Unsupervised Concept Learning for Sentiment Analysis

Gabriel Mukobi, Liana Patel, Atharva Amdekar
Stanford University
{gmukobi, lianapat, aamdekar}@stanford.edu

Abstract

This work aims to address the problem of concept-based explainability in natural language understanding tasks. Concept-based interpretability methods are important to helping users understand model predictions in human-level intuitive concepts. While prior works in concept-based explainabilty have largely focused on the vision domain, we explore challenges in adapting such methods to natural langauge by building from recent works (Shi et al., 2020) in this area and applying concept-based explainability methods to a sentiment classification task. We compare the concept-based interpretable methods to black-box model predictions and find that performance on the task is comparable, however the coherency of discovered concepts in the interpretable model suffers from lack of coherency. Code to reproduce our results is at github.com/mukobi/Unsupervised-Concept-Based-Explanations-For-Sentiment-Analysis.

1 Introduction

State-of-the-art deep neural networks provide impressive performance, following a trend toward larger and more complex models. However, such blackbox models are often difficult to interpret and reason about. Explaining these models helps to address the safety and ethical concerns of deploying the models in practice. In particular, methods in concept-based post-hoc explainablity have gained recent traction (Kim et al., 2017) due to their ability to provide feedback through higher-level ideas or concepts that reflect how humans intuitively reason.

While much current concept-based explainability work focuses on applying techniques in the computer vision domain, transferring these methods to the natural language domain introduces interesting questions. In this work we build upon recent works (Shi et al., 2020) to create an efficient way of larning concepts in an unsupervised manner for one of the most common NLP tasks, sentiment analysis. We demonstrate that an explainable, concept-based Attention Abstraction Network fine-tuned for sentiment analysis performs on par with state-of-the-art non-interpretable methods. In addition we show that the Attention Abstraction Network produces concept explanations that are quantitateively diverse and qualitatively interpretable to humans

2 Related Work

2.1 Concept-Based Interpretability

Recent works introduce principles in concept-based explainability in order to address the challenges of interpretability in deep learning models by proposing methodos that operate on high-level human concepts rather than low-level features (Kim et al., 2017) . While other interpretability techniques, such as saliency maps, focus on understanding a model's decision with respect to individual feature inputs, such as pixels of an image, concept-based explainability, by contrast, focuses on higher-level, human-friendly concepts through the proposed use of Concept Activation Vectors. Specifically, Concept Activation Vectors (CAVs) provide a translation between E_m, the vector space represented by an ML model's internal values (e.g. neural activations), and E_h, a separate vector space which humans think in and which is spanned by implicit vectors e_h, corresponding to an unknown set of human-interpretable concepts.

Building from this technique, Testing with CAVs (TCAV) provides a method to quantify the degree to which a model prediction is sensitive to a user-defined concept, learned by a CAV (Kim et al., 2017). Specifically the TCAV score is the fraction of training samples whose model classifier scores increase when the input is infinitesimally moved in the direction of the concept. A concept is then considered to be related to a class label k if the TCAV score is significantly different from TCAV scores

with random concepts. CAV and TCAV are primary methods in the concept-based explainability literature, however these methods primarily apply and have been used in the computer vision domain, where concept-examples can be provided from images. In contrast, this work focuses on applying concept-based explainability methods in the NLP domain.

2.2 Unsupervised Concept Discovery

One central challenge with the aforementioned concept-based explainability methods is that users must choose and define concepts, which is subject to human bias. Thus, an important subproblem in concept-based explainability is unsupervised concept-discovery, whereby relevant concepts are automatically found by the system, as opposed to being explicitly defined and probed for by the user, as the initial work on CAV and TCAV proposes to do.

Recent works make progress towards this goal. One approach proposed algorithm, ACE automatically extracts visual concepts using an unsupervised approach for discovering concepts in a dataset of images based on super-pixel segmentation followed by k-means clustering (Ghorbani et al., 2019). This additionally lays out desirable requirements that concept-based explanations should satisfy, namely 1) Meaningfulness - an example of a concept should be semantically meaningful on its own. 2) Coherency - examples of a concept should be perceptually similar to each other while being different from examples of other concepts. 3) Importance - a concept is "important" for the prediction of a class if its presence is necessary for the true prediction of samples in that class.

Other approaches in unsupervised concept-based explainability focus on the challenge of effectively disentangling important concepts, relating to the challenge of coherency of concepts (Ghandeharioun et al., 2021). Additionally, a recent work (Yeh et al., 2019) also tackles the problem of unsupervised concept discovery with a specific focus of finding concepts that meet a notion of "sufficiency" or "completeness", which has not been addressed by prior works.

In this work we focus leverage unsupervised concept discovery, specifically in the NLP domain for a sentiment classification task.

2.3 Concept-Based Interpretability for NLP Tasks

While, much of the focus in concept-based explainability centers on the visual domain, some recent work proposes concept-based interpretability methods in natural language tasks. In the natural language domain, attention mechanisms have been widely studied because they achieve state-of-the-art performance, and they can also provide interpretability since attention weights can be visualized using heat-maps over the text input. However, attention mechanisms cannot provide interpretability in terms of higher-level concepts that are important for model predictions (Shi et al., 2020).

Towards the goal of addressing the drawbacks of attention visualization and providing concept-based interpretability in natural language, one recent work proposes a corpus-level explanation approach, which aims to capture causal relationships between keywords and model predictions by learning the importance of keywords for predicted labels across a training corpus using attention weights (Shi et al., 2020). Extending this idea further, the work also proposes a concept-based explanation method which automatically learns higher level concepts and their importance to model predictions. The method introduces a novel architecture called an Abstraction-Aggregation Network (AAN) for interpretable document classification. An AAN consists of an encoder (typically word embeddings followed by an LSTM or a pretrained transformer encoder), a pooling layer, and a classification layer. The proposed Abstraction-Aggregation Network uses two stacked attention layers for its pooling. In this work, we build from the aforementioned ideas and proposed methods to specifically develop and evaluate a concept-based explanability method in the task of sentiment classification.

3 Data

In our implementation and evaluation, we use DynaSent (Potts et al., 2020), an English-language benchmark task for ternary (positive/negative/neutral) sentiment analysis. DynaSent has a total of 121,634 sentences, each validated by five crowdworkers. The dataset combines naturally occurring sentences with sentences created on the Dynabench Platform, an opensource platform which allows for human-and-model-in-the-loop dataset creation. DynaSent is intended to be a dynamic benchmark that expands in response

to new models, new modeling goals and new adversarial attacks. The dataset creation is based on a series of rounds, with the goal of the final dataset fooling a top-performance sentiment model (Potts et al., 2020). Specifically, we use the Round 2 dataset for training and for our evaluation we use the dataset from both Round 1 and Round 2.

4 Model

We two categories of models with slight variations in their constructions: Uninterpretable black-box baseline models, and interpretable concept-based models.

4.1 Baseline Black-Box Models

Our black box model is a simple but effective construction for the task of sentiment analysis. We use an encoder that transforms our input sequence into a hidden latent space, perform some pooling operation to extract a fixed-dimension representation of the sequential encoder output, and then feed that representation to a small classification network. This resulting network is then fine-tuned end-to-end on the sentiment analysis training data and evaluated.

We test a variety of encoders to explore for their varying complexities and pretrained capabilities. They include:

- RoBERTa-base (Liu et al., 2019), a 125M parameter transformer model (Vaswani et al., 2017) pretrained on the masked language modeling (MLM) objective.
- DynaSent Model 1 (Potts et al., 2020), a version of RoBERTa-base that was fine-tuned for sentiment analysis on a wide variety of product and service reviews.

