Let the CAT out of the bag: Contrastive Attributed explanations for Text

Saneem A. Chemmengath[1*] **Amar Prakash Azad**[2] **Ronny Luss**[2] **Amit Dhurandhar**[2]
[1]Microsoft [2]IBM Research
schemmengath@microsoft.com, amarazad@in.ibm.com, {rluss, adhuran}@us.ibm.com

Abstract

Contrastive explanations for understanding the behavior of black box models has gained a lot of attention recently as they provide potential for recourse. In this paper, we propose a method **C**ontrastive **A**ttributed explanations for **T**ext (CAT) which provides contrastive explanations for natural language text data with a novel twist as we build and exploit attribute classifiers leading to more semantically meaningful explanations. To ensure that our contrastive generated text has the fewest possible edits with respect to the original text, while also being fluent and close to a human generated contrastive, we resort to a minimal perturbation approach regularized using a BERT language model and attribute classifiers trained on available attributes. We show through qualitative examples and a user study that our method not only conveys more insight because of these attributes, but also leads to better quality (contrastive) text. Quantitatively, we show that our method outperforms other state-of-the-art methods across four data sets on four benchmark metrics.

1 Introduction

Explainable AI (XAI) has seen an explosion of interest over the last five years, not just in research (Molnar, 2019; Arya et al., 2019), but also in the real world where governments (Yannella and Kagan, 2018; Gunning, 2017) and industry have made sizeable investments. The primary driver for this level of interest has been the inculcation of deep learning technologies (Goodfellow et al., 2016), which are inherently black box, into decision making systems that affect billions of people. Trust thus has become a central theme in relying on these black box systems, and one way to achieve it is seemingly through obtaining explanations.

Although many feature-based (Ribeiro et al., 2016; Lundberg and Lee, 2017; Simonyan et al., 2013) and exemplar-based methods (Gurumoorthy et al., 2019; Koh and Liang, 2017; Kim et al., 2016) have been proposed to explain local instance level decisions of black box models, contrastive/-counterfactual explanations have seen a surge of interest recently (Wachter et al., 2017; Dhurandhar et al., 2018; Madaan et al., 2021; Luss et al., 2021; Ross et al., 2021). One reason for this is that contrastive explanations are viewed as one of the main tools to achieve recourse (Karimi et al., 2021). For example, companies commonly use chatbots to communicate with customers, which starts by passing customer text through a classifier that decides which support department should handle the customer. A common problem is locating bias in these classifier models since "the chatbot will continue to show the behavior" due to biased knowledge bases(Brown, 2021); recourse in terms of removing bias could be achieved by identifying examples that drive the bias.

Given this surge of interest and its importance in recourse, in this paper, we propose a novel method **C**ontrastive **A**ttributed explanations for **T**ext (CAT) which provides contrastive explanations for natural language data, a modality that has received comparatively less attention when it comes to these type of explanations. We show that our method produces fluent contrasts and possesses an additional novel twist not seen in prior works. As such, our method also outputs a minimal set of semantically meaningful attributes that it thinks led to the final contrast. These attributes could be subtopics in a dataset, different from the class labels, that characterize a piece of text or the attributes could even be obtained from a different dataset. Our approach is to leverage these attributes by building models (viz. classifiers) for them and then using these classifiers to guide the search for contrasts in the original task. Regarding the motivating biased chatbot example above, learning attributes that lead to contrasts can be particularly useful. Gender bias, for instance, is

[*]Work done while at IBM Research.

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 7190–7206
December 7-11, 2022

Table 1: Contrastive explanations are shown for state-of-the-art methods GYC and MICE along with our method CAT to explain predictions on two sentences from the AGNews dataset. Red highlighting indicates text that has changed in the contrastive explanation. Both inputs were classified as Sci-Tech while all contrastive sentences were classified as Business. In addition to these changes, CAT outputs the attributes (or subtopics) that were added/removed from the input to create the contrast. In the first example, “tax" could correspond to “Entertainment" or “Politics" which are attributes added, while “file" corresponds to software files that are often encrypted and hence to “Cryptography" (denoted Crypt.) which is removed. In the second example it is easy to see that the added word “Healthcare" relates to "Medicine" which is added while the removed word “Search" is related to "Windows" (denoted Wndws) and "Cryptography" which are removed attributes.

Input	**GYC**	**MICE**	**CAT**	
			Contrast	Attributes
Movie Studios to sue illegal film file traders	Movie Studios to sue juveniles psychiatrically	Movie Studios to sue illegal film - Investors	Movie Studios to sue illegal film tax traders	+Entertainment, +Politics, −Crypt.
Search providers seek video find challenges	Search providers seek video find videos,	seek opportunities find challenges	Healthcare providers seek to find challenges	+Medicine, −Wndws, −Crypt.

hard to locate because contrasts can modify gender on individual chats in different ways (removing pronouns, replacing gendered words with neutral, etc.). Such bias can be more easily identified if the different modifications all affect a (latent) attribute that is gendered such as motherhood, which our CAT explanations should highlight.

To better understand what CAT offers, consider the examples provided in Table 1. Here we show two example sentences from the AG News dataset (Zhang et al., 2015) which were classified as Sci-Tech by our neural network black box (BB) model (details in Section 4). Each explanation method generates a contrast in the class Business where the other choices were World and Sports. As can be seen, our method CAT produces closer contrasts than two recent methods: Generate Your Counterfactuals (GYC) (Madaan et al., 2021) and Minimal Contrastive Editing (MICE) (Ross et al., 2021).

A key novelty is that CAT provides additional information in terms of characteristic attributes it thinks the contrast sentence belongs to (indicated by + sign), while also indicating characteristics it thinks are no longer applicable (indicated by − sign). Our method picks a few relevant attributes that guide generation of the contrast; the attributes themselves provide additional insight into the functioning of the black box model. This is confirmed through two separate user studies conducted with a total of 75 participants for which the results are reported in section 4.4. Users found it easier to predict the class of the input sentence given our explanation over GYC, MICE, MICE with no fine tuning (MICE-nft), and an ablation of our method, called CAT with no attributes (CAT-na), and moreover, users qualitatively preferred our method in terms of understandability, sufficiency, satisfiability and completeness on a five point Likert scale.

It is important to note that there are various ways to generate text using different language models (GPT-2, BERT, T5, etc.) and even different techniques on how infilling might be performed (forward or bi-directional). The key idea of guiding the generation of contrasts through attributes can be considered for other recent methods, whether the contrast is learned through differentiable optimization (Madaan et al., 2021) or through combinatorial search (Ross et al., 2021).

We note that author provided implementations of GYC and MICE require models that use specific text embeddings and are not easily adaptable to other embeddings. Our method CAT, however, is easily adaptable, and we thus compare against the methods GYC and MICE using the embeddings they are respectively implemented for in order to get stringent and fair comparisons for CAT with previous state-of-the-art methods. This means we must compare against GYC and MICE on different text classification models, and hence require two separate user studies, one comparing CAT with GYC and the other comparing CAT with MICE (and we suspect this is why no comparisons are found in the literature).

