Implementing Concept Bottlenecks for NLP Tasks

Josh Ludan, Matthew Pressimone, Saurabh Shah

Abstract

Concept bottleneck models allow for end-to-end deep learning systems to have an added level of interpretability at the concept level. In this paper, we propose a generalization on concept bottleneck models: concept spaces, which iteratively leverages a large language model to come up with new concepts to improve performance on the given task (be it supervised or unsupervised). The concepts are generated by sampling training examples where performance was previously worst, adding the examples to an LLM prompt that includes the task, and outputting the new concept. We use concept space embeddings on a variety of datasets to test its performance on different clustering, classification, and regression tasks.

Background

In concept bottleneck models, Koh et al. bring up the issue that the end to end nature of most deep learning systems currently make it difficult to interpret how they work. This makes it hard for researchers to interact with these systems to perform things like counterfactual reasoning or debiasing. To solve this, they propose creating “concept bottlenecks” where the inputs get transformed into human interpretable features which practitioners can analyze and intervene on.

Much of the work with respect to concept bottleneck models in the past few years have focused on the domain of image classification. In terms of work focused in the domain of natural language processing, in Goal Driven Discovery of Distributional Differences via Language Descriptions, Zhong et al. formulate the D5 task where the goal is to automatically discover differences between two different text corpora in service of some goal. These differences are expressed in natural language making them human interpretable. In one example task, the system is shown side effects from two different drugs and the goal is to discover side effects present in one drug but not the other (ex: drug A tends to mention symptoms associated with paranoia). We can describe the D5 task as a concept bottleneck in the restricted domain of binary classification (predicting whether a text belongs to corpus A vs corpus B) where the concepts are the differences discovered. Beyond this work, little else has been published with regards to concept bottlenecks in natural language processing tasks in the current deep

learning paradigm. The most significant one seen so far seems to be CHiLL: Zero-shot Custom Interpretable Feature Extraction from Clinical Notes with Large Language Models, in which McInerney et al. come up with the concepts and the LLM only computes its values on each example in the dataset.

We note that part of the reason why little work has been done in this space is because the problems being solved by concept bottlenecks in NLP are being solved via other methods. In terms of interpretability, you have attention-based methods and concept probing. In the domain of classifying texts, *topic modeling* tools allow users to know which specific tokens increase the likelihood of a given classification for a topic. However, we note that the unit of explanation in these models comes at the token level rather than at the concept level. This means that these explanations are hard to fully comprehend given that they are not *sparse*. Additionally, in CEBaB: Estimating the Causal Effects of Real-World Concepts on NLP Model Behavior, Abraham et al. evaluate how many different nlp interpretation methods that work on the concept level perform and find that many of them fail to truly capture the effect of concepts.

In this paper, we introduce the notion of *concept spaces* which generalize the notion of concept bottlenecks to several NLP task domains which include classification, regression, and clustering. Under our system, we use a language model to iteratively come up with concepts to grow the *concept space*. We note that in our system, users are theoretically given arbitrary control over how specific and fine-grained their concepts can be given that they can adjust the prompting which controls the concept generation. In our current implementation, we configure the system to generate sparse, high-level concepts. As an example, in the domain of restaurant rating, we generate concepts to measure aspects such as *ambiance* and *service quality* which are evaluated by language models.

Currently, we evaluate our system on a suite of NLP tasks ranging from poetry sentiment analysis to amazon product reviews.

Methodology

Our system starts off with an empty concept space which we grow iteratively. In each iteration, the concepts generated are designed to be distinct or provide additional information relative to the other concepts. Additionally, we steer this concept generation process so that the concepts generated are useful for downstream tasks. Our concept generation process is based around prompting a large language model.