To get a fixed-dimension representation from the encoder outputs, we use the `CLS` token for the RoBERTa-based models. We do this for simplicity, but one could instead use an attention layer or mean/median/max pooling across all of the hidden representations. Consequently, all of our encoders output vectors of dimension 768 (the default word embedding size in BERT-base (Devlin et al., 2018) and thus RoBERTa-base (Liu et al., 2019)).

Our classification layer is a simple two-layer feed forward linear network. That is, Linear $\rightarrow$ ReLU $\rightarrow$ Linear $\rightarrow$ Softmax, or:

$$y = \text{softmax}(W_2 \cdot (\text{ReLU}(W_1 \cdot x + b_1)) + b_2)$$

where x is the encoder output/classifier input, y is the vector of output probabilities, and W_1, W_2, b_1, and b_2 are parameters learned by fine-tuning on the training data.

Additionally, we test a random classifier baseline that simply uniformly samples one of the output classes in order to establish a minimum performance threshold for the dataset.

4.2 Concept-Based Interpretable Models

We follow heavily from section 3 of (Shi et al., 2020), which proposes a novel architecture called an Abstraction-Aggregation Network (AAN) for interpretable document classification. Very similarly to our baseline models, an AAN consists of an encoder (typically word embeddings followed by an LSTM or just a pretrained transformer encoder), a pooling layer, and a classification layer. For experimental purposes, we maintain the same encoders and classification layer architecture as described above for our baseline models, but manipulate the pooling layer.

The proposed Abstraction-Aggregation Network uses two stacked attention layers for its pooling. These are called an *abstraction-attention (abs)* with k attention units and an *aggregation-attention (agg)* layer with 1 attention unit. A diversity penalty and dropout are added one to each layer to encourage learning diverse concepts. The k *abs* units thus represent k learned concepts, and we can use statistics to automatically compute both the scores of each concept and the keywords from the vocabulary most relevant to the concept for a given example in order to evaluate how the model uses each concept and get a sense of the semantic meaning of each concept.

5 Methods

5.1 Training

We train all our models on the DynaSent Round 2 Train set (since it is smaller than Round 1 but more challenging). We train each model with the Adam optimizer for 10 epochs without early stopping using a learning rate of $4e-5$, a batch size of 16, and 2 gradient accumulation steps. We split the training data into an 80/20 train/validation split in order to track model performance over training.

Our black-box models are trained with cross-entropy loss only. Our AAN models are trained with a special loss function:

$$\textit{total loss} = \textit{cross entropy loss} + \beta * \textit{concept diversity penalty}$$

where *concept diversity penalty* is the abstraction-attention weight loss as described in (Shi et al., 2020). This is the same loss function as in (Shi et al., 2020), except we also introduce the β parameter which allows weighting the diversity penalty and balancing the tradeoff between model performance and learning more diverse concepts (however, we mostly only tested $\beta = 1$).

Our AAN models use $K = 10$ concepts, $\beta = 1$, and a dropout of 0.02 on the aggregation-attention weights, and all our models use a dropout of 0.1 before the final linear layer of the classification layer (these parameters are the same as in (Shi et al., 2020)).

Code to reproduce our results is available at github.com/mukobi/Unsupervised-Concept-Based-Explanations-For-Sentiment-Analysis.

5.2 Metrics

Our metrics aim to measure both how our interpretable concept-based model performs compared to the baseline as well as the quality of the learned concept explanations.

5.2.1 Quantitative Model Performance

Quantitatively, how does the interpretable concept model compare to a baseline? Hopefully, our AAN-based sentiment classifier models can perform about as well as their equivalent black-box models–that is, it is desirable that adding interpretability functionality like AAN attention layers and loss to a model will not significantly change its performance.

To evaluate this metric, we compare an interpretable model to a similar but uninterpretable black-box model on some performance metric on a test dataset; in our case, macro-F1 score across the 3 classes.

Additionally, we wish to not just calculate the raw performance metrics for each model, but to determine to what degree the models predictions are statistically significantly different. For this, we use McNemar's test (Lachenbruch, 2014) to determine whether there's a statistically significant difference between the predictions of one model compared to another and referenced against the gold labels for the test set. If the result of one such test was statistically significant (i.e. $p < 0.01$), then we can say the difference between the two models tested can not be attributed to chance and the models' performance is sufficiently different.

5.2.2 Qualitative Concept Explainability

Qualitatively, how good are the generated concepts at making the model's behavior interpretable? We aim to subjectively evaluate this with the following techniques:

- Showing examples of concept explanations for classifier decisions and commenting on whether they make semantic sense to us as humans.
- Finding mis-classified evaluation examples, showing the concept explanations, and attempting to determine why the model was wrong.
- See how changing certain hyperparameters results in different concepts for a given input.

To generate our concept explanations, we build on the Attention-Aggregation Network explanation methods proposed in (Shi et al., 2020) (note: we only use the concept-based explanation methods proposed there, not the corpus-based explanation methods).

To compute the score for each concept, we follow (Shi et al., 2020) by reading the attention weights of the aggregation-attention head. These represent a weight per concept attention head from the abstraction-attention layer to the aggregation-attention layer.

For the concept keywords, we diverge from previous work and propose a new method. The methods in (Shi et al., 2020) only find keywords that are a subset of a given input document. This works fine for documents with hundreds of words, but in the case of sentiment classification, input sentences only have 10s of words which is too few words to draw keywords directly from their text as concepts. Instead, we propose a novel way of comparing the internal abstraction-attention context representations of a given sentence to a precomputed library of contextual word representations in order to select the n most relevant keywords for a given concept.

That is, as a precomputation step we first use a library of about 25,000 common English words, forward these through the model, and for each

word store the context outputs of the abstraction-attention layer (10 concepts by 768-dim word vectors per embedding, or 1.5 GB compressed). Then, at explanation-time, we can forward a sentence through the model, read the context outputs of the abstraction-attention layer, and compare those outputs via a distance metric (in our case, cosine distance) to each word in the library.

6 Results

6.1 Quantitative Model Performance

First, we evaluate the evaluate the macro-F1 classification score for each model on two held-out test sets:

On DynaSent Round 1 Test (R1 Test, Figure 1), the macro-f1 scores are 0.695, 0.692, 0.740, 0.710 for RoBERTa-Base (Baseline), RoBERTa-Base (AAN), DynaSent-M1 (Baseline), and DynaSent-M1 (AAN), respectively.

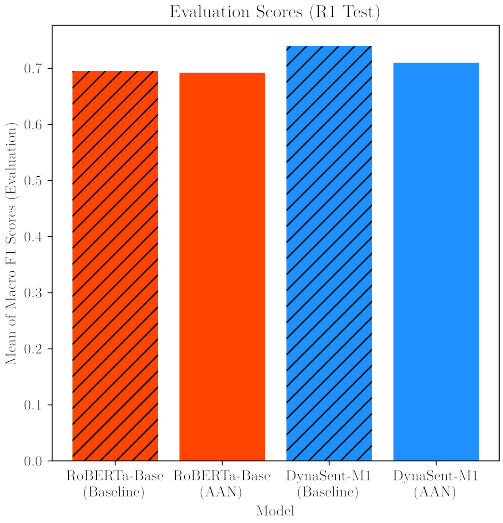


Figure 1: DynaSent Round 1 Test scores for each model.

On DynaSent Round 2 Test (R2 Test, Figure 2), the macro-f1 scores are 0.683, 0.645, 0.662, 0.722 for RoBERTa-Base (Baseline), RoBERTa-Base (AAN), DynaSent-M1 (Baseline), and DynaSent-M1 (AAN), respectively.