As such, our contributions are as follows: 1) CAT introduces the idea of using attributes to drive the generation of text contrastives. This contribution is both conceptual as it brings new insight to the user as well as methodological as it leads to user-preferred contrasts as seen in the user study. 2) The CAT implementation is easily adaptable to classifiers with different embeddings. 3) We qualitatively evaluate CAT through examples on four different datasets from different domain tasks. 4) We quantitatively evaluate CAT over other methods in terms

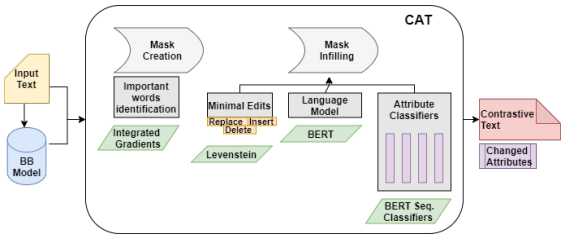


Figure 1: Above we see how CAT creates contrastive explanations. It first finds important words in the text which it then (minimally) alters by either replacing/deleting a subset of them or by inserting new ones in the spaces around them so as to i) change the black box model's prediction, ii) maintain fluency and iii) change a minimal number of attribute classifier predictions (by a significant amount). The output is the contrastive text along with a list of the attributes added/removed relative to the input text.

of flip rate, content preservation, fluency, Levenstein distance and efficiency. 5) We demonstrate the utility of CAT through two user studies that ask users to determine a model's prediction on an instance given an explanation; CAT is compared with GYC, MICE, MICE with no fine tuning, and an ablation of our method CAT-na.

2 Related Literature

Regarding the explanations of machine learning predictions on text data, a recent survey (Danilevsky et al., 2020) considered 50 recent explainability works for natural language processing, and moreover only methods that "justify predictions" rather than understanding "a model's behavior in general". Our intention is also to explain individual predictions. Little work has been done for global explainability with respect to text classification; (Ribeiro et al., 2016) suggests using various "representative" local predictions to get a global understanding of model behavior. The vast majority of explainability techniques found by (Danilevsky et al., 2020) fall under local explainability.

Local methods can be divided among post-hoc methods that explain a fixed model's prediction and self-explainable methods where the model is itself understandable; our focus is on the former. One large group of explainability methods are feature based where the explanation outputs some form of feature importance (i.e., ranking, positive/negative relevance, contributions, etc.) of the words in text (Wallace et al., 2018; Papernot and Patrick, 2018; Feng et al., 2018; Harbecke et al., 2018; Ribeiro et al., 2016; Alvarez-Melis and Jaakkola, 2017). Other types of local post-hoc explanations include exemplar based (Gurumoorthy et al., 2019; Koh and Liang, 2017; Kim et al., 2016) that output similar instances to the input.

Amongst local methods our focus is on contrastive/counterfactual methods (Dhurandhar et al., 2018; Madaan et al., 2021; Luss et al., 2021) that modify the input such that the class changes and explains the prediction as "If the sample were modified by X, the prediction would have been Y instead," where X is a change to the input and Y is a new class. Such explanations are complementary to the other methods discussed above. They offer different types of intuition and thus should be used in conjunction with rather than instead of counterfactuals. (Luss et al., 2021) learn contrasts using latent attributes, but specifically for color images. Their approach does not readily translate to text as humans do not perceive minor unrealistic parts of a generated image, whereas any nuance in text is easily noticed. Thus, generating (fluent) contrasts for text models is inherently much more challenging.

The most relevant comparisons to our work are GYC (Madaan et al., 2021) and MICE (Ross et al., 2021). GYC builds on the text generation of (Dathathri et al., 2020) by learning perturbations to the history matrix used for language modeling that are also trained to reconstruct input from the generated counterfactuals. A diversity regularization term ensures that this does not result in counterfactuals that are identical to the input. A more recent work MICE masks and replaces important words selected by (Simonyan et al., 2013) in order to flip the prediction. MICE requires fine-tuning their language model to each dataset, which is a significant overhead especially given the fact that we are simply generating local explanations versus our CAT framework. Not to mention in many real applications (sufficient) data may not be available to fine tune explanations (Dhurandhar et al., 2019). Both GYC and MICE works are targeted, i.e., the user must decide what class the counterfactual should be in as opposed to CAT which automatically decides the contrast class.

Other recent methods are POLYJUICE (Wu et al., 2021) and a contrastive latent space method (Jacovi et al., 2021). The former is a human-in-the-loop method requiring supervision about the type of modification to be performed to the text such as negation, word replacement, insertion, deletion, and is not catered towards explaining a specific

classifier by automatically finding the appropriate edits. The latter does not generate contrastive text but rather highlights (multiple) words in the input text that are most likely to alter the prediction if changed, where again the target class has to be provided. Further, (Jacovi et al., 2021) assume access to encodings from the second-to-last layer of the model being explained, and is thus not a black box method like CAT. Our focus being automated contrastive explanation generation, we compare with GYC and MICE, i.e., methods designed towards explaining a classifier, where a valid contrast is also generated. The value of CAT versus GYC and MICE comes from the output of (hidden) subtopics that are added/removed from the original text to create the contrast, giving important intuition that is missing from these other methods. This is confirmed through two user studies we conduct and qualitative examples we provide. The attribute classifiers built to predict these subtopics also aid in creating better contrasts. Moreover, as will be evident, such attribute classifiers can be used across datasets, thus precluding the need for each dataset to contain such attributes. Furthermore, topic models or autoencoders could be used to divulge such attributes.

3 Proposed Approach

We now describe our **C**ontrastive **A**ttributed explanations for **T**ext (**CAT**) method. Contrastive explanations, convey why the model classified a certain input instance to a class p, and not another class q. This is achieved by creating contrastive examples (also called contrasts) from input instance which get predicted as q. Contrastive examples are created by minimally perturbing the input such that the model prediction changes. In the case of text data, perturbations can be of three types: (1) inserting a new word, (2) replacing a word with another, and (3) deleting a word. In addition to keeping the number of such perturbations small, constrastive explainers also try to maintain grammatical correctness and fluency of the contrasts (Madaan et al., 2021; Ross et al., 2021).

As an example, take the case of a black box model trained on the AG News dataset that predicts which category a certain news headline falls under. Given a headline, "*Many technologies may be a waste of time and money, researcher says*" which is predicted as Sci-Tech, a contrastive explainer will try to explain why this headline wasn't predicted as, say, *Business* by generating a contrastive example,"*Many ~~technologies~~ jobs may be a waste of time and money, researcher says*" which is predicted as Business. Observe that a single word replacement achieves a prediction change. Such contrastive explanations can help users test the robustness of black box classification models.

We observed that even with constraints for minimal perturbation and fluency on a given black box model and an instance, there are multiple contrastive examples to choose from and, very often, many are less informative than others. For example, another possible contrast is, "*Many technologies may be a waste of ~~time~~ investment and money , researcher says"* which also gets predicted as Business. However, this particular explanation is not as intuitive as the previous one as "money" is a form of "investment" and the nature of the sentence has not changed in an obvious sense with the word "technologies" still present in the sentence.

To alleviate this problem, we propose to construct and use a set of *attribute classifiers*, where the attributes could be tags/subtopics relevant to the classification task obtained from the same or a related dataset used to build the original classifier. Attribute classifiers indicate the presence/absence of a certain subtopic in the text and confidence scores from these classifiers could be used as a regularization to create a contrast. We thus prefer contrasts which change attribute scores measurably as opposed to those contrasts which do not. However, at the same time, we want a minimal number of attribute scores to change so as to have crisp explanations. Hence, our regularization not only creates more intuitive contrasts, but also provides additional information to the user in terms of changed subtopics which, as confirmed through our user study in Section 4.4, provide better understanding of the model behavior. The important steps in our method are depicted in Figure 1.