We define a concept via a “concept json” which contains the following elements. Below we show an example of what these look like for a concept created for evaluating product reviews:

- Concept name (`good build quality`)

- Concept description (`build quality refers to the craftsmanship, durability, and overall construction of a product. It encompasses aspects such as materials used, design, manufacturing techniques, and attention to detail. A product with good build quality is typically considered to be well-made, sturdy, and long-lasting, while a product with poor build quality may be prone to defects or wear out quickly.`)
- Concept measurement question (`Does the review mention the build quality in a positive way?`)
- Valid responses to measurement (`["positive", "negative", "unknown"]`)
- Response Guide (`{`
  `"positive": "The review explicitly mentions or implies the product has good build quality, such as well-made, sturdy, or long-lasting construction.",`
  `"negative": "The review explicitly mentions or implies the product has poor build quality, such as being prone to defects or wearing out quickly.",`
  `"unknown": "The review does not mention anything about the build quality nor is there any information about the build quality."}`)

To generate these concepts for a given dataset task, we generate a prompt that takes in the following elements as an input (an example of this full concept generation prompt is attached in the appendix)

- Instruction set: This is where we specify the nature of the dataset and the goal we want to achieve. If our dataset contains a label, this is where we explain the labels.
- Example concepts: We are performing concept generation in a semi-few shot manner where we demonstrate what valid outputs look like. In the concept above, we have a concept that has “positive”,”negative”, and “unknown” as possible values for the concept. It is also possible to prompt the creation of concepts that are purely binary, or are based on a rating scale instead by loading in concept examples with the response flexibility we desire. We note that changing the example concepts heavily influences the types of concepts that the system will search over.
- Examples from the dataset: Similar to Zhong et al, we prompt the language model to find differences between two groups of text from the dataset to make sure that the concepts being generated by the system are valid on the dataset. We ask it to identify differences in the text associated with each group. We also have the notion of “informative sampling” where we prioritize feeding in examples that are identical in the concept space yet highly misclassified to bias the system towards generating concepts that are unique in each iteration yet useful for the end goal. For regression and classification, we have two ways of doing this:
  - Tree based method: We construct a decision tree using the features in the current concept space, we then sample the leaf which has the highest value of (gini impurity x number of items in leaf)

- Model agnostic method: We sample for points which are highly misclassified, for each point we then sample its neighbors in the concept space to also see if they are misclassified. If they are, we sample from this neighborhood

For clustering tasks, we also use the tree-based method. However, since we do not have labels, we cannot use the gini impurity to identify a leaf to sample from. Instead, we use a sentence transformer to encode each example, and then create two clusters in this embedding space. We use these two clusters as pseudo-labels and ask the language model to identify differences between the clusters. We evaluate the clustering using the purity of each cluster, where labels are hidden from the system until evaluation.

For supervised learning tasks, instead of using the tree-based method, we used a version of the model agnostic method in which we consider each example grouped with its neighbors using K-nearest neighbors in the concept space ("neighborhood"). In the prompt, our samples come from neighborhoods with the highest average error. After the new concept is generated, its values are computed for each example, and a new classifier is fit with the concepts as features. We can then repeat the above process as needed. Evaluation (aside from evaluating the generated concepts, see below) is simple, as it only consists of evaluating a metric of choice on the test set.

Experiments & Results

Supervised Setting

We perform our evaluation on 6 different datasets:

1. Yelp dataset - Yelp reviews for sentiment analysis. This dataset contains user-generated Yelp reviews. The goal is to predict the review rating (1 to 5 stars) based on the text of the review.
2. Snli dataset - Stanford Natural Language Inference (SNLI) dataset for textual entailment. This dataset contains pairs of sentences (premise and hypothesis) and their entailment labels. The goal is to predict the relationship between the premise and hypothesis based on their text.
3. Financial phrasebank - Finance phrases for sentiment analysis. This dataset contains financial phrases and their sentiment labels. The goal is to predict the sentiment of a financial phrase based on its text.
4. Rotten tomatoes - Rotten Tomatoes movie reviews for sentiment analysis. This dataset contains movie reviews from Rotten Tomatoes. The goal is to predict the binary sentiment of a review based on its text, determining if the reviewer enjoyed the movie.
5. Yahoo answers - Yahoo Answers for topic classification. This dataset contains questions and answers from Yahoo Answers. The goal is to predict the topic of a question based on its text and the text of the best answer.

