mathbf{44.0}_{1.3}$	$51.1_{2.0}$	$43.8_{6.5}$
LaGoNN$_{exp}$	$30.3_{0.7}$	$40.7_{2.9}$	$49.1_{4.4}$	$42.2_{6.2}$
RoBERTa$_{freeze}$	$30.2_{1.4}$	$33.5_{3.1}$	$34.4_{3.4}$	$33.1_{1.4}$
kNN	$\mathbf{31.5}_{1.2}$	$35.9_{2.7}$	$37.4_{2.0}$	$35.8_{1.7}$
SetFit	$30.3_{0.8}$	$38.4_{2.5}$	$41.1_{1.5}$	$37.8_{3.3}$
LaGoNN	$30.3_{0.7}$	$35.7_{2.6}$	$39.1_{2.4}$	$35.6_{2.7}$
Probe	$29.0_{0.2}$	$34.7_{1.5}$	$40.1_{2.1}$	$35.1_{3.8}$
LaGoNN$_{cheap}$	$29.0_{0.1}$	$36.9_{1.8}$	$40.5_{2.1}$	$36.2_{3.7}$

Table 14: kNN is the strongest performer on the first step, while SetFit$_{exp}$ is on the 5^{th}, and RoBERTa$_{full}$ is the strongest on the 10^{th} while also being strongest overall performer for all steps. LaGoNN-based methods are generally beaten by ST/SetFit-based baselines, with the exception of LaGoNN$_{cheap}$ which consistently outperforms Probe.

Method		**Hate Speech Offensive**		
Imbalanced	1^{st}	5^{th}	10^{th}	Average
RoBERTa$_{full}$	$50.6_{3.0}$	$65.2_{3.9}$	$\mathbf{70.3}_{1.2}$	$64.2_{5.3}$
SetFit$_{exp}$	$54.4_{4.3}$	$66.3_{1.8}$	$68.9_{2.0}$	$64.3_{4.5}$
LaGoNN$_{exp}$	$\mathbf{57.0}_{5.2}$	$\mathbf{67.0}_{4.4}$	$69.8_{2.1}$	$\mathbf{64.9}_{4.6}$
RoBERTa$_{freeze}$	$50.6_{3.0}$	$54.1_{1.6}$	$55.3_{2.3}$	$54.1_{1.3}$
kNN	$55.6_{4.8}$	$57.3_{2.3}$	$58.8_{3.6}$	$57.4_{1.1}$
SetFit	$54.4_{4.3}$	$57.0_{3.9}$	$58.2_{3.8}$	$57.2_{1.1}$
LaGoNN	$\mathbf{57.0}_{5.2}$	$58.2_{4.1}$	$58.3_{3.4}$	$58.3_{0.6}$
Probe	$46.5_{2.2}$	$57.8_{1.7}$	$60.3_{1.2}$	$56.5_{4.5}$
LaGoNN$_{cheap}$	$47.1_{1.3}$	$56.5_{2.2}$	$59.5_{2.5}$	$55.6_{3.8}$

Table 15: LaGoNN and LaGoNN$_{exp}$ are the strongest performers on the first step, with LaGoNN$_{exp}$ being the strongest on the 5^{th} and RoBERTa$_{full}$ taking over on the 10^{th}. LaGoNN$_{exp}$ is the strongest performer overall based on the average over all steps.

Method		**Hate Speech Offensive**		
Moderate	1^{st}	5^{th}	10^{th}	Average
RoBERTa$_{full}$	$61.9_{3.4}$	$70.8_{1.0}$	$\mathbf{72.5}_{1.4}$	$69.9_{3.2}$
SetFit$_{exp}$	$\mathbf{64.3}_{4.2}$	$70.6_{2.4}$	$72.4_{0.5}$	$69.8_{2.8}$
LaGoNN$_{exp}$	$63.8_{4.9}$	$\mathbf{71.0}_{2.1}$	$72.3_{1.0}$	$\mathbf{70.0}_{3.0}$
RoBERTa$_{freeze}$	$61.9_{3.4}$	$63.2_{4.1}$	$64.1_{4.5}$	$63.2_{0.6}$
kNN	$\mathbf{64.3}_{4.0}$	$63.3_{2.9}$	$63.9_{2.5}$	$63.7_{0.4}$
SetFit	$\mathbf{64.3}_{4.2}$	$67.3_{3.2}$	$67.6_{2.3}$	$66.9_{1.1}$
LaGoNN	$63.8_{4.9}$	$65.0_{5.3}$	$66.7_{5.9}$	$65.3_{0.9}$
Probe	$55.6_{1.7}$	$63.8_{0.8}$	$66.1_{0.3}$	$63.2_{3.0}$
LaGoNN$_{cheap}$	$56.0_{3.6}$	$62.2_{1.4}$	$66.0_{0.9}$	$62.3_{2.9}$

Table 16: kNN, SetFit, and SetFit$_{exp}$ start the strongest, but are overtaken by LaGoNN$_{exp}$ on the 5^{th} step, which is in turn overtaken by RoBERTa$_{full}$ on the 10^{th} step. Overall LaGoNN$_{exp}$ is the strongest performer based on the average.

Method		**Hate Speech Offensive**		
Balanced	1^{st}	5^{th}	10^{th}	Average
RoBERTa$_{full}$	$59.7_{3.5}$	$\mathbf{66.9}_{1.2}$	$\mathbf{69.2}_{1.8}$	$\mathbf{66.4}_{2.7}$
SetFit$_{exp}$	$60.7_{1.3}$	$66.3_{1.6}$	$67.5_{0.9}$	$65.9_{2.2}$
LaGoNN$_{exp}$	$\mathbf{61.5}_{1.7}$	$66.4_{1.4}$	$67.7_{0.9}$	$66.1_{1.8}$
RoBERTa$_{freeze}$	$59.7_{3.5}$	$60.4_{2.7}$	$63.1_{2.3}$	$61.0_{1.3}$
kNN	$60.7_{1.3}$	$59.6_{2.8}$	$59.5_{2.5}$	$59.5_{0.5}$
SetFit	$60.7_{1.3}$	$62.5_{0.7}$	$63.4_{1.0}$	$62.3_{1.0}$
LaGoNN	$\mathbf{61.5}_{1.7}$	$62.8_{1.5}$	$64.2_{1.0}$	$63.0_{0.9}$
Probe	$54.9_{1.4}$	$58.5_{0.9}$	$60.9_{0.4}$	$58.7_{1.7}$
LaGoNN$_{cheap}$	$54.2_{2.3}$	$58.6_{0.6}$	$60.6_{0.5}$	$58.5_{1.8}$

Table 17: LaGoNN and LaGoNN$_{exp}$ are the strongest performers on the first step, but are overtaken by RoBERTa$_{full}$ on later steps, which also is the strongest overall classifier. We note that LaGoNN and LaGoNN$_{exp}$ consistently outperform SetFit and SetFit$_{exp}$, respectively.

Method		**Liar**		
Extreme	1^{st}	5^{th}	10^{th}	Average
RoBERTa$_{full}$	$\mathbf{32.0}_{2.7}$	$\mathbf{34.7}_{2.9}$	$35.1_{4.3}$	$33.7_{1.0}$
SetFit$_{exp}$	$31.2_{3.8}$	$30.4_{3.1}$	$31.8_{2.9}$	$31.5_{0.7}$
LaGoNN$_{exp}$	$30.6_{4.7}$	$30.3_{2.0}$	$31.3_{2.0}$	$31.1_{0.6}$
RoBERTa$_{freeze}$	$\mathbf{32.0}_{2.7}$	$32.8_{4.5}$	$34.2_{5.0}$	$33.2_{0.7}$
kNN	$27.0_{0.5}$	$27.3_{0.8}$	$27.9_{0.8}$	$27.4_{0.3}$
SetFit	$31.2_{3.8}$	$33.7_{5.1}$	$\mathbf{35.7}_{5.1}$	$\mathbf{34.3}_{1.6}$
LaGoNN	$30.6_{4.7}$	$32.0_{4.6}$	$33.7_{5.4}$	$32.6_{0.9}$
Probe	$30.7_{2.0}$	$30.6_{3.9}$	$31.7_{2.9}$	$31.1_{0.4}$
LaGoNN$_{cheap}$	$30.7_{2.0}$	$30.5_{3.8}$	$31.4_{2.6}$	$31.0_{0.4}$

Table 18: RoBERTa$_{freeze}$ and RoBERTa$_{full}$ start out as the strongest performers but are eventually overtaken by SetFit on the 10^{th} step, and SetFit ends up being the strongest performer over all steps based on the average.

Method		**Liar**		
Imbalanced	1^{st}	5^{th}	10^{th}	Average
RoBERTa$_{full}$	$31.4_{3.2}$	$35.8_{2.6}$	$\mathbf{40.0}_{4.3}$	$36.2_{2.4}$
SetFit$_{exp}$	$\mathbf{32.3}_{4.5}$	$35.9_{3.1}$	$36.4_{2.2}$	$35.2_{1.1}$
LaGoNN$_{exp}$	$\mathbf{32.3}_{4.6}$	$35.7_{3.4}$	$36.5_{2.3}$	$35.7_{1.4}$
RoBERTa$_{freeze}$	$31.4_{3.2}$	$34.1_{2.6}$	$35.6_{3.2}$	$34.0_{1.4}$
kNN	$27.0_{0.2}$	$28.5_{1.0}$	$29.0_{1.0}$	$28.7_{0.7}$
SetFit	$\mathbf{32.3}_{4.5}$	$\mathbf{36.5}_{3.1}$	$38.5_{3.4}$	$\mathbf{36.3}_{2.0}$
LaGoNN	$\mathbf{32.3}_{4.6}$	$34.9_{2.2}$	$36.9_{2.5}$	$35.3_{1.4}$
Probe	$30.7_{3.0}$	$32.8_{1.8}$	$35.0_{1.6}$	$33.5_{1.5}$
LaGoNN$_{cheap}$	$30.4_{3.0}$	$32.9_{1.8}$	$35.4_{1.7}$	$33.5_{1.7}$

Table 19: SetFit, SetFit$_{exp}$, LaGoNN, and LaGoNN$_{exp}$ start out as the strongest performers. On the 5^{th} step, SetFit is overtaken the other systems, but is eventually overtaken by RoBERTa$_{full}$. Overall SetFit is the strongest system, but we note that LaGoNN$_{exp}$ outperforms SetFit$_{exp}$.

Method *Moderate*	1^{st}	**Liar** 5^{th}	10^{th}	Average
RoBERTa$_{full}$	$33.9_{3.1}$	$38.4_{2.7}$	$\mathbf{43.9}_{2.2}$	$\mathbf{39.5}_{3.0}$
SetFit$_{exp}$	$33.0_{2.6}$	$37.2_{1.8}$	$38.7_{1.5}$	$37.4_{1.6}$
LaGoNN$_{exp}$	$\mathbf{34.1}_{3.4}$	$\mathbf{38.7}_{2.3}$	$39.0_{1.8}$	$37.8_{1.5}$
RoBERTa$_{freeze}$	$33.9_{3.1}$	$35.3_{2.6}$	$36.8_{2.2}$	$35.4_{1.0}$
kNN	$29.2_{0.8}$	$29.7_{1.5}$	$30.0_{0.6}$	$29.8_{0.3}$
SetFit	$33.0_{2.6}$	$37.2_{3.9}$	$39.4_{3.5}$	$37.0_{1.8}$
LaGoNN	$\mathbf{34.1}_{3.4}$	$37.0_{3.1}$	$38.6_{3.0}$	$36.8_{1.3}$
Probe	$31.6_{1.1}$	$34.7_{2.5}$	$37.0_{2.5}$	$34.9_{1.7}$
LaGoNN$_{cheap}$	$31.4_{0.9}$	$35.3_{2.3}$	$37.6_{2.0}$	$35.3_{1.9}$

Table 20: LaGoNN and LaGoNN$_{exp}$ start out as the strongest performers and LaGoNN$_{exp}$ continues to be strong, until the 10^{th} step where it is overtaken by RoBERTa$_{full}$, which ends up as the most performant classifier over all steps based on the average.

Method *Balanced*	1^{st}	**Liar** 5^{th}	10^{th}	Average
RoBERTa$_{full}$	$33.8_{2.1}$	$\mathbf{39.4}_{2.4}$	$\mathbf{43.5}_{1.7}$	$\mathbf{40.2}_{3.2}$
SetFit$_{exp}$	$\mathbf{34.4}_{2.3}$	$36.7_{1.7}$	$37.0_{1.3}$	$36.5_{1.1}$
LaGoNN$_{exp}$	$33.8_{1.8}$	$34.2_{2.7}$	$37.2_{1.9}$	$36.2_{1.4}$
RoBERTa$_{freeze}$	$33.8_{2.1}$	$36.6_{1.6}$	$38.6_{1.5}$	$36.7_{1.5}$
kNN	$30.1_{0.4}$	$31.3_{2.1}$	$30.6_{1.1}$	$30.9_{0.4}$
SetFit	$\mathbf{34.4}_{2.3}$	$38.3_{2.5}$	$40.0_{2.0}$	$37.9_{1.6}$
LaGoNN	$33.8_{1.8}$	$38.3_{1.3}$	$40.6_{0.6}$	$38.1_{2.0}$
Probe	$32.1_{1.9}$	$35.2_{1.4}$	$37.2_{2.5}$	$35.2_{1.7}$
LaGoNN$_{cheap}$	$31.9_{1.9}$	$36.0_{1.0}$	$37.5_{2.5}$	$35.7_{1.8}$

Table 21: SetFit and SetFit$_{exp}$ are the most performant systems on the first step, but are overtaken by RoBERTa$_{full}$, the strongest overall classifier. We note that LaGoNN outperforms SetFit after the first step and in aggregate.

A.3 Additional results for secondary experiments

Here, we provide additional results from our second set of experiments that, due to space limitations, could not be included in the main text. We note that a version of LaGoNN outperforms or has the same performance of all methods, including our upper bound RoBERTa$_{full}$, on 60% of all displayed results, and is the best performer relative to Sentence Transformer-based methods on 65%. This excludes LaGoNN$_{cheap}$. This method showed strong performance on the Insincere Questions dataset, but hurts performance in other cases.

In cases when SetFit-based methods do outperform our system, the performances are comparable, usually within one point, yet they can be quite different when LaGoNN-based methods are the strongest. Below, we report the mean average precision $\times 100$ for all methods over five seeds with the standard deviation, except in the case of Hate Speech Offensive, where the evaluation metric is the macro-F1. Each table shows the results for a given dataset and a given label-balance distribution on the first, fifth, and tenth step followed by the average for all ten steps. In the table caption we provide a summary/interpretation of the results for a given setting. Liar appears to be the most difficult dataset for all methods. This is expected because it likely does not include enough context to determine the truth of a statement.

Method *Extreme*	1^{st}	**Insincere Questions** 5^{th}	10^{th}	Average
RoBERTa$_{full}$	$19.9_{8.4}$	$30.9_{7.9}$	$42.0_{7.4}$	$33.5_{6.7}$
SetFit$_{exp}$	$24.1_{6.3}$	$29.2_{6.7}$	$36.7_{7.3}$	$31.7_{3.4}$
LaGoNN$_{exp}$	$\mathbf{30.7}_{8.9}$	$37.6_{6.1}$	$39.0_{6.1}$	$36.1_{2.3}$
SetFit$_{lite}$	$24.1_{6.3}$	$38.1_{6.3}$	$41.1_{6.5}$	$35.6_{5.5}$
LaGoNN$_{lite}$	$\mathbf{30.7}_{8.9}$	$\mathbf{41.8}_{8.3}$	$\mathbf{43.4}_{8.5}$	$\mathbf{39.3}_{4.4}$
RoBERTa$_{freeze}$	$19.9_{8.4}$	$34.1_{5.4}$	$37.9_{5.2}$	$32.5_{5.4}$
kNN	$6.8_{0.4}$	$15.9_{3.4}$	$16.9_{4.3}$	$14.4_{3.0}$
SetFit	$24.1_{6.3}$	$31.7_{4.9}$	$36.1_{5.4}$	$31.8_{3.6}$
LaGoNN	$\mathbf{30.7}_{8.9}$	$39.3_{4.9}$	$41.2_{4.7}$	$38.4_{3.0}$
Probe	$24.3_{8.4}$	$39.8_{5.6}$	$44.8_{4.2}$	$38.3_{6.2}$
LaGoNN$_{cheap}$	$23.6_{7.8}$	$40.7_{5.9}$	$45.3_{4.4}$	$38.6_{6.6}$

Table 22: LaGoNN, LaGoNN$_{lite}$, and LaGoNN$_{exp}$ start out as the strongest models, but LaGoNN$_{lite}$ remains the most performant by the 10^{th} step. It is also the overall strongest performer based on the average. We note the strength of LaGoNN$_{cheap}$ relative to far more expensive methods.

Method *Imbalanced*	1^{st}	**Insincere Questions** 5^{th}	10^{th}	Average
RoBERTa$_{full}$	$39.8_{5.5}$	$53.1_{4.6}$	$55.7_{1.2}$	$50.6_{4.4}$
SetFit$_{exp}$	$43.7_{2.7}$	$52.2_{1.9}$	$53.8_{0.9}$	$51.4_{2.9}$
LaGoNN$_{exp}$	$\mathbf{44.5}_{4.5}$	$52.7_{2.4}$	$55.4_{2.0}$	$51.8_{3.0}$
SetFit$_{lite}$	$43.7_{2.7}$	$52.9_{2.6}$	$55.8_{1.8}$	$52.2_{3.4}$
LaGoNN$_{lite}$	$\mathbf{44.5}_{4.5}$	$\mathbf{53.5}_{2.7}$	$\mathbf{55.9}_{2.4}$	$\mathbf{52.6}_{3.5}$
RoBERTa$_{freeze}$	$39.8_{5.5}$	$44.1_{3.6}$	$46.3_{2.4}$	$44.0_{2.0}$
kNN	$23.9_{2.2}$	$30.3_{3.0}$	$31.6_{2.4}$	$30.0_{2.1}$
SetFit	$43.7_{2.7}$	$47.6_{1.6}$	$50.1_{2.1}$	$47.6_{1.8}$
LaGoNN	$\mathbf{44.5}_{4.5}$	$48.1_{2.2}$	$50.3_{1.7}$	$48.1_{1.9}$
Probe	$40.4_{4.2}$	$49.4_{2.3}$	$52.3_{1.7}$	$49.0_{3.3}$
LaGoNN$_{cheap}$	$40.8_{4.3}$	$51.1_{2.4}$	$54.5_{1.4}$	$50.4_{4.0}$

Table 23: LaGoNN, LaGoNN$_{lite}$, and LaGoNN$_{exp}$ start out as the strongest models, but LaGoNN$_{lite}$ remains the most performant by the 10^{th} step. It is also the overall strongest performer based on the average. We note the strength of LaGoNN$_{cheap}$ relative to far more expensive methods.

Method		**Insincere Questions**		
Moderate	1^{st}	5^{th}	10^{th}	Average
RoBERTa$_{full}$	$48.1_{2.3}$	$54.7_{1.9}$	$57.5_{1.5}$	$53.9_{2.9}$
SetFit$_{exp}$	$48.9_{1.7}$	$53.9_{0.7}$	$54.2_{1.5}$	$52.3_{1.6}$
LaGoNN$_{exp}$	$\mathbf{49.8}_{1.6}$	$52.2_{1.9}$	$53.2_{3.3}$	$52.0_{1.4}$
SetFit$_{lite}$	$48.9_{1.7}$	$\mathbf{56.5}_{1.4}$	$\mathbf{58.7}_{0.6}$	$\mathbf{55.0}_{3.5}$
LaGoNN$_{lite}$	$\mathbf{49.8}_{1.6}$	$56.1_{2.8}$	$58.3_{1.5}$	$54.6_{3.5}$
RoBERTa$_{freeze}$	$48.1_{2.3}$	$50.2_{2.2}$	$52.0_{1.4}$	$50.2_{1.4}$
kNN	$28.0_{2.4}$	$33.9_{2.8}$	$33.6_{2.0}$	$33.5_{1.9}$
SetFit	$48.9_{1.7}$	$53.6_{1.9}$	$55.8_{1.7}$	$53.3_{2.2}$
LaGoNN	$\mathbf{49.8}_{1.6}$	$54.4_{1.3}$	$56.9_{0.5}$	$54.2_{2.2}$
Probe	$45.7_{2.1}$	$52.3_{1.8}$	$54.4_{1.1}$	$51.4_{2.5}$
LaGoNN$_{cheap}$	$45.7_{2.2}$	$54.4_{1.6}$	$56.4_{0.6}$	$53.2_{3.2}$

Table 24: LaGoNN, LaGoNN$_{lite}$, and LaGoNN$_{exp}$ start out as the strongest models, but SetFit$_{lite}$ overtakes the other methods by the 5^{th} step and is the strongest performer based on the average. We note the strength of LaGoNN$_{cheap}$ relative to far more expensive methods.

Method		**Insincere Questions**		
Balanced	1^{st}	5^{th}	10^{th}	Average
RoBERTa$_{full}$	$47.1_{4.2}$	$52.1_{3.6}$	$55.7_{2.6}$	$52.5_{2.9}$
SetFit$_{exp}$	$43.5_{4.2}$	$47.1_{4.6}$	$48.5_{3.9}$	$48.0_{1.7}$
LaGoNN$_{exp}$	$42.8_{5.3}$	$47.6_{2.9}$	$47.0_{1.7}$	$46.2_{2.0}$
SetFit$_{lite}$	$43.5_{4.2}$	$\mathbf{54.6}_{2.4}$	$\mathbf{59.6}_{0.9}$	$53.6_{5.8}$
LaGoNN$_{lite}$	$42.8_{5.3}$	$53.5_{3.7}$	$58.6_{2.5}$	$52.2_{6.4}$
RoBERTa$_{freeze}$	$47.1_{4.2}$	$52.1_{0.4}$	$53.3_{1.1}$	$51.5_{2.1}$
kNN	$22.3_{2.3}$	$30.2_{2.3}$	$30.9_{1.8}$	$29.5_{2.5}$
SetFit	$43.5_{4.2}$	$53.8_{2.2}$	$55.5_{1.6}$	$52.8_{3.5}$
LaGoNN	$42.8_{5.3}$	$54.1_{2.9}$	$56.3_{1.3}$	$53.4_{3.7}$
Probe	$47.5_{1.6}$	$52.4_{1.7}$	$55.3_{1.1}$	$52.2_{2.5}$
LaGoNN$_{cheap}$	$\mathbf{49.3}_{2.6}$	$54.4_{1.4}$	$57.6_{0.7}$	$\mathbf{54.2}_{2.7}$

Table 25: LaGoNN$_{cheap}$, starts out as the strongest model, but SetFit$_{lite}$ overtakes the other methods on the 5^{th} and 10^{th} step. Overall LaGoNN$_{cheap}$ is the strongest model despite being one of the least expensive.