Formally, given an input text $x \in \mathcal{X}$, and a text classification model $f(\cdot)$ which predicts $y = f(x) \in \mathcal{Y}$, we aim to create a perturbed instance x' such that the predictions $f(x) \neq f(x')$ and x' is "minimally" different from x. We use a set of m attribute classifiers $\zeta_i : \mathcal{X} \rightarrow \mathbb{R}, \forall i \in \{1, \ldots, m\}$, which produce scores indicative of presence (higher scores) or absence (lower scores) of corresponding attributes in the text. We say that attribute i is added to the perturbed sentence if $\zeta_i(x') - \zeta_i(x) > \tau$ and removed when $\zeta_i(x') - \zeta_i(x) < -\tau$, for a fixed

$\tau > 0$. Word-level Levenshtein distance between original and perturbed instance $d_{Lev}(x', x)$, which is the minimum number of deletions, substitutions, or insertions required to transform x to x' is used to keep the perturbed instance close to the original. The naturalness (fluency) of a generated sentence x' is quantified by the likelihood of sentence x' as measured by the language model used for generation; we denote this likelihood by $p_{\text{LM}}(x')$. For a predicate ϕ, we denote $\mathbb{1}_\phi$ the indicator of ϕ, which takes the value 1 if ϕ is true and 0 otherwise. Given this setup, we propose to find contrastive examples by solving the following optimization problem:

$$\begin{aligned}\max_{x' \in \mathcal{X}} \; & ||\zeta(x') - \zeta(x)||_\infty - \beta \sum_i \mathbb{1}_{|\zeta_i(x') - \zeta_i(x)| > \tau} \\ & + \lambda \cdot \max_{j \in \mathcal{Y} \setminus y} \{ \, [f(x')]_j - [f(x)]_y \, \} \\ & + \eta \cdot p_{\text{LM}}(x') \; - \; \nu \cdot d_{Lev}(x', x), \end{aligned} \tag{1}$$

where $\zeta(x)$ is a vector such that $[\zeta(x)]_i = \zeta_i(x)$, $\beta, \lambda, \eta, \nu > 0$ are hyperparameters that trade-off different aspects, and $|| \cdot ||_\infty$ is the l_∞ norm. The first term in the objective function encourages to pick an x' where at least one attribute is either added/removed from x. The second term minimizes the number of such attributes for ease of interpretation. The third term is the contrastive score, which encourages the perturbed instance to be predicted different than the original instance. Fourth and fifth terms ensure that the contrast is fluent and close to the original instance, respectively.

The above objective function defines a controlled natural language generation problem. Earlier methods for controlled generation that shift the latent representation of a language model (such as GPT-2) (Madaan et al., 2021; Dathathri et al., 2020) have resulted in generated sentences being very different from the original sentence. We thus adopt a different strategy where we first take the original sentence and identify locations where substitutions/insertions need to be made using available feature attribution methods such as Integrated Gradients (Sundararajan et al., 2017). These words are ordered by their attribution and greedily replaced with a `[MASK]` token. An MLM pre-trained BERT model (Vaswani et al., 2017; Devlin et al., 2018) is then used to fill these masks. We take the top k such replacements ranked by BERT likelihood. For insertions, a mask token is inserted to the right and left of important words in order to generate a set of perturbations similar to the input example. The attribute classifiers are applied to each generated candidate contrast, and the best x' is selected as evaluated by Eq. 1. For m token perturbations, the above process is repeated m times, where at each round, the top k perturbed texts are ranked and selected according to Eq. 1, and the above perturbation process is applied to all selected perturbed texts from the previous round. Note that we perform the hyperparameter tuning for Eq. 1 only once per dataset. Details on hyperparameter tuning and optimizing Eq. 1 are in Appendix A.

Regarding generalizability of our approach, as already noted, the attribute classifiers can be derived from other sources of data and are not necessarily dependent on the data and model being explained. Furthermore, other methods could be used to obtain attributes; unsupervised methods such as LDA, VAEs, GANs could be leveraged to ascertain semantically meaningful attributes. The attribute classifiers that appear in the loss function of Eq. 1 could be replaced by disentangled representations learned by VAEs (Kumar et al., 2018) or by topic models. Hence, CAT is generalizable beyond annotated datasets.

4 Experimental Study

4.1 Setup Details

We use an MLM pre-trained BERT[1] model from Huggingface (Wolf et al., 2019) to generate text perturbations. For attributes, classes from the Huffpost News-Category (Misra, 2018) and 20 Newsgroups (Newsgroup, 2008) datasets were used. The Huffpost dataset has 200K news headlines split into 41 classes. We merged similar classes and removed those which weren't a standard topic; 22 classes remained. The 20 Newsgroups dataset has 18000 newsgroup posts with 20 topic classes. Together, we obtained 42 attributes. For 22 classes from Huffpost, we trained 22 1-vs-all binary classifiers with a distilbert (Sanh et al., 2019) base, so that the same sentence can have multiple classes. For 20 Newsgroups, we trained multiclass classifiers on the other 20 classes. More details on attribute classifiers are provided in Appendix B. Note that attribute classifiers are transferable as they need not depend on the dataset and model being explained.

We evaluate our explanation method on models trained on AgNews (Zhang et al., 2015), DBPedia (Lehmann et al., 2015), Yelp (Shen et al., 2017),

[1] https://huggingface.co/bert-base-uncased

and NLI (Bowman et al., 2015). For an apples-to-apples comparison of our methods with GYC on AgNews, DBPedia and Yelp, we trained models with the same architecture as the ones in their work: an Embedding Bag layer followed by a linear layer. For MICE the Roberta based model was used for all datasets as that is what the publicly provided implementation naturally applies to. MICE uses a two-step framework to generate conterfactual explanations, with the generator being T5 (Raffel et al., 2019) fine tuned on the task-specific dataset. More details on model training are provided in Appendix C and on datasets in Appendix D.

4.2 Qualitative Evaluations

We now provide qualitative examples from two datasets, AgNews and NLI, with additional examples for Yelp, and DBpedia in the Appendix G.

AgNews. The dataset is from a real-world news domain which contains short news headlines and bodies from four news categories - world, business, sports, and sci-tech. Our experiments focus on explanations for predicting the class from headlines.

Table 2 (top) shows results of applying CAT to five headlines in the AgNews dataset. The first row explains that the headline is predicted as sci-tech because if the headline was more related to topics such as things for sale, baseball, hockey, and less about computer-related topics, it would have been predicted sports, which is achieved in the contrast by replacing “File" with “Salary". It is important to consider the interaction of words; here, the black box model considers Kazaa a sports team because of the change. The second and third rows offer intuitive examples as to how the attribute changes relate to the contrasts. The insight is that we learn what topics the black box model finds most relevant to the prediction, as opposed to only knowing the single word and needing to figure out why that caused the change. In the fourth row, adding politics and removing money leads to changing the input from business to sci-tech as “factory growth" has more relationship to money while “population growth" is related to politics. The fifth row shows the opposite as adding money and removing politics changes the input from world to business. This last example illustrates that, for longer text, single perturbations are often insufficient to flip the label and multiple changes are needed. These last two examples offer the insight that the classifier associates business with politics, which is not obvious a priori.

NLI. The Natural Language Inference (Bowman et al., 2015) dataset contains samples of two short ordered texts, and the labels are either contradiction if the second text contradicts the first, neutral if the two texts do not contradict or imply one another, or entailment if the second text is a logical consequence of the first text.

Table 2 (bottom) illustrates CAT applied to five example texts from the NLI dataset. The first row shows an entailment that was modified to a contradiction by replacing the word “break" with “lawn". While this would be the explanation offered by a typical counterfactual method, CAT additionally shows that the topic of electronics (often associated with the word “break") was removed. Such insight can offer more clarity as to why the contrast was predicted a contradiction rather than neutral, which seems more likely until we learn that the electronics topic that has nothing to do with the hypothesis was removed from the text. In row two, the change of “chrome" to “rust" is attributed to adding the space topic (as rust is an important issue in space exploration) and removed the graphics and electronics topics associated with chrome (books or web browsers). In row three, the difference of children being a part of a tournament versus simply watching the tournament is attributed to a reduction of the entertainment topic (and similarly in row four for playing a guitar versus stealing a guitar). In row five, the change of “naps" to “photos" is attributed to adding the graphics topic, which makes sense and helps build trust in the model.