For McNemar's test, we calculate the following comparisons between the models:

DynaSent Round 1 Test

- p=0.384 (NOT significant) RoBERTa Baseline vs RoBERTa AAN
- p=3.9e-05 (SIGNIFICANT) DynaSent-M1 Baseline vs DynaSent-M1 AAN

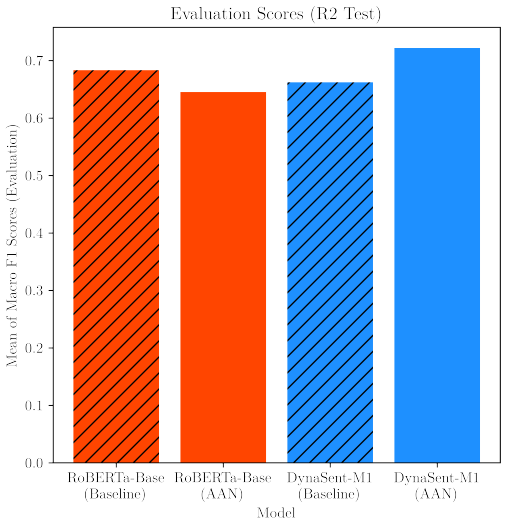


Figure 2: DynaSent Round 1 Test scores for each model.

- p=1.63e-08 (SIGNIFICANT) RoBERTa Baseline vs DynaSent-M1 Baseline
- p=0.0109 (NOT significant) RoBERTa AAN vs DynaSent-M1 AAN

DynaSent Round 2 Test

- p=0.0717 (NOT significant) RoBERTa Baseline vs RoBERTa AAN
- p=0.000231 (SIGNIFICANT) DynaSent-M1 Baseline vs DynaSent-M1 AAN
- p=0.232 (NOT significant) RoBERTa Baseline vs DynaSent-M1 Baseline
- p=0.000222 (SIGNIFICANT) RoBERTa AAN vs DynaSent-M1 AAN

6.2 Qualitative Concept Explainability

We generate concept explanations for a few different types of sentences. For simplicity, we only evaluated the explanations of the DynaSent-M1 AAN model. Here is an explanation of an example sentence from the training set:

Prediction: POSITIVE
Sentence: The food look really bad, it looked as it was over cooked, to my surprise I was highly mistaken.

CID	Score	Keywords
2	0.265	succeeds, exceedingly, enjoyed, enjoys, mesmerizing, succeeded
0	0.196	intrigues, scintillating, brilliance, flawless, glamorous, vibrating
5	0.179	succeeds, scintillating, unconscionable, exhilarated, exhilaration, penetrating
7	0.155	incapacitated, uplink, disinherit, exceedingly, entertained, succeeded
6	0.129	incapacitate, incapacitated, incompetence, incompetent, incomprehensible, whatchamacallit
4	0.068	succeeds, hallelujah, succeeded, exhilaration, enthralled, succeed
8	0.006	souvlaki, horseshit, sauerkraut, dummkopf, whatchamacallit, sarcoidosis
9	0.002	cholinesterase, eucalyptus, disgraced, extraterrestrials, vertebrae, electrolyte
3	0.000	disgraced, incompetence, incapacitated, disgustingly, inconsiderate, disproportionate
1	0.000	suckers, crucifixion, infuriates, inconsiderate, disappoints, wastebasket

6.2.1 Explaining Correct Classifications

Here we explain a few randomly sampled classifications where the model was correct, one for each output class:

Output exceeds the size limit. Open the full output data in a text editor Examples the model correctly predicted 'positive':

Prediction: POSITIVE
Sentence: As for the service I found the staff to do their job appropriately and they were sociable as well.

CID	Score	Keywords
0	0.274	intrigues, fantasizing, scintillating, fantasized, evangelical, fantasize
2	0.263	succeeds, enjoys, exceedingly, mesmerizing, enjoyed, enjoying
5	0.182	scintillating, exhilaration, exhilarated, penetrating, gallivanting, enthusiastic
7	0.137	uplink, entertained, shorthanded, supervising, entertaining, improvising
6	0.072	whatchamacallit, extraordinaire, mausoleum, enthusiastic, fantasizing, enthusiast
4	0.047	hallelujah, succeeds, extravaganza, enthralled, enthusiasm, enthusiastic
9	0.016	cholinesterase, shorthanded, egomaniacal, disgraced, exceedingly, freshening
8	0.008	souvlaki, horseshit, whatchamacallit, applesauce, cacciatore, sauerkraut
1	0.001	suckers, succeeds, motherfuckers, wastebasket, tiresome, boredom
3	0.000	disgraced, whatchamacallit, disproportionate, incapacitated, incompetence, inconveniencing

Prediction: NEUTRAL
Sentence: My kids had a long christmas gift list.

CID	Score	Keywords
5	0.257	preeclampsia, claustrophobia, cacophony, aphrodisiac, toxoplasmosis, smorgasbord
0	0.235	floodgates, cappuccino, gazebo, worshippers, flapjacks, caboose
2	0.162	pinochle, prednisone, honeysuckle, jezebel, mincemeat, rhinoceros
7	0.137	proprietor, blindfolded, cataloging, plowed, hologram, disguised
6	0.094	sauerkraut, hippopotamus, chimpanzee, chihuahua, souvlaki, doorknobs
4	0.066	iambic, telekinesis, disregard, insignia, maneuver, wisecracks
9	0.040	patrolmen, spartan, lunatic, mathematician, littering, diabetics
1	0.005	snowballed, tiresome, aargh, heavyset, foolhardy, labored
8	0.005	souvlaki, sauerkraut, toxoplasmosis, hippopotamus, maggot, sarcoidosis
3	0.000	sauerkraut, toxoplasmosis, sarcoidosis, caterwauling, horseshit, dummkopf

Prediction: NEGATIVE
Sentence: People say she's a great friend to everyone but they don't know her true nature.

CID	Score	Keywords
5	0.278	godforsaken, claustrophobic, untraceable, unconscionable, eviscerate, befall
2	0.229	crucifixion, sourpuss, inaccuracies, devastatingly, unwittingly, cummerbund
0	0.219	shoplifters, loses, uncontrollable, shoplifter, dowager, scrutiny
7	0.132	sacrificing, motherfuckers, misogynistic, disgrace, smugness, unfunny
6	0.091	malfeasance, shoplifters, gobbledygook, murderess, souvenirs, muckraker
4	0.028	revenge, predecessors, arbitration, aargh, outrage, bawl
9	0.022	godforsaken, shoplifters, disgraced, boredom, uncontrollable, incompetence
8	0.002	sauerkraut, dummkopf, souvlaki, bobcat, horseshit, doofus
1	0.000	suckers, crucifixion, inconsiderate, infuriates, wastebasket, pissed
3	0.000	disgraced, incompetence, crucifixion, inconsiderate, disgustingly, mauled

6.2.2 Explaining Opposite Errors

We also explain a few examples where the model produced the *opposite* of the true label. That is, where the model incorrectly predicts 'positive'/'negative' and the gold label is 'negative'/'positive'. Note these are subsets of all false 'positives'/'negatives' since there is a 'neutral' class.

Prediction: POSITIVE
Sentence: The cupcakes were very special for my birthday girl, they would have been better if they were not smashed when they arrived.