Method		**Amazon Counterfactual**		
Extreme	1^{st}	5^{th}	10^{th}	Average
RoBERTa$_{full}$	$21.8_{6.6}$	$63.9_{10.2}$	$72.3_{3.0}$	$59.6_{16.8}$
SetFit$_{exp}$	$22.3_{8.8}$	$64.2_{3.3}$	$68.6_{4.6}$	$56.8_{14.9}$
LaGoNN$_{exp}$	$\mathbf{26.1}_{17.5}$	$\mathbf{68.4}_{4.4}$	$\mathbf{74.9}_{2.9}$	$\mathbf{63.2}_{16.7}$
SetFit$_{lite}$	$22.3_{8.8}$	$62.4_{5.1}$	$67.5_{5.2}$	$56.5_{14.7}$
LaGoNN$_{lite}$	$\mathbf{26.1}_{17.5}$	$68.3_{4.3}$	$68.9_{4.3}$	$60.6_{15.1}$
RoBERTa$_{freeze}$	$21.8_{6.6}$	$41.0_{12.7}$	$51.3_{10.7}$	$40.6_{8.9}$
kNN	$10.3_{0.2}$	$15.3_{4.2}$	$18.4_{3.7}$	$15.6_{2.4}$
SetFit	$22.3_{8.8}$	$32.4_{11.5}$	$42.3_{8.8}$	$34.5_{5.9}$
LaGoNN	$\mathbf{26.1}_{17.5}$	$31.1_{19.4}$	$33.0_{19.1}$	$30.9_{2.3}$
Probe	$24.2_{9.0}$	$46.3_{4.4}$	$54.6_{2.0}$	$45.1_{10.3}$
LaGoNN$_{cheap}$	$20.1_{6.9}$	$38.3_{4.9}$	$47.8_{3.4}$	$38.2_{9.5}$

Table 26: LaGoNN, LaGoNN$_{lite}$, and LaGoNN$_{exp}$ are the most performant models on the first step, but only LaGoNN$_{exp}$ remains the most performant on subsequent steps, also being the strongest overall method based on the average over all steps.

Method		**Amazon Counterfactual**		
Imbalanced	1^{st}	5^{th}	10^{th}	Average
RoBERTa$_{full}$	$68.2_{4.5}$	$\mathbf{81.0}_{1.7}$	$\mathbf{82.2}_{1.0}$	$79.2_{3.9}$
SetFit$_{exp}$	$72.0_{2.1}$	$78.4_{2.8}$	$78.8_{1.2}$	$78.0_{2.1}$
LaGoNN$_{exp}$	$\mathbf{74.3}_{3.8}$	$80.1_{1.4}$	$79.0_{1.6}$	$79.5_{1.9}$
SetFit$_{lite}$	$72.0_{2.1}$	$79.1_{1.4}$	$81.6_{1.3}$	$79.1_{2.7}$
LaGoNN$_{lite}$	$\mathbf{74.3}_{3.8}$	$79.2_{1.7}$	$81.9_{1.1}$	$\mathbf{80.2}_{2.2}$
RoBERTa$_{freeze}$	$68.2_{4.5}$	$75.0_{2.2}$	$77.0_{2.4}$	$74.2_{2.6}$
kNN	$51.0_{4.1}$	$60.0_{3.1}$	$61.3_{2.1}$	$59.7_{3.0}$
SetFit	$72.0_{2.1}$	$74.4_{2.3}$	$76.7_{1.8}$	$74.8_{1.4}$
LaGoNN	$\mathbf{74.3}_{3.8}$	$76.1_{3.6}$	$77.3_{3.2}$	$76.1_{1.0}$
Probe	$46.6_{2.8}$	$60.3_{1.4}$	$64.2_{1.2}$	$59.2_{5.2}$
LaGoNN$_{cheap}$	$38.2_{3.2}$	$55.3_{1.8}$	$61.0_{1.2}$	$54.4_{6.7}$

Table 27: On the first step, LaGoNN, LaGoNN$_{lite}$, and LaGoNN$_{exp}$ start out the strongest but LaGoNN$_{lite}$ performs slightly worse than RoBERTa$_{full}$ on the 5^{th} and 10^{th} step. However, LaGoNN$_{lite}$ is the best overall method based on the average.

Method		**Amazon Counterfactual**		
Moderate	1^{st}	5^{th}	10^{th}	Average
RoBERTa$_{full}$	$73.9_{2.5}$	$80.0_{1.0}$	$80.1_{2.3}$	$79.1_{2.1}$
SetFit$_{exp}$	$76.5_{1.6}$	$77.0_{2.4}$	$74.7_{0.5}$	$76.5_{1.0}$
LaGoNN$_{exp}$	$\mathbf{78.6}_{2.2}$	$78.0_{2.1}$	$76.3_{4.9}$	$78.2_{1.0}$
SetFit$_{lite}$	$76.5_{1.6}$	$80.4_{3.8}$	$\mathbf{83.5}_{0.8}$	$80.3_{2.8}$
LaGoNN$_{lite}$	$\mathbf{78.6}_{2.2}$	$80.8_{1.9}$	$83.1_{0.7}$	$\mathbf{81.0}_{1.7}$
RoBERTa$_{freeze}$	$73.9_{2.5}$	$76.6_{1.4}$	$78.5_{0.7}$	$76.4_{1.7}$
kNN	$54.5_{3.1}$	$64.2_{1.9}$	$66.6_{1.3}$	$64.7_{3.5}$
SetFit	$76.5_{1.6}$	$80.6_{0.5}$	$81.2_{0.3}$	$80.0_{1.4}$
LaGoNN	$\mathbf{78.6}_{2.2}$	$\mathbf{81.2}_{1.4}$	$81.6_{1.1}$	$80.8_{0.9}$
Probe	$52.3_{2.0}$	$64.1_{1.8}$	$67.2_{1.4}$	$63.1_{4.3}$
LaGoNN$_{cheap}$	$47.3_{3.4}$	$60.7_{1.5}$	$65.2_{1.4}$	$59.5_{5.2}$

Table 28: On the first step, LaGoNN, LaGoNN$_{lite}$, and LaGoNN$_{exp}$ start out the strongest. On the 5^{th} step, LaGoNN is the most performant method while on the 10^{th} step it is SetFit$_{lite}$. However, LaGoNN$_{lite}$ is the best overall method based on the average.

Method		**Amazon Counterfactual**		
Balanced	1^{st}	5^{th}	10^{th}	Average
RoBERTa$_{full}$	$73.6_{2.1}$	$78.6_{3.9}$	$82.4_{1.1}$	$78.9_{2.2}$
SetFit$_{exp}$	$73.8_{4.4}$	$69.8_{4.0}$	$64.1_{4.6}$	$69.6_{3.6}$
LaGoNN$_{exp}$	$\mathbf{76.0}_{3.0}$	$73.4_{2.6}$	$72.3_{2.9}$	$72.5_{3.4}$
SetFit$_{lite}$	$73.8_{4.4}$	$\mathbf{80.4}_{1.8}$	$82.4_{0.8}$	$78.3_{4.3}$
LaGoNN$_{lite}$	$\mathbf{76.0}_{3.0}$	$80.0_{1.3}$	$\mathbf{82.5}_{0.9}$	$79.2_{3.2}$
RoBERTa$_{freeze}$	$73.6_{2.1}$	$76.8_{1.6}$	$77.9_{1.0}$	$76.5_{1.3}$
kNN	$41.7_{3.4}$	$57.9_{3.3}$	$58.3_{3.3}$	$56.8_{5.1}$
SetFit	$73.8_{4.4}$	$79.2_{1.9}$	$80.1_{1.0}$	$78.6_{1.8}$
LaGoNN	$\mathbf{76.0}_{3.0}$	$80.1_{2.0}$	$81.4_{1.1}$	$\mathbf{79.8}_{1.4}$
Probe	$52.4_{3.4}$	$64.7_{2.5}$	$67.5_{0.4}$	$63.4_{4.4}$
LaGoNN$_{cheap}$	$48.1_{3.4}$	$62.0_{2.0}$	$65.3_{0.8}$	$60.5_{5.0}$

Table 29: On the first step, LaGoNN, LaGoNN$_{lite}$, and LaGoNN$_{exp}$ start out the strongest. On the 5^{th} step, SetFit$_{lite}$ pulls ahead slightly, yet on the 10^{th} step LaGoNN$_{lite}$ is the best performer. Overall, LaGoNN is the best method based on the average.

Method *Extreme*	**Toxic Conversations** 1^{st}	5^{th}	10^{th}	Average
RoBERTa$_{full}$	$7.9_{0.5}$	$21.2_{3.7}$	$33.8_{5.5}$	$21.9_{9.3}$
SetFit$_{exp}$	$8.8_{1.2}$	$18.1_{3.4}$	$24.7_{4.1}$	$17.6_{5.5}$
LaGoNN$_{exp}$	$8.9_{1.7}$	$17.4_{6.6}$	$26.4_{5.2}$	$17.9_{6.0}$
SetFit$_{lite}$	$8.8_{1.2}$	$15.9_{4.8}$	$18.0_{3.9}$	$14.9_{3.2}$
LaGoNN$_{lite}$	$8.9_{1.7}$	$16.1_{5.9}$	$19.8_{6.0}$	$15.5_{3.7}$
RoBERTa$_{freeze}$	$7.9_{0.5}$	$12.8_{2.4}$	$19.1_{3.2}$	$13.5_{3.5}$
kNN	$7.9_{0.0}$	$8.7_{0.4}$	$8.7_{0.2}$	$8.5_{0.3}$
SetFit	$8.8_{1.2}$	$13.1_{2.5}$	$16.3_{3.0}$	$13.0_{2.6}$
LaGoNN	$8.9_{1.7}$	$13.8_{3.9}$	$17.1_{4.8}$	$13.4_{2.6}$
Probe	$\mathbf{13.1}_{2.8}$	$\mathbf{24.6}_{2.6}$	$\mathbf{30.1}_{2.1}$	$\mathbf{23.9}_{5.6}$
LaGoNN$_{cheap}$	$11.3_{2.2}$	$21.7_{2.7}$	$27.4_{2.3}$	$21.3_{5.3}$

Table 30: Probe is most performant method on all steps and the overall strongest performer. We note, however, that LaGoNN-based methods tend to outperform their SetFit-based counterparts.

Method *Imbalanced*	**Toxic Conversations** 1^{st}	5^{th}	10^{th}	Average
RoBERTa$_{full}$	$\mathbf{24.1}_{5.6}$	$43.1_{3.4}$	$52.1_{2.5}$	$42.4_{8.2}$
SetFit$_{exp}$	$21.8_{6.6}$	$44.5_{4.1}$	$51.4_{1.9}$	$42.1_{9.3}$
LaGoNN$_{exp}$	$22.7_{9.8}$	$\mathbf{49.1}_{5.6}$	$\mathbf{53.4}_{2.3}$	$\mathbf{45.6}_{9.8}$
SetFit$_{lite}$	$21.8_{6.6}$	$41.4_{4.4}$	$44.8_{3.1}$	$39.0_{7.0}$
LaGoNN$_{lite}$	$22.7_{9.8}$	$47.0_{6.3}$	$50.2_{5.4}$	$43.7_{8.6}$
RoBERTa$_{freeze}$	$\mathbf{24.1}_{5.6}$	$31.2_{4.4}$	$34.0_{4.0}$	$30.5_{3.1}$
kNN	$11.5_{2.5}$	$14.7_{4.0}$	$15.3_{3.2}$	$14.6_{1.1}$
SetFit	$21.8_{6.6}$	$26.7_{5.3}$	$30.2_{4.0}$	$26.6_{2.7}$
LaGoNN	$22.7_{9.8}$	$27.6_{8.9}$	$30.3_{8.7}$	$27.4_{2.4}$
Probe	$23.3_{2.7}$	$33.0_{2.8}$	$37.1_{1.8}$	$32.5_{4.2}$
LaGoNN$_{cheap}$	$20.5_{3.2}$	$31.1_{3.2}$	$35.6_{1.8}$	$30.5_{4.6}$

Table 31: RoBERTa$_{full}$ and RoBERTa$_{freeze}$ start out as the strongest classifiers on the first step, but are overtaken on subsequent steps by LaGoNN$_{exp}$, which ends up as strongest method overall.

Method *Moderate*	**Toxic Conversations** 1^{st}	5^{th}	10^{th}	Average
RoBERTa$_{full}$	$34.2_{3.4}$	$45.5_{1.9}$	$52.4_{3.3}$	$45.7_{5.6}$
SetFit$_{exp}$	$33.6_{2.9}$	$47.2_{2.2}$	$46.6_{3.3}$	$44.3_{4.3}$
LaGoNN$_{exp}$	$\mathbf{36.6}_{4.2}$	$48.2_{2.7}$	$49.9_{3.7}$	$48.0_{4.4}$
SetFit$_{lite}$	$33.6_{2.9}$	$52.6_{2.0}$	$55.1_{1.6}$	$48.8_{7.3}$
LaGoNN$_{lite}$	$\mathbf{36.6}_{4.2}$	$\mathbf{56.1}_{1.5}$	$\mathbf{57.7}_{1.4}$	$\mathbf{52.3}_{6.8}$
RoBERTa$_{freeze}$	$34.2_{3.4}$	$38.4_{2.1}$	$39.5_{1.8}$	$38.0_{1.5}$
kNN	$19.4_{1.9}$	$21.5_{3.4}$	$22.4_{2.9}$	$21.6_{0.8}$
SetFit	$33.6_{2.9}$	$39.2_{2.9}$	$41.6_{2.7}$	$38.6_{2.4}$
LaGoNN	$\mathbf{36.6}_{4.2}$	$42.7_{3.7}$	$45.0_{3.5}$	$42.0_{2.5}$
Probe	$29.0_{2.7}$	$36.1_{1.2}$	$39.1_{1.5}$	$35.5_{3.3}$
LaGoNN$_{cheap}$	$26.1_{2.7}$	$34.3_{1.3}$	$37.5_{1.8}$	$33.6_{3.6}$

Table 32: On the first step, LaGoNN, LaGoNN$_{lite}$, and LaGoNN$_{exp}$ start out the strongest, but it is LaGoNN$_{lite}$ that remains performant for all other steps. LaGoNN$_{lite}$ is also the strongest overall method based on the average.

Method *Balanced*	**Toxic Conversations** 1^{st}	5^{th}	10^{th}	Average
RoBERTa$_{full}$	$32.3_{1.1}$	$42.7_{1.8}$	$54.1_{3.4}$	$43.8_{6.3}$
SetFit$_{exp}$	$35.7_{3.4}$	$32.6_{6.2}$	$37.4_{2.7}$	$36.5_{1.9}$
LaGoNN$_{exp}$	$\mathbf{40.4}_{4.4}$	$40.2_{6.6}$	$39.8_{7.5}$	$40.0_{1.2}$
SetFit$_{lite}$	$35.7_{3.4}$	$52.7_{2.5}$	$53.9_{2.2}$	$46.8_{7.8}$
LaGoNN$_{lite}$	$\mathbf{40.4}_{4.4}$	$\mathbf{52.9}_{2.6}$	$\mathbf{54.0}_{2.3}$	$\mathbf{48.3}_{6.4}$
RoBERTa$_{freeze}$	$32.3_{1.1}$	$39.2_{1.5}$	$41.0_{0.6}$	$38.5_{2.4}$
kNN	$17.4_{0.8}$	$23.7_{2.6}$	$24.3_{2.7}$	$23.1_{2.0}$
SetFit	$35.7_{3.4}$	$44.5_{2.9}$	$46.1_{2.8}$	$43.6_{2.9}$
LaGoNN	$\mathbf{40.4}_{4.4}$	$46.6_{2.7}$	$48.1_{2.2}$	$46.1_{2.2}$
Probe	$29.5_{2.4}$	$35.9_{0.9}$	$40.2_{0.9}$	$36.1_{3.5}$
LaGoNN$_{cheap}$	$26.8_{2.7}$	$34.5_{1.3}$	$38.5_{0.8}$	$34.4_{3.7}$

Table 33: On the first step, LaGoNN, LaGoNN$_{lite}$, and LaGoNN$_{exp}$ start out the strongest, but it is LaGoNN$_{lite}$ that remains performant for all other steps. LaGoNN$_{lite}$ is also the strongest overall method based on the average.

Method *Extreme*	**Hate Speech Offensive** 1^{st}	5^{th}	10^{th}	Average
RoBERTa$_{full}$	$30.2_{1.4}$	$43.5_{2.5}$	$\mathbf{51.2}_{2.2}$	$\mathbf{44.3}_{7.4}$
SetFit$_{exp}$	$30.3_{0.8}$	$\mathbf{44.0}_{1.3}$	$51.1_{2.0}$	$43.8_{6.5}$
LaGoNN$_{exp}$	$30.3_{0.7}$	$40.7_{2.9}$	$49.1_{4.4}$	$42.2_{6.2}$
SetFit$_{lite}$	$30.3_{0.8}$	$43.4_{2.5}$	$45.5_{3.4}$	$41.6_{4.6}$
LaGoNN$_{lite}$	$30.3_{0.7}$	$40.9_{3.4}$	$41.5_{4.8}$	$39.1_{3.6}$
RoBERTa$_{freeze}$	$30.2_{1.4}$	$33.5_{3.1}$	$34.4_{3.4}$	$33.1_{1.4}$
kNN	$\mathbf{31.5}_{1.2}$	$35.9_{2.7}$	$37.4_{2.0}$	$35.8_{1.7}$
SetFit	$30.3_{0.8}$	$38.4_{2.5}$	$41.1_{1.5}$	$37.8_{3.3}$
LaGoNN	$30.3_{0.7}$	$35.7_{2.6}$	$39.1_{2.4}$	$35.6_{2.7}$
Probe	$29.0_{0.2}$	$34.7_{1.5}$	$40.1_{2.1}$	$35.1_{3.8}$
LaGoNN$_{cheap}$	$29.0_{0.1}$	$36.9_{1.8}$	$40.5_{2.1}$	$36.2_{3.7}$

Table 34: kNN is the strongest method at first, but is overtaken by SetFit$_{exp}$ on the 5^{th} step, which is then overtaken by RoBERTa$_{full}$ on the 10^{th} step. RoBERTa$_{full}$ is overall most performant system based on the average.

Method *Imbalanced*	**Hate Speech Offensive** 1^{st}	5^{th}	10^{th}	Average
RoBERTa$_{full}$	$50.6_{3.0}$	$65.2_{3.9}$	$\mathbf{70.3}_{1.2}$	$64.2_{5.3}$
SetFit$_{exp}$	$54.4_{4.3}$	$66.3_{1.8}$	$68.9_{2.0}$	$64.3_{4.5}$
LaGoNN$_{exp}$	$\mathbf{57.0}_{5.2}$	$\mathbf{67.0}_{4.4}$	$69.8_{2.1}$	$\mathbf{64.9}_{4.6}$
SetFit$_{lite}$	$54.4_{4.3}$	$65.5_{3.0}$	$65.9_{3.5}$	$63.5_{3.9}$
LaGoNN$_{lite}$	$\mathbf{57.0}_{5.2}$	$66.6_{2.6}$	$66.6_{1.9}$	$64.3_{4.1}$
RoBERTa$_{freeze}$	$50.6_{3.0}$	$54.1_{1.6}$	$55.3_{2.3}$	$54.1_{1.3}$
kNN	$55.6_{4.8}$	$57.3_{2.3}$	$58.8_{3.6}$	$57.4_{1.1}$
SetFit	$54.4_{4.3}$	$57.0_{3.9}$	$58.2_{3.8}$	$57.2_{1.1}$
LaGoNN	$\mathbf{57.0}_{5.2}$	$58.2_{4.1}$	$58.3_{3.4}$	$58.3_{0.6}$
Probe	$46.5_{2.2}$	$57.8_{1.7}$	$60.3_{1.2}$	$56.5_{4.5}$
LaGoNN$_{cheap}$	$47.1_{1.3}$	$56.5_{2.2}$	$59.5_{2.5}$	$55.6_{3.8}$

Table 35: On the first step, LaGoNN, LaGoNN$_{lite}$, and LaGoNN$_{exp}$ start out the strongest, and LaGoNN$_{exp}$ continues to be performant, but is overtaken on the 10^{th} step by RoBERTa$_{full}$. LaGoNN$_{exp}$ is the strongest overall method based on the average.

Method	Hate Speech Offensive			
Moderate	1^{st}	5^{th}	10^{th}	Average
RoBERTa$_{full}$	$61.9_{3.4}$	$70.8_{1.0}$	$\mathbf{72.5}_{1.4}$	$69.9_{3.2}$
SetFit$_{exp}$	$\mathbf{64.3}_{4.2}$	$70.6_{2.4}$	$72.4_{0.5}$	$69.8_{2.8}$
LaGoNN$_{exp}$	$63.8_{4.9}$	$\mathbf{71.0}_{2.1}$	$72.3_{1.0}$	$\mathbf{70.0}_{3.0}$
SetFit$_{lite}$	$64.3_{4.2}$	$70.3_{2.2}$	$71.2_{2.1}$	$69.3_{2.3}$
LaGoNN$_{lite}$	$63.8_{4.9}$	$70.7_{1.4}$	$71.4_{1.0}$	$69.4_{2.5}$
RoBERTa$_{freeze}$	$61.9_{3.4}$	$63.2_{4.1}$	$64.1_{4.5}$	$63.2_{0.6}$
kNN	$\mathbf{64.3}_{4.0}$	$63.3_{2.9}$	$63.9_{2.5}$	$63.7_{0.4}$
SetFit	$\mathbf{64.3}_{4.2}$	$67.3_{3.2}$	$67.6_{2.3}$	$66.9_{1.1}$
LaGoNN	$63.8_{4.9}$	$65.0_{5.3}$	$66.7_{5.9}$	$65.3_{0.9}$
Probe	$55.6_{1.7}$	$63.8_{0.8}$	$66.1_{0.3}$	$63.2_{3.0}$
LaGoNN$_{cheap}$	$56.0_{3.6}$	$62.2_{1.4}$	$66.0_{0.9}$	$62.3_{2.9}$

Table 36: Similar to the imbalanced setting, on the first step, LaGoNN, LaGoNN$_{lite}$, and LaGoNN$_{exp}$ start out the strongest, and LaGoNN$_{exp}$ continues to be performant, but is overtaken on the 10^{th} step by RoBERTa$_{full}$. LaGoNN$_{exp}$ is the strongest overall method based on the average.