4.3 Quantitative Evaluations

We evaluate the explanation methods on 500 randomly selected test instances from each dataset. We do not report results on NLI for GYC as it did not produce valid contrasts possibly because of the longer length of the texts. For each dataset, we measure the following properties: i) *Flip rate* (Flip), ii) *Edit distance* (Dist), iii) *Content Preservation* (Cont), and iv) *Fluency*. Flip rate is a measure of the model’s efficacy to generate contrastive sentences and is defined as the fraction of inputs for which an edit successfully flips the prediction. Edit Distance is the number of edits as measured by the word-level Levenstein distance between input and contrastive sentences, i.e., the minimum number of deletions, insertions, or substitutions required to transform one into the other. We report a normalized version given by the Levenshtein dis-

Table 2: Five examples of CAT applied to the AgNews and NLI datasets. Modifications to/in original/contrasts are shown by strikeout/red highlighting. Attribute changes offer insight into how the black box model views certain words with multiple possible meanings. For NLI, the </s> delimiter separates text from hypothesis.

AgNews

Input/CAT	Attribute Changes	Input Pred	Contrast Pred
Kazaa Owner Cheers ~~File~~ Salary-Swapping Decision (AP)	+4sale, +bball, +hockey, -arts, -wndws, -cryptgphy	sci-tech	sports
New Human Species ~~Discovered~~ influenza	+medicine, -cryptgphy	sci-tech	world
US shows flexibility on ~~Israeli~~ virtual settlements	+arts, +cryptgphy, -mideast	world	sci-tech
Pace of U.S. ~~Factory~~ population Growth Climbs in Dec	+politics, -money, -travel	business	sci-tech
It may take 146 years for Nigeria to wipe out ~~corruption~~ funds from ~~its~~ bank system going by the latest report ...	+ money, - politics	world	business

NLI

Input/CAT	Attribute Changes	Input Pred	Contrast Pred
Two outdoor workers conversing while on ~~break~~ lawn. </s> people on sidewalk	-electronics	entail -ment	contra -diction
The double sink is freshly polished chrome. </s> The sink is ~~chrome~~ rust colored	+space, -graphics -electronics	entail -ment	contra -diction
A group of children, wearing white karate shirts, look at the American flag.</s> The children ~~are~~ looked at a karate tournament.	-entertainment	neutral	contra -diction
A man is about to ~~play~~ steal his guitar.</s> a man is performing for school children	-entertainment	neutral	contra -diction
Two women are giving each other a hug while a man holding a glass is looking at the camera. </s> The people are all taking ~~naps~~ photos during the hottest part of the day.	+graphics, -electronics	contra -diction	neutral

tance divided by the number of words in the input; this metric ranges from 0 to 1. Content preservation measures how much input content is preserved while generating a contrastive sentence in a latent embedding space. For this, we compute the cosine similarity between input and contrastive sentence embeddings obtained from a pre-trained sentence BERT (Reimers and Gurevych, 2019) model. Fluency measures the alignment of the contrastive and input sentence distributions. We evaluate fluency by calculating masked language modeling loss on both original and edited sentences using a pre-trained GPT-2 model and compute fluency as the ratio of the loss of the contrast to the loss of the input sentence. A value of 1.0 indicates the contrast is as fluent as the original sentence. Table 3 reports means of each metric across all instances obtained from generated contrastive sentences.

Observations: In Table 3 we compare the performance of CAT with two state-of-the-art contrastive methods GYC and MICE. MICE-nft denotes MICE without fine tuning. As can be seen CAT produces contrasts with perfect flip rate, retains highest content relative to the original sentence, that too with fewest changes, and maintains best language fluency in all cases, but one. This can be accredited to the core details of the CAT approach which is based on minimal but relevant perturbations through a controlled local greedy search procedure guided by attribute classifiers. The single case where CAT is not best performing is possibly because fine tuning helps create more natural contrasts; also confirmed by the similar performance of MICE-nft and CAT.

We also estimated the efficiency of CAT by computing the time it takes to obtain a contrastive explanation normalized by the input length. CAT, MICE, and GYC were evaluated on a NVIDIA V100 GPU for 10 contrastive explanations. The mean times taken by CAT, MICE and GYC were 2.37, 1.17 and 10.69 seconds respectively. The efficiency of CAT over GYC can be credited to the controlled local greedy search approach, although guiding the search with attribute classifiers makes it slightly more expensive than MICE, where for the latter we are ignoring the time for fine tuning.

4.4 Human Evaluation

We now describe two user studies we conducted to ascertain the value of our attributed explanations. User studies have become ubiquitous in explainable AI literature (Ribeiro et al., 2016; Singh et al., 2019; Ramamurthy et al., 2020; Luss et al., 2021) because they illustrate the benefit of these tools to the end user. We follow in the direction of most user studies by ascertaining benefit by requiring participants to perform a task given our explanations.

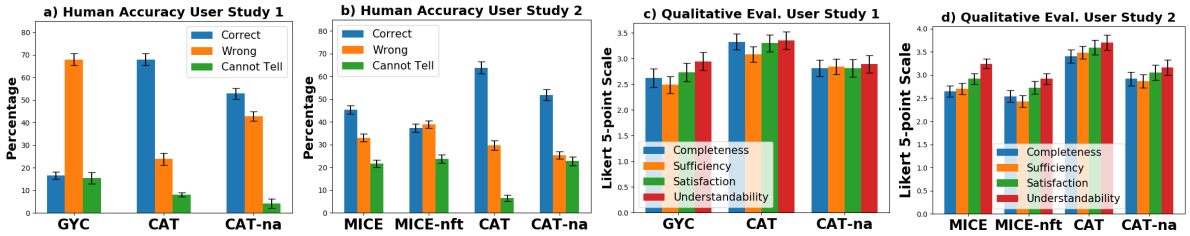


Figure 2: Figures a) and b) show the percentage (human) accuracy in predicting the label of the input sentence based on the contrastive explanations. Users perform significantly better using our CAT explanations showcasing the benefit of the attributes. Figures c) and d) show a 5-point Likert scale (higher better) for four qualitative metrics used in previous studies (Madumal et al., 2020) compared across the methods. Here too, the difference is noticeable for all four metrics. Error bars are one standard error.

Table 3: We evaluate CAT on four properties: 1) **Flip** rate, ii) **Dist**(distance), iii) **Fluency**, iv) **Cont** (content preservation). We report mean values for each metric. ↑ (↓) indicates the higher (lower) desired score whereas ≈ 1 indicates desired value be close to 1. Best results are bolded and NR implies no result. Differences are statistically significant; see Appendix E.