CID	Score	Keywords
0	0.273	brilliance, lovesick, flawless, scintillating, intrigues, vibrating
2	0.247	inaccuracies, crucifixion, unbearably, sourpuss, abrasive, incompetence
7	0.172	exceedingly, misogynistic, motherfuckers, succeeded, disgrace, patronize
5	0.151	unconscionable, claustrophobic, succeeds, preeclampsia, inconsolable, inconsiderate
4	0.088	revenge, arbitration, predecessors, unimpressed, disgraced, aargh
6	0.055	malfeasance, whatchamacallit, shoplifters, mausoleum, horseshit, gobbledygook
8	0.007	souvlaki, horseshit, sauerkraut, sarcoidosis, dummkopf, toxoplasmosis
9	0.006	cholinesterase, eucalyptus, extraterrestrials, vertebrae, disgraced, electrolyte
1	0.000	suckers, crucifixion, infuriates, inconsiderate, disappoints, wastebasket
3	0.000	disgraced, incompetence, inconsiderate, disgustingly, crucifixion, motherfuckers

Prediction: NEGATIVE **Sentence**: There is nothing that can be done to make this place better.		
CID	**Score**	**Keywords**
0	0.300	talentless, inconsiderate, disappoints, untraceable, unimaginative, scrutiny
2	0.211	crucifixion, inaccuracies, nausea, sourpuss, incompetence, disgrace
7	0.165	disgrace, motherfuckers, misogynistic, exasperating, infuriates, nausea
6	0.141	incompetence, murderess, shoplifters, misdemeanors, inaccuracies, inconveniencing
5	0.135	unconscionable, inconsiderate, talentless, untraceable, claustrophobic, inconsolable
4	0.025	disgrace, aargh, unimpressed, overhear, briar, incapacitate
9	0.013	disgraced, cholinesterase, eucalyptus, extraterrestrials, disheartening, interfering
8	0.010	horseshit, souvlaki, bungalows, cummerbund, shenanigans, sarcoidosis
1	0.001	suckers, crucifixion, wastebasket, inconsiderate, pissed, boredom
3	0.000	disgraced, inconsiderate, crucifixion, incompetence, motherfuckers, disgustingly

7 Analysis

7.1 Quantitative Model Performance

When analyzing the quantitative performance of our models on the test sets, we find they all perform relatively favorably.

For DynaSent Round 1, going from a black-box model to an AAN model is only a significant decrease in performance for the DynaSent-M1 model, and we also notice that the DynaSent-M1 models perform better than than RoBERTa-Base models (presumably because the original DynaSent Model 1 which we fine-tuned off of was pre-trained on DynaSent Round 1).

For DynaSent Round 2, going from a black-box model to an AAN model is not a significant decrease in performance for RoBERTa-Base, and interestingly it is a significant *increase* in performance for the DynaSent-M1 model. This could be due to random initialization, or less likely, it's possible that learning concepts could act as a form of regularization and improve out-of-distribution generalization.

7.1.1 Qualitative Concept Explainability

Subjectively, our model's explanations are quite bad. It is not immediately clear after reading a concept explanation table how the model came to a specific decision which itself is a failure in interpretability. Furthermore, many of the keywords in each concept table seem unrelated to the input sentence, and each concept's keywords tend to not describe a single cohesive semantic concept. This perhaps implies the models explanations are still quite entangled and not yet human-interpretable.

However, there are a few promising signs from the explanations. Although the concept keywords don't seem fully related to the input sentence, most of them reflect the sentiment of the output decision. We were also able to get a decent amount of diversity in the keywords chosen for each concept and the distributions of concept scores, and a test of increasing the β parameter which weights the diversity loss term from 1 to 2 seems to encourage learning concept diversity sooner in training (about $1/2$ vs $1/4$ of the way through training) and sharing fewer keywords between different concepts. And on occasion, some keywords seem to hint at the internal functionality of the model, e.g. the highest-scoring concept for the sentence "She would refill a drink without us even seeing her do it" (incorrectly classified as 'negative') yields keywords "shoplifters, shoplifter, loses, troublemakers, dowager, farts," which perhaps implies that the model isn't thinking of a waiter who is quick to refill customers' drinks but rather a ruffian who furtively takes free drinks from the soda machine at a restaurant.

8 Conclusion

We build on the Attention-Aggregation Network architectures proposed in (Shi et al., 2020) by adapting them to the task of sentiment classification. In order to address the problem of sentiment classification input sentences having too few words to draw keywords directly from their text as concepts, we propose a novel way of comparing the internal abstraction-attention context representations of a given sentence to a precomputed library of contextual word representations in order to select the n most relevant keywords for a given concept.

Generally, we find that our AAN models maintain favorable quantitative performance but are lacking in the quality and interpretability of their concept-based explanations. Concept-based interpretability for text domains is an important research direction, and future research directions include developing methods for increased coherency and meaningfulness of discovered concepts.

8.1 Quantitative Concept Variety

Quantitatively, how good are the generated concepts at being distinct, varied, and useful for prediction? We aim to empirically evalute this with the following techniques:

- Calculating how disentangled different concepts are, e.g. via comparing cross-concept word co-occurrences.

- Graphing the distribution of tokens frequency use in the concepts in train, development and test sets to determine how well the concept distributions generalize.

Known Project Limitations

Our results conclude that performance using a concept-based interpretable model is comparable to a black-box state-of-the-art model. However, notably, the concepts seem to have limited coherency. Prior works (Ghorbani et al., 2019; Yeh et al., 2019) evaluate the meaningfulness of discovered according to various criteria, including sufficiency, coherency, semantic meaningfulness, and importance of the concepts. Our results do not concretely evaluate the discovered text concepts with respect to these criteria, and the results sections suggests that some of the discovered concepts may be incoherent.

Authorship Statement

Gabe worked on the model, methods and evaluation. Liana worked on the introduction, related works section and the evaluation. Atharva worked on the implementation of the methods.

References

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

Asma Ghandeharioun, Been Kim, Chun-Liang Li, Brendan Jou, Brian Eoff, and Rosalind W. Picard. 2021. DISSECT: disentangled simultaneous explanations via concept traversals. *CoRR*, abs/2105.15164.

Amirata Ghorbani, James Wexler, James Zou, and Been Kim. 2019. Towards automatic concept-based explanations.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory Sayres. 2017. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). *ICML*, pages 2668–2677.

Peter A Lachenbruch. 2014. Mcnemar test. *Wiley StatsRef: Statistics Reference Online*.

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.

Christopher Potts, Zhengxuan Wu, Atticus Geiger, and Douwe Kiela. 2020. DynaSent: A dynamic benchmark for sentiment analysis. *arXiv preprint arXiv:2012.15349*.

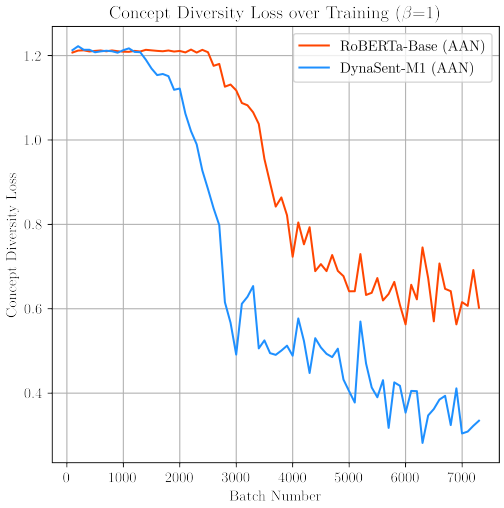
Tian Shi, Xuchao Zhang, Ping Wang, and Chandan K. Reddy. 2020. Corpus-level and concept-based explanations for interpretable document classification.

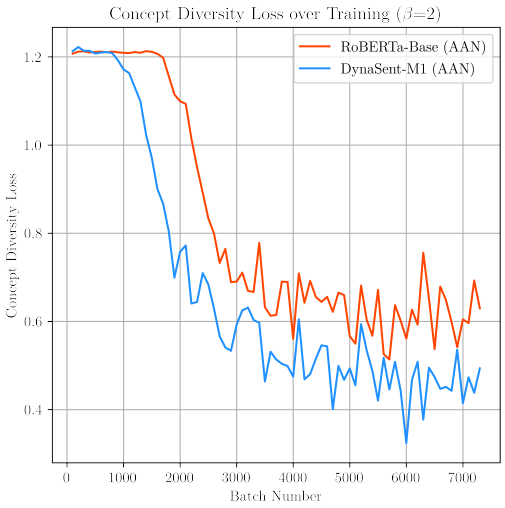
Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

Chih-Kuan Yeh, Been Kim, Sercan O. Arik, Chun-Liang Li, Tomas Pfister, and Pradeep Ravikumar. 2019. On completeness-aware concept-based explanations in deep neural networks.