6. Amazon reviews - Amazon reviews for sentiment analysis. This dataset contains product reviews from Amazon. The goal is to predict the binary sentiment of a review based on its text, determining if the reviewer had a positive or negative experience with the product.

For each dataset, we generate 5 concepts from a training set of size 100 and evaluate on a test set of size 200. These smaller sizes are due to the fact that the concept scoring was done using the davinci model of GPT-3 which leads to a cost per test of around ~$5. For each dataset, we report the following in the appendix:

- a sample of the test set (including concept labels)
- a description of the concepts generated
- the accuracy of different models built on top of the systems
- the correlation of the concepts to one another
- the coefficients of each dataframe

Evaluating the results over all six datasets in terms of accuracy when compared to the baseline (predicting the average training label for regression, predicting the modal label for classification), we get the following table:

dataset	model	train_accuracy	test_accuracy
amazon_reviews	Baseline	0.5100	0.545
amazon_reviews	Logistic Regression	0.8500	0.890
amazon_reviews	XGBoost	0.8500	0.890
finance_phrases	Baseline	0.5500	0.595
finance_phrases	Logistic Regression	0.7700	0.775
finance_phrases	XGBoost	0.7700	0.720
rotten_tomatoes	Baseline	0.5700	0.445
rotten_tomatoes	Logistic Regression	0.8400	0.830
rotten_tomatoes	XGBoost	0.8400	0.825
snli	Baseline	0.3900	0.325
snli	Logistic Regression	0.5800	0.435
snli	XGBoost	0.6200	0.455
yahoo_answers	Baseline	0.1600	0.100
yahoo_answers	Logistic Regression	0.2600	0.140
yahoo_answers	XGBoost	0.3600	0.090
yelp	Baseline	1.8384	1.955
yelp	LinearRegression	0.467333	0.670892
yelp	XGBoost	0.332831	0.838257

Interpreting the results of this table, we get that the system is able to do a good job on datasets related to sentiment (amazon reviews, rotten tomatoes, yelp, and finance phrases) and poorly on the SNLI task and the yahoo answers task which is a 20-way topic classification.

Investigating the feature importances of xgboost models trained on top of the generated concept spaces for each dataset, we also see that in the SNLI and yahoo answers task we see a high prevalence of “useless concepts” which get a near-zero feature importance.

One plausible explanation for why this is the case is because the system is naturally biased towards discovering concepts related to sentiment classification. This bias comes from the “example concepts” loaded into the discovery prompt (which can be seen in the appendix)

Overall, we can see that the system shows a lot of promise. In terms of diversity, qualitatively, none of the concepts discovered were ever repeated for any dataset and each concept is relatively unique and provides additional information relative to the others. The decent accuracy on the tasks also indicate that the system is good at discovering relevant concepts and then subsequently measuring those concepts.

Unsupervised setting

For the task of clustering, we evaluate our system on two different datasets: jokes and blogs. We use text-davinci-003 for concept generation and text-curie-001 for concept measurement.

The jokes dataset contains 500 examples from r/cleanjokes and 500 from r/darkjokes. We use these subreddits as labels for evaluation, but they are hidden to the system.