Method	Hate Speech Offensive			
Balanced	1^{st}	5^{th}	10^{th}	Average
RoBERTa$_{full}$	$59.7_{3.5}$	$66.9_{1.2}$	$\mathbf{69.2}_{1.8}$	$\mathbf{66.4}_{2.7}$
SetFit$_{exp}$	$60.7_{1.3}$	$66.3_{1.6}$	$67.5_{0.9}$	$65.9_{2.2}$
LaGoNN$_{exp}$	$\mathbf{61.5}_{1.7}$	$66.4_{1.4}$	$67.7_{0.9}$	$66.1_{1.8}$
SetFit$_{lite}$	$60.7_{1.3}$	$66.3_{2.0}$	$66.5_{0.9}$	$65.1_{1.7}$
LaGoNN$_{lite}$	$\mathbf{61.5}_{1.7}$	$\mathbf{67.1}_{1.1}$	$67.3_{0.8}$	$66.0_{1.7}$
RoBERTa$_{freeze}$	$59.7_{3.5}$	$60.4_{2.7}$	$63.1_{2.3}$	$61.0_{1.3}$
kNN	$60.7_{1.3}$	$59.6_{2.8}$	$59.5_{2.5}$	$59.5_{0.5}$
SetFit	$60.7_{1.3}$	$62.5_{0.7}$	$63.4_{1.0}$	$62.3_{1.0}$
LaGoNN	$\mathbf{61.5}_{1.7}$	$62.8_{1.5}$	$64.2_{1.0}$	$63.0_{0.9}$
Probe	$54.9_{1.4}$	$58.5_{0.9}$	$60.9_{0.4}$	$58.7_{1.7}$
LaGoNN$_{cheap}$	$54.2_{2.3}$	$58.6_{0.6}$	$60.6_{0.5}$	$58.5_{1.8}$

Table 37: Similar to the moderate setting, on the first step, LaGoNN, LaGoNN$_{lite}$, and LaGoNN$_{exp}$ start out the strongest, but RoBERTa$_{full}$ overtakes LaGoNN$_{lite}$ by the 10^{th} step. RoBERTa$_{full}$ slightly outperforms LaGoNN$_{lite}$ and LaGoNN$_{exp}$ as the overall strongest method based on the average.

Method		**Liar**		
Extreme	1^{st}	5^{th}	10^{th}	Average
RoBERTa$_{full}$	$\mathbf{32.0}_{2.7}$	$\mathbf{34.7}_{2.9}$	$35.1_{4.3}$	$33.7_{1.0}$
SetFit$_{exp}$	$31.2_{3.8}$	$30.4_{3.1}$	$31.8_{2.9}$	$31.5_{0.7}$
LaGoNN$_{exp}$	$30.6_{4.7}$	$30.3_{2.0}$	$31.3_{2.0}$	$31.1_{0.6}$
SetFit$_{lite}$	$31.2_{3.8}$	$32.7_{3.8}$	$33.5_{4.2}$	$32.7_{0.8}$
LaGoNN$_{lite}$	$30.6_{4.7}$	$31.8_{3.9}$	$32.4_{2.7}$	$31.6_{0.6}$
RoBERTa$_{freeze}$	$\mathbf{32.0}_{2.7}$	$32.8_{4.5}$	$34.2_{5.0}$	$33.2_{0.7}$
kNN	$27.0_{0.5}$	$27.3_{0.8}$	$27.9_{0.8}$	$27.4_{0.3}$
SetFit	$31.2_{3.8}$	$33.7_{5.1}$	$\mathbf{35.7}_{5.1}$	$\mathbf{34.3}_{1.6}$
LaGoNN	$30.6_{4.7}$	$32.0_{4.6}$	$33.7_{5.4}$	$32.6_{0.9}$
Probe	$30.7_{2.0}$	$30.6_{3.9}$	$31.7_{2.9}$	$31.1_{0.4}$
LaGoNN$_{cheap}$	$30.7_{2.0}$	$30.5_{3.8}$	$31.4_{2.6}$	$31.0_{0.4}$

Table 38: RoBERTa$_{freeze}$ and RoBERTa$_{full}$ start out performant and RoBERTa$_{full}$ continues to be until the 10^{th} step where it is overtaken by SetFit, which ends up being the strongest overall method.

Method		**Liar**		
Imbalanced	1^{st}	5^{th}	10^{th}	Average
RoBERTa$_{full}$	$31.4_{3.2}$	$35.8_{2.6}$	$\mathbf{40.0}_{4.3}$	$36.2_{2.4}$
SetFit$_{exp}$	$\mathbf{32.3}_{4.5}$	$35.9_{3.1}$	$36.4_{2.2}$	$35.2_{1.1}$
LaGoNN$_{exp}$	$\mathbf{32.3}_{4.6}$	$35.7_{3.4}$	$36.5_{2.3}$	$35.7_{1.4}$
SetFit$_{lite}$	$\mathbf{32.3}_{4.5}$	$35.6_{2.7}$	$37.4_{2.6}$	$35.8_{1.6}$
LaGoNN$_{lite}$	$\mathbf{32.3}_{4.6}$	$35.2_{2.4}$	$36.6_{2.7}$	$35.5_{1.3}$
RoBERTa$_{freeze}$	$31.4_{3.2}$	$34.1_{2.6}$	$35.6_{3.2}$	$34.0_{1.4}$
kNN	$27.0_{0.2}$	$28.5_{1.0}$	$29.0_{1.0}$	$28.7_{0.7}$
SetFit	$\mathbf{32.3}_{4.5}$	$\mathbf{36.5}_{3.1}$	$38.5_{3.4}$	$\mathbf{36.3}_{2.0}$
LaGoNN	$\mathbf{32.3}_{4.6}$	$34.9_{2.2}$	$36.9_{2.5}$	$35.3_{1.4}$
Probe	$30.7_{3.0}$	$32.8_{1.8}$	$35.0_{1.6}$	$33.5_{1.5}$
LaGoNN$_{cheap}$	$30.4_{3.0}$	$32.9_{1.8}$	$35.4_{1.7}$	$33.5_{1.7}$

Table 39: LaGoNN, LaGoNN$_{lite}$, LaGoNN$_{exp}$, SetFit, SetFit$_{lite}$, and SetFit$_{exp}$ start out as the most performant, but SetFit is the strongest on the 5^{th} step and RoBERTa$_{full}$ on the 10^{th}. Overall, SetFit is strongest method based on the average over all steps.

Method		**Liar**		
Moderate	1^{st}	5^{th}	10^{th}	Average
RoBERTa$_{full}$	$33.9_{3.1}$	$38.4_{2.7}$	$\mathbf{43.9}_{2.2}$	$\mathbf{39.5}_{3.0}$
SetFit$_{exp}$	$33.0_{2.6}$	$37.2_{1.8}$	$38.7_{1.5}$	$37.4_{1.6}$
LaGoNN$_{exp}$	$\mathbf{34.1}_{3.4}$	$\mathbf{38.7}_{2.3}$	$39.0_{1.8}$	$37.8_{1.5}$
SetFit$_{lite}$	$33.0_{2.6}$	$38.5_{1.3}$	$40.4_{2.0}$	$38.2_{2.1}$
LaGoNN$_{lite}$	$\mathbf{34.1}_{3.4}$	$38.4_{2.0}$	$39.6_{1.5}$	$37.9_{1.6}$
RoBERTa$_{freeze}$	$33.9_{3.1}$	$35.3_{2.6}$	$36.8_{2.2}$	$35.4_{1.0}$
kNN	$29.2_{0.8}$	$29.7_{1.5}$	$30.0_{0.6}$	$29.8_{0.3}$
SetFit	$33.0_{2.6}$	$37.2_{3.9}$	$39.4_{3.5}$	$37.0_{1.8}$
LaGoNN	$\mathbf{34.1}_{3.4}$	$37.0_{3.1}$	$38.6_{3.0}$	$36.8_{1.3}$
Probe	$31.6_{1.1}$	$34.7_{2.5}$	$37.0_{2.5}$	$34.9_{1.7}$
LaGoNN$_{cheap}$	$31.4_{0.9}$	$35.3_{2.3}$	$37.6_{2.0}$	$35.3_{1.9}$

Table 40: LaGoNN, LaGoNN$_{lite}$, and LaGoNN$_{exp}$ are the most performant classifiers on the first step, while LaGoNN$_{exp}$ remains strong until the 10^{th} step where it is overtaken by RoBERTa$_{full}$. RoBERTa$_{full}$ is the overally strongest method if we aggregate over all steps.

Method		**Liar**		
Balanced	1^{st}	5^{th}	10^{th}	Average
RoBERTa$_{full}$	$33.8_{2.1}$	$\mathbf{39.4}_{2.4}$	$\mathbf{43.5}_{1.7}$	$\mathbf{40.2}_{3.2}$
SetFit$_{exp}$	$\mathbf{34.4}_{2.3}$	$36.7_{1.7}$	$37.0_{1.3}$	$36.5_{1.1}$
LaGoNN$_{exp}$	$33.8_{1.8}$	$34.2_{2.7}$	$37.2_{1.9}$	$36.2_{1.4}$
SetFit$_{lite}$	$\mathbf{34.4}_{2.3}$	$38.7_{2.3}$	$40.3_{2.8}$	$38.0_{2.1}$
LaGoNN$_{lite}$	$33.8_{1.8}$	$37.6_{2.0}$	$39.4_{2.8}$	$37.2_{1.9}$
RoBERTa$_{freeze}$	$33.8_{2.1}$	$36.6_{1.6}$	$38.6_{1.5}$	$36.7_{1.5}$
kNN	$30.1_{0.4}$	$31.3_{2.1}$	$30.6_{1.1}$	$30.9_{0.4}$
SetFit	$\mathbf{34.4}_{2.3}$	$38.3_{2.5}$	$40.0_{2.0}$	$37.9_{1.6}$
LaGoNN	$33.8_{1.8}$	$38.3_{1.3}$	$40.6_{0.6}$	$38.1_{2.0}$
Probe	$32.1_{1.9}$	$35.2_{1.4}$	$37.2_{2.5}$	$35.2_{1.7}$
LaGoNN$_{cheap}$	$31.9_{1.9}$	$36.0_{1.0}$	$37.5_{2.5}$	$35.7_{1.8}$

Table 41: SetFit, SetFit$_{lite}$, and SetFit$_{exp}$ start out the strongest on the first step, but are overtaken by RoBERTa$_{full}$ on the 5^{th} which remains the most performant on the 10^{th} step and if we consider the average over all steps.

A.4 Ablations

In this Appendix section, we perform ablation studies with LAGONN to support our findings in the main text.

A.4.1 Ablation: LAGONN configurations

Here, we provide an in-depth comparison between all LAGONN configurations, LABEL, DIST, LABDIST, TEXT, and ALL (see Table 1) for all datasets, balances, and levels of expense. The evaluation metric is the mean average precision ($\times 100$) over five seeds in all cases except for Hate Speech Offensive where the metric is the macro-F1.

Below, Figures 5 through 9 are the results for the LAGONN$_{cheap}$ training strategy, Figures 10 through 14 are the results for LAGONN, Figures 15 through 19 are the results for LAGONN$_{lite}$, and Figures 20 through 24 are the results for LAGONN$_{exp}$. We place the figures on a new page for ease of viewing.

In the case of LAGONN$_{cheap}$, if we do not fine-tune the embedding model we see little variation in the standard deviation bands, with the exception of the LIAR dataset, which seems to be a very difficult dataset. When we do fine-tune, we see a great deal of variation, especially in cases of label imbalance, which is expected as the representations are altered more. The performance of TEXT and ALL is very unstable, often being the worst performers, while sometimes being the best. Interestingly, we note that DIST, LABEL, and LABDIST often show very similar performance. In our opinion. LABDIST seems to be the most consistent and stable performer, especially in cases when the embedding model is fine-tuned, LAGONN, LAGONN$_{lite}$, and LAGONN$_{exp}$.

Overall, we believe that LABDIST is the most performant/stable configuration of LAGONN, and it is about this version that we present results in the main text. We note that we could have presented the best performer for each evaluation scenario, however, this is not in the spirit of our work as it adds yet another hyperparameter to configure, standing in the way of practical usage and convoluting our analysis. However, in our codebase, we hope that we have made it easy for one to change these configurations for their own usage, be it scientific or otherwise.

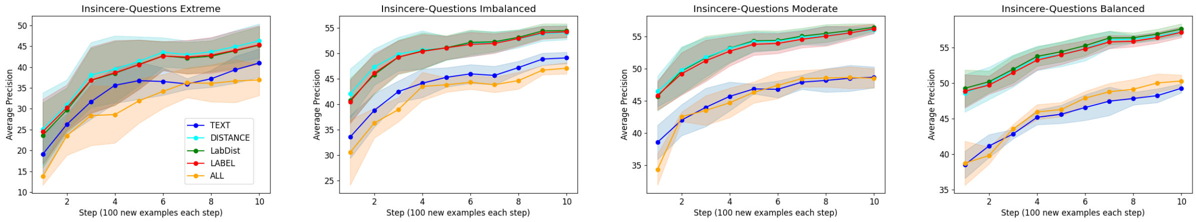

Figure 5: LAGONN$_{cheap}$ performance for all configurations and balance regimes on the Insincere Questions dataset. The relevant balance is in the title of each panel.

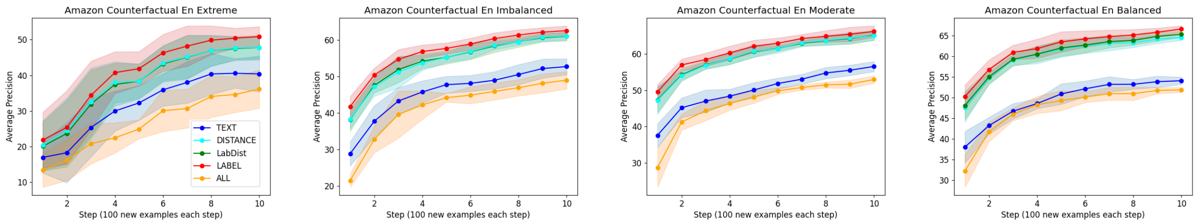

Figure 6: LAGONN$_{cheap}$ performance for all configurations and balance regimes on the Amazon Counterfactual dataset. The relevant balance is in the title of each panel.

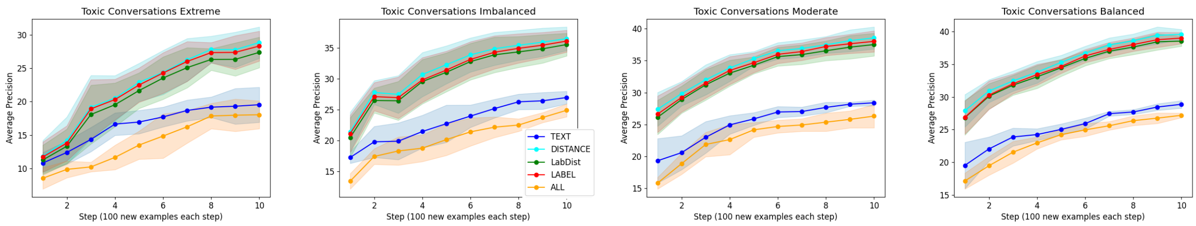

Figure 7: LAGONN$_{cheap}$ performance for all configurations and balance regimes on the Toxic Conversations dataset. The relevant balance is in the title of each panel.

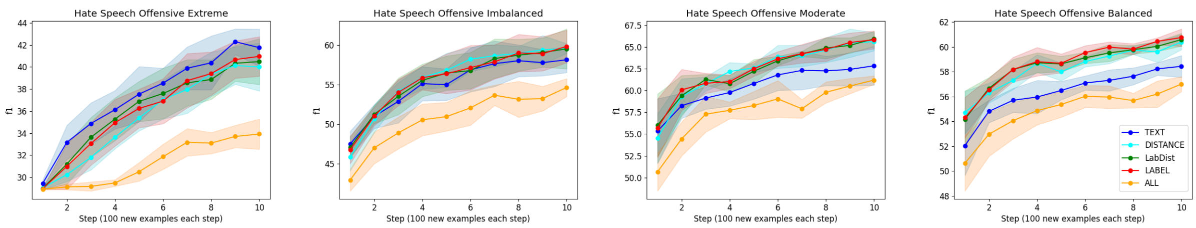

Figure 8: LAGONN$_{cheap}$ performance for all configurations and balance regimes on the Hate Speech Offensive dataset. The relevant balance is in the title of each panel.

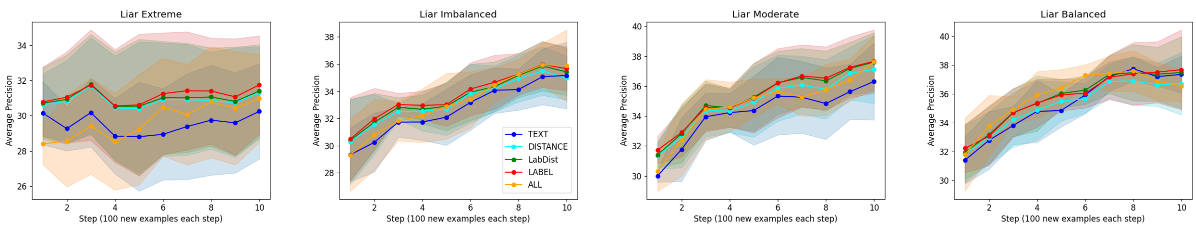

Figure 9: LAGONN$_{cheap}$ performance for all configurations and balance regimes on the Liar dataset. The relevant balance is in the title of each panel.

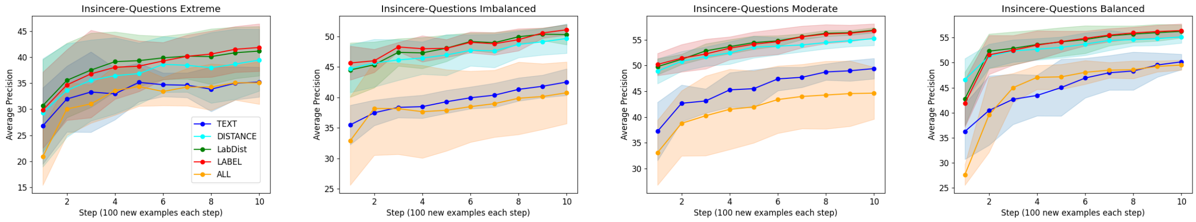

Figure 10: LAGONN performance for all configurations and balance regimes on the Insincere Questions dataset. The relevant balance is in the title of each panel.

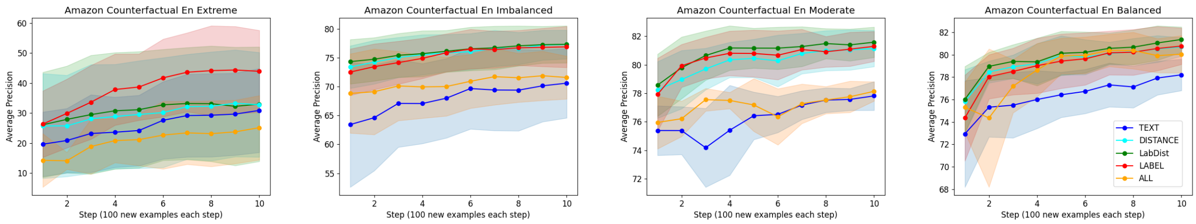

Figure 11: LAGONN performance for all configurations and balance regimes on the Amazon Counterfactual dataset. The relevant balance is in the title of each panel.

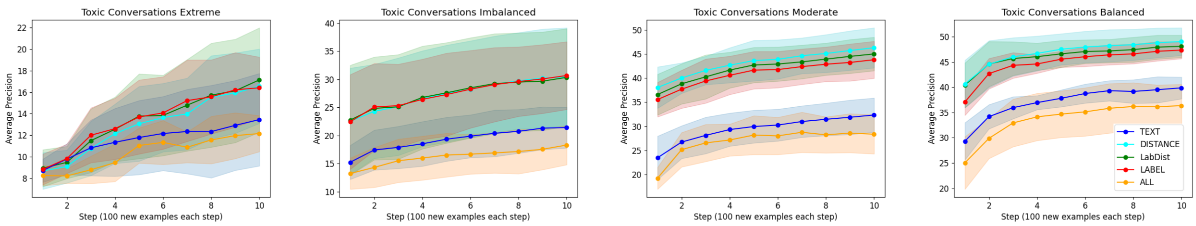

Figure 12: LAGONN performance for all configurations and balance regimes on the Toxic Conversations dataset. The relevant balance is in the title of each panel.

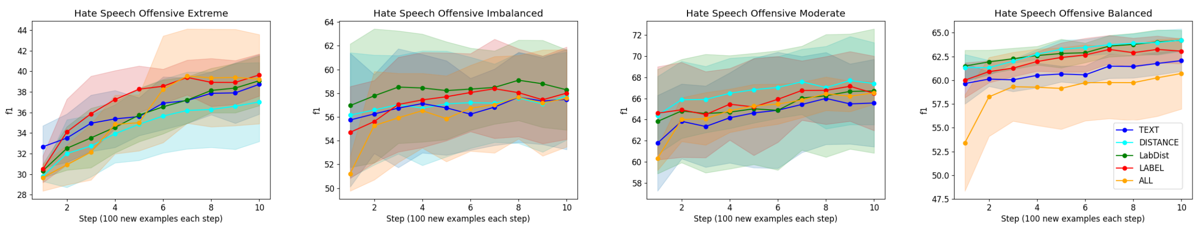

Figure 13: LAGONN performance for all configurations and balance regimes on the Hate Speech Offensive dataset. The relevant balance is in the title of each panel.

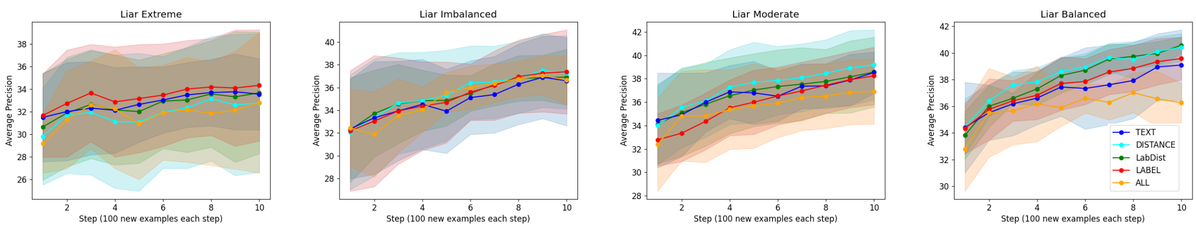

Figure 14: LAGONN performance for all configurations and balance regimes on the Liar dataset. The relevant balance is in the title of each panel.

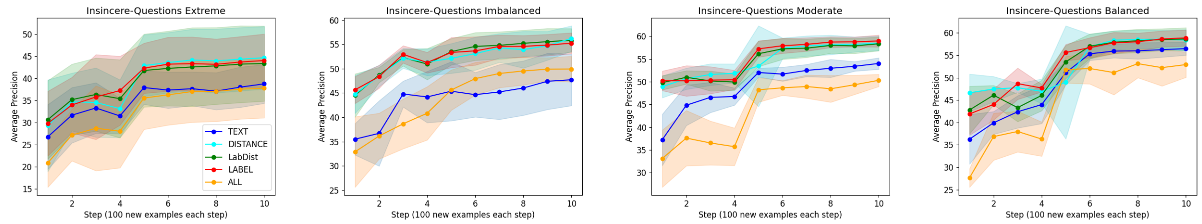

Figure 15: LAGONN$_{lite}$ performance for all configurations and balance regimes on the Insincere Questions dataset. The relevant balance is in the title of each panel.