Embedding Bag based Classifier

Dataset	Method	Fine Tuning	↑ Flip	↓ Dist	↑ Cont	≈ 1 Fluency
AgNews	GYC	No	0.42	0.40	0.77	1.13
	CAT	No	**1**	**0.22**	**0.87**	**1.01**
Yelp	GYC	No	0.70	0.49	0.61	1.32
	CAT	No	**1**	**0.23**	**0.88**	**1.09**
Dbpedia	GYC	No	0.72	0.52	0.55	1.33
	CAT	No	**1**	**0.16**	**0.89**	**1.05**
NLI	GYC	No	NR			
	CAT	No	**1**	**0.08**	**0.98**	**1.03**

Roberta based Classifier

Dataset	Method	Fine Tuning	↑ Flip	↓ Dist	↑ Cont	≈ 1 Fluency
AgNews	MICE	Yes	1	0.42	0.79	**1.01**
	MICE-nft	No	0.98	0.52	0.70	0.90
	CAT	No	**1**	**0.19**	**0.90**	0.90
Yelp	MICE	Yes	0.99	0.35	0.86	1.09
	MICE-nft	No	0.97	0.41	0.82	1.06
	CAT	No	**1**	**0.11**	**0.96**	**1.04**
Dbpedia	MICE	Yes	0.81	0.28	0.90	1.23
	MICE-nft	No	0.75	0.34	0.84	1.19
	CAT	No	**1**	**0.09**	**0.96**	**1.06**
NLI	MICE	Yes	1	0.25	0.85	1.14
	MICE-nft	No	0.99	0.29	0.90	1.12
	CAT	No	**1**	**0.07**	**0.98**	**1.03**

Methods: We consider five different explanation methods: 1) CAT, 2) CAT-na (i.e. CAT without attributes), 3) GYC, 4) MICE and 5) MICE-nft (i.e. MICE without fine-tuning). Comparison of CAT with CAT-na provides an ablation, while its comparison with GYC, MICE, MICE-nft showcases its value relative to other state-of-the-art contrastive text explanations. Again, here we ran two separate user studies, one comparing with GYC (User Study 1) and the other with MICE (User Study 2), since each implementation was more suited to a specific type of model and embedding. CAT, however, was amenable to either setting.

Setup: We chose the AgNews dataset for the study since this was the only dataset where we underperformed on one of the benchmark metrics (fluency) w.r.t. the competitors. We thus wanted to see if this had any effect in terms of the quality of insight provided by our explanations to typical users of such explanations. For each study, we built a four class neural network black box model (described earlier) to predict articles as either belonging to Business, Sci-Tech, World or Sports categories. The task given to users was to determine the classification of the article based on a contrastive explanation from one of the respective methods. Five options were provided to the user which included the four classes along with a "Can't Tell" option. For each method, seven (User Study 1) or five (User Study 2) randomly chosen sentence-explanation pairs were provided to the user where the users were blinded to the exact method producing the explanation. For each (anonymized) explanation method, we additionally asked the users for their qualitative assessment along four dimensions; completeness, sufficiency, satisfaction and understandability based on a 5 point Likert scale. A total of 24 questions were answered by users in both user studies. Users were also allowed to leave optional comments which we provide in the appendix, along with screen shots from the user study.

We hosted our survey on Google Forms. A total of 75 participants with backgrounds in data science, engineering and business analytics voluntarily took part in user studies (37 in study 1 and 38 in study 2). We chose this demographic as recent studies show that typical users of such explanations have these backgrounds (Bhatt et al., 2020).

Observations: Figure 2 depicts the results from our user studies. Figure 2a demonstrates that both

our methods CAT and its ablation CAT-na (which does not use attributes) significantly outperform GYC in terms of usability towards the task. Figure 2b shows similar performance for CAT and CAT-na, although MICE is a much stronger competitor than GYC. It seems that the embeddings used by MICE lead to better contrasts than GYC, not to mention the mechanistic difference between them where GYC biases towards changing the end of sentences which may not be preferable in many cases. Additionally, the fine tuning done in MICE seems to further help elevate the quality of the contrasts. Nonetheless, CAT still outperforms MICE without the (expensive) need to fine tune.

Our method performs much better than GYC, MICE and CAT-na from a qualitative perspective, as well, as seen in Figures 2c and 2d, where CAT seems to score highest in understandability. These qualitative preferences of CAT are further confirmed through the (optional) comments written by some participants, e.g., "... explainer B was very good, and explainer C was reasonably good", where "explainer B" refers to CAT and "explainer C" to CAT-na in user study 1; or "the additional info in explanation C was useful" in user study 2, where "explanation C" here refers to CAT.

5 Conclusion

In this paper, we proposed a contrastive explanation method, CAT, with a novel twist where we construct attribute classifiers from relevant subtopics available in the same or different dataset used to create the black box model, and leverage them to produce high quality contrasts. The attributes themselves provide additional semantically meaningful information that can help gain further insight into the behavior of the black box model. We have provided evidence for this through diverse qualitative examples on multiple datasets, a carefully conducted user study and through showcasing superior performance on four benchmark quantitative metrics.

In the future, it would be interesting to test CAT on other applications with the appropriate attribute classifiers. Another useful direction is generating counterfactually-augmented data (CAD) using CAT. If used to generate multiple contrasts, CAT could offer attributional/topic diversity. Since diversity in types is key to producing robust models, CAT could potentially to a large degree alleviate "the lack of perturbation diversity that limits CAD's effectiveness" (N. Joshi, 2022).

Limitations

Although our work has the potential to have a positive impact on discovering models that have unknown discrimination, a nefarious agent could potentially provide accurate but purposely incorrectly labeled attribute classifiers in order to drive certain misleading or simply incorrect insights. Another unethical act by a developer could be to hide sensitive attributes so that biases could not be discovered. In order to use the explanation system in a beneficial manner we do assume the developer can be trusted. It is also possible that we may not uncover the globally minimal contrast since the optimization is non-convex given the complexity of classifiers we are trying to explain. Thus, smaller edits may be possible that change the output of a classifier that go unnoticed. This however, is a concern for even other contrastive/counterfactual explainability methods.

References

David Alvarez-Melis and Tommi S. Jaakkola. 2017. A causal framework for explaining the predictions of black-box sequence-to-sequence models. In *Proceedings of EMNLP*.

Vijay Arya, Rachel K. E. Bellamy, Pin-Yu Chen, Amit Dhurandhar, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Q. Vera Liao, Ronny Luss, Aleksandra Mojsilović, Sami Mourad, Pablo Pedemonte, Ramya Raghavendra, John Richards, Prasanna Sattigeri, Karthikeyan Shanmugam, Moninder Singh, Kush R. Varshney, Dennis Wei, and Yunfeng Zhang. 2019. One explanation does not fit all: A toolkit and taxonomy of AI explainability techniques. arXiv:1909.03012.

Umang Bhatt, Alice Xiang, Shubham Sharma, Adrian Weller, Yunhan Jia Ankur Taly, Joydeep Ghosh, Ruchir Puri, José M. F. Moura, and Peter Eckersley. 2020. Explainable machine learning in deployment. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of EMNLP*.

Annie Brown. 2021. Brilliance knows no gender: Eliminating bias in chatbot development.

Marina Danilevsky, Kun Qian, Ranit Aharonov, Yannis Katsis, Ban Kawas, and Prithviraj Sen. 2020. A survey of the state of explainable AI for natural language processing. In *AACL-IJCNLP*.

S. Dathathri, A. Madotto, J. Lan, J. Hung, E. Frank, P. Molino, J. Yosinski, and R. Liu. 2020. Plug and play language models: A simple approach to controlled text generation. In *ICLR*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Amit Dhurandhar, Pin-Yu Chen, Ronny Luss, Chun-Chen Tu, Paishun Ting, Karthikeyan Shanmugam, and Payel Das. 2018. Explanations based on the missing: Towards contrastive explanations with pertinent negatives. In *Advances in Neural Information Processing Systems*, pages 592–603.

Amit Dhurandhar, Tejaswini Pedapati, Avinash Balakrishnan, Kartik Ahuja Pin-Yu Chen, Karthikeyan Shanmugam, and Ruchir Puri. 2019. Model agnostic contrastive explanations for structured data. https://arxiv.org/abs/1906.00117.