A Appendix

A.1 Changing Beta-Weight





A.2 Additional Concept Explanations

A.2.1 Training/Handwritten Examples

Prediction: NEGATIVE
Sentence: If they said it would be excellent, then why did the food taste that way?

CID	Score	Keywords
5	0.222	inconsiderate, talentless, untraceable, incapacitate, disappoints, unimaginative
7	0.181	incapacitate, incompetent, uncooperative, inconsiderate, shoplifters, disappoints
6	0.169	incompetence, incompetent, misdemeanors, murderess, shoplifters, scrutiny
0	0.168	talentless, disappoints, shoplifters, scrutiny, frustrates, flaws
2	0.162	crucifixion, inaccuracies, sourpuss, nausea, incompetence, infuriating
4	0.091	sordid, aargh, incessantly, incapacitate, bogus, disgrace
9	0.004	disgraced, cholinesterase, boredom, interfering, disheartening, irritated
8	0.003	horseshit, souvlaki, bobcat, disinherit, doofus, dummkopf
3	0.000	disgraced, crucifixion, inconsiderate, incompetence, disgustingly, motherfuckers
1	0.000	suckers, crucifixion, infuriates, inconsiderate, disappoints, wastebasket

Prediction: POSITIVE
Sentence: Customer service is fast!

CID	Score	Keywords
0	0.200	lovesick, vibrating, brilliance, dynamic, flawless, grinning
2	0.151	loved, liked, excels, exceeded, nailed, appreciates
6	0.124	excellent, genius, awesome, profound, phenomena, immense
8	0.124	glamour, pamper, sophistication, glamor, enchantment, cutesy
7	0.123	whirlwind, gorgeous, vivid, stunningly, wicked, triumphs
4	0.101	exceed, blockbusters, awol, miraculous, enthusiasm, thanking
5	0.095	perfecting, addictive, best, perfectly, stylish, perfection
9	0.068	sublime, stunning, excellent, spectacular, trusting, superior
1	0.014	succeeds, magnificent, triumphs, succeeded, excels, hallelujah
3	0.001	seamless, triumphs, magnificent, gorgeous, breathtaking, excels

Prediction: NEUTRAL
Sentence: My doctor is a woman.

CID	Score	Keywords
0	0.196	floodgates, caboose, gazebo, flapjacks, worshippers, tattoos
5	0.179	chimpanzee, lidocaine, crocodile, hippopotamus, bachelorette, rhinoceros
6	0.176	coconut, wrestler, nocturnal, sixpence, fibre, biochemist
7	0.171	rattlesnake, puppeteer, newlyweds, wrestler, nocturnal, coconut
2	0.146	prednisone, journeyed, sleet, jezebel, rhinoceros, honeysuckle
4	0.056	insignia, disregard, smashes, unquote, relieved, disbelief
9	0.040	cholinesterase, aforementioned, remainder, nocturnal, extraterrestrials, unattached
8	0.022	bobcat, souvlaki, angioplasty, pantyhose, hippopotamus, hayloft
1	0.014	boogeyman, snowballed, unscheduled, inconspicuous, skewed, malfeasance
3	0.000	toxoplasmosis, caterwauling, sarcoidosis, sauerkraut, dummkopf, horseshit

Prediction: NEUTRAL
Sentence: My doctor is a man.

CID	Score	Keywords
9	0.192	undisclosed, aforementioned, unattached, remainder, nearby, dissident
0	0.188	floodgates, caboose, gazebo, flapjacks, tattoos, worshippers
7	0.151	diabetics, biochemist, closest, mathematician, snowmobile, wrestler
6	0.150	biochemist, snowmobile, mathematician, anesthesiologist, stereotype, venetian
2	0.141	prednisone, journeyed, sleet, jezebel, rhinoceros, bridegroom
5	0.073	puppeteer, hayloft, psychotherapist, evolution, windsurfing, zebra
4	0.064	insignia, smashes, unquote, disregard, relieved, disbelief
8	0.027	bobcat, souvlaki, hippopotamus, braggart, angioplasty, maggot
1	0.013	boogeyman, snowballed, inconspicuous, skewed, malfeasance, unscheduled
3	0.000	caterwauling, toxoplasmosis, sarcoidosis, dummkopf, sauerkraut, horseshit

Prediction: NEGATIVE
Sentence: The egg drop soup was as delicious as the long-forgotten, untouched fruitcake that my mother made for me decades ago!

CID	Score	Keywords
0	0.247	talentless, shoplifters, disappoints, uncontrollable, frustrates, scrutiny
2	0.214	crucifixion, inaccuracies, sourpuss, nausea, incompetence, opportune
5	0.195	unconscionable, inconsiderate, talentless, inconsolable, claustrophobic, untraceable
7	0.163	misogynistic, motherfuckers, disgrace, infuriates, nausea, exasperating
6	0.104	shoplifters, malfeasance, murderess, incompetence, inconveniencing, misdemeanors
4	0.067	sordid, aargh, unimpressed, incapacitate, revenge, disgrace
8	0.007	horseshit, souvlaki, sarcoidosis, bobcat, angioplasty, sauerkraut
9	0.004	cholinesterase, eucalyptus, extraterrestrials, disgraced, vertebrae, electrolyte
3	0.000	disgraced, incompetence, crucifixion, inconsiderate, disgustingly, motherfuckers
1	0.000	crucifixion, infuriates, suckers, inconsiderate, disappoints, disgustingly

Prediction: POSITIVE
Sentence: The food look really bad, it looked as it was over cooked, to my surprise I was highly mistaken.

CID	Score	Keywords
2	0.265	succeeds, exceedingly, enjoyed, enjoys, mesmerizing, succeeded
0	0.196	intrigues, scintillating, brilliance, flawless, glamorous, vibrating
5	0.179	succeeds, scintillating, unconscionable, exhilarated, exhilaration, penetrating
7	0.155	incapacitated, uplink, disinherit, exceedingly, entertained, succeeded
6	0.129	incapacitate, incapacitated, incompetence, incompetent, incomprehensible, whatchamacallit
4	0.068	succeeds, hallelujah, succeeded, exhilaration, enthralled, succeed
8	0.006	souvlaki, horseshit, sauerkraut, dummkopf, whatchamacallit, sarcoidosis
9	0.002	cholinesterase, eucalyptus, disgraced, extraterrestrials, vertebrae, electrolyte
3	0.000	disgraced, incompetence, incapacitated, disgustingly, inconsiderate, disproportionate
1	0.000	suckers, crucifixion, infuriates, inconsiderate, disappoints, wastebasket

Prediction: POSITIVE
Sentence: The staff has always been friendly and willing to help with any questions I've had.