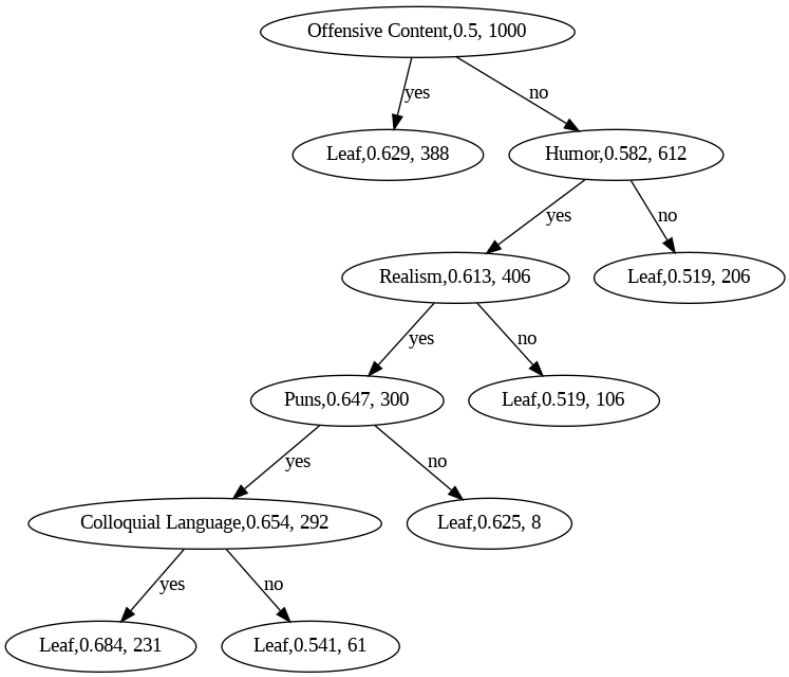

Figure 1: *Results from joke dataset. Purities of each node and number of samples below that node are shown*

The following are interesting to us:

- All of the concepts except Humor are intuitively discriminatory for clean and dark jokes. Despite this, leaves are very impure (all purities between 0.55 and 0.7). This suggests the concept measurement may be flawed.
- Order matters: from manually looking over the data, many of the posts from r/cleanjokes contain puns, however most do not employ realism. That is, puns could have been a much more discriminatory concept if it was generated earlier, however out of 300 examples only 8 did not contain puns (according to the concept scorer)
- The concept scorer has a tendency to answer “yes”. We would expect more concepts to be sparse, but we see the opposite. Again, suggesting concept measurement may be flawed.

The blogs dataset contains 500 examples from blogs labeled “Technology” and 500 examples from blogs labeled “Law”. Again, these labels are hidden from the system and only used to calculate purities of each node.

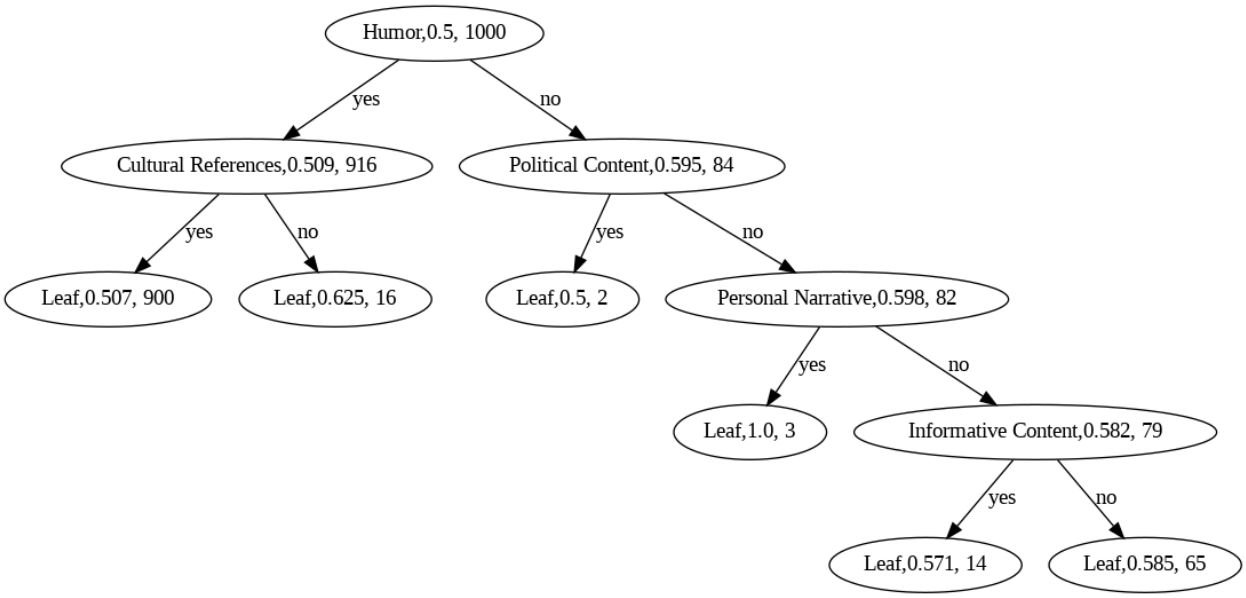

Figure 2: *Results from blog dataset for clustering*

We find the following interesting:

- Since we are only sampling 5 examples from each cluster, it is very possible to get “unlucky” with sampling and choose a non-discriminatory concept, e.g. humor as the first concept generated: most blog posts are humorous, both intuitively and according to the concept measurer.