Shi Feng, Eric Wallace, Alvin Grissom II, Mohit Iyyer, Pedro Rodriguez, and Jordan Boyd-Graber. 2018. Pathologies of neural models make interpretations difficult. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3719–3728, Brussels, Belgium. Association for Computational Linguistics.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. Allennlp: A deep semantic natural language processing platform. *arXiv preprint arXiv:1803.07640*.

I. Goodfellow, Y. Bengio, and A. Courville. 2016. *Deep Learning.* MIT Press.

David Gunning. 2017. Explainable artificial intelligence (xai). In *Defense Advanced Research Projects Agency*.

Karthik Gurumoorthy, Amit Dhurandhar, Guillermo Cecchi, and Charu Aggarwal. 2019. Efficient data representation by selecting prototypes with importance weights. In *Proceedings of the IEEE International Conference on Data Mining*.

David Harbecke, Robert Schwarzenberg, and Christoph Alt. 2018. Learning explanatiosn from language data. In *EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neureal Networks for NLP*.

Alon Jacovi, Swabha Swayamdipta, Shauli Ravfogel, Yanai Elazar, Yejin Choi, and Yoav Goldberg. 2021. Contrastive explanations for model interpretability.

Amir-Hossein Karimi, Bernhard Schölkopf, and Isabel Valera. 2021. Algorithmic recourse: from counterfactual explanations to interventions. In *ACM conference on Fairness, Accountability and Transparency (FAccT)*.

Been Kim, Rajiv Khanna, and Oluwasanmi Koyejo. 2016. Examples are not enough, learn to criticize! Criticism for interpretability. In *In Advances of Neural Inf. Proc. Systems*.

Pang Wei Koh and Percy Liang. 2017. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning*.

Abhishek Kumar, Prasanna Sattigeri, and Avinash Balakrishnan. 2018. Variational inference of disentangled latent concepts from unlabeled observations. In *Proceedings of the International Conference on Learning Representations*.

Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, D. Kontokostas, Pablo N. Mendes, Sebastian Hellmann, M. Morsey, Patrick van Kleef, S. Auer, and C. Bizer. 2015. Dbpedia - a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web*, 6:167–195.

Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, et al. 2021. Datasets: A community library for natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 175–184.

Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, pages 4765–4774.

Ronny Luss, Pin-Yu Chen, Amit Dhurandhar, Prasanna Sattigeri, Karthik Shanmugam, and Chun-Chen Tu. 2021. Leveraging latent features for local explanations. In *ACM KDD*.

Nishtha Madaan, Inkit Padhi, Naveen Panwar, and Diptikalyan Saha. 2021. Generate your counterfactuals: Towards controlled counterfactual generation for text. *AAAI*.

Prashan Madumal, Tim Miller, Liz Sonenberg, and Frank Vetere. 2020. Explainable reinforcement learning through a causal lens. In *AAAI*.

Rishabh Misra. 2018. News category dataset.

Christoph Molnar. 2019. Interpretable machine learning.

H. He N. Joshi. 2022. An investigation of the (in)effectiveness of counterfactually augmented data. In *ACL*.

20 Newsgroup. 2008. Kaggle.

Nicolas Papernot and McDaniel Patrick. 2018. Deep k-nearest neighbors: Towards confident, interpretable and robust deep learning. In *arXiv:1803.04765*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *CoRR*, abs/1910.10683.

Karthikeyan Natesan Ramamurthy, Bhanukiran Vinzamuri, Yunfeng Zhang, and Amit Dhurandhar. 2020. Model agnostic multilevel explanations. In *Advances in Neural Information Processing Systems*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144.

Alexis Ross, Ana Marasovic, and Matthew E. Peters. 2021. Explaining NLP models via minimal contrastive editing (mice). *ACL*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6833–6844.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv:1312.6034.

Chandan Singh, W. James Murdoch, and Bin Yu. 2019. Hierarchical interpretations for neural network predictions. In *Intl. Conference on Learning Representations (ICLR 2019)*.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, pages 3319–3328. PMLR.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Sandra Wachter, Brent Mittelstadt, and Luciano Floridi. 2017. Why a right to explanation of automated decision-making does not exist in the General Data Protection Regulation. *Int. Data Privacy Law*, 7(2):76–99.

Eric Wallace, Shi Feng, and Jordan Boyd-Graber. 2018. Interpreting neural networks with nearest neighbors. In *EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neureal Networks for NLP*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

T. Wu, M. T. Ribeiro, J. Heer, and D. S. Weld. 2021. Polyjuice: Generating counterfactuals for explaining, evaluating, and improving models. In *ACL*.

Philip N. Yannella and Odia Kagan. 2018. Analysis: Article 29 working party guidelines on automated decision making under gdpr. www.cyberadviserblog.com/2018/01/analysis-article-29-working-party-guidelines-on-automated-decision-making-under-gdpr/.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, page 649–657, Cambridge, MA, USA. MIT Press.

Let the CAT out of the bag: Contrastive Attributed explanations for Text (Appendix)

A Hyperparemeter tuning for CAT

Since no ground truth exists for our contrastive explanations, hyperparameters for CAT objective (given in Eq.1) were tuned qualitatively by observation. It is important to consider that, in practice a user will only need to tune the hyperparameters once at the beginning and will then be able to generate explanations for an arbitrary number of instances from the dataset. Our tuning led to the following values which we used across datasets: We kept highest weights for $\lambda = 5.0$ to make sure that perturbed sentences have a label different from that of original sentence. We used mean of BERT logits of words inserted/replaced as a surrogate for $p_{LM}(x')$ with regularization parameter η set to be 1.0 and regularization parameter ν for Levenshtein distance set to 2.0. β was set to 3.0. The threshold for predicting addition or removal of attributes τ, we used two values. For binary attribute classifiers from the Huffpost News dataset, we set $\tau = 0.3$, and for multiclass attribute classifiers trained on the 20 Newsgroup dataset, we set $\tau = 0.05$.

We took 50 randomly selected examples from the training set, generated explanations and evaluated them manually to choose hyperparameters which is similar to prior works (Luss et al., 2021; Madaan et al., 2021). The boundary values for hyperparameter search were $\lambda \in [4, 10]$, $\beta \in [1, 5]$, $\eta \in [0.5, 2]$, $\nu \in [1, 4]$, and $\tau \in [0.01, 0.5]$. With more examples for tuning we would expect better explanations, although we found this number to be sufficient.

B Attribute classifiers

In our experiments we created 42 attributes from Huffpost News-Category dataset(Misra, 2018) and 20 Newsgroups (Newsgroup, 2008).

News-Category dataset [2] has 41 classes of which we merged similar classes and removed those which weren't standard topic. For instance, classes "*food & drink*" and "*taste*" was merged and labelled as a new class "*food*". At the end we obtained the following 22 classes: *education, money, world, home & living, comedy, food, black voices, parenting, travel, sports, women, religion, Latino voices, weddings, entertainment, crime, queer voices, arts, politics, science, fifty*, and *environment*. We created 22 1-vs-all binary classifiers, so that the same sentence can have multiple classes. To alleviate the class imbalance issue in training these binary classifiers we sampled fewer instances uniformly at random from the negative classes making sure that the number of negative instances are no more than 80% of the training data.

The 20 Newsgroup dataset [3] has the following 20 classes: *atheism, graphics, ms-windows.misc, computer, mac.hardware, ms-windows.x, forsale, autos, motorcycles, baseball, hockey, cryptography, electronics, medicine, space, christian, guns, mideast, politics*, and *religion*. A distilbert multi-class classifier was trained on this data.