CID	Score	Keywords
2	0.273	succeeds, exceedingly, enjoys, enjoyed, mesmerizing, enamored
0	0.222	brilliance, scintillating, intrigues, glamorous, vibrating, glamour
5	0.174	scintillating, succeeds, exhilaration, exhilarated, penetrating, gallivanting
7	0.132	exceedingly, succeeded, succeeds, wondrous, intrigue, sophisticated
8	0.089	whatchamacallit, hallelujah, exhilaration, extraordinaire, extravaganza, uplink
6	0.059	whatchamacallit, extraordinaire, mausoleum, enthusiastic, enthusiast, congeniality
4	0.034	hallelujah, succeeds, extravaganza, enthusiasm, enthralled, enthusiastic
9	0.017	shorthanded, disgraced, evangelical, comprehensive, enamored, unforgiving
1	0.001	suckers, succeeds, motherfuckers, tiresome, wastebasket, boredom
3	0.000	disgraced, disproportionate, incapacitated, incompetence, whatchamacallit, inconveniencing

Prediction: POSITIVE
Sentence: My boy Anthony good with them clippers if you go here ask for him. He is cool.

CID	Score	Keywords
2	0.292	mesmerizing, succeeds, enchantment, engrossed, enjoys, enjoyed
5	0.245	scintillating, succeeds, gallivanting, exhilaration, admirable, penetrating
0	0.207	brilliance, lovesick, glamour, scintillating, vibrating, glamorous
7	0.105	exceedingly, succeeded, sophisticated, wondrous, bombshell, succeeds
4	0.081	enthusiasm, hallelujah, extravaganza, entrapment, rebuttal, admirably
6	0.064	whatchamacallit, extraordinaire, mausoleum, enthusiastic, enthusiast, applesauce
8	0.003	souvlaki, whatchamacallit, horseshit, sauerkraut, hallelujah, dumdum
9	0.002	eucalyptus, cholinesterase, extraterrestrials, shorthanded, egomaniacal, disgraced
1	0.002	suckers, succeeds, motherfuckers, tiresome, wastebasket, boredom
3	0.000	disgraced, disproportionate, incapacitated, incompetence, inconveniencing, motherfuckers

A.2.2 Explaining Correct Classifications

Prediction: POSITIVE
Sentence: Friendly staff, both front and back office. nice place

CID	Score	Keywords
2	0.351	enjoyed, enjoying, enriching, embracing, uplink, enchanting
0	0.208	vibrating, inventive, intrigues, addictive, glamour, charming
7	0.166	wondrous, admiration, glamorous, inspirational, brilliant, wonderfully
5	0.103	scintillating, exhilaration, exhilarated, penetrating, gallivanting, succeeds
4	0.101	hallelujah, extravaganza, succeeds, enthusiastic, enthralled, uplink
6	0.050	whatchamacallit, mausoleum, extraordinaire, enthusiastic, fantasizing, enthusiast
8	0.012	whatchamacallit, hallelujah, souvlaki, applesauce, horseshit, mausoleum
9	0.006	cholinesterase, eucalyptus, extraterrestrials, egomaniacal, electrolyte, vertebrae
1	0.004	succeeds, motherfuckers, suckers, tiresome, succeeded, cribbage
3	0.000	whatchamacallit, disgraced, succeeds, disproportionate, succeeded, impeccable

Prediction: POSITIVE
Sentence: Don't come here if you're looking for a terrible experience with staff.

CID	Score	Keywords
5	0.265	scintillating, exhilaration, exhilarated, penetrating, exhilarating, enthusiastic
2	0.202	mesmerizing, whatchamacallit, extraordinaire, enchantment, enthusiastic, entertaining
0	0.170	geniuses, intrigues, uplifting, extraordinaire, extravaganza, inspiration
6	0.122	inconveniencing, whatchamacallit, extraordinaire, enthusiastic, shoplifters, incompetence
7	0.113	motherfuckers, disgrace, misogynistic, sacrificing, imperfection, smugness
9	0.106	disgraced, freshening, disfiguring, shorthanded, egomaniacal, disheartening
4	0.020	outsmarted, arbitration, revenge, predecessors, unimpressed, disgraced
8	0.002	souvlaki, horseshit, sauerkraut, dummkopf, sarcoidosis, toxoplasmosis
1	0.000	suckers, crucifixion, inconsiderate, wastebasket, pissed, pisses
3	0.000	disgraced, incompetence, inconsiderate, incapacitated, motherfuckers, disgustingly

Prediction: POSITIVE
Sentence: If I'm ever back in the Phoenix/Scottsdale area I'll definitely pay it a visit again. love it

CID	Score	Keywords
2	0.352	scintillating, enjoyed, mesmerizing, enjoying, enriching, entertained
0	0.265	intrigues, scintillating, glamorous, brilliance, fantasizing, flawless
7	0.156	succeeded, succeeds, exceedingly, intrigue, gorgeous, bombshell
5	0.148	succeeds, scintillating, exhilaration, exhilarated, enriching, uplink
4	0.038	succeeds, hallelujah, extravaganza, enthralled, exhilaration, enthusiastic
6	0.027	whatchamacallit, extraordinaire, mausoleum, enthusiastic, enthusiast, malfeasance
9	0.011	cholinesterase, egomaniacal, eucalyptus, shorthanded, freshening, disgraced
8	0.003	whatchamacallit, souvlaki, horseshit, hallelujah, sauerkraut, applesauce
1	0.001	suckers, wastebasket, tiresome, motherfuckers, boredom, pissed
3	0.000	disgraced, disproportionate, incompetence, incapacitated, inconveniencing, motherfuckers

Prediction: POSITIVE
Sentence: It was too good

CID	Score	Keywords
0	0.169	vibrating, lovesick, brilliance, glamour, grinning, flawless
2	0.151	loved, exceedingly, liked, succeeds, appreciates, amuses
6	0.139	enamored, fascinated, invigorated, fantastic, informal, fascinate
7	0.123	treasured, beloved, excellent, distinguished, tasteful, superb
4	0.110	enamored, exhilaration, terrifically, wondrous, entertaining, fascinated
8	0.103	enamored, brilliance, nourishing, entertaining, wondrous, funniest
5	0.098	phenomena, brilliance, flattering, mastermind, delights, phenomenon
9	0.092	fascinated, invigorated, fantastic, intrigued, satisfying, tasteful
1	0.014	succeeds, magnificent, excels, succeeded, triumphs, hallelujah
3	0.000	terrifically, breathtaking, impeccable, magnificent, excels, effortless

Prediction: NEUTRAL
Sentence: The closest thing to it is an average shoe release that happens not very often but not too often.

CID	Score	Keywords
0	0.270	floodgates, worshippers, sharpshooters, gazebo, cacophony, cappuccino
2	0.259	rattlesnake, honeysuckle, sauerkraut, pinochle, jezebel, neighborhoods
5	0.254	preeclampsia, claustrophobia, cacophony, aphrodisiac, anorexic, claustrophobic
7	0.150	blindfolded, somerset, proprietor, cataloging, sanctioned, beheading
4	0.034	iambic, doppelganger, telekinesis, revenge, chimpanzee, disbelief
6	0.028	sauerkraut, gobbledygook, horseshit, muckraker, chihuahua, souvlaki
9	0.003	cholinesterase, eucalyptus, extraterrestrials, vertebrae, electrolyte, reprobate
8	0.001	souvlaki, toxoplasmosis, sauerkraut, sarcoidosis, dummkopf, horseshit
1	0.001	suckers, crucifixion, wastebasket, tiresome, boredom, pissed
3	0.000	disgraced, disproportionate, incompetence, inconveniencing, betrayed, sauerkraut

Prediction: NEUTRAL
Sentence: She told me to wait to get the charge reversed as the guy was in the bathroom.