Conclusion

This paper successfully demonstrates the proof of concept for a “concept space” generation system which automatically constructs a concept bottleneck for sentiment related classification and regression tasks. Compared to other works such as Zhong et. al and McInerney et. al, our system demonstrates greater flexibility in terms of the task domains and topic domains we can apply our system to. Additionally, we show qualitatively the diversity of our automatic concept generation and we also show its performance on a wide set of tasks. Given the promising results in this report, we are aiming to perform further research, with the goal of publishing at a conference (tentatively, EMNLP 2023)

Future work

Dynamic changes in sample concepts

A notable limitation in our system is that it only performed well on sentiment related tasks given that the sample concepts provided in the concept generation prompt are all related to sentiment analysis. It may be possible for our system to adapt this section of the prompt to be better suited to the target task. This is important for two reasons. The first is that it allows us to perform more tasks, and the second is that we may be able to steer the system to discover general concepts at initialization which could apply to most examples, and then iteratively discover more and more specific concepts which are only applicable (and only measured on) specific examples.

Scalability

Our system is fairly expensive to run even on a small scale. One reason for this is because we are using openai’s instruction finetuned davinci model to perform the concept labeling. We also initially explored using the text-curie-001 gpt-3 model which is also instruction finetuned but found that all the concept answers it generated were highly correlated with each other (which indicates that the model often devolved into measuring general sentiment). Some directions to explore in terms of scalability are 1) figuring out how to use a smaller model for concept scoring 2) figuring out the feasibility of finetuning an end to end bert-like model to measure the concepts after they have discovered, and 3) implementing a filtering stage before a concept is measured to evaluate if an example is “within the domain” of the concept. This also has the added benefit of allowing us to generate highly specific concepts.

More evaluations

To have a better understanding of the system, it is necessary to have a more robust set of evaluations. In the supervised setting, we plan on evaluating our system on three different levels. We evaluate the performance based on the individual concepts generated, on the

concept set generated as a whole (system level), and the performance of our model for potential use cases. In performing this evaluation, we plan on using a test suite of NLP tasks collected and on the openD5 dataset.

Concept generation and concept measurement level metrics

- Consistency: Are the concepts coherent and consistent enough to be used as a reliable metric? In other words, can we different raters consistently agree on how to evaluate a given text? (human-rated), (also potentially automatically rated - if we paraphrase the prompt a bit how much do the responses shift?)
- Measurability: Are the concepts being measured accurately by the language model?
- Validity: Is the concept valid on the specific domain of texts we have? (ex, do all texts return the same value when measured or can the concept be discriminative)

System level metrics

- Concept orthogonality:
    - What is the correlation of the concepts with respect to one another?
    - If we train a simple linear model based on the concepts, are all concepts being "used" by the classifier, what proportion gets a negligible weight?
- End to end accuracy:
    - (Supervised & unsupervised) If we constructed a model using the concepts in the concept space, how accurate is the system?
    - Baselines:
        - GPT-3 Fewshot
        - Bag of words
        - Rationale-based interpretable systems
- Sample efficiency
    - GPT-3 Fewshot
    - Bag of words
    - Rationale-based interpretable systems
    - Sample efficiency curve (cost permitting)

- Simulatability:
    - (Supervised only) If we construct a simple model based on the features in the concept space, can a human more reliably predict how the system will evaluate a given example?