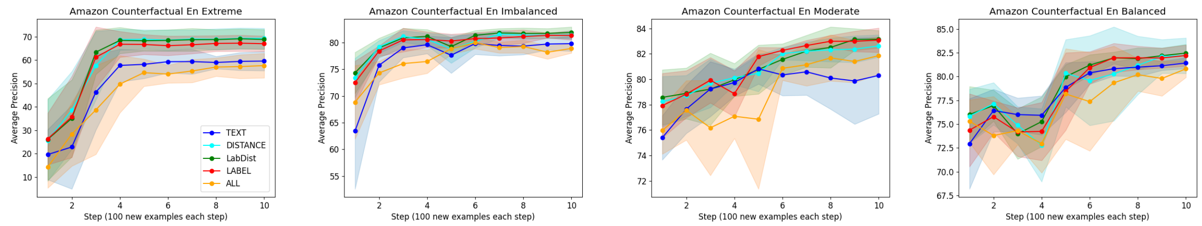

Figure 16: LAGONN$_{lite}$ performance for all configurations and balance regimes on the Amazon Counterfactual dataset. The relevant balance is in the title of each panel.

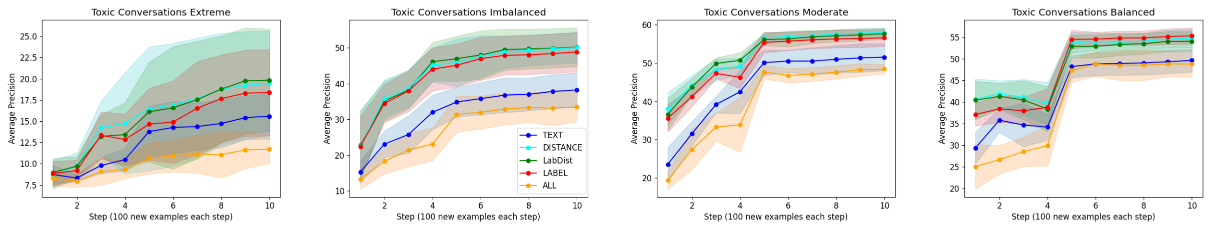

Figure 17: LAGONN$_{lite}$ performance for all configurations and balance regimes on the Toxic Conversations dataset. The relevant balance is in the title of each panel.

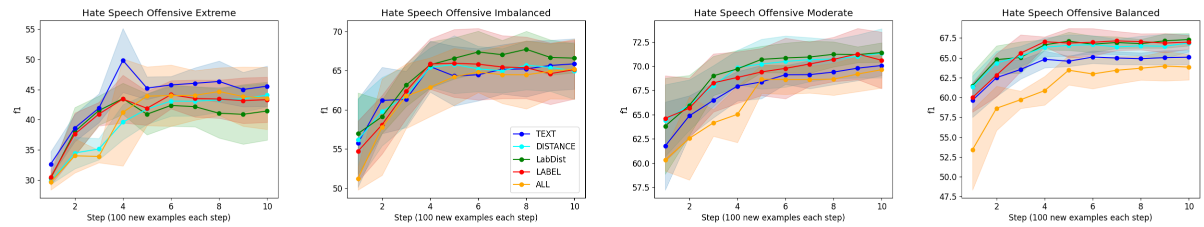

Figure 18: LAGONN$_{lite}$ performance for all configurations and balance regimes on the Hate Speech Offensive dataset. The relevant balance is in the title of each panel.

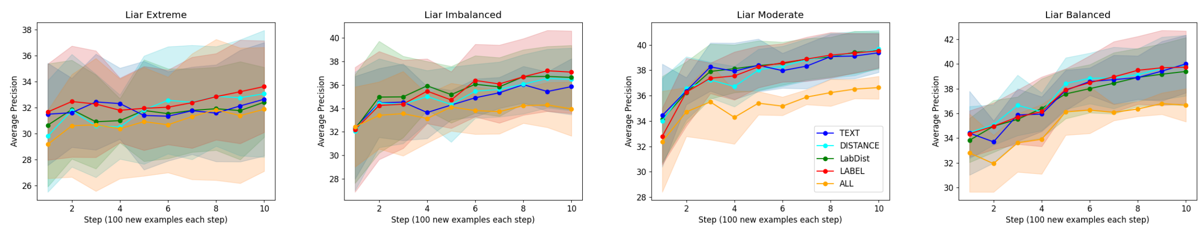

Figure 19: LAGONN$_{lite}$ performance for all configurations and balance regimes on the Liar dataset. The relevant balance is in the title of each panel.

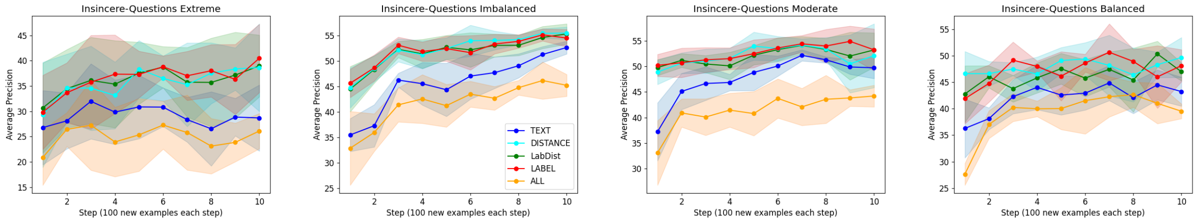

Figure 20: LAGONN$_{exp}$ performance for all configurations and balance regimes on the Insincere Questions dataset. The relevant balance is in the title of each panel.

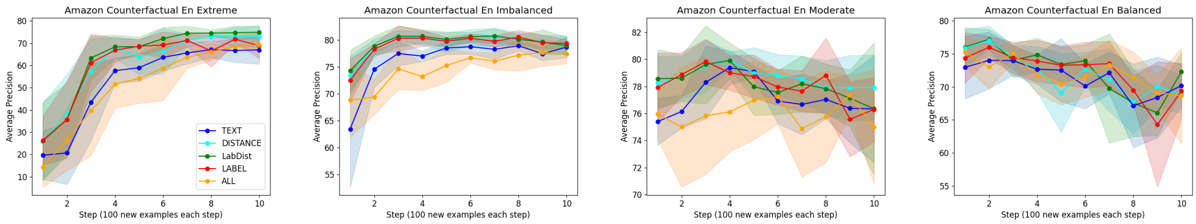

Figure 21: LAGONN$_{exp}$ performance for all configurations and balance regimes on the Amazon Counterfactual dataset. The relevant balance is in the title of each panel.

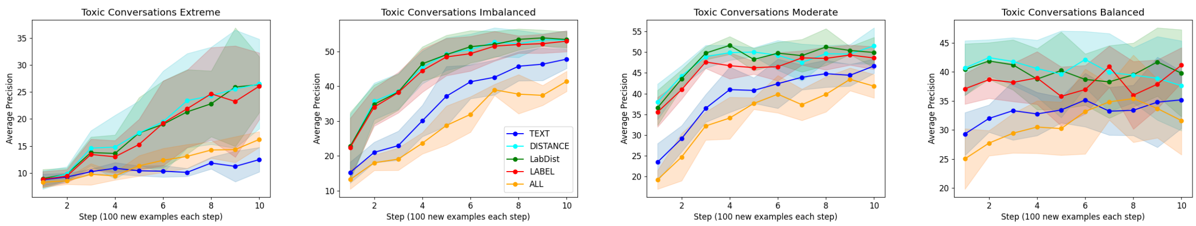

Figure 22: LAGONN$_{exp}$ performance for all configurations and balance regimes on the Toxic Conversations dataset. The relevant balance is in the title of each panel.

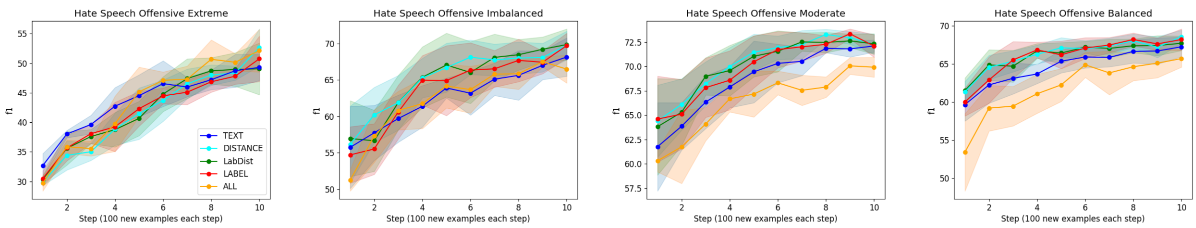

Figure 23: LAGONN$_{exp}$ performance for all configurations and balance regimes on the Hate Speech Offensive dataset. The relevant balance is in the title of each panel.

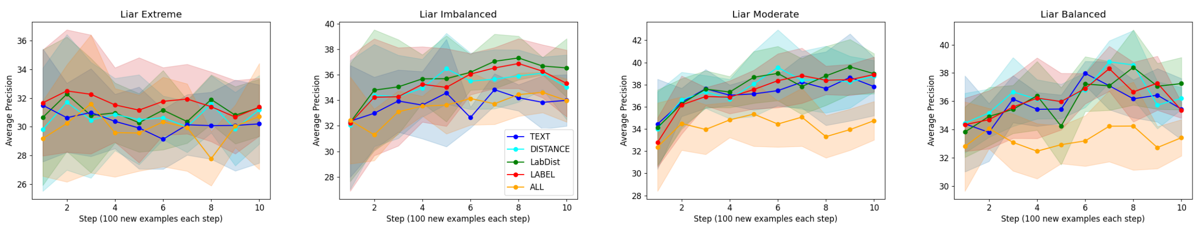

Figure 24: LAGONN$_{exp}$ performance for all configurations and balance regimes on the Liar dataset. The relevant balance is in the title of each panel.

A.4.2 Ablation: the effect of encoding distance

Here, at the suggestion of an anonymous reviewer, we present ablation results and analysis of how encoding distance affects LAGONN, because PLMs often struggle to understand numbers. Note that during our development stage, we ensured that our tokenizer was capable of encoding floats with trailing digits. To examine the effect of trailing digits on LAGONN, we consider the DIST configuration (see Table 1), where we append only the Euclidean distance to the input text. In this ablation, however, we round to different levels of precision. For example, if the distance were a float of 0.123456789, we round it to the nearest whole number, 0.0, single digit float, 0.1, three digit float, 0.123, six digit float, 0.123457, and finally keep it unrounded, that is, the original DIST configuration, 0.123456789. The below results are only for the LAGONN$_{lite}$ training strategy. We chose LAGONN$_{lite}$ for this ablation because it provides insight into both how distance affects full-model fine-tuning and only refitting the classification head. The results can be seen below in Figures 25 through 29. We place the figures on a new page for ease of viewing.

Interestingly, we tend to observe very similar performance curves for all rounding precisions. The exceptions to this would perhaps be Amazon Counterfactual and Hate Speech Offensive in the balanced regime where DIST and rounding to the third trailing digit respectively exhibit large instability.

Although not always the case, it appears that providing the model with the distance rounded to the nearest whole number tends to result in the strongest and stablest performer, however, we emphasize that in general there does not seem to a dramatic difference between the rounding precisions we considered. Longer digits slightly worsen model performance and the model might learn the most from simpler or abbreviated representations of distance. This finding motivated us to consider the ablation in Appendix A.4.3.

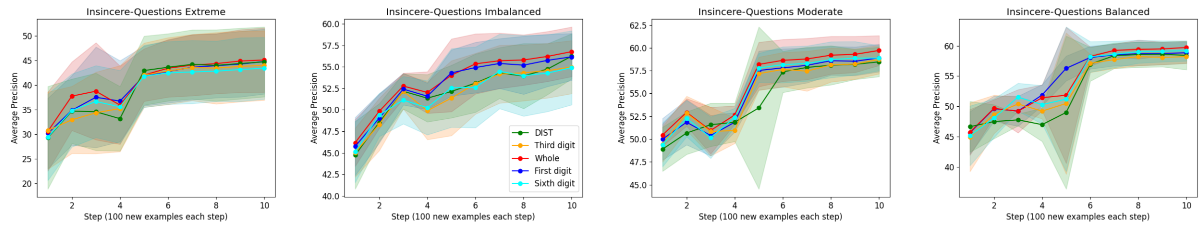

Figure 25: LAGONN$_{lite}$ performance when considering different rounding precisions for the Euclidean distance before appending it to a modified instance. We consider all balance regimes on the Insincere Questions dataset and the relevant balance is in the title of each panel.

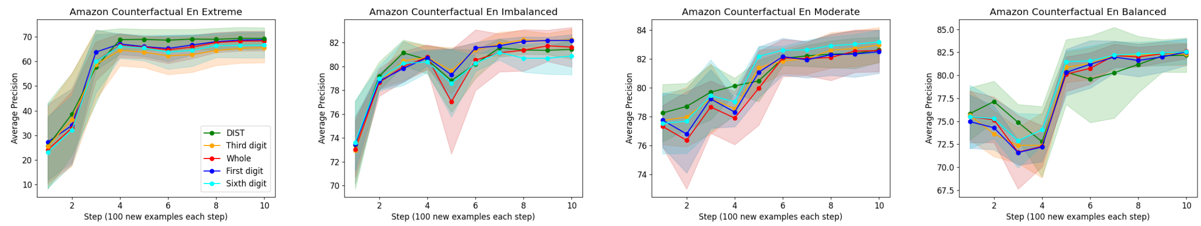

Figure 26: LAGONN$_{lite}$ performance when considering different rounding precisions for the Euclidean distance before appending it to a modified instance. We consider all balance regimes on the Amazon Counterfactual dataset and the relevant balance is in the title of each panel.

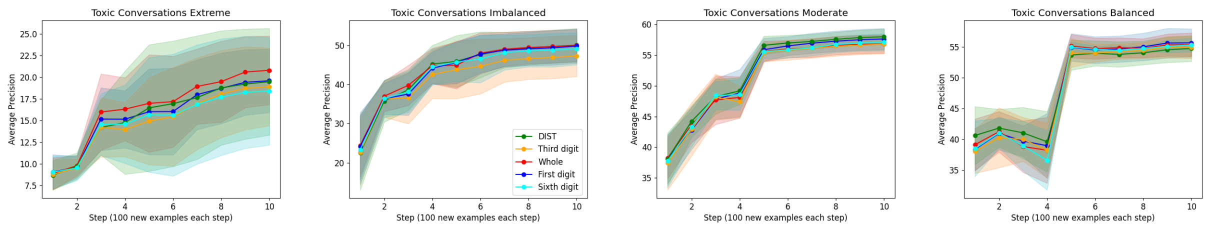

Figure 27: LAGONN$_{lite}$ performance when considering different rounding precisions for the Euclidean distance before appending it to a modified instance. We consider all balance regimes on the Toxic Conversations dataset and the relevant balance is in the title of each panel.

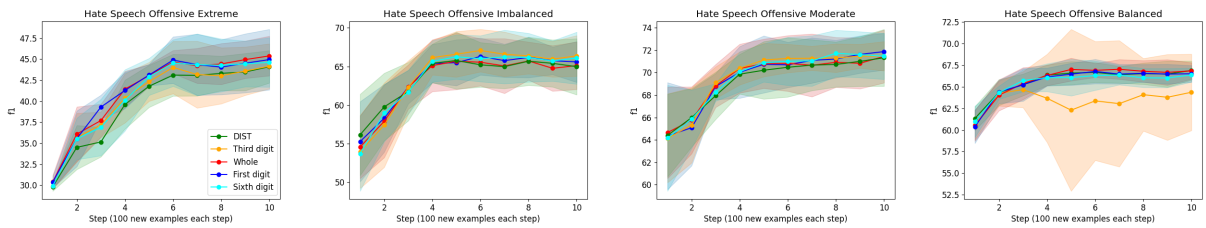

Figure 28: LAGONN$_{lite}$ performance when considering different rounding precisions for the Euclidean distance before appending it to a modified instance. We consider all balance regimes on the Hate Speech Offensive dataset and the relevant balance is in the title of each panel.

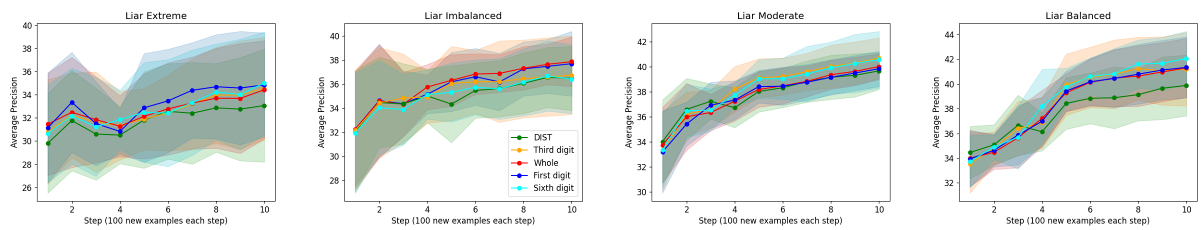

Figure 29: LAGONN$_{lite}$ performance when considering different rounding precisions for the Euclidean distance before appending it to a modified instance. We consider all balance regimes on the Liar dataset and the relevant balance is in the title of each panel.

A.4.3 Ablation: support for LABDIST

The results from the ablation in Appendix A.4.2 suggest that rounding the distance to the nearest whole number results in a stronger classifier than appending the unrounded distance. Thus far, we have asserted that LABDIST, where we append both the gold label of the NN and unrounded distance is the most performant version of LAGONN (see Table 1). To demonstrate that this is reasonable, in this ablation study, we compare the original LABDIST configuration against three models, namely the LABEL configuration, distance rounded to near whole number (Whole), and finally a new configuration similar to LABDIST, but where we append the gold label and distance rounded to a whole number, which we refer to as LABROUND. As in Appendix A.4.2, in this ablation we consider only the LAGONN$_{lite}$ fine-tuning strategy. We chose for this ablation because it provides insight into both how the different configurations affect full-model fine-tuning and only re-fitting the classification head. The results can be seen below in Figures 30 through 34. We place the figures on a new page for ease of viewing.

In general, we note very similar performance curves for these four models. In the case of Insincere Questions, appending the distance after rounding it to the nearest whole number (Whole, the red curve), is a strong model, except in the balanced regime where we note large instability. The results for Amazon Counterfactual tell a different story, where rounding the Euclidean distance to the nearest whole number causes large instability and even degrades performance on the fifth step.

For the other evaluation scenarios, it is unclear what is the strongest method as sometimes LABDIST is the best performer and sometimes it is Whole (the red curve). However, we believe that in general LABDIST is the most stable model while also often being the most performant. We therefore choose it as our default LAGONN configuration as a compromise between strength and stability. It is about this configuration which we report results in the main text. Our interpretation of this is that passing the model both a discrete prediction (the gold label of the NN) and a truly continuous measure of similarity (the unrounded Euclidean distance) gives it the most consistent and dependable reasoning ability.

We note, as we did in Appendix A.4.1, that we could have presented the best performer for each evaluation scenario, however, it is not the goal of our work to create even more hyperparameters that must be iterated over. However, we hope that our codebase has made it easy for one to change these configurations for their own purposes.

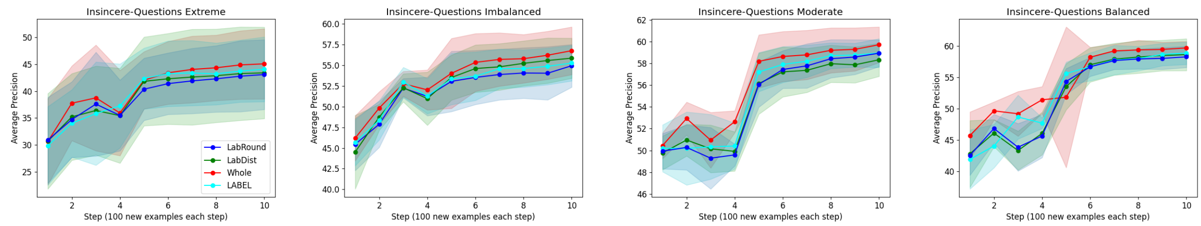

Figure 30: LAGONN$_{lite}$ performance where we compare the LABDIST against LABEL, LABROUND, and rounding the distance to the nearest whole number. We consider all balance regimes on the Insincere Questions dataset and the relevant balance is in the title of each panel.

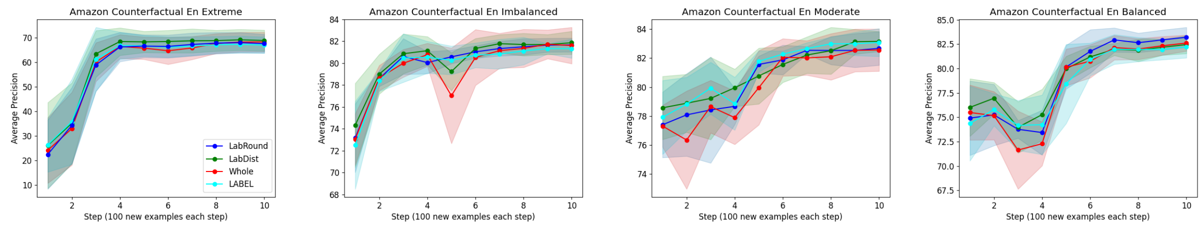

Figure 31: LAGONN$_{lite}$ performance where we compare the LABDIST against LABEL, LABROUND, and rounding the distance to the nearest whole number. We consider all balance regimes on the Amazon Counterfactual dataset and the relevant balance is in the title of each panel.

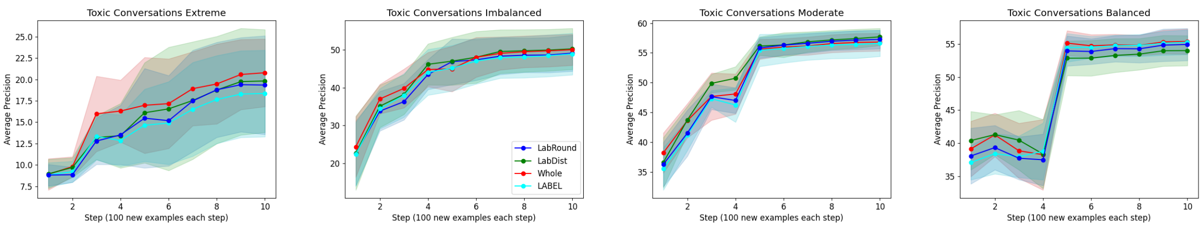

Figure 32: LAGONN$_{lite}$ performance where we compare the LABDIST against LABEL, LABROUND, and rounding the distance to the nearest whole number. We consider all balance regimes on the Toxic Conversations dataset and the relevant balance is in the title of each panel.