CID	Score	Keywords
5	0.265	tomfoolery, aphrodisiac, boogeyman, chimpanzee, cacophony, journeyed
0	0.256	thoracotomy, floodgates, focussing, cacophony, worshippers, caboose
2	0.165	rattlesnake, prednisone, journeyed, apprehended, somerset, claustrophobia
7	0.148	blindfolded, cordoned, beheading, cataloging, hiding, disguised
6	0.082	sauerkraut, claustrophobia, gobbledygook, muckraker, lidocaine, dummkopf
4	0.048	flinching, iambic, revenge, doppelganger, chimpanzee, disbelief
9	0.035	cholinesterase, extraterrestrials, eucalyptus, aforementioned, vertebrae, electrolyte
8	0.001	toxoplasmosis, souvlaki, sauerkraut, dummkopf, sarcoidosis, horseshit
3	0.000	toxoplasmosis, sauerkraut, disgraced, disproportionate, snowballed, eviscerated
1	0.000	infuriates, crucifixion, suckers, disgustingly, inconsiderate, disgraced

Prediction: NEUTRAL
Sentence: My husband and I left without any strong thoughts or feelings about this place.

CID	Score	Keywords
2	0.304	powdered, explanatory, transcends, moisturizer, nutcracker, applauded
0	0.284	cappuccino, videotaped, floodgates, shoelaces, seashells, heirlooms
7	0.205	barbecued, proprietor, electrocuted, unharmed, heirlooms, theoretically
4	0.110	iambic, squealed, eucalyptus, telekinesis, souvlaki, lunchroom
6	0.047	sauerkraut, mausoleum, gobbledygook, muckraker, cacciatore, angioplasty
9	0.034	eucalyptus, extraterrestrials, electrolyte, cholinesterase, vertebrae, reprobate
5	0.014	preeclampsia, claustrophobic, claustrophobia, unconscionable, smorgasbord, anorexic
8	0.002	souvlaki, sarcoidosis, toxoplasmosis, sauerkraut, horseshit, dummkopf
1	0.001	suckers, crucifixion, wastebasket, tiresome, boredom, pissed
3	0.000	disgraced, disproportionate, incompetence, inconveniencing, incapacitated, motherfuckers

Prediction: NEUTRAL
Sentence: It is actually located a good distance from the Pacific Ocean.

CID	Score	Keywords
5	0.297	anorexic, aphrodisiac, thespian, hydrochloride, cacophony, ventriloquist
0	0.245	floodgates, cappuccino, worshippers, gazebo, sharpshooters, flapjacks
2	0.223	pinochle, honeysuckle, jezebel, neighborhoods, rattlesnake, prednisone
7	0.136	proprietor, chardonnay, unharmed, tangerine, unbeknownst, barbecued
4	0.046	iambic, telekinesis, doppelganger, squealed, revenge, disregard
6	0.046	sauerkraut, horseshit, souvlaki, gobbledygook, chimpanzee, dummkopf
1	0.003	suckers, tiresome, wastebasket, motherfuckers, warped, foolhardy
9	0.002	cholinesterase, eucalyptus, extraterrestrials, vertebrae, electrolyte, nocturnal
8	0.002	souvlaki, sauerkraut, toxoplasmosis, dummkopf, smorgasbord, horseshit
3	0.000	sauerkraut, disgraced, disproportionate, toxoplasmosis, inconveniencing, incompetence

Prediction: NEUTRAL
Sentence: I'll take less pics then.

CID	Score	Keywords
0	0.242	floodgates, caboose, gazebo, focussing, flapjacks, remodeled
2	0.161	prednisone, sleet, journeyed, pinochle, somerset, jezebel
5	0.141	beige, previous, pronounced, unmarried, presumed, disservice
7	0.134	restructuring, hologram, diabetics, mongoose, sentinels, dissident
9	0.134	disguised, aforementioned, overhead, redhead, clutched, nonchalant
6	0.128	stereotype, diabetics, diazepam, restructuring, mongoose, lobotomy
4	0.052	disbelief, smashes, disregard, revenge, insignia, unquote
1	0.005	snowballed, foolhardy, tiresome, labored, aargh, shoplifters
8	0.003	souvlaki, toxoplasmosis, angioplasty, shenanigans, sarcoidosis, pantyhose
3	0.000	toxoplasmosis, caterwauling, sarcoidosis, dummkopf, horseshit, angioplasty

A.2.3 Explaining Opposite Errors

Examples the model predicted 'positive' but were truly 'negative':

Prediction: POSITIVE
Sentence: The regular items were a bit pricey. WE like cheap.

CID	Score	Keywords
5	0.213	unconscionable, gallivanting, intimidating, exhilarated, penetrating, believable
2	0.199	succeeds, exceedingly, patronized, unbearably, disillusioned, inundated
0	0.192	lovesick, brilliance, flawless, vibrating, scintillating, glamour
7	0.136	bombshell, unappreciated, blessed, exceedingly, brilliant, amazingly
6	0.135	whatchamacallit, malfeasance, congeniality, applesauce, extraordinaire, unscheduled
4	0.106	predecessors, killjoy, arbitration, unappreciated, idolized, revenge
8	0.010	souvlaki, horseshit, cacciatore, shenanigans, sugarplum, bungalows
9	0.007	cholinesterase, disgraced, eucalyptus, extraterrestrials, disheartening, egomaniacal
1	0.002	suckers, tiresome, motherfuckers, wastebasket, boredom, warped
3	0.000	disgraced, disproportionate, incompetence, incapacitated, inconveniencing, motherfuckers

Prediction: POSITIVE

Sentence: Add bacon and then drowned them in butter, which was lovely for a vegan.

CID	Score	Keywords
2	0.297	succeeds, exceedingly, enjoys, mesmerizing, enjoyed, engrossed
0	0.259	scintillating, brilliance, intrigues, glamorous, glamour, vibrating
5	0.161	scintillating, succeeds, exhilaration, exhilarated, penetrating, gallivanting
7	0.133	succeeded, exceedingly, succeeds, wondrous, intrigue, bombshell
6	0.072	whatchamacallit, extraordinaire, mausoleum, enthusiastic, enthusiast, fantasizing
4	0.069	hallelujah, succeeds, extravaganza, enthusiasm, enthralled, enthusiastic
8	0.006	whatchamacallit, hallelujah, souvlaki, horseshit, sauerkraut, applesauce
9	0.002	shorthanded, unforgiving, evangelical, unappreciated, unimpressed, comprehensive
1	0.001	suckers, wastebasket, motherfuckers, tiresome, succeeds, boredom
3	0.000	disgraced, disproportionate, incompetence, incapacitated, inconveniencing, whatchamacallit

Prediction: POSITIVE

Sentence: We were expecting wonderful friendly staff since it is minutes away from my home but we were surprised it was the oppositie

CID	Score	Keywords
2	0.200	succeeds, exceedingly, disgraced, inaccuracies, enjoys, incapacitate
4	0.195	succeeds, hallelujah, succeeded, exhilaration, succeed, breathtaking
5	0.192	unconscionable, scintillating, exhilaration, exhilarated, penetrating, extraordinaire
0	0.157	scintillating, intrigues, brilliance, flawless, glamorous, lovesick
9	0.131	disgraced, freshening, exceedingly, shorthanded, enthralled, impeccable
7	0.091	disinherit, disgrace, unimpressed, motherfuckers, imperfection, sacrificing
6	0.032	whatchamacallit, malfeasance, mausoleum, extraordinaire, shoplifters, horseshit
8	0.002	souvlaki, sauerkraut, horseshit, dummkopf, whatchamacallit, toxoplasmosis
3	0.000	disgraced, incompetence, incapacitated, inconsiderate, disgustingly, crucifixion
1	0.000	suckers, crucifixion, infuriates, inconsiderate, disappoints, wastebasket

Prediction: POSITIVE

Sentence: I tried Cousins Maine Lobster for the first time during the Farmers Market at Tivoli Village, and I was completely blown away by how expensive it was..