Use case evaluation

- Bias auditing/spurious cue detection
    - If we create a dataset with synthetic biases such that the biases are correlated with the label, will the system be able to pick up on this potential bias?
- Automated research

- Evaluate on the openD5 dataset, for each problem, we tally how many unique concepts the system can discover and the “validity” of those discoveries
- Improvement to E2E system
  - (Supervised only) Can the original E2E system be improved by the concepts discovered?

Appendix

Yelp dataset results

	text	label	Customer Service	Price Value	mentions local ingredients	mentions atmosphere	mentions quality
	the food wasn't very good, and the organ player didn't seem to know any of the songs we requested.	1	0	0	-1	-1	-1
	Very cozy little place. Good service and fairly good food for a good price. And you get to shop Vintage clothing afterwards! My Mom adores this place, I take her here on Sundays for lunch sometimes, it always makes her day. :) Which makes mine!	3	0	0	-1	1	1
	I find Harmony House Music a great place to shop for almost all my music needs. The review before me I find to be a nervous shopper. The staff watches you to make sure they meet ALL your needs. I found the staff very friendly and helpful. I bought a New Fender Ultimate Chorus Amp, So for him to say what he did about their line is just not true. Yes they do allot of catering to students, it's called Bread and Butter, and I fully understand that. No they can't compete with the large chain stores with allot of the famous brands, But they do carry some. They will go out of their way to make sure you are a satisfied customer. As for the antique side of the store, I have found many little trinkets of my likening. If you don't go look you wont know it's there. Over all I find Harmony House Music a great place to shop.	4	1	0	-1	-1	1
	Instead of writing a review I'm going to write a warning. Avoid this place at all costs. If you want a lesson on how to run a successful Mexican restaurant do not come here. From the kid at the hostess stand to the sub par food this was truly a horrible experience.\n\nOf course you start with complimentary chips and salsa. Chips were room temp and not a single grain of salt. Both salsas, one mild chunky and another hot and smooth were both lacking salt, garlic and heat, essentially tomato sauce with some red pepper flakes added for good measure.\n\nWe ordered a quesadilla with green chile ($6.00) and what they call a Hal's Sampler For Two ($15.95) which is a sampling of all their house specials. Pollo Fundido, mini chimi's, flauta's, tostada's and a green corn tamale served with rice and beans. Sounds promising and a no brainer.\n\nAs soon as the quesadilla arrived we knew what we were in for. The green chile's were of the canned variety and from the first bite, of the old variety....	0	-1	-1	-1	-1	-1
	As my first time to Vegas, I really enjoyed this Hotel & Casino. I thought that it was perfect for my age group (mid 20's). After visiting many other casino's on the strip...I think our casino was definitely the hippest...and most fun! Double plus---Miracle Mile Shops!! I can seriously see people not leaving the hotel! It has so much to offer! Our rooms were pretty decent...wish there was more lighting tho, room seemed dark...bathroom was amazing!! I thought the hollywood movie theme to each room was cute too. The free bottle of premium vodka was a plus and saved us a good $45! Wish we had a coffee maker in the room tho. \n\nThe location on the strip is perfect. I feel like it's in the heart of the strip, right across the street from the new City Center. \n\nI would recommend this hotel definitely!! I feel as if all the hotels in Vegas will put any normal hotel to shame. Hard to be really disappointed! :)	3	0	0	-1	-1	1