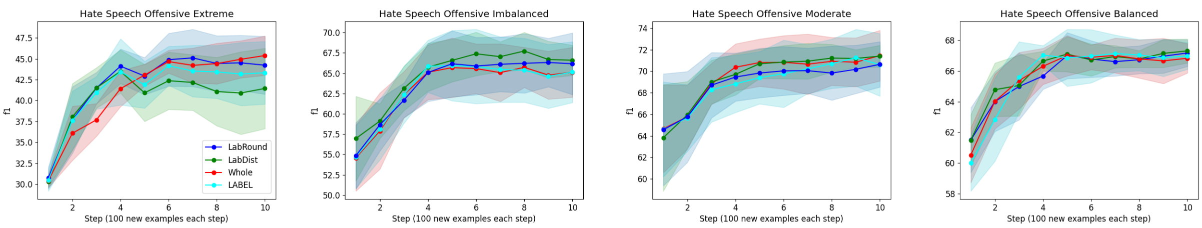

Figure 33: LAGONN$_{lite}$ performance where we compare the LABDIST against LABEL, LABROUND, and rounding the distance to the nearest whole number. We consider all balance regimes on the Hate Speech Offensive dataset and the relevant balance is in the title of each panel.

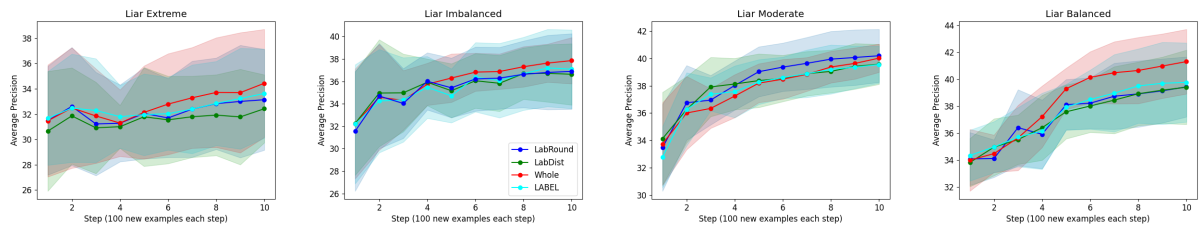

Figure 34: LAGONN$_{lite}$ performance where we compare the LABDIST against LABEL, LABROUND, and rounding the distance to the nearest whole number. We consider all balance regimes on the Liar dataset and the relevant balance is in the title of each panel.

A.5 Examples of LaGoNN modified text

WARNING: Some of the examples below are of an offensive nature. Please view with caution.

In this section, we provide examples of how LaGoNN$_{exp}$ modifies test text from the datasets we studied under the ALL configuration. We choose this configuration because the information it appends from a NN in the training data to a test instance encapsulates all configurations. LaGoNN$_{exp}$ was trained under a balanced distribution and five examples per label were chosen randomly on the first, fifth, and tenth step to demonstrate how the same test instance might be decorated with different training examples as the training data grow. We recognize that some the images below are difficult to see and have made the .csv files available with our code and data files. Note that MPNET's separator token is </s>, not [SEP].

Test Modified	Gold Label
What rapper still relevant and popular today has the best rhyme schemes? </s> <insincere question 3.859471321105957> What would be a good nickname for Trump, Donald Dumbck, and President Spankovich? </s> <valid question 4.124274253845215> What are after class 12 courses in commerce stream to choose from? I have completed my class 12 (expexted 90+) and aim to do business (not aim to do job).	valid question
Which books do you suggest to someone who get a free time and will help him stay motivated? </s> <valid question 3.9509353637695312> What are the best online courses to learn data science? </s> <insincere question 4.300448417663574> What are the more steps in Career Oriented Education?	valid question
How will you feel if someone talks badly about Kunti? </s> <insincere question 3.5063605308532715> Why are the UK government and the media (especially the BBC and the Guardian) demonising ordinary British people, manipulating buzz words like â€œalt-rightâ€, â€œIslamaphobiaâ€, â€œracistâ€ to suppress legitimate outrage at Muslim grooming gangs? </s> <valid question 3.6699037551879883> How do Israelis and Palestinians view Nuseir Yassin?	valid question
Why is equine HYPP inherited? </s> <valid question 4.066534996032715> Can you share some of the pics of hostel of Indira Gandhi medical college, Shimla, Himachal Pradesh? </s> <insincere question 4.231775760650635> I am an experienced programmer and in my high school my teacher tried to make me use python so I said, "No; Trust me, python is just a language for beginners, thereby making it not for me." I got sent out. Did I do anything wrong?	valid question
How do the Valerie Stevens leather jackets achieve their quality during the manufacturing process? </s> <valid question 3.9721384048461914> How are the Lancaster leather sofas manufactured? </s> <insincere question 4.3559441566467285> I am an experienced programmer and in my high school my teacher tried to make me use python so I said, "No; Trust me, python is just a language for beginners, thereby making it not for me." I got sent out. Did I do anything wrong?	valid question
Is Ariana Grande really as mean and bitchy as she seems? </s> <insincere question 3.572277545928955> Why is Alia Bhatt so dumb? </s> <valid question 3.924571990966797> Do you agree with Congressman Steve King's comments on immigrant children in detention centers?	insincere question
Do you guys know that aliens are real and all those satellites we send up in space work as a sort of tracking device for them so in a few years it will be too late for Earth? </s> <insincere question 3.6094439029693604> Have you noticed how conservatives are capturing the English language and modifying the definitions of political words? </s> <valid question 3.6901655197143555> Do you agree with Congressman Steve King's comments on immigrant children in detention centers?	insincere question
Is it politically incorrect to say female privilege, but it is a more accurate term to say, white female privilege? </s> <insincere question 3.323280096054077> Why are the UK government and the media (especially the BBC and the Guardian) demonising ordinary British people, manipulating buzz words like â€œalt-rightâ€, â€œIslamaphobiaâ€, â€œracistâ€ to suppress legitimate outrage at Muslim grooming gangs? </s> <valid question 3.986680269241333> Do you agree with Congressman Steve King's comments on immigrant children in detention centers?	insincere question
On Motherâ€™s Day, is it reasonable to reflect there is some truth in the unfashionable notion than women are more driven by emotion and men more driven by reason? </s> <insincere question 3.499204158782959> Why are the UK government and the media (especially the BBC and the Guardian) demonising ordinary British people, manipulating buzz words like â€œalt-rightâ€, â€œIslamaphobiaâ€, â€œracistâ€ to suppress legitimate outrage at Muslim grooming gangs? </s> <valid question 3.771740198135376> Do you agree with Congressman Steve King's comments on immigrant children in detention centers?	insincere question
If the U.S. president is a role model, is it acceptable for children to say "go fuck yourself," and use the word "pussy" to describe women? </s> <insincere question 3.497847080230713> Why are the UK government and the media (especially the BBC and the Guardian) demonising ordinary British people, manipulating buzz words like â€œalt-rightâ€, â€œIslamaphobiaâ€, â€œracistâ€ to suppress legitimate outrage at Muslim grooming gangs? </s> <valid question 3.845909357070923> Do you agree with Congressman Steve King's comments on immigrant children in detention centers?	insincere question

Figure 35: Insincere Questions, step 1.

Test Modified	Gold Label
What rapper still relevant and popular today has the best rhyme schemes? </s> <insincere question 3.871907949447632> What would be a good nickname for Trump, Donald Dumbck, and President Spankovich? </s> <valid question 4.028958797454834> Why does Dancing with the Stars not include Bachata as one their dance styles?	valid question
Which books do you suggest to someone who get a free time and will help him stay motivated? </s> <valid question 3.6081225872039795> What is a good degree to get at community college if you want to explore different subjects and figure out your career path? </s> <insincere question 3.8502604961395264> What are the more steps in Career Oriented Education?	valid question
How will you feel if someone talks badly about Kunti? </s> <valid question 3.5355563163757324> How do I stop feeling bad after a girl had a crush on me? </s> <insincere question 3.689171075820923> Why Indian girls go crazy about marrying Shri. Rahul Gandhi ji?	valid question
Why is equine HYPP inherited? </s> <insincere question 3.6035702228546143> Can female animals with male humans sex? </s> <valid question 3.7413032054901123> How long do guinea pigs live for?	valid question
How do the Valerie Stevens leather jackets achieve their quality during the manufacturing process? </s> <valid question 2.747288227081299> How are the Lancaster leather sofas manufactured? </s> <insincere question 3.944884777069092> Why don't all Trump supporters buy only made in USA goods, e.g. many of them have their cars of Asian/European companies, shop in places where more than 70% of items are not made in USA, eat multi-national cuisine or otherwise stop their hypocrisy?	valid question
Is Ariana Grande really as mean and bitchy as she seems? </s> <insincere question 3.3252298831939697> Why is Alia Bhatt so dumb? </s> <valid question 3.7413415908813477> How do I stop feeling bad after a girl had a crush on me?	insincere question
Do you guys know that aliens are real and all those satellites we send up in space work as a sort of tracking device for them so in a few years it will be too late for Earth? </s> <insincere question 3.0673365592956543> Isn't it obvious now that walking on the moon by the Americans was a hoax, because walking on the bright side of the moon, even in a space suit would be fatal? </s> <valid question 3.1978228092193604> Why do we weunch satellites?	insincere question
Is it politically incorrect to say female privilege, but it is a more accurate term to say, white female privilege? </s> <insincere question 2.917681271936035> How does the privilege of being attractive compare to the privilege of being White in the US? </s> <valid question 3.112481117248535> Is the media wrong for enforcing gender stereotypes?	insincere question
On Motherâ€™s Day, is it reasonable to reflect there is some truth in the unfashionable notion than women are more driven by emotion and men more driven by reason? </s> <insincere question 3.102355811264038> Do women look down on men who are single, even if the man is more successful in other aspects of his life? </s> <valid question 3.1890125274658203> Why are some women uninterested in sex?	insincere question
If the U.S. president is a role model, is it acceptable for children to say "go fuck yourself," and use the word "pussy" to describe women? </s> <insincere question 3.163693904876709> Is it wrong to take your retarded son to a hooker for his 21st birthday? </s> <valid question 3.456286907196045> Do you agree with Congressman Steve King's comments on immigrant children in detention centers?	insincere question

Figure 36: Insincere Questions, step 5.

Test Modified	Gold Label
What rapper still relevant and popular today has the best rhyme schemes? </s> <valid question 3.7103171348571777> What is the oldest fashion trends running yet? </s> <insincere question 3.871907949447632> What would be a good nickname for Trump, Donald Dumbck, and President Spankovich?	valid question
Which books do you suggest to someone who get a free time and will help him stay motivated? </s> <valid question 3.1401429176330566> How can I stay motivated when learning something new? </s> <insincere question 3.7235560417175293> I'm hungry and I'm too lazy too get out of bed, should I get a psychologist or ask you questions?	valid question
How will you feel if someone talks badly about Kunti? </s> <insincere question 3.4893462657928467> Does Tamil Isai Soundarajan support Vijayendra for disrespecting the Tamil Anthem? </s> <valid question 3.5355563163757324> How do I stop feeling bad after a girl had a crush on me?	valid question
Why is equine HYPP inherited? </s> <valid question 3.5067965984344482> What disadvantages do animals that don't have bones face? </s> <insincere question 3.6035702228546143> Can female animals with male humans sex?	valid question
How do the Valerie Stevens leather jackets achieve their quality during the manufacturing process? </s> <valid question 2.747288227081299> How are the Lancaster leather sofas manufactured? </s> <insincere question 3.9087233543395996> Are Newport cigarettes designed to selectively destroy black people's DNA?	valid question
Is Ariana Grande really as mean and bitchy as she seems? </s> <valid question 3.183567762374878> I like this girl who used to be quite rude and would run through boyfriends very fast. But now that school started again, she seems to have gotten a lot nicer throughout Summer. Is she faking her politeness, and is it worth pursuing her? </s> <insincere question 3.3253660202026367> Why is Alia Bhatt so dumb?	insincere question
Do you guys know that aliens are real and all those satellites we send up in space work as a sort of tracking device for them so in a few years it will be too late for Earth? </s> <insincere question 3.0673365592956543> Isn't it obvious now that walking on the moon by the Americans was a hoax, because walking on the bright side of the moon, even in a space suit would be fatal? </s> <valid question 3.1978228092193604> Why do we weunch satellites?	insincere question
Is it politically incorrect to say female privilege, but it is a more accurate term to say, white female privilege? </s> <insincere question 2.9176158905029297> How does the privilege of being attractive compare to the privilege of being White in the US? </s> <valid question 3.112481117248535> Is the media wrong for enforcing gender stereotypes?	insincere question
On Motherâ€™s Day, is it reasonable to reflect there is some truth in the unfashionable notion than women are more driven by emotion and men more driven by reason? </s> <insincere question 2.9901626110076904> Do you agree that females think with their brains and males with their testicles? </s> <valid question 3.1890125274658203> Why are some women uninterested in sex?	insincere question
If the U.S. president is a role model, is it acceptable for children to say "go fuck yourself," and use the word "pussy" to describe women? </s> <insincere question 2.994286298751831> Why do feminists let their daughters have sex with their boyfriendâ€™s at home? </s> <valid question 3.456286907196045> Do you agree with Congressman Steve King's comments on immigrant children in detention centers?	insincere question

Figure 37: Insincere Questions, step 10.

Test Modified	Gold Label
Clings to the wall, doesn't flop around when a bag is pulled out, the mess of bags falling out is gone. </s> <not-counterfactual 3.6492726802825928> Hopes that it will keep it's shape after washing. </s> <counterfactual 4.012346267700195> "\\\\\Had I reviewed this immediately I would have given this product five stars because it worked."""	not-counterfactual
I like these jeans they sit low enough without being inappropriate when you sit or bend over. </s> <counterfactual 3.402600049972534> "But oddly enough, the bottoms are a little too loose in the waist (37\) and could have used another inch or two in the inseam (I normally take a 35\"" or 36\"" in jeans, depending on the brand if this helps).""" </s> <not-counterfactual 3.4201438426971436> These boxer-briefs are very soft, very comfortable, and fit like high-end underwear the likes of which you might get at, oh, say, Calvin Klein for example, but for about half the price.	not-counterfactual
He was very professional and wish all transactions I make through Amazon were this good. </s> <counterfactual 3.4319906816927> I wish I had had him as an instructor at college. </s> <not-counterfactual 4.054030895233154> I worried that it would be cheap or not fit or...whatever...But WOW!	not-counterfactual
Well written with a twist I didn't expect. </s> <not-counterfactual 3.3257973194122314> "The crossover from the characters from one novel to others keeps me interested; after all, I do hate to miss a \Dee-Ann or Eggie\"" appearance.""" </s> <counterfactual 3.6820030212402344> "\\\\\Had I reviewed this immediately I would have given this product five stars because it worked."""	not-counterfactual
Doesn't feel like the quality levi's I am used to. </s> <not-counterfactual 3.2773308753967285> However, the fabric is not that great, it's cheap scratchy cotton. </s> <counterfactual 3.746659755706787> The blanket is nice and soft but it is white, so if it doesn't light up it isn't much use!	not-counterfactual
If we had wall studs, I believe the enclosed hardware would have been sufficient. </s> <counterfactual 3.433864355072803> I wish the storage compartment was a little bigger and opened up instead of slidding on and off. </s> <not-counterfactual 3.9785308837890625> I worried that it would be cheap or not fit or...whatever...But WOW!	counterfactual
If this ever turns into a film, I hope they do it justice! </s> <not-counterfactual 3.5291523933410645> "The crossover from the characters from one novel to others keeps me interested; after all, I do hate to miss a \Dee-Ann or Eggie\"" appearance.""" </s> <counterfactual 3.751143217086792> "\\\\\Had I reviewed this immediately I would have given this product five stars because it worked."""	counterfactual
If you don't want a prominent display this rack is too large for most bed or living rooms, it is wider and taller than my tall Broyhill wardrobe style dresser which was the largest piece in the room until this shoe rack. </s> <not-counterfactual 3.865670680999756> "It also validates \\\\the incorrect\\\\"" assumption that we are alone in the feelings we suppress when we sense the complete garbage that is thrown out into society.""" </s> <counterfactual 4.063361167907715> The blanket is nice and soft but it is white, so if it doesn't light up it isn't much use!	counterfactual
I wish I could have seen all of the places he recommends! </s> <counterfactual 3.5627076625823975> I wish I had had him as an instructor at college. </s> <not-counterfactual 4.141315937042236> I worried that it would be cheap or not fit or...whatever...But WOW!	counterfactual
I wish I could replace just that small stupid piece, since there's nothing wrong with the rest of the hose assembly. </s> <counterfactual 3.6057372093200684> I wish the storage compartment was a little bigger and opened up instead of slidding on and off. </s> <not-counterfactual 4.064871311187744> I worried that it would be cheap or not fit or...whatever...But WOW!	counterfactual

Figure 38: Amazon Counterfactual, step 1.

Test Modified	Gold Label
Clings to the wall, doesn't flop around when a bag is pulled out, the mess of bags falling out is gone. </s> <not-counterfactual 3.161406993865967> And the dvd cases were tightly packed to ensure they didn't move around. </s> <counterfactual 3.308583974838257> The case is small, cord seems to always want to stay kinked and coiled, plug should be angled and not straight...which are all items that others have pointed out.	not-counterfactual
I like these jeans they sit low enough without being inappropriate when you sit or bend over. </s> <counterfactual 2.606198310852051> "But oddly enough, the bottoms are a little too loose in the waist (37\) and could have used another inch or two in the inseam (I normally take a 35\"" or 36\"" in jeans, depending on the brand if this helps).""" </s> <not-counterfactual 2.6380045413970947> A tad loose but I rather have it fit this way than too tight.	not-counterfactual
He was very professional and wish all transactions I make through Amazon were this good. </s> <not-counterfactual 3.3291680812835693> This new speaker was just what the doctor ordered and I couldn't be more pleased. </s> <counterfactual 3.4589436054229736> Had the person handling the shipping of this item been at all concerned with the use of the product at the end of the mailing process, the slightest bit of care could have been taken to ensure it's proper delivery.	not-counterfactual
Well written with a twist I didn't expect. </s> <not-counterfactual 2.651658535003662> The book had some interesting twists that I did see coming and I look forward to reading part two of this series. </s> <counterfactual 2.837316274629443> Fun read Could have been a little longer with more detail.	not-counterfactual
Doesn't feel like the quality levi's I am used to. </s> <counterfactual 2.733877182006836> It has the same great comfortable & flattering features plus the great denim texture that Lee has perfected- smoothing and stretchy without the excessive cling- but I think it must have been designed for people who have a greater surplus of belly fat than I. </s> <not-counterfactual 2.856729745864868> Will keep but won't be that casual sexy top you always want to turn to.	not-counterfactual
If we had wall studs, I believe the enclosed hardware would have been sufficient. </s> <not-counterfactual 2.6638145446777344> It was a little tricky to find the center of the studs using my stud finder but once I felt comfortable with the lines I had drawn, I drilled the pilot holes and bolted this thing to the wall. </s> <counterfactual 2.879924774169922> The only thing I would have like for it to have a hole in the middle so I can put the stopper in without removing the mat.	counterfactual
If this ever turns into a film, I hope they do it justice! </s> <not-counterfactual 2.671574354171753> I read this book because of the motion picture that is coming out soon. </s> <counterfactual 3.1458709239959717> Was a good story, though there could have been more to it.	counterfactual
If you don't want a prominent display this rack is too large for most bed or living rooms, it is wider and taller than my tall Broyhill wardrobe style dresser which was the largest piece in the room until this shoe rack. </s> <counterfactual 2.7353768348693848> I bought this mount because I wanted one that would sit on three studs instead of two because my TV is quite heavy and I would have had a hard time centering it on my wall if I didn't have the wide hanging rail that this one has. </s> <not-counterfactual 2.873617172241211> Good for under the bed shoe storage, IF the wife wants to use it.	counterfactual
I wish I could have seen all of the places he recommends! </s> <counterfactual 2.799947738647461> I wish I had had him as an instructor at college. </s> <not-counterfactual 3.3013432025909424> And as the ole man isn't any version of slender it was good that he got to try on some shirts before hand.	counterfactual
I wish I could replace just that small stupid piece, since there's nothing wrong with the rest of the hose assembly. </s> <counterfactual 2.628289222717285> The only thing I would have like for it to have a hole in the middle so I can put the stopper in without removing the mat. </s> <not-counterfactual 2.9200568199157715> The only downside is my laptop does not have the screw holes on it and the screws do not retract far enough back for me to push the connector all the way in, but a simple smash will rid that issue (this thing is durable!)	counterfactual

Figure 39: Amazon Counterfactual, step 5.

Test Modified	Gold Label
Clings to the wall, doesn't flop around when a bag is pulled out, the mess of bags falling out is gone. </s> <not-counterfactual 3.161406993865967> And the dvd cases were tightly packed to ensure they didn't move around. </s> <counterfactual 3.289605140686035> If I had to come up with anything negative, I would say that the attachments donâ€™t seem to stay on the vacuum cleaner when not in use - but that could be me not putting them on properly!	not-counterfactual
I like these jeans they sit low enough without being inappropriate when you sit or bend over. </s> <not-counterfactual 2.447404623031616> These shorts fit really well and look good too. </s> <counterfactual 2.550638198852539> The top fits great just wish the bottoms fit too.	not-counterfactual
He was very professional and wish all transactions I make through Amazon were this good. </s> <not-counterfactual 3.3291127681732178> This new speaker was just what the doctor ordered and I couldn't be more pleased. </s> <counterfactual 3.3897111415863037> But the author alleviated my concerns quickly with a few well-timed comments about how it was the man could have known that the arrangement was something Jack wanted.	not-counterfactual
Well written with a twist I didn't expect. </s> <not-counterfactual 2.557446002960205> "A bit workmanlike, not up to Lord's high standard of \A Night to Remember,\"" but well-detailed, and a story that not many now know.""" </s> <counterfactual 2.792485475540161> Wow I am really glad I didn't read these reviews BEFORE I read this book because I would have passed on the book and missed a really great start to a series that captured my attention and made me laugh all the while using my imagination and painting a clear picture of the author's world she was building for us.	not-counterfactual
Doesn't feel like the quality levi's I am used to. </s> <counterfactual 2.5612902641296387> i was hoping the pants would be thicker but being that it's not too expensive it's understandable. </s> <not-counterfactual 2.572395086288452> But it doesn't have a lining like the last couple models I bought.	not-counterfactual
If we had wall studs, I believe the enclosed hardware would have been sufficient. </s> <not-counterfactual 2.6638145446777344> It was a little tricky to find the center of the studs using my stud finder but once I felt comfortable with the lines I had drawn, I drilled the pilot holes and bolted this thing to the wall. </s> <counterfactual 2.771395206451416> Wish it had a little more padding, otherwise just as advertised.	counterfactual
If this ever turns into a film, I hope they do it justice! </s> <not-counterfactual 2.671574354171753> I read this book because of the motion picture that is coming out soon. </s> <counterfactual 3.141676187515259> Wish this story would have been longer and turned into a book, with some gut wrenching action, love/hate lovers quarrels scenes, with a happy ending at the end...	counterfactual
If you don't want a prominent display this rack is too large for most bed or living rooms, it is wider and taller than my tall Broyhill wardrobe style dresser which was the largest piece in the room until this shoe rack. </s> <counterfactual 2.7353768348693848> I bought this mount because I wanted one that would sit on three studs instead of two because my TV is quite heavy and I would have had a hard time centering it on my wall if I didn't have the wide hanging rail that this one has. </s> <not-counterfactual 2.873617172241211> Good for under the bed shoe storage, IF the wife wants to use it.	counterfactual
I wish I could have seen all of the places he recommends! </s> <counterfactual 2.7999041080474854> I wish I had had him as an instructor at college. </s> <not-counterfactual 3.2604622840881348> I wanted to order him a few afordable hats I wouldn't mind him loosing.	counterfactual
I wish I could replace just that small stupid piece, since there's nothing wrong with the rest of the hose assembly. </s> <counterfactual 2.474032402038574> I wish I could just hook up a hose rather than connecting and routing an tube internally </s> <not-counterfactual 2.9200568199157715> The only downside is my laptop does not have the screw holes on it and the screws do not retract far enough back for me to push the connector all the way in, but a simple smash will rid that issue (this thing is durable!)	counterfactual

Figure 40: Amazon Counterfactual, step 10.