All attribute classifiers were trained with DistliBERT base [4] for 10 epochs using the Adam optimizer with a learning rate of 5×10^{-5}, weight decay of 0.01, and a batch size of 16. Each of the models were trained using NVIDIA V100 GPUs in under 12 hours.

C Text classifiers

We conducted experiments on two sets of text classifiers trained on four datasets: One set to compare CAT with GYC and another to compared CAT with MICE. All experiments were run on NVIDIA V100 GPUs.

For experiments comparing CAT with GYC(Madaan et al., 2021), we trained model with the same architecture as the one used in their paper. This model is composed of an *EmbeddingBag* layer followed by a linear layer, a *ReLU* and another linear layer to obtain the logit. The logit is provided to a classifier we trained with cross entropy loss. A GPT2 tokenizer [5] was used to convert text into bag of words. The models were trained using the Adam optimizer with a learning rate of 1×10^{-3}, weight decay of 0.01, and a batch size of 32.

For experiments comparing CAT with MICE(Ross et al., 2021), we used author provided implementation[6] to train models with `AllenNLP` (Gardner et al., 2018). This architecture is composed of a *RoBERTa-base* model with a linear

[2] https://www.kaggle.com/rmisra/news-category-dataset

[3] https://scikit-learn.org/stable/modules/generated/sklearn.datasets.fetch_20newsgroups.html

[4] https://huggingface.co/distilbert-base-uncased

[5] https://huggingface.co/gpt2-medium

[6] https://github.com/allenai/mice

Table 4: We evaluate CAT on three metrics: i) **Dist**(distance), ii) **Fluency**, iii) **Cont** (content preservation). We report p-values from a pairwise t-test of the difference of means between GYC with CAT and MICE with CAT for all metrics. As can be seen by the negligible p-values, CAT is (statistically) significantly better than GYC as well than MICE. We also report standard deviation (std. dev.) for all metrics for CAT, GYC and MICE, which together with the means reported in the main paper, are used to run the t-tests.

Embedding Bag based Classifier

	AgNews			Yelp		
Stat	Dist	Cont	Fluency	Dist	Cont	Fluency
p-value	$< 1e^{-8}$	$< 1e^{-8}$	$< 1e^{-8}$	$< 1e^{-8}$	$< 1e^{-8}$	$< 1e^{-8}$
std. dev. (GYC)	0.164	0.110	0.286	0.140	0.105	0.467
std. dev. (CAT)	**0.098**	**0.084**	**0.105**	**0.100**	**0.076**	**0.134**

	Dbpedia			NLI		
Stat	Dist	Cont	Fluency	Dist	Cont	Fluency
p-value	$< 1e^{-8}$	$< 1e^{-8}$	$< 1e^{-8}$	$< 1e^{-8}$	$< 1e^{-8}$	$< 1e^{-8}$
std. dev. (GYC)	0.132	0.115	0.344		NR	
std. dev. (CAT)	**0.059**	**0.069**	**0.064**	**0.323**	**0.010**	**0.033**

Roberta based Classifier

	AgNews			Yelp		
Stat	Dist	Cont	Fluency	Dist	Cont	Fluency
p-value	$< 1e^{-8}$	$< 1e^{-8}$	$< 1e^{-8}$	$< 1e^{-8}$	$< 1e^{-8}$	$< 1e^{-8}$
std. dev. (MICE)	**0.182**	**0.121**	0.196	0.174	0.109	0.208
std. dev. (MICE-nft)	0.307	0.209	0.224	0.240	0.154	0.212
std. dev. (CAT)	0.307	0.209	**0.144**	**0.084**	**0.044**	**0.098**

	Dbpedia			NLI		
Stat	Dist	Cont	Fluency	Dist	Cont	Fluency
p-value	$< 1e^{-8}$	$< 1e^{-8}$	$< 1e^{-8}$	$< 1e^{-8}$	$< 1e^{-8}$	$< 1e^{-8}$
std. dev.(MICE)	0.151	0.077	0.168	0.132	0.109	0.131
std. dev.(MICE-nft)	0.200	0.160	0.160	**0.164**	0.088	0.143
std. dev.(CAT)	**0.058**	**0.039**	**0.071**	**0.037**	**0.013**	**0.054**

layer. For all four datasets the models were trained for 5 epochs with batch size of 8 using Adam optimizer with a learning rate of $4e-05$, weight decay of 0.1, and slanted triangular learning rate scheduler with cut frac 0.06.

D Datasets

We performed experiments on 4 datasets: AgNews(Zhang et al., 2015), NLI(Bowman et al., 2015), DBpedia (Lehmann et al., 2015), and Yelp(Shen et al., 2017). AgNews dataset was taken from Kaggle website [7] and rest three datasets from *huggingface datasets*(Lhoest et al., 2021).

AgNews. The dataset is from a real-world news domain which contains short news headlines and bodies from four new categories - world, business, sports, and sci-tech. It contains 30K training and 1.9K test examples per class, comprising of 128K samples. Our experiments focus on explanations for predicting the class from the headlines.

NLI. The Natural Language Inference (Bowman et al., 2015) dataset[8] contains samples of two short ordered texts, and the labels are either *contradiction* if the second text contradicts the first, *neutral* if the two texts do not contradict or imply one another, or *entailment* if the second text is a logical consequence of the first text. The dataset contains 550K training, 10K test and 10K validation examples.

DBpedia. This dataset is a subset of original DBpedia data which is a crowd-sourced community effort to extract structured information from Wikipedia. This dataset[9] is constructed by picking 14 non-overlapping classes from DBpedia 2014 with 40K training and 5K test examples per class. Task here is to predict the class a DBpedia entry belong to. In our experiments we only use *content* and drop the *title* field provided with with the dataset.

Yelp. This is a binary sentiment classification dataset[10] containing 560K highly polar reviews for training and 38K for testing. The dataset consists of reviews from Yelp which is extracted from the Yelp Dataset Challenge 2015.

E Quantitative Evaluation Statistics

Table 3 (in the main paper) shows the performance evaluation of our proposed approach, CAT, on five different datasets and 2 classifier models, namely Embedding Bag based classifier and Roberta based classifier models. It was noted that CAT outperformed GYC and MICE over AgNews, Yelp and Dbpedia by statistically significant margins on respective model. Recall that GYC usage Embedding Bag based classifier model and MICE usage Roberta based classifier. Therefore, we implemented both the classifier models with CAT to compare both GYC and MICE. We verify this statement in Table 4 where we report pairwise t-tests comparing the means for CAT with GYC and CAT with MICE and standard deviation of CAT, GYC, MICE for each metric and dataset. The improvement of CAT over GYC and MICE is observed to be statistically significant across all metrics. We do not report additional statistics for the flip rate metric as CAT always produces a contrastive sentence with a flipped class label unlike GYC or MICE which sometimes fails to flip.