CID	Score	Keywords
0	0.244	scintillating, intrigues, brilliance, glamorous, flawless, vibrating
2	0.241	succeeds, inaccuracies, exceedingly, crucifixion, disgraced, incompetence
4	0.222	succeeds, hallelujah, succeeded, exhilaration, exhilarated, succeed
7	0.173	succeeded, exceedingly, succeeds, disgrace, unimpressed, exceeded
6	0.046	whatchamacallit, mausoleum, extraordinaire, malfeasance, shoplifters, horseshit
5	0.044	unconscionable, preeclampsia, succeeds, claustrophobic, claustrophobia, scintillating
9	0.027	disgraced, cholinesterase, exceedingly, disheartening, eucalyptus, egomaniacal
8	0.004	souvlaki, horseshit, sauerkraut, dummkopf, whatchamacallit, sarcoidosis
3	0.000	disgraced, incompetence, inconsiderate, incapacitated, motherfuckers, disgustingly
1	0.000	suckers, crucifixion, infuriates, inconsiderate, disappoints, wastebasket

Examples the model predicted 'negative' but were truly 'positive':

Prediction: NEGATIVE

Sentence: Last time we were here was a year ago and the waitress was horrible and the food was just as horrible. Now they have new staff and everything was peaches and cream.

CID	Score	Keywords
0	0.267	talentless, disappoints, unimaginative, shoplifters, inconsiderate, scrutiny
2	0.246	crucifixion, inaccuracies, nausea, incompetence, incapacitate, disgrace
5	0.178	unconscionable, inconsiderate, talentless, untraceable, claustrophobic, inconsolable
7	0.165	motherfuckers, disgrace, misogynistic, nausea, infuriates, exasperating
6	0.104	incompetence, shoplifters, murderess, inconveniencing, misdemeanors, malfeasance
4	0.024	unimpressed, incapacitate, revenge, disgrace, sordid, aargh
9	0.012	disgraced, cholinesterase, eucalyptus, interfering, disheartening, butchered
8	0.005	horseshit, souvlaki, sauerkraut, bobcat, dummkopf, sarcoidosis
1	0.000	suckers, crucifixion, infuriates, inconsiderate, disappoints, wastebasket
3	0.000	disgraced, inconsiderate, incompetence, crucifixion, disgustingly, motherfuckers

Prediction: POSITIVE

Sentence: Only people who lack good taste would eat at the new restaurant.

CID	Score	Keywords
5	0.243	unconscionable, inconsiderate, scintillating, exhilaration, exhilarated, incomprehensible
0	0.242	geniuses, intrigues, flawlessly, flawless, disgraced, scintillating
2	0.232	crucifixion, inaccuracies, disgraced, incapacitate, incompetence, succeeds
7	0.165	disgrace, misogynistic, imperfection, motherfuckers, unimpressed, infuriates
4	0.059	predecessors, disgrace, arbitration, unimpressed, displeasure, revenge
6	0.043	whatchamacallit, malfeasance, shoplifters, mausoleum, murderess, extraordinaire
8	0.010	souvlaki, horseshit, sarcoidosis, cacciatore, shenanigans, sugarplum
9	0.004	disgraced, eucalyptus, boredom, cholinesterase, disheartening, extraterrestrials
1	0.001	suckers, crucifixion, wastebasket, pissed, inconsiderate, boredom
3	0.000	disgraced, incompetence, inconsiderate, motherfuckers, incapacitated, disgustingly

Prediction: NEGATIVE

Sentence: Sorry - you won't be gifted with the sight of hair from previous guests in the sink at this B&B.

CID	Score	Keywords
7	0.289	godforsaken, disinherit, incapacitated, sacrificing, disappointments, motherfuckers
0	0.253	shoplifters, uncontrollable, scrutiny, talentless, disappoints, loses
2	0.222	crucifixion, inaccuracies, sourpuss, devastatingly, unbearably, incompetence
5	0.142	unconscionable, claustrophobic, claustrophobia, preeclampsia, inconsiderate, talentless
6	0.051	shoplifters, malfeasance, whatchamacallit, mausoleum, horseshit, sauerkraut
4	0.031	revenge, arbitration, predecessors, unimpressed, aargh, incapacitate
9	0.008	cholinesterase, eucalyptus, extraterrestrials, vertebrae, disgraced, electrolyte
8	0.004	souvlaki, horseshit, sauerkraut, dummkopf, sarcoidosis, toxoplasmosis
1	0.000	suckers, crucifixion, inconsiderate, infuriates, wastebasket, pissed
3	0.000	disgraced, crucifixion, incompetence, inconsiderate, disgustingly, motherfuckers

Prediction: NEGATIVE

Sentence: They treat you like you are the only and most important patient.

CID	Score	Keywords
2	0.231	crucifixion, inaccuracies, sourpuss, nausea, incompetence, unbearably
7	0.204	sacrificing, motherfuckers, disgrace, misogynistic, imperfection, sordid
5	0.204	unconscionable, claustrophobic, claustrophobia, talentless, inconsiderate, eviscerate
0	0.201	shoplifters, loses, lacks, flaws, frustrates, talentless
6	0.102	malfeasance, shoplifters, souvenirs, sabotage, murderess, incompetence
4	0.041	revenge, predecessors, arbitration, aargh, outrage, briar
8	0.009	horseshit, disinherit, bobcat, souvlaki, cummerbund, inconveniencing
9	0.007	disgraced, cholinesterase, eucalyptus, disheartening, boredom, interfering
1	0.000	suckers, crucifixion, wastebasket, inconsiderate, pissed, pisses
3	0.000	disgraced, incompetence, inconsiderate, crucifixion, disgustingly, motherfuckers

Prediction: NEGATIVE

Sentence: It was so crazy, I was super scared, what a haunted house to go to on Holloween!

CID	Score	Keywords
0	0.258	loses, shoplifters, talentless, scrutiny, lacks, flaws
2	0.189	crucifixion, inaccuracies, sourpuss, misdemeanors, devastatingly, opportune
5	0.181	unconscionable, claustrophobic, claustrophobia, preeclampsia, talentless, inconsiderate
7	0.167	motherfuckers, misogynistic, sacrificing, marshmallows, smugness, disgrace
6	0.105	souvenirs, malfeasance, shoplifters, sabotage, murderess, horseshit
4	0.084	aargh, revenge, sordid, briar, predecessors, incessantly
9	0.014	cholinesterase, disgraced, eucalyptus, extraterrestrials, vertebrae, boredom
8	0.003	souvlaki, dummkopf, horseshit, sauerkraut, toxoplasmosis, sarcoidosis
1	0.000	suckers, crucifixion, inconsiderate, infuriates, wastebasket, disappoints
3	0.000	crucifixion, disgraced, inconsiderate, incompetence, disgustingly, motherfuckers

Prediction: NEGATIVE

Sentence: I would hate to not come out here again!

CID	Score	Keywords
5	0.255	inconsiderate, talentless, incapacitate, disappoints, incapacitated, disappointing
2	0.191	inaccuracies, crucifixion, nausea, infuriating, sourpuss, incompetence
6	0.187	incompetence, inconveniencing, inaccuracies, misdemeanors, murderess, shoplifters
0	0.184	talentless, frustrates, disappoints, flaws, unimaginative, shoplifters
7	0.155	exasperating, infuriates, misogynistic, nausea, boredom, motherfuckers
4	0.014	revenge, unimpressed, incapacitate, disgrace, arbitration, sordid
9	0.009	disgraced, cholinesterase, eucalyptus, interfering, boredom, irritated
8	0.004	horseshit, souvlaki, sarcoidosis, shenanigans, sauerkraut, cummerbund
1	0.001	suckers, crucifixion, inconsiderate, wastebasket, pissed, pisses
3	0.000	inconsiderate, crucifixion, disgraced, disgustingly, incompetence, motherfuckers