Test Modified	Gold Label
Homos demand that you accept their fudge packing. But none of us ever will- </s> <not toxic 3.5619421005249023> Sounds just awful, working for the state that is. So it begs the question, why work for the state if the pay is so bad versus the private sector? Seems logical to just make the switch? </s> <toxic 4.056504249572754> I would imagine it is a miserable existence. Being a whining sore loser is no way to go through life.	not toxic
I don't think anyone likes this health care bill, it stinks for everyone. 50 years and older are going to get hammered with higher premiums. People with preexisting conditions will also see their premiums go through the roof. Eventually no one will be able to afford it. They're still not addressing the reason and the solution. Prescription drugs. Why are Americans paying 2 to 3 times higher for prescription drugs than Canadians are? They should sit down with Democrats and fix it not play musical chairs with our health care. Here is the add that AARP put out, and it has a fact check at the end. It's worth watching. http://www.thedenverchannel.com/news/politics/aarp-ad-says-house-gop-health-care-bill-would-boost-insurance-rates-for-older-americans </s> <toxic 3.6495330333709717> Calling everybody that disagrees with you a racist and anti-Muslim is pretty uncivil. Also intellectually bankrupt... </s> <not toxic 3.8587634563446045> Sounds just awful, working for the state that is. So it begs the question, why work for the state if the pay is so bad versus the private sector? Seems logical to just make the switch?	not toxic
Rather than call you a liar, I'll just post a link to HART's helicopter footage of this "most densely populated" ag land. https://www.youtube.com/watch?v=opJzRWWVDdw And if you've been following a little closer, you'd know about the luxury towers and boutique hotels that developers are planning on fulfilling their TOD requirement with "cash gifts" to the city instead of actual affordable TOD. </s> <not toxic 3.8009092807769775> Sounds just awful, working for the state that is. So it begs the question, why work for the state if the pay is so bad versus the private sector? Seems logical to just make the switch? </s> <toxic 4.463983535766602> I would imagine it is a miserable existence. Being a whining sore loser is no way to go through life.	not toxic
Get the "real news" from Breitbart and Drudge Report? Ahahahahaha! Are you for real? </s> <not toxic 3.559427499771118> Sounds just awful, working for the state that is. So it begs the question, why work for the state if the pay is so bad versus the private sector? Seems logical to just make the switch? </s> <toxic 4.239322662353516> I would imagine it is a miserable existence. Being a whining sore loser is no way to go through life.	not toxic
Wud luv to see one of those American Boats with all the planes on it at the bottom of the sea. </s> <not toxic 3.8268911838531494> Sounds just awful, working for the state that is. So it begs the question, why work for the state if the pay is so bad versus the private sector? Seems logical to just make the switch? </s> <toxic 4.241443157196045> I would imagine it is a miserable existence. Being a whining sore loser is no way to go through life.	not toxic
Wtf, nowhere in scripture does it say that Elizabeth was a kid when married to Zachariah and nowhere does it say that Mary was a child when married to Joseph. That's a complete and utter fabrication. In fact, it says in Luke 1:42, "And she spake out with a loud voice, and said, Blessed art thou among women, and blessed is the fruit of thy womb." It says blessed art thou among WOMEN. </s> <not toxic 3.6674933433532715> Sounds just awful, working for the state that is. So it begs the question, why work for the state if the pay is so bad versus the private sector? Seems logical to just make the switch? </s> <toxic 4.0054121017456055> I would imagine it is a miserable existence. Being a whining sore loser is no way to go through life.	toxic
Angela Merkel and all other European political leaders who have aided and abetted the ongoing invasion of Europe by the forces of the crescent moon death cult should be tried as accessories to Mr. Urban's murder. </s> <toxic 3.262410879135132> Calling everybody that disagrees with you a racist and anti-Muslim is pretty uncivil. Also intellectually bankrupt... </s> <not toxic 3.8916428089141846> It's always important to remember what can happen when you have sociopaths as leaders and also have compliant followers. Some of the younger posters on this site might want to Google "Jim Jones and Jonestow"... There were no "checks and balances" in Jonestown; I fear there are none in North Korea....and I can only hope those in our country are firmly in place and functioning. Gary Crum	toxic
I hope you don't have kids if you see this woman's actions as acceptable. And I applaud the den for kicking the kid out. She brought unwanted negative attention upon them. However, she will, and is already likely, pay the the price for her stupid stunt. </s> <toxic 3.0406124591827393> Calling everybody that disagrees with you a racist and anti-Muslim is pretty uncivil. Also intellectually bankrupt... </s> <not toxic 4.094666481018066> Christ never said he would give grace, mercy, and acceptance to those who determinedly violate Scripture. In fact, he often spoke of hell.	toxic
no one cares what a paid liberal trolling hack like you believes lunatic,, </s> <toxic 2.8411786556243896> Calling everybody that disagrees with you a racist and anti-Muslim is pretty uncivil. Also intellectually bankrupt... </s> <not toxic 4.034884929656982> Christ never said he would give grace, mercy, and acceptance to those who determinedly violate Scripture. In fact, he often spoke of hell.	toxic
Ok all you NDP "LEAP" manifesto types, where is your hero Naomi Klein? Her fawning adoration of Chavez and Venezuelan thuggery knows no bounds. I'm sure she's awfully hysterical over the thought that such a pathetic dictstorship could ever be sanctioned. </s> <toxic 3.3616135120391846> Calling everybody that disagrees with you a racist and anti-Muslim is pretty uncivil. Also intellectually bankrupt... </s> <not toxic 3.903903007507324> I have very high respect for teachers that get the job done. Teaching is an extremely difficult and important job. And it is quite apparent that we are in desperate need of teachers that can actually do the job. Pride of workmanship would have teachers wanting to have their students periodically evaluated and tested to show how well they have done their job. We have some very competent teachers that get the job done and welcome student testing (in spite of sorry admin and unfair union). But the majority of teachers here instead of doing their jobs band together wear purple shirts and mob the government for a better contract, and no accountability in the form of testing students. Many teachers are middle class kids that took the path of least resistance in what was expected by their parents (college) and because they lacked drive ended up teachers. That lack of drive shows by what the private sector taxpayers get for their money. Your degrees mean nothing if you don't do your job.	toxic

Figure 41: Toxic Conversations, step 1.

Test Modified	Gold Label
Homos demand that you accept their fudge packing. But none of us ever will- </s> <toxic 3.1383495330810547> So you admit you would exterminate inferior humans. </s> <not toxic 3.2954952716827393> Mark MacKinnon and the interests he work for would like us to 'get used to it', because they don't want to do anything practical to stop it.	not toxic
I don't think anyone likes this health care bill, it stinks for everyone. 50 years and older are going to get hammered with higher premiums. People with preexisting conditions will also see their premiums go through the roof. Eventually no one will be able to afford it. They're still not addressing the reason and the solution. Prescription drugs. Why are Americans paying 2 to 3 times higher for prescription drugs than Canadians are? They should sit down with Democrats and fix it not play musical chairs with our health care. Here is the add that AARP put out, and it has a fact check at the end. It's worth watching. http://www.thedenverchannel.com/news/politics/aarp-ad-says-house-gop-health-care-bill-would-boost-insurance-rates-for-older-americans </s> <not toxic 2.5086519718170166> so in the mean time tens of thousands of Oregonians go without health insurance which will now be unaffordable to them. And sorry, the republicans have had 8 years to figure out a better system, they aren't going to do it anytime soon. Neither party has any desire to actually find a solution to all this. Hopefully Trump will also soon eliminate the tax penalty for not having insurance so all us folks who buy our own insurance don't get penalized for not being able to afford to buy the insurance we are required to have. Of course he is probably clueless that detail even exists </s> <toxic 2.801957607269287> reducing number of brackets. Another is lowering corporate tax rates which would be OK if all loopholes, including tax-free political donations for wealthy people only. Another is doubling the earned income tax credit which would help families with children but for people like me, would not make up for loss of the state tax deduction. Essentially the proposed tax "reform bill as it stands is a huge wealth transfer, from working people and the treasury to the offshore accounts of wealthy people and corporations. Median earners like me will be screwed. Poor people with children will be screwed. The rich will get a huge bundle of cash, and the Republicans will lead us further into deep debt. That's if Ryan and McConnell can pass this huge sack of excrement.	not toxic
Rather than call you a liar, I'll just post a link to HART's helicopter footage of this "most densely populated" ag land. https://www.youtube.com/watch?v=opJzRWWVDdw And if you've been following a little closer, you'd know about the luxury towers and boutique hotels that developers are planning on fulfilling their TOD requirement with "cash gifts" to the city instead of actual affordable TOD. </s> <not toxic 3.0351996421813965> So are you saying we should build the road. Lots of doom and gloom but I'm missing your point. </s> <toxic 3.1464040279388428> Isn't that terrible? Caldwell and Inouye do not know the amount a 10-year extension would raise yet they are proposing it. Caldwell asked for two years and he got it. Two years later he is now asking for ten years. Now, which is only weeks later, HART director is claiming 10 years is insufficient. This scenario shows our mayor and HART do not know the scope of this project. Pathetic!	not toxic
Get the "real news" from Breitbart and Drudge Report? Ahahahahaha! Are you for real? </s> <toxic 2.634126901626587> "If one read the Dispatch one would think Trump is the most evil person on the planet." Not evil, just idiotic. And it would be easy to give his behaviour a pass if he wasn't POTUS. ".....five to eight anti-Trump stories per day. Never any good one's or one's that just stuck to the facts." Well, when there's good a news Trump story to print, I'm sure ADN will be all over it, problem is, there's been a dearth of those since his election. Facts? Ok Rich, give us a list of incorrect facts in the above story. </s> <not toxic 3.0248677730560303> With dismay I noticed that ADN had printed yet another column from Fox commentator Charles Krauthammer but after reading it I'm glad the editors chose it as the feature article on the opinion page. Krauthammer is also a psychiatrist so his analyses of Trump as a man who has never emotionally, intellectually developed beyond adolescence holds some weight. But what does it say about Trumps supporters that so many millions of them can't see through the boorish, confrontational attitude of the man? How can so many Americans have devolved into such anger, fear and irrationality that they would/could find redemption in Trump after how he has exposed his true narcissistic self for all to see. When you've lost the Jennifer Rubins and Charles Krauthammer's of the media world you've lost the battle yet the Trumpian cult members will soldier on and then become even angrier and more full of fear after the election. Something to do with their choice of "information" sources no doubt.	not toxic
Wud luv to see one of those American Boats with all the planes on it at the bottom of the sea. </s> <toxic 3.3901054859161377> I bet Regent Seven Seas will never offer Mr Hammond another trip. Wow, what a snarky article. He makes , I assume, some valid points about food and atmosphere. However, after discovering the treats available on his "massive deck" he "blew off" his remaining restaurant reservations , donned his comfy bathrobe and ordered-in. He was certainly not an ideal passenger and, for one floating on a freebie, he's a total ingrate! </s> <not toxic 3.409156084060669> Now replaced by the sexy EA-18G Growler! Using a preexisting Military Operating Area! Get over it!!!!!!	not toxic
Wtf, nowhere in scripture does it say that Elizabeth was a kid when married to Zachariah and nowhere does it say that Mary was a child when married to Joseph. That's a complete and utter fabrication. In fact, it says in Luke 1:42, "And she spake out with a loud voice, and said, Blessed art thou among women, and blessed is the fruit of thy womb." It says blessed art thou among WOMEN. </s> <not toxic 2.769857406616211> I was informed that my first grandchild had been conceived the evening of the day when I had inserted a prayer note in the Wailing Wall in Jerusalem that asked God to help my daughter conceive after a year of frustrated attempts. Maybe Elizabeth did the same thing? After all, she was in the same neighborhood. :-) </s> <toxic 3.286393880844116> Christians who support Trump are the most mind-boggling to me. I just don't see how they square the circle between Trump and their moral foundations. â€œBeware then of useless grumbling, and keep your tongue from slander; because no secret word is without result, and a lying mouth destroys the soul.â€ (Wisdom 1:11) If that is the case, then Trump's soul was utterly destroyed decades ago.	toxic
Angela Merkel and all other European political leaders who have aided and abetted the ongoing invasion of Europe by the forces of the crescent moon death cult should be tried as accessories to Mr. Urban's murder. </s> <toxic 3.2037758827209473> that's what happens when you betray the people of your country for foreign bs. let's go Le Pen, Geert Wilders. If the media refuses to mention the muslim crisis the total incompatibility of primitive, uneducated muslim males swarming countries and turning them into misogynistic fundamentalist religious areas then we need these people to save us from YOU! </s> <not toxic 3.2326502799987793> your first mistake is believing what a politician says because generally it has nothing to do with what they do. The Libs will be happy to let this die because Monsef is now a very poor salesman given her own immigration dishonesty. That said if the election prospects sour significantly for the Libs I have no doubts that PM Butts will ram through Ranked Ballot	toxic
I hope you don't have kids if you see this woman's actions as acceptable. And I applaud the den for kicking the kid out. She brought unwanted negative attention upon them. However, she will, and is already likely, pay the the price for her stupid stunt. </s> <toxic 2.8730525970458984> Ms. Van Brocklin: You state that there is Payne's conduct is inexcusable, then proceed to use the rest of your space to justify him and his actions. You have denigrated the proud courage of countless people who took up causes via civil disobedience. I marched in the non-violent peaceful Civil Rights protest. So the dogs and the firehoses used by a certain southern sheriff were justified, by your logic. So were the citizens beaten by Chicago police during the Democratic Convention Police riots. Resolved in 20 minutes? Nonsense. If Miss Wubbels hadn't protested as she did, she'd likely have ended up in a back room, somewhere, and who know when she would have been granted the presence of a lawyer and what she would have gone through prior to that. You are completely ignorant of the shortage of nurses in this country - in some cases, critical shortages. And why would anyone want to be a nurse when they are disrespected by a former state and federal prosecutor such as you. </s> <not toxic 2.92986798286438> Acquit her, then commit her. This womens cheese has slid so far off the cracker she's a danger to herself and to others. Animal rights activism is a just cause, but her and her group have gone off the deep end into radical extremist territory.	toxic
no one cares what a paid liberal trolling hack like you believes lunatic,, </s> <toxic 2.6076736450195312> aa another hate filled left winger again! save the stupid nonsense sheep, trump is not causing anything, our weak leadership is. </s> <not toxic 2.810284376144409> Ouch... didnâ€™t see that one coming. A liberal stealing my own line... just like they take everything else they like.	toxic
Ok all you NDP "LEAP" manifesto types, where is your hero Naomi Klein? Her fawning adoration of Chavez and Venezuelan thuggery knows no bounds. I'm sure she's awfully hysterical over the thought that such a pathetic dictstorship could ever be sanctioned. </s> <toxic 2.941824436187744> Ms. Van Brocklin: You state that there is Payne's conduct is inexcusable, then proceed to use the rest of your space to justify him and his actions. You have denigrated the proud courage of countless people who took up causes via civil disobedience. I marched in the non-violent peaceful Civil Rights protest. So the dogs and the firehoses used by a certain southern sheriff were justified, by your logic. So were the citizens beaten by Chicago police during the Democratic Convention Police riots. Resolved in 20 minutes? Nonsense. If Miss Wubbels hadn't protested as she did, she'd likely have ended up in a back room, somewhere, and who know when she would have been granted the presence of a lawyer and what she would have gone through prior to that. You are completely ignorant of the shortage of nurses in this country - in some cases, critical shortages. And why would anyone want to be a nurse when they are disrespected by a former state and federal prosecutor such as you. </s> <not toxic 3.0424022674560547> So anyone who doesn't share your opinion is a "raving loony?" And as for my post being "speculation? - which part - that the Liberals are the party in power, or that this involves money? As for me not knowing what is going on, you are correct, I am not a member of the Liberal party insider clique, as you apparently are.	toxic

Figure 42: Toxic Conversations, step 5.

Test Modified	Gold Label
Homos demand that you accept their fudge packing. But none of us ever will- </s> <toxic 3.1383132934570312> So you admit you would exterminate inferior humans. </s> <not toxic 3.295428514480591> Mark MacKinnon and the interests he work for would like us to 'get used to it', because they don't want to do anything practical to stop it.	not toxic
I don't think anyone likes this health care bill, it stinks for everyone. 50 years and older are going to get hammered with higher premiums. People with preexisting conditions will also see their premiums go through the roof. Eventually no one will be able to afford it. They're still not addressing the reason and the solution. Prescription drugs. Why are Americans paying 2 to 3 times higher for prescription drugs than Canadians are? They should sit down with Democrats and fix it not play musical chairs with our health care. Here is the add that AARP put out, and it has a fact check at the end. It's worth watching. http://www.thedenverchannel.com/news/politics/aarp-ad-says-house-gop-health-care-bill-would-boost-insurance-rates-for-older-americans </s> <not toxic 2.5086519718170166> so in the mean time tens of thousands of Oregonians go without health insurance which will now be unaffordable to them. And sorry, the republicans have had 8 years to figure out a better system, they aren't going to do it anytime soon. Neither party has any desire to actually find a solution to all this. Hopefully Trump will also soon eliminate the tax penalty for not having insurance so all us folks who buy our own insurance don't get penalized for not being able to afford to buy the insurance we are required to have. Of course he is probably clueless that detail even exists </s> <toxic 2.5147175788879395> hate to bust the bubble but over 60% of people including all those trump voters never liked or wanted obamacare and dont want it now. , trump has NOT told the gop to back off you lying sack of bs.. he wants it gone and replaced period and the gop are doing just that together WITH trump. , you will continue to turn reality into stupidity	not toxic
Rather than call you a liar, I'll just post a link to HART's helicopter footage of this "most densely populated" ag land. https://www.youtube.com/watch?v=opJzRWWVDdw And if you've been following a little closer, you'd know about the luxury towers and boutique hotels that developers are planning on fulfilling their TOD requirement with "cash gifts" to the city instead of actual affordable TOD. </s> <not toxic 2.9202661514282227> I suppose you just support urban sprawl then with that logic. </s> <toxic 2.9730660915374756> Why don't you go and live in one of their buildings and see what they're like? "Deadbeats" - you're an idiot. They're my neighbours.	not toxic
Get the "real news" from Breitbart and Drudge Report? Ahahahahaha! Are you for real? </s> <toxic 2.634126901626587> "If one read the Dispatch one would think Trump is the most evil person on the planet." Not evil, just idiotic. And it would be easy to give his behaviour a pass if he wasn't POTUS. ".....five to eight anti-Trump stories per day. Never any good one's or one's that just stuck to the facts." Well, when there's good a news Trump story to print, I'm sure ADN will be all over it, problem is, there's been a dearth of those since his election. Facts? Ok Rich, give us a list of incorrect facts in the above story. </s> <not toxic 2.9902079105377197> "a gift from the political gods when the struggling effort to pass a health bill dominates the headlines." It was a gift from media that isn't worried about the actual news, they are more worried about trying to influence soft heads. CNN has been screwing up a lot when it comes to Trump, same with the NYT that is now being sued for libel. Every stupid mistake they make gives an even larger advantage to Trump and strengthens his supporters that already believe the MSM is biased against him and makes fence sitters begin to question what's news and what's crap. Like I've been saying, the MSM is slitting it's own throat.	not toxic
Wud luv to see one of those American Boats with all the planes on it at the bottom of the sea. </s> <toxic 3.390166759490967> I bet Regent Seven Seas will never offer Mr Hammond another trip. Wow, what a snarky article. He makes , I assume, some valid points about food and atmosphere. However, after discovering the treats available on his "massive deck" he "blew off" his remaining restaurant reservations , donned his comfy bathrobe and ordered-in. He was certainly not an ideal passenger and, for one floating on a freebie, he's a total ingrate! </s> <not toxic 3.409156084060669> Now replaced by the sexy EA-18G Growler! Using a preexisting Military Operating Area! Get over it!!!!!!	not toxic
Wtf, nowhere in scripture does it say that Elizabeth was a kid when married to Zachariah and nowhere does it say that Mary was a child when married to Joseph. That's a complete and utter fabrication. In fact, it says in Luke 1:42, "And she spake out with a loud voice, and said, Blessed art thou among women, and blessed is the fruit of thy womb." It says blessed art thou among WOMEN. </s> <not toxic 2.769857406616211> I was informed that my first grandchild had been conceived the evening of the day when I had inserted a prayer note in the Wailing Wall in Jerusalem that asked God to help my daughter conceive after a year of frustrated attempts. Maybe Elizabeth did the same thing? After all, she was in the same neighborhood. :-) </s> <toxic 2.9830095767974854> I don't know if the bishops ever told priests that if they rape a child and celebrate Mass they are committing sacrilege. But even if they didn't, the priest should know that...it is pretty basic Catholicism that receiving Communion with a mortal sin on your soul is sacrilege. By the logic of your second paragraph, there can never be a mortally sinful sexual act, since sex acts by definition occur in a state of passion. Which begs the question, why in this case would the Scriptures go through the trouble of condemning sexual immorality? This sounds like something your example of a rapist priest could say to himself to justify himself taking Communion in a state of mortal sin.	toxic
Angela Merkel and all other European political leaders who have aided and abetted the ongoing invasion of Europe by the forces of the crescent moon death cult should be tried as accessories to Mr. Urban's murder. </s> <toxic 3.024935007095337> About 415 million Europeans cannot compete with Google, Amazon, Facebook, Oracle, Intel, Apple, etc. and the socialist European welfare states need more revenue because they are running out of other peoples' money. Thus the Euro-socialist-bureaucrats pick the low-hanging fruit with litigious persecution of American firms which dominate because unlike their pathetic Euro-competitors, the U.S. firms are clever, hard-working, and well-capitalized. If the the Europeans wish to engage in this transparent financial inquisition, then the US should respond with counter litigation for $trillions against corrupt scofflaws like VW (think diesel fiddle!) as well as UBS/Credit Suisse/HSBC/Credit Lyonnaise (think tax cheats!)and sue/litigate them out of existence. If the lazy, corrupt, incompetent Euros want to play with fire, then let them be financially incinerated! </s> <not toxic 3.2237966060638428> Rome should never have made such inane pronouncements at Trent in their attempt to define the substance of holy Eucharist. Most reasonable people understand that perfectly well. That Rome also made their pronouncements (faith and morals) "infallible" is equally tragic, for the simple reason that so-called infallible statements cannot be retracted without calling into question other so-called infallible statements. Sincere question for you: If Jesus and his followers celebrated Eucharist as a communal meal seated around a table, what gives Rome the right to alter this simple act of worship (perhaps "fellowship" is a better word--more suited toward love of God and neighbor), given to us by the Lord himself?	toxic
I hope you don't have kids if you see this woman's actions as acceptable. And I applaud the den for kicking the kid out. She brought unwanted negative attention upon them. However, she will, and is already likely, pay the the price for her stupid stunt. </s> <toxic 2.873025417327881> Ms. Van Brocklin: You state that there is Payne's conduct is inexcusable, then proceed to use the rest of your space to justify him and his actions. You have denigrated the proud courage of countless people who took up causes via civil disobedience. I marched in the non-violent peaceful Civil Rights protest. So the dogs and the firehoses used by a certain southern sheriff were justified, by your logic. So were the citizens beaten by Chicago police during the Democratic Convention Police riots. Resolved in 20 minutes? Nonsense. If Miss Wubbels hadn't protested as she did, she'd likely have ended up in a back room, somewhere, and who know when she would have been granted the presence of a lawyer and what she would have gone through prior to that. You are completely ignorant of the shortage of nurses in this country - in some cases, critical shortages. And why would anyone want to be a nurse when they are disrespected by a former state and federal prosecutor such as you. </s> <not toxic 2.8961446285247803> Well, I can't very well respect or fear an imaginary sky-being. As for my concept of character, it was good enough for the Alaska Judicial Council and Governor Knowles. But that was long ago. I've gotten older and, crikey, maybe I am going downhill. You're right about the inappropriateness of my comment. First Lady Walker's piece is very laudable and I shouldn't have taken it as an occasion to rant. (But look on the bright side: my misplaced comment gave you yet another occasion to rant about how your Fosterism is saving civilization.)	toxic
no one cares what a paid liberal trolling hack like you believes lunatic,, </s> <toxic 2.5876824855804443> It always amuses me when a troll gets on, they like their own comments and simply assert everyone else is wrong. Never any evidence to rebut it just blind assertions. </s> <not toxic 2.810284376144409> Ouch... didnâ€™t see that one coming. A liberal stealing my own line... just like they take everything else they like.	toxic
Ok all you NDP "LEAP" manifesto types, where is your hero Naomi Klein? Her fawning adoration of Chavez and Venezuelan thuggery knows no bounds. I'm sure she's awfully hysterical over the thought that such a pathetic dictstorship could ever be sanctioned. </s> <toxic 2.883418321609497> Um, no. The major left-wing Labour party was decimated; Mr. Rutte lost 8 seats; and Mr. Wilders Freedom party GAINED 4 seats. Now Mr. Rutte will have to garner favour among RIGHT-WING parties to cobble together his coalition. And as to your glib little comment about people "embracing left-leaning parties", you need only look to Canada to see the fiasco that results... The corrupt and divisive Trudeau government was elected on a slate of fuzzy, alt-left promises; only to renege on most of them ! </s> <not toxic 3.0424022674560547> So anyone who doesn't share your opinion is a "raving loony?" And as for my post being "speculation? - which part - that the Liberals are the party in power, or that this involves money? As for me not knowing what is going on, you are correct, I am not a member of the Liberal party insider clique, as you apparently are.	toxic