F Additional Information for User Study

Figure 3 shows screenshots of user study 1, including the instructions and example questions for the three different methods discussed in the study. The same instructions and format was used for user study 2, except that we asked five task oriented questions (rather than seven) per explainer as there were four explainers (as opposed to 3) keeping the total number of questions to be 24, and thus keeping the overall effort of both user studies roughly the same for participants. For the users the methods were named as Explainer A, B and C, where they correspond to GYC, CAT, and CAT-na respectively for user study 1. For user study 2, Explainer A, B, C and D corresponded to MICE, MICE-nft, CAT and CAT-na. Some users also left optional comments at the end of the user study which we list here:

- "I found the questions confusing... What is complete, sufficient and understandable explanation in a word change? Also, in each example, what was I supposed to guess, the category of the article or the possible prediction of the method? Are those things different?" (Study 1)
- "Explainer A was pretty bad, explainer B was very good, and explainer C was reasonably good." (Study 1)
- "Explainer b was the best. The q&a at the end of each page allows multiple choices per row,

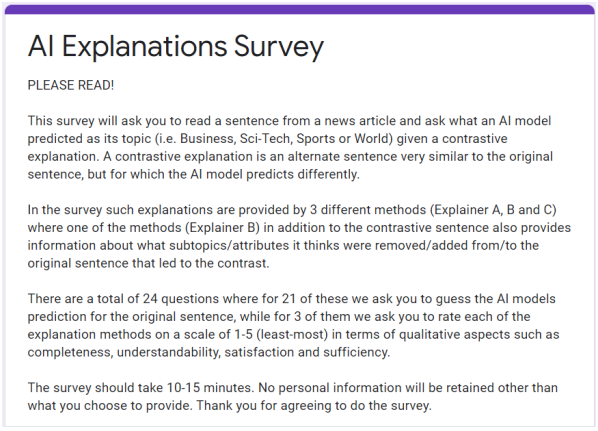
[7] https://www.kaggle.com/datasets/amananandrai/ag-news-classification-dataset
[8] https://huggingface.co/datasets/snli
[9] https://huggingface.co/datasets/dbpedia_14
[10] https://huggingface.co/datasets/yelp_polarity

which is bad. Each question should remove the modified classification as an option (if the modified sentence is World, I shouldn't be allowed to choose World). Early on the survey should state clearly that there are only 4 categories (business/sci-tech/sports/world) which I didn't know right away, I expected each question have 4 different options, and knowing this might have helped me understand what we're doing here better. The opening text was very confusing." (Study 1)

- "I found B to be the best one." (Study 1)
- "nice survey" (Study 1)
- "the additional info in explanation C was useful" (Study 2)

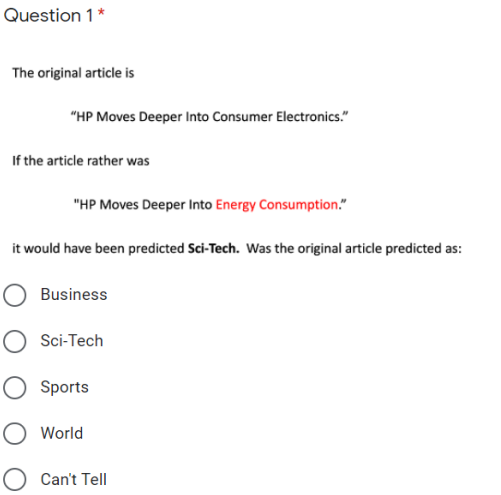
G Additional Qualitative Examples

Table 5 offers at least five more examples of contrastive explanations from each dataset. Additional insight can be gained from the attributes; in the first row, adding travel-related text can flip the class from business to world, i.e., the article was predicted business because if it was more about travel it would have been predicted as world. This type of explanation adds extra intuition as opposed to only being given the replacement of words, "oil" to "winds" in this case. As can also been seen, insight is also a function of using attributes that have a relationship to the dataset and task. For example, attributes derived from news sources are often not helpful for explaining sentiment classifiers (e.g., Yelp) for which sentiment is often flipped by the negation of the text. This is to be expected; good explanations require good attributes.

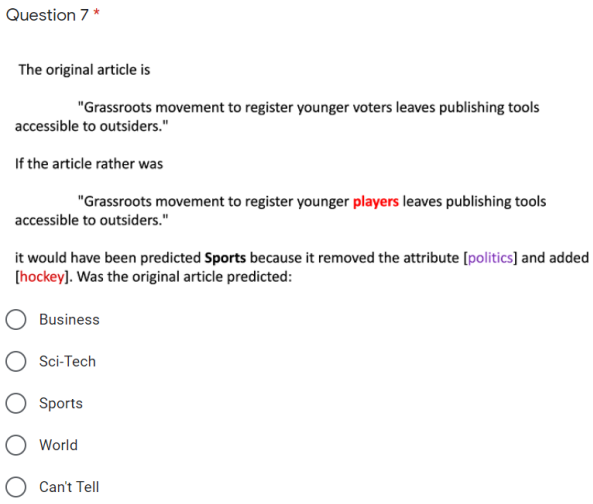


(a)

(b)



(c)



(d)

Figure 3: In (a) we show the instructions for the survey. In (b) we see the qualitative questions. In (c) and (d) we see the task oriented questions for GYC and CAT respectively.

Table 5: Extra contrastive explanation examples from four datasets: AgNews, NLI, DBPedia, and Yelp.

Input/CAT	Attribute Changes	Input Pred	Contrast Pred	Dataset
~~Oil~~ Winds Up from 7-Week Lows on U.S. Weather	+travel	business	world	AgNews
New Hummer Is Smaller, Cheaper and Less ~~Gas~~ cpu Hungry	+cryptography, -travel, -space	business	sci-tech	AgNews
Perfect start for France in ~~Federation~~ Microsoft Cup	+cryptography, -motorcycles, -hockey, -space, -politics	sports	sci-tech	AgNews
Will sinking Lowe resurface in ~~playoffs~~ 2020?	+space, -hockey	sports	sci-tech	AgNews
~~Kuwait~~ Source: Fundamentalists Recruiting Teens (AP)	+guns, religion, -mideast	world	sci-tech	AgNews
Harry ~~in nightclub~~ inflight scuffle	+sports, +space, -religion, -cryptography, -electronics	world	sci-tech	AgNews
Five soldiers hold and aim their ~~weapons~~ blades. </s> The soldiers are ~~eating~~ armed.	-guns	contra -diction	entail -ment	NLI
Skateboarder jumps of off dumpster. </s> a ~~bike~~ young rider in a race	+world	contra -diction	neutral	NLI
Two people walking down a dirt trail with backpacks on looking at ~~items~~ map they are carrying. </s> One is holding a map.	+space, -motorcyles, -electronics	neutral	entail -ment	NLI
A brown and black dog is jumping to catch a red ball. </s> Two dogs are not playing catch with their owner.	-arts	neutral	contra -diction	NLI
A girl is standing in a field pushing up her hat with one finger and her hand is covering most of her face. </s> A ~~girl~~ boy covers her face.	+parenting, -medicine	entail -ment	contra -diction	NLI
Channel Chaos is a 1985 Australian ~~film~~ production set at a TV station	+electronics, -travel, -space	film	company	DBPedia
Dinik is a ~~village~~ lake in Croatia	+medicine, +space, -mideast	village	natural place	DBPedia
Air Cargo Mongolia is a Mongolian ~~airline~~ newspaper	+politics, -space	plant	written work	DBPedia
ACTION is a bus service ~~operator~~ based in Canberra Australia	+autos, -electronics	company	transport -ation	DBPedia
Jagjaguwar is an indie rock record ~~label~~ musician based in Bloomington Indiana	+space, +politics, -world -forsale, -electronics	company	artist	DBPedia
he ~~really~~ hardly made our anniversary dinner entertaining memorable	+world, +sports, +cryptography -home&living, -electronics	positive	negative	Yelp
they also have deep fried desserts if you're ~~brave~~ poor enough	+medicine, -travel -religion, -politics	positive	negative	Yelp
i ~~definitely~~ never will go back	-entertainment	positive	negative	Yelp
jesse is completely ~~rude~~ perfect	+cryptography, -entertainment, -medicine, -space	negative	positive	Yelp
the pizza ~~sauce~~ heaven is also way too sweet	+world, +travel, +arts, +atheism +religion, -medicine, -politics	negative	positive	Yelp