Figure 43: Toxic Conversations, step 10.

Test Modified	Gold Label
Afscme says In labor negotiations with city employees, Milwaukee Mayor Tom Barrett demanded concessions that went beyond those mandated by Gov. Scott Walkers collective bargaining law </s> a letter to members </s> <true statement 3.833270311355591> Donald Trump says Libya Ambassador (Christopher) Stevens sent 600 requests for help in Benghazi. </s> the second 2016 presidential debate </s> <false statement 4.013778209686279> Donald Trump says The federal government is sending refugees to states with governors who are Republicans, not to the Democrats. </s> an interview on Laura Ingraham's radio show	true statement
Rick Scott says All Aboard Florida is a 100 percent private venture. There is no state money involved. </s> a TV interview </s> <false statement 3.664231777191162> Donald Trump says The federal government is sending refugees to states with governors who are Republicans, not to the Democrats. </s> an interview on Laura Ingraham's radio show </s> <true statement 3.831820011138916> Patrick Murphy says Marco Rubio opposes immigration reform. Worse, Rubio supports Donald Trump. His plan would deport 800,000 children, destroying families. </s> a TV ad	true statement
Julie Pace says The Obama administration is using as its legal justification for these airstrikes (on the Islamic State), an authorization for military force that the president himself has called for repeal of. </s> a question to White House Press Secretary Josh Earnest </s> <false statement 3.5803754329681396> Donald Trump says Hillary Clinton invented ISIS with her stupid policies. She is responsible for ISIS. </s> an interview on 60 Minutes </s> <true statement 3.869307518005371> Donald Trump says Libya Ambassador (Christopher) Stevens sent 600 requests for help in Benghazi. </s> the second 2016 presidential debate	true statement
John Kasich says We are now eighth in the nation in job creation . . . we are No. 1 in the Midwest. </s> a news conference </s> <true statement 3.851958990097046> Jorge Elorza says In the last six years of Ciancis administration violent crime was down in the United States. It was down in the region. It was down in Rhode Island. But it was up in Providence. </s> a debate </s> <false statement 4.010262966156006> Donald Trump says The federal government is sending refugees to states with governors who are Republicans, not to the Democrats. </s> an interview on Laura Ingraham's radio show	true statement
Mike Pence says It was Hillary Clinton who left Americans in harms way in Benghazi and after four Americans fell said, What difference at this point does it make? </s> the Republican national convention </s> <true statement 3.7440342903137207> Jorge Elorza says In the last six years of Ciancis administration violent crime was down in the United States. It was down in the region. It was down in Rhode Island. But it was up in Providence. </s> a debate </s> <false statement 3.746598958969116> Donald Trump says You will learn more about Donald Trump by going down to the Federal Elections to see the financial disclosure form than by looking at tax returns. </s> a Presidential debate at Hofstra University	true statement
Rand Paul says Of the roughly 15 percent of Americans who dont have health insurance, half of them made more than $50,000 a year. </s> an interview on Comedy Central's "The Daily Show" </s> <true statement 3.7997491359710693> Bernie S says We have the highest rate of childhood poverty of any major country on Earth. </s> an interview on CNN </s> <false statement 3.963353872299194> Donald Trump says The federal government is sending refugees to states with governors who are Republicans, not to the Democrats. </s> an interview on Laura Ingraham's radio show	false statement
Barack Obama says Stimulus tax cuts "began showing up in paychecks of 4.8 million Indiana households about three months ago." </s> a speech in Wakarusa, Ind. </s> <true statement 3.8199117183685303> Jorge Elorza says In the last six years of Ciancis administration violent crime was down in the United States. It was down in the region. It was down in Rhode Island. But it was up in Providence. </s> a debate </s> <false statement 3.916092872619629> Donald Trump says The federal government is sending refugees to states with governors who are Republicans, not to the Democrats. </s> an interview on Laura Ingraham's radio show	false statement
Allen West says If you look at the application for a security clearance, I have a clearance that even the president of the United States cannot obtain because of my background. </s> a candidate forum </s> <false statement 3.760773181915283> Rush Limbaugh says 11 straight years of no major hurricanes striking land in the United States bores a hole right through the whole climate change argument. </s> a radio show broadcast </s> <true statement 3.77760648727417> Arizona Citizens Defense League says a gun bill before the Senate would make it a federal felony to leave town for more than seven days, and leave someone else at home with your firearms. </s> an email to supporters	false statement
Bernie S says We now work the longest hours of any people around the world. </s> a C-SPAN interview </s> <true statement 3.7155606746673584> Bernie S says We have the highest rate of childhood poverty of any major country on Earth. </s> an interview on CNN </s> <false statement 4.0561442375183105> Rush Limbaugh says 11 straight years of no major hurricanes striking land in the United States bores a hole right through the whole climate change argument. </s> a radio show broadcast	false statement
Sarah Palin says Donald Trumps conversion to pro-life beliefs are akin to Justin Biebers, who said in the past that abortion was no big deal to him. </s> an interview on CNN </s> <false statement 3.7367687225341797> Donald Trump says The federal government is sending refugees to states with governors who are Republicans, not to the Democrats. </s> an interview on Laura Ingraham's radio show </s> <true statement 3.7425951957702637> Patrick Murphy says Marco Rubio opposes immigration reform. Worse, Rubio supports Donald Trump. His plan would deport 800,000 children, destroying families. </s> a TV ad	false statement

Figure 44: LIAR, step 1.

Test Modified	Gold Label
Afscme says In labor negotiations with city employees, Milwaukee Mayor Tom Barrett demanded concessions that went beyond those mandated by Gov. Scott Walkers collective bargaining law </s> a letter to members </s> <false statement 3.131746292114258> Tom Barrett says Gov. Scott Walker said no to equal pay for equal work for women. </s> a TV ad </s> <true statement 3.1800825595855713> Scott Walker says If public employees dont pay more for benefits starting April 1, 2011, the equivalent is 1,500 state employee layoffs by June 30, 2011 and 10,000 to 12,000 state and local government employee layoffs in the next two years. </s> a news conference	true statement
Rick Scott says All Aboard Florida is a 100 percent private venture. There is no state money involved. </s> a TV interview </s> <true statement 3.0582425594329834> Charlie Crist says All Aboard Florida is receiving millions in Florida taxpayer dollars. </s> a fundraising email </s> <false statement 3.1522974967956543> Corey Lewandowski says Mr. Trump is self-financing his campaign, so we dont have any donors. </s> a radio interview.	true statement
Julie Pace says The Obama administration is using as its legal justification for these airstrikes (on the Islamic State), an authorization for military force that the president himself has called for repeal of. </s> a question to White House Press Secretary Josh Earnest </s> <true statement 2.9627556800842285> Martha Raddatz says The Obama administration originally wanted 10,000 troops to remain in Iraq -- not combat troops, but military advisers, special operations forces, to watch the counterterrorism effort. </s> comments on ABC's "This Week" </s> <false statement 3.246009588241577> Rick Perry says Obama has chosen to deny the vicious anti-Semitic motivation of the attack on a kosher Jewish grocery in Paris. </s> a statement	true statement
John Kasich says We are now eighth in the nation in job creation . . . we are No. 1 in the Midwest. </s> a news conference </s> <false statement 2.610369920730591> Ted Strickland says Gov. John Kasich incorrectly claimed Ohios economy was 38th in the nation when he took office. We were sixth in the nation in terms of economic job growth. </s> an interview on CNN </s> <true statement 3.028876543045044> Terry Mcauliffe says If you take the population growth here in Virginia, we are net zero on job creation since (Bob McDonnell) became governor. </s> a speech.	true statement
Mike Pence says It was Hillary Clinton who left Americans in harms way in Benghazi and after four Americans fell said, What difference at this point does it make? </s> the Republican national convention </s> <true statement 2.5875017642974854> Hillary Clinton says When terrorists killed more than 250 Americans in Lebanon under Ronald Reagan, the Democrats didnt make that a partisan issue. </s> a CNN town hall </s> <false statement 2.9331557750701904> Facebook Posts says Hillary Clinton refuses to testify before Congress about the 2012 attack in Benghazi. </s> a meme on social media	true statement
Rand Paul says Of the roughly 15 percent of Americans who dont have health insurance, half of them made more than $50,000 a year. </s> an interview on Comedy Central's "The Daily Show" </s> <true statement 2.932455062866211> Joe Biden says Among the money spent on health care in the United States, "46 cents on every dollar spent is through Medicare and Medicaid." </s> an interview on NBC's 'Meet the Press' </s> <false statement 3.02447247505188> Trent Franks says The top 1 percent pay over half of the entire revenue for this country. </s> an interview on MSNBC's 'The Dylan Ratigan Show'	false statement
Barack Obama says Stimulus tax cuts "began showing up in paychecks of 4.8 million Indiana households about three months ago." </s> a speech in Wakarusa, Ind. </s> <false statement 2.8908281326293945> Paul Broun says Stimulus money funded a government board that made recommendations that would cost 378,000 jobs and $28.3 billion in sales. </s> a tweet </s> <true statement 2.9225375652313232> Sarah Palin says "One state even spent a million bucks to put up signs that advertise that they were spending on the federal stimulus projects." </s> an address at the Tea Party convention	false statement
Allen West says If you look at the application for a security clearance, I have a clearance that even the president of the United States cannot obtain because of my background. </s> a candidate forum </s> <false statement 3.050549268722534> Ted Cruz says One of the most troubling aspects of the Rubio-Schumer Gang of Eight bill was that it gave President Obama blanket authority to admit refugees, including Syrian refugees, without mandating any background checks whatsoever. </s> a Republican presidential debate in Las Vegas </s> <true statement 3.196129560470581> David Shuster says Said former U.S. Ambassador to Kenya Scott Gration was forced to resign two years ago because of his personal use of emails. </s> a Hillary Clinton press conference	false statement
Bernie S says We now work the longest hours of any people around the world. </s> a C-SPAN interview </s> <true statement 3.08957576751709> Jim Sensenbrenner says We have the highest corporate tax rate in the world. Its 35 percent. </s> an interview </s> <false statement 3.3488667011260986> Mitt Romney says Today there are more men and women out of work in America than there are people working in Canada. </s> a speech to the Conservative Political Action Conference	false statement
Sarah Palin says Donald Trumps conversion to pro-life beliefs are akin to Justin Biebers, who said in the past that abortion was no big deal to him. </s> an interview on CNN </s> <false statement 3.1018259525299072> Herman Cain says Said Planned Parenthoods early objective was to help kill black babies before they came into the world. </s> a talk at a conservative think tank </s> <true statement 3.1297004222869873> Greg Abbott says After Texas defunded Planned Parenthood, both the unintended pregnancy and abortion rates dropped. </s> a tweet	false statement

Figure 45: LIAR, step 5

Test Modified	Gold Label
Afscme says In labor negotiations with city employees, Milwaukee Mayor Tom Barrett demanded concessions that went beyond those mandated by Gov. Scott Walkers collective bargaining law </s> a letter to members </s> <false statement 3.131746292114258> Tom Barrett says Gov. Scott Walker said no to equal pay for equal work for women. </s> a TV ad </s> <true statement 3.1403446197509766> Portland Association Teachers says Did you know that if you accepted the Districts proposal today you would have NO pay increase for 4 years? Seven years of frozen wages = Disrespect. </s> a newsletter	true statement
Rick Scott says All Aboard Florida is a 100 percent private venture. There is no state money involved. </s> a TV interview </s> <true statement 3.0582022666931152> Charlie Crist says All Aboard Florida is receiving millions in Florida taxpayer dollars. </s> a fundraising email </s> <false statement 3.152191162109375> Corey Lewandowski says Mr. Trump is self-financing his campaign, so we dont have any donors. </s> a radio interview.	true statement
Julie Pace says The Obama administration is using as its legal justification for these airstrikes (on the Islamic State), an authorization for military force that the president himself has called for repeal of. </s> a question to White House Press Secretary Josh Earnest </s> <true statement 2.962770462036133> Martha Raddatz says The Obama administration originally wanted 10,000 troops to remain in Iraq -- not combat troops, but military advisers, special operations forces, to watch the counterterrorism effort. </s> comments on ABC's "This Week" </s> <false statement 2.9962246417999268> Rand Paul says The president is advocating a drone strike program in America. </s> a tweet	true statement
John Kasich says We are now eighth in the nation in job creation . . . we are No. 1 in the Midwest. </s> a news conference </s> <false statement 2.610369920730591> Ted Strickland says Gov. John Kasich incorrectly claimed Ohios economy was 38th in the nation when he took office. We were sixth in the nation in terms of economic job growth. </s> an interview on CNN </s> <true statement 2.896986246109009> John Kasich says We are in the bottom 10 in dollars in the classroom and the top 10 in dollars in the bureaucracy and red tape. </s> an interview on Fox News	true statement
Mike Pence says It was Hillary Clinton who left Americans in harms way in Benghazi and after four Americans fell said, What difference at this point does it make? </s> the Republican national convention </s> <true statement 2.5874826908111572> Hillary Clinton says When terrorists killed more than 250 Americans in Lebanon under Ronald Reagan, the Democrats didnt make that a partisan issue. </s> a CNN town hall </s> <false statement 2.849807024002075> Donald Trump says Sidney Blumenthal wrote that the Benghazi attack was almost certainly preventable. Clinton was in charge of the State Department, and it failed to protect U.S. personnel and an American consulate in Libya. </s> a rally in Wilkes-Barre, Pa.	true statement
Rand Paul says Of the roughly 15 percent of Americans who dont have health insurance, half of them made more than $50,000 a year. </s> an interview on Comedy Central's "The Daily Show" </s> <false statement 2.9004263877868652> Rand Paul says Over half of the young people in medical, dental and law schools are women. </s> an interview with CNN </s> <true statement 2.932455062866211> Joe Biden says Among the money spent on health care in the United States, "46 cents on every dollar spent is through Medicare and Medicaid." </s> an interview on NBC's 'Meet the Press'	false statement
Barack Obama says Stimulus tax cuts "began showing up in paychecks of 4.8 million Indiana households about three months ago." </s> a speech in Wakarusa, Ind. </s> <false statement 2.8908281326293945> Paul Broun says Stimulus money funded a government board that made recommendations that would cost 378,000 jobs and $28.3 billion in sales. </s> a tweet </s> <true statement 2.898074150085449> Chain Email says Having an entirely Democrat congressional delegation in 2009, when the [federal stimulus] bill passed, increases the per capita stimulus dollars that the state receives per person by $460. </s> a message via the Internet	false statement
Allen West says If you look at the application for a security clearance, I have a clearance that even the president of the United States cannot obtain because of my background. </s> a candidate forum </s> <false statement 3.02140736579895> Steve Southerland says 92 percent of President Barack Obamas administration has never worked outside government. </s> comments at the Liberty County Chamber of Commerce annual dinner. </s> <true statement 3.1747167110443115> John Mccain says "The fact is it's not amnesty." </s> a debate in Manchester, N.H.	false statement
Bernie S says We now work the longest hours of any people around the world. </s> a C-SPAN interview </s> <false statement 3.0254147052764893> Bernie S says We spend twice as much per capita on health care as any other nation on Earth. </s> an appearance on the Rachel Maddow Show </s> <true statement 3.08957576751709> Jim Sensenbrenner says We have the highest corporate tax rate in the world. Its 35 percent. </s> an interview	false statement
Sarah Palin says Donald Trumps conversion to pro-life beliefs are akin to Justin Biebers, who said in the past that abortion was no big deal to him. </s> an interview on CNN </s> <false statement 2.7887768745422363> Donald Trump says Public support for abortion is actually going down a little bit, polls show. </s> comments on CNN's "State of the Union" </s> <true statement 3.1297004222869873> Greg Abbott says After Texas defunded Planned Parenthood, both the unintended pregnancy and abortion rates dropped. </s> a tweet	false statement

Figure 46: LIAR, step 10

Test Modified	Gold Label
If this ugly ass niggah can be with her , I can too . Tf she with Chuu Baka for anyways this niggah look like a... http://t.co/lVNDRDPpQT </s> <hate speech 3.680654525756836> @AtBlackface @MrMooncricket @UncleUnco @BobbyBotsods @FuckTheMallcops @RacistNegro69 ape? Fuck u broke ass racist </s> <offensive language 3.9933786392211914> @Tamela_X But a wannabe ghetto white girl will deal with an unaccomplished black man for more than sex. </s> <neither 4.20903205871582> @VNNForum These Marxist kikes promoted Anti-Semitism by creating this film, what was their incentive? They just promote their enemies.	hate speech
“@iamkrause: No need to thank me, killing niggers is it's own reward.” Dis my pal from our graduation when we was 26 </s> <neither 3.5232555866241455> @LanceTHESPOKEN Statement of the Day: Ppl who RT @ChiefsatWar are even dumber than he is. The ole birds of a feather type of deal. </s> <hate speech 3.5796432495117188> @Brenddan_ they said some bullshit. Damn niggers </s> <offensive language 3.7068209648132324> BRUH lmaoo “@NoBeeetch: Them hoes was nasty but I kept eating them "@roblockk: What the fuck happened to these? http://t.co/G9w10SmQdr"”	hate speech
RT @IsaidNick: niggers are so ignorant http://t.co/P9jDdVsRVb </s> <hate speech 3.2627384662628174> @AtBlackface @MrMooncricket @UncleUnco @BobbyBotsods @FuckTheMallcops @RacistNegro69 ape? Fuck u broke ass racist </s> <neither 3.7804527282714844> @VNNForum These Marxist kikes promoted Anti-Semitism by creating this film, what was their incentive? They just promote their enemies. </s> <offensive language 3.9285027980804443> @Tamela_X But a wannabe ghetto white girl will deal with an unaccomplished black man for more than sex.	hate speech
RT @RosieZaya1: Ur fucking white trash </s> <hate speech 2.951470136642456> @AtBlackface @MrMooncricket @UncleUnco @BobbyBotsods @FuckTheMallcops @RacistNegro69 ape? Fuck u broke ass racist </s> <offensive language 3.6144936084747314> @Tamela_X But a wannabe ghetto white girl will deal with an unaccomplished black man for more than sex. </s> <neither 3.7668633460998535> @VNNForum These Marxist kikes promoted Anti-Semitism by creating this film, what was their incentive? They just promote their enemies.	hate speech
@mike_ray7 congratulations, you are officially fucking retarded. </s> <neither 3.4077796936035156> RT @JakeG_BasedGod: "Never go full retard" </s> <hate speech 3.479813575744629> @Brenddan_ they said some bullshit. Damn niggers </s> <offensive language 3.601623773574829> BRUH lmaoo “@NoBeeetch: Them hoes was nasty but I kept eating them "@roblockk: What the fuck happened to these? http://t.co/G9w10SmQdr"”	hate speech
gonna have them pussies mix up some concrete today. teach them to pose like me. I am a badass motherfucker. and I will let you be too (: </s> <hate speech 3.227602243423462> @AtBlackface @MrMooncricket @UncleUnco @BobbyBotsods @FuckTheMallcops @RacistNegro69 ape? Fuck u broke ass racist </s> <offensive language 3.5520639419555664> BRUH lmaoo “@NoBeeetch: Them hoes was nasty but I kept eating them "@roblockk: What the fuck happened to these? http://t.co/G9w10SmQdr"” </s> <neither 3.809976816177368> RT @Venus_Lynn: “@dylxnl: look ghetto but it work http://t.co/chrvW9dPca”😭😭😭😭😭😭😭😭😭😭	offensive language
@maddieevaans u call ur bestie a bitch I'm guessing she's a dog that barks too much </s> <offensive language 3.2484242916107178> RT @Im_Amy_Bitches: I'm not always a bitch, sometimes I sleep </s> <hate speech 3.3403964042663574> Women who are feminist are the ugly bitches who cant find a man for themselves </s> <neither 3.8210768699645996> Gary's girl was a tranny. RT @Em_99car: I wonder what would have become of @rickspringfield if he'd written 'Gary's Girl'.	offensive language
No less than 3 bad bitches in my bed at a time... </s> <offensive language 3.2010550498962402> Then again... i shoot bitches in the ass for doing some stupid shit like key my car or bust my windows. </s> <hate speech 3.666656732559204> bitch kill yoself, go on to the bathroom and eat the pills bitch, all of em </s> <neither 4.249817848205566> ★ BEST ASIAN MASSAGE ON THE park slope --TOP RATED SPA ★ 718-622-0221 - 24 http://t.co/ZsAAzFL0p5	offensive language
RT @TheDrugTribe: mary isn't a backstabbing bitch that lies and deceives me </s> <offensive language 3.4536943435668945> RT @Im_Amy_Bitches: I'm not always a bitch, sometimes I sleep </s> <hate speech 3.6065785884857178> @vinny2vicious faggot I knew you weren't really my friend. </s> <neither 3.638406753540039> Gary's girl was a tranny. RT @Em_99car: I wonder what would have become of @rickspringfield if he'd written 'Gary's Girl'.	offensive language
#porn,#android,#iphone,#ipad,#sex,#xxx, | #CloseUp | Squirting pussy and fingered asshole http://t.co/bKYeoUwWv2 </s> <offensive language 3.5574071407318115> BRUH lmaoo “@NoBeeetch: Them hoes was nasty but I kept eating them "@roblockk: What the fuck happened to these? http://t.co/G9w10SmQdr"” </s> <neither 3.6928675174713135> #DegenerateArtist Sniffs whiffy balls involuntary, cuz a FAIRY walks DOWNTOWN HAIRY, #climatechange all our scary sense of hillbilly sneeze> </s> <hate speech 3.7489092350006104> Wake up you stoopid dyke @RandyPawlik	offensive language
My baby guinea pig is so cute #Adorable </s> <hate speech 3.82438588142395> @whiteangelss84 y dont u end us monkeys then? U hate us because were better than crackers & I love it. Deep down u know weve done nothin 2 u </s> <offensive language 3.8302881717681885> I always wanted a bull dog them hoes clean fuck a pit </s> <neither 3.8650126457214355> Breakfast fried chicken jerk chicken Tater tots white rice nd press yellow rice nd beans Mac nd cheese http://t.co/Usz8gJnZl0	neither
RT @Kick_Man: Giants- Pitiful .. Jets-Pitiful .. Mets- Pitiful .. Yankees-Pitiful .. Nets- Pitiful .. Knicks-Pitiful ... Ny sports- Pitiful </s> <neither 3.8152754306793213> You know I'm not big on the NFL, but I'm so sick of hearing all of this "Black and yellow" shit. LOL @ bandwagon fans and hell, GO PACKERS! </s> <offensive language 4.071953773498535> BRUH lmaoo “@NoBeeetch: Them hoes was nasty but I kept eating them "@roblockk: What the fuck happened to these? http://t.co/G9w10SmQdr"” </s> <hate speech 4.095180511474609> @whiteangelss84 y dont u end us monkeys then? U hate us because were better than crackers & I love it. Deep down u know weve done nothin 2 u	neither
@jesstoth we could get matching burner phones and be ghetto fab for a few months </s> <hate speech 3.4667954444885254> @whiteangelss84 y dont u end us monkeys then? U hate us because were better than crackers & I love it. Deep down u know weve done nothin 2 u </s> <offensive language 3.595543622970581> RT @NickBratton3: I wish my parents bought me a car man.. People bitch about not getting what car they want when they want it, and its free… </s> <neither 3.6075007915496826> RT @Venus_Lynn: “@dylxnl: look ghetto but it work http://t.co/chrvW9dPca”😭😭😭😭😭😭😭😭😭😭	neither
Thw White Iron Band plays this weekend in Fargo,ND at the Aquarium(21+) ,Friday(10-29-10) with Charlie Parr. The next night,Saturday... </s> <hate speech 3.8393898010253906> 😂😂😂RT @kwagiheath: Them 1st 48 Charlotte goon niccas gon Fkkk around and burn Bank Of America stadium down🔥 </s> <offensive language 3.9061098098754883> I be telling Mcgirt music ain't enough.You gotta have a non music related agenda.Them crackers want to sell something with your face. </s> <neither 3.980140447616577> You know I'm not big on the NFL, but I'm so sick of hearing all of this "Black and yellow" shit. LOL @ bandwagon fans and hell, GO PACKERS!	neither
RT @dsrtvet: @FoxNews @tjoy7 And I don't have any confidence NONWHATSOEVER in you Barack! You're the sole reason why this country is in thi… </s> <neither 3.78818678855896> RT @lachlan: So who wants to tell the Washington Post that Charlie Crist is a Democrat? http://t.co/LGzzYusEKZ http://t.co/2jCVv7qxqf </s> <hate speech 3.942119598388672> @whiteangelss84 y dont u end us monkeys then? U hate us because were better than crackers & I love it. Deep down u know weve done nothin 2 u </s> <offensive language 3.984135150909424> BRUH lmaoo “@NoBeeetch: Them hoes was nasty but I kept eating them "@roblockk: What the fuck happened to these? http://t.co/G9w10SmQdr"”	neither

Figure 47: Hate Speech Offensive, step 1

Test Modified	Gold Label
If this ugly ass niggah can be with her , I can too . Tf she with Chuu Baka for anyways this niggah look like a... http://t.co/lVNDRDPpQT </s> <hate speech 2.7132835388183594> RT @WaavyLee: His balls ashy RT @Yattabing: @Trelaire1st: Real women do this http://t.co/VG5DBqH8aT” and real faggots let em do that.. Smh </s> <offensive language 2.7854363918304443> RT @JayyMfCarter: If you gotta nigga or bitch.... PLEASE STAY OUT MY DM's I'm tired of y'all taken girls </s> <neither 3.0996105670928955> @NeonTreezz @PoCBeauty ......so.....white people selling #African art? NOT COOL? Okay. #blackish #redskins #doublestandard	hate speech
“@iamkrause: No need to thank me, killing niggers is it's own reward.” Dis my pal from our graduation when we was 26 </s> <hate speech 2.6058554649353027> @samzbikowski some negro amigo pulled a gun on Nate and I a few weeks ago. I was STOKED!🔫🔫🔫 </s> <neither 2.670192003250122> RT @jennaferjenbub: “@BarryClerjuste: "Anything below a A+ and we disown you ling ling" http://t.co/m1QiWK4xZg” @AustinBedsaul </s> <offensive language 2.740609645843506> “@Alondra_Lu: Ain't that a bitch”	hate speech
RT @IsaidNick: niggers are so ignorant http://t.co/P9jDdVsRVb </s> <hate speech 2.057800769805908> RT @WhitesOnly_1: #niggers! http://t.co/Hb3uJaLky2 </s> <neither 2.7749483585357666> & thots are wearing Uggs RT @BigBootyJudy814: #ItsFallBecause negros are pulling out their Timbs" </s> <offensive language 2.926729440689087> RT @Jayy_Gee96: Dumb bitches	hate speech
RT @RosieZaya1: Ur fucking white trash </s> <hate speech 2.422173500061035> @FrankieJGrande fugly queer white trash </s> <offensive language 2.6756434440612793> RT @Jayy_Gee96: Dumb bitches </s> <neither 2.783188819885254> RT @BeardedNixon: Poont gotta be trash	hate speech
@mike_ray7 congratulations, you are officially fucking retarded. </s> <hate speech 2.4854748249053955> @darthdanaa Yes you do retard. </s> <offensive language 2.851564645767212> Lol!! “@ItzSweetz_Bitch: Ooop! QT @TiFFANY_PORSCHE: You little twats.” </s> <neither 2.8971688747406006> RT @jennaferjenbub: “@BarryClerjuste: "Anything below a A+ and we disown you ling ling" http://t.co/m1QiWK4xZg” @AustinBedsaul	hate speech
gonna have them pussies mix up some concrete today. teach them to pose like me. I am a badass motherfucker. and I will let you be too (: </s> <offensive language 2.7589027881622314> @40oz_VAN IYCMI. I can't get any work done if you keep showin off your bitches. </s> <hate speech 2.8690829277038574> #SlightlyAdjusted RT @CapoToHeaven Alls niggers wanna do is fuck, tweet, and drink pineapple soda all day </s> <neither 3.0193798542022705> @cakedjake We're laying rock around our lake. You're welcome to join a redneck workout. #muscles 💪😁	offensive language
@maddieevaans u call ur bestie a bitch I'm guessing she's a dog that barks too much </s> <offensive language 3.0170230865478516> Lol!! “@ItzSweetz_Bitch: Ooop! QT @TiFFANY_PORSCHE: You little twats.” </s> <hate speech 3.0850884914398193> @Princesslexii16 Fucking coon </s> <neither 3.21132493019104> lmaoooo RT @ComedyTruth: Girls, don't let a guy treat you like a yellow starburst. You are a pink starburst.	offensive language
No less than 3 bad bitches in my bed at a time... </s> <offensive language 3.023017406463623> Don't lose sleep bout these bitches bc they come and go 👌 </s> <hate speech 3.2786214351654053> bitch kill yoself, go on to the bathroom and eat the pills bitch, all of em </s> <neither 3.4171059131622314> Keep those away frm Charlie Day RT @JhonenV: Just once in my life I'd like for someone's favorite part of my body to be my disgusting knees.	offensive language
RT @TheDrugTribe: mary isn't a backstabbing bitch that lies and deceives me </s> <offensive language 2.991457939147949> RT @StevStiffler: If her bio says "Only God can judge me" she's a hoe. </s> <hate speech 3.098494529724121> RT @sorryimalex: I got called a faggot for buying girl toms so now I'm gonna fuck that person in the ass </s> <neither 3.3266849517822266> lmaoooo RT @ComedyTruth: Girls, don't let a guy treat you like a yellow starburst. You are a pink starburst.	offensive language
#porn,#android,#iphone,#ipad,#sex,#xxx, | #CloseUp | Squirting pussy and fingered asshole http://t.co/bKYeoUwWv2 </s> <neither 1.5677733421325684> #porn,#android,#iphone,#ipad,#sex,#xxx, | #Desi | paki http://t.co/XxcdQvzI9t </s> <hate speech 2.8566393852233887> RT @mitchmancuso: @BrantPrintup:What straight guys take a picture of themselves naked in a hot tun.What fags @jakesiwy @Ryan_Murphy3 @Randy… </s> <offensive language 2.932191848754883> Lol!! “@ItzSweetz_Bitch: Ooop! QT @TiFFANY_PORSCHE: You little twats.”	offensive language
My baby guinea pig is so cute #Adorable </s> <neither 3.1643435955047607> Our female guinea pig is pregnant 🎊🎉🎁😍😍😍 </s> <offensive language 3.4907007217407227> I impress da young white girl next doe by taking out my giant negro thang and usin it to flip da hamburgers for da #KoolQueefTribute </s> <hate speech 3.5861661434173584> What a wetback looks like when he gets caught crossing the border. @Ilovebamf http://t.co/j3Uf1TYubO	neither
RT @Kick_Man: Giants- Pitiful .. Jets-Pitiful .. Mets- Pitiful .. Yankees-Pitiful .. Nets- Pitiful .. Knicks-Pitiful ... Ny sports- Pitiful </s> <hate speech 3.3251242637634277> RT @__J_R: Smh nigga is mildly retarded RT @Thotcho: LMFAO RT @JustDoJ: If Griff wasn’t injuries we’d legit be 6-1 </s> <neither 3.328580856323242> Don't follow the astros they said. They're trash they said. Well now look at them #astros </s> <offensive language 3.3433356285095215> Them shits ugly hoe. RT @SirRocObama: .... RT @BurgerKing: All these nuggets & u still actin chicken. http://t.co/tRy8Lvyo9O	neither
@jesstoth we could get matching burner phones and be ghetto fab for a few months </s> <offensive language 3.122525930404663> @JZolly23 @JBilinovich we need to grow mullets together so we can get all the bitches and @HannahKubiak can hate on us </s> <hate speech 3.291858434677124> RT @NoWomanIsRight: You can be a good girl all you want and those hoes still gonna get us niggas attention from time to time </s> <neither 3.3428289890289307> RT @Venus_Lynn: “@dylxnl: look ghetto but it work http://t.co/chrvW9dPca”😭😭😭😭😭😭😭😭😭😭	neither
Thw White Iron Band plays this weekend in Fargo,ND at the Aquarium(21+) ,Friday(10-29-10) with Charlie Parr. The next night,Saturday... </s> <neither 3.4018728733062744> Lmaooo naw man RT @DipOnline Yo want in RT @HumbltonBanks: U serious bro?? lol RT @CheezMoeJenkinz 2-3:10am early bird special </s> <hate speech 3.544551372528076> 😂😂😂RT @kwagiheath: Them 1st 48 Charlotte goon niccas gon Fkkk around and burn Bank Of America stadium down🔥 </s> <offensive language 3.711003065109253> I be telling Mcgirt music ain't enough.You gotta have a non music related agenda.Them crackers want to sell something with your face.	neither
RT @dsrtvet: @FoxNews @tjoy7 And I don't have any confidence NONWHATSOEVER in you Barack! You're the sole reason why this country is in thi… </s> <hate speech 2.696760654449463> RT @veeveeveeveevee: If I was Obama Id call a press conference & slit joe bidens neck on live tv just 2 show these crackers I mean business… </s> <neither 2.762817144393921> RT @jennaferjenbub: “@BarryClerjuste: "Anything below a A+ and we disown you ling ling" http://t.co/m1QiWK4xZg” @AustinBedsaul </s> <offensive language 2.9660327434539795> RT @CoffyBrownChi: If he don't believe you, no refunds hoe.	neither

Figure 48: Hate Speech Offensive, step 5

Test Modified	Gold Label
If this ugly ass niggah can be with her , I can too . Tf she with Chuu Baka for anyways this niggah look like a... http://t.co/lVNDRDPpQT </s> <offensive language 2.6535706520080566> RT @CurrenSy_Spitta: And if a bitch can't respect a nigga wit some paper and a fresh pair of bball shorts then she was raised terribly.. </s> <hate speech 2.7132835388183594> RT @WaavyLee: His balls ashy RT @Yattabing: @Trelaire1st: Real women do this http://t.co/VG5DBqH8aT” and real faggots let em do that.. Smh </s> <neither 3.0996105670928955> @NeonTreezz @PoCBeauty ......so.....white people selling #African art? NOT COOL? Okay. #blackish #redskins #doublestandard	hate speech
“@iamkrause: No need to thank me, killing niggers is it's own reward.” Dis my pal from our graduation when we was 26 </s> <hate speech 2.545886278152466> RT @Tae_Rhodes: “@kim92493: “@Tae_Rhodes: @kim92493 @patpatbush uhhh you've been judged” it happens. #whitepower...I'll hang you nigger” wo… </s> <offensive language 2.6044790744781494> “@NoRapist: on my way to fuck ur bitch http://t.co/SgVBBrwOg2” @mckinley719 </s> <neither 2.670259714126587> RT @jennaferjenbub: “@BarryClerjuste: "Anything below a A+ and we disown you ling ling" http://t.co/m1QiWK4xZg” @AustinBedsaul	hate speech
RT @IsaidNick: niggers are so ignorant http://t.co/P9jDdVsRVb </s> <hate speech 2.057800769805908> RT @WhitesOnly_1: #niggers! http://t.co/Hb3uJaLky2 </s> <neither 2.7749483585357666> & thots are wearing Uggs RT @BigBootyJudy814: #ItsFallBecause negros are pulling out their Timbs" </s> <offensive language 2.8298287391662598> This Uncle Tom mother fucking wants to invoke 3/5 a man in his speech? Dude you ain't white no matter how much... http://t.co/3yrcyC9ezc	hate speech
RT @RosieZaya1: Ur fucking white trash </s> <hate speech 2.422173500061035> @FrankieJGrande fugly queer white trash </s> <offensive language 2.6756434440612793> RT @Jayy_Gee96: Dumb bitches </s> <neither 2.783188819885254> RT @BeardedNixon: Poont gotta be trash	hate speech
@mike_ray7 congratulations, you are officially fucking retarded. </s> <hate speech 2.4854748249053955> @darthdanaa Yes you do retard. </s> <offensive language 2.8516175746917725> Lol!! “@ItzSweetz_Bitch: Ooop! QT @TiFFANY_PORSCHE: You little twats.” </s> <neither 2.8972203731536865> RT @jennaferjenbub: “@BarryClerjuste: "Anything below a A+ and we disown you ling ling" http://t.co/m1QiWK4xZg” @AustinBedsaul	hate speech
gonna have them pussies mix up some concrete today. teach them to pose like me. I am a badass motherfucker. and I will let you be too (: </s> <offensive language 2.758687734603882> @40oz_VAN IYCMI. I can't get any work done if you keep showin off your bitches. </s> <hate speech 2.8036365509033203> Just to get u mad go on your search bar on here and search up "stupid niggers" & hop on somebodys head then mention me lol @stonethegreat23 </s> <neither 3.012741804122925> @charloosss @keepitplur @nicoleariel_ I'll chug my tall can . but homegirl won't approve lol	offensive language
@maddieevaans u call ur bestie a bitch I'm guessing she's a dog that barks too much </s> <hate speech 2.8469488620758057> @RylannWilliams whooooo? Chelsey? Fuck her lol. She juss a bitch </s> <offensive language 2.8842358589172363> RT @_Ezzzylove: She a bad bitch, let's get to it right away . </s> <neither 3.0819990634918213> @charliesheen Charlie, im an old lady. don't EVER SAY UGLY THINGS ABOUT UR CHILDRENS MOM.. I GET IT!!!, JUS DONT! BIG HUG	offensive language
No less than 3 bad bitches in my bed at a time... </s> <offensive language 2.8522520065307617> Bad bitches in the pen make my toes curl </s> <hate speech 3.2539432048797607> I didn't forsake all other bitches for my wife to be getting fucked on by another nigga. and you know she married? you gotta die. </s> <neither 3.4170782566070557> Keep those away frm Charlie Day RT @JhonenV: Just once in my life I'd like for someone's favorite part of my body to be my disgusting knees.	offensive language
RT @TheDrugTribe: mary isn't a backstabbing bitch that lies and deceives me </s> <offensive language 2.9916186332702637> RT @StevStiffler: If her bio says "Only God can judge me" she's a hoe. </s> <hate speech 3.02489972114563> @triple6em96 @Hunglikerobby_ bitch you watch your fucking mouth you dirty whore. I swear to god that's a thin line </s> <neither 3.1058743000030518> RT @shakiraevanss: Criticize Amanda for saying the n word, sure, but don't make jokes about her sexual assault, don't be trash.	offensive language
#porn,#android,#iphone,#ipad,#sex,#xxx, | #CloseUp | Squirting pussy and fingered asshole http://t.co/bKYeoUwWv2 </s> <neither 1.5677733421325684> #porn,#android,#iphone,#ipad,#sex,#xxx, | #Desi | paki http://t.co/XxcdQvzI9t </s> <offensive language 2.8408925533294678> RT @FunnyPicsDepot: bitches be like "I'm a virgin" http://t.co/mFDwXmg8ic </s> <hate speech 2.8566393852233887> RT @mitchmancuso: @BrantPrintup:What straight guys take a picture of themselves naked in a hot tun.What fags @jakesiwy @Ryan_Murphy3 @Randy…	offensive language
My baby guinea pig is so cute #Adorable </s> <neither 3.1643435955047607> Our female guinea pig is pregnant 🎊🎉🎁😍😍😍 </s> <offensive language 3.4907007217407227> I impress da young white girl next doe by taking out my giant negro thang and usin it to flip da hamburgers for da #KoolQueefTribute </s> <hate speech 3.5861661434173584> What a wetback looks like when he gets caught crossing the border. @Ilovebamf http://t.co/j3Uf1TYubO	neither
RT @Kick_Man: Giants- Pitiful .. Jets-Pitiful .. Mets- Pitiful .. Yankees-Pitiful .. Nets- Pitiful .. Knicks-Pitiful ... Ny sports- Pitiful </s> <neither 3.0366640090942383> @Buster_ESPN Huh.....last 10 games..Tampa 8-2/Balt 7-3/Yanks 6-4...and they lost their best pitcher. Please explain your logic. </s> <hate speech 3.3251242637634277> RT @__J_R: Smh nigga is mildly retarded RT @Thotcho: LMFAO RT @JustDoJ: If Griff wasn’t injuries we’d legit be 6-1 </s> <offensive language 3.3433356285095215> Them shits ugly hoe. RT @SirRocObama: .... RT @BurgerKing: All these nuggets & u still actin chicken. http://t.co/tRy8Lvyo9O	neither
@jesstoth we could get matching burner phones and be ghetto fab for a few months </s> <hate speech 3.034785270690918> @SAMMI_boyden bruh we can finally roll like rednecks (': ((drug dealers)) </s> <offensive language 3.122525930404663> @JZolly23 @JBilinovich we need to grow mullets together so we can get all the bitches and @HannahKubiak can hate on us </s> <neither 3.271000623703003> RT @sassytbh: a girl tweeted "you might be ghetto if u bring food from outside into the movies" no u might be stupid if u pay 4.99 for a b…	neither
Thw White Iron Band plays this weekend in Fargo,ND at the Aquarium(21+) ,Friday(10-29-10) with Charlie Parr. The next night,Saturday... </s> <neither 3.2462401390075684> RT @toddknife: Full @weakenednachos set (except the last song) from Southern Darkness Fest last month. Who's the ape on guitar? https://t.c… </s> <hate speech 3.3524651527404785> Eagles fuck around & lose it'll be kill the cracker at the Sophi crib smfh </s> <offensive language 3.511016368865967> My dawg @ceomiamimike told me it's a must I be @901k2lounge this Saturday ROCKIN that bitch wit Tha… http://t.co/0NV9cHtwOs	neither
RT @dsrtvet: @FoxNews @tjoy7 And I don't have any confidence NONWHATSOEVER in you Barack! You're the sole reason why this country is in thi… </s> <hate speech 2.696760654449463> RT @veeveeveeveevee: If I was Obama Id call a press conference & slit joe bidens neck on live tv just 2 show these crackers I mean business… </s> <neither 2.762908458709717> RT @jennaferjenbub: “@BarryClerjuste: "Anything below a A+ and we disown you ling ling" http://t.co/m1QiWK4xZg” @AustinBedsaul </s> <offensive language 2.881894588470459> “@LongMoneyTonyy: @vintage_monroe_ DONT Say Shit Else ! Just Stfu !”bitch we can do a lot more off this Twitter shit you can come see me	neither

Figure 49: Hate Speech Offensive, step 10.