Concept Name	Concept Description	Concept Question	Possible Responses	Response Guide
Customer Service	Customer service refers to the quality of service provided by a business to its customers. It includes aspects such as attentiveness, responsiveness, and helpfulness of staff, as well as the overall experience of the customer.	Does the review mention the customer service in a positive way?	['positive', 'negative', 'unknown']	{'positive': 'The review explicitly mentions or implies the customer service was good, such as attentive, responsive, or helpful staff.', 'negative': 'The review explicitly mentions or implies the customer service was poor, such as unhelpful or unresponsive staff.', 'unknown': 'The review does not mention anything about the customer service nor is there any information about the customer service.'}
Price Value	Price value refers to the perceived value of the product or service in relation to its price. A product or service with good price value is considered to be a good deal, while one with poor price value is considered to be overpriced.	Does the text mention that the product or service is a good value for its price?	['positive', 'negative', 'unknown']	{'positive': 'The text explicitly mentions or implies that the product or service is a good value for its price.', 'negative': 'The text explicitly mentions or implies that the product or service is not a good value for its price.', 'unknown': 'The text does not mention anything about the price value of the product or service.'}
mentions local ingredients	mentions local ingredients refers to the discussion or mention of locally-sourced ingredients within a given text, conversation, or communication.	Does the text mention any local ingredients?	['positive', 'negative', 'unknown']	{'positive': 'The text includes words or phrases related to locally-sourced ingredients, such as specific local produce, locally-made products, or locally-sourced ingredients.', 'negative': 'The text is does not mention any locally-sourced ingredients, or any words related to locally-sourced ingredients.', 'unknown': 'The text does not have any information about locally-sourced ingredients.'}
mentions atmosphere	mentions atmosphere refers to the discussion or mention of the atmosphere of a place within a given text, conversation, or communication.	Does the text mention the atmosphere of the place?	['positive', 'negative', 'unknown']	{'positive': 'The text explicitly mentions or implies the atmosphere of the place, such as being peaceful, lively, or cozy.', 'negative': 'The text does not mention the atmosphere of the place.', 'unknown': 'The text does not have any information about the atmosphere of the place.'}
mentions quality	mentions quality refers to the discussion or mention of the quality of a product or service within a given text, conversation, or communication.	Does the text mention the quality of the product or service in a positive way?	['positive', 'negative', 'unknown']	{'positive': 'The text mentions the quality of the product or service in a positive way, such as being good, excellent, or high-quality.', 'negative': 'The text mentions the quality of the product or service in a negative way, such as being bad, poor, or low-quality.', 'unknown': 'The text does not mention anything about the quality of the product or service.'}

	Unnamed: 0	Customer Service	Price Value	mentions local ingredients	mentions atmosphere	mentions quality
0	Customer Service	1.000000	0.559238	0.274920	0.439248	0.647435
1	Price Value	0.559238	1.000000	0.218986	0.453944	0.688562
2	mentions local ingredients	0.274920	0.218986	1.000000	0.368835	0.287147
3	mentions atmosphere	0.439248	0.453944	0.368835	1.000000	0.476672
4	mentions quality	0.647435	0.688562	0.287147	0.476672	1.000000

	Feature	LinearRegression Coefficient	XGBoost Importance
0	Customer Service	0.144325	0.002620
1	Price Value	0.172197	0.029988
2	mentions local ingredients	-0.142826	0.008791
3	mentions atmosphere	0.468775	0.050010
4	mentions quality	1.189038	0.908592

	Model	Train MSE	Test MSE
0	LinearRegression	0.467333	0.670892
1	XGBoost	0.332831	0.838257

Snli dataset results

label	text	Contextual Relevance	Action	Location	Object Mention	Sentiment
0	Woman in a gold jacket walking on a New York subway platform. [PREMISE] A woman walks. [HYPOTHESIS]	1	1	-1	0	0
0	A girl in a black bikini is getting out of a swimming pool. [PREMISE] The girl is wet. [HYPOTHESIS]	1	1	-1	1	0
0	Children laugh and smile. [PREMISE] Kids smile and laugh. [HYPOTHESIS]	1	0	-1	0	1
2	A woman in blue jeans and a creme shirt walking down the road with a rake and hoe in her right hand and gloves in her left hand. [PREMISE] A person is running with a stolen TV. [HYPOTHESIS]	0	1	-1	1	0
0	A man in a black shirt is playing percussion on a set of empty five gallon buckets. [PREMISE] A man is making noise [HYPOTHESIS]	1	1	-1	0	0

Concept Name	Concept Description	Concept Question	Possible Responses	Response Guide
Contextual Relevance	Contextual Relevance refers to the degree to which the premise and hypothesis are related to each other in the context of the given text.	Does the premise and hypothesis have a high degree of contextual relevance?	['positive', 'negative', 'unknown']	{'positive': 'The premise and hypothesis are strongly related to each other in the context of the given text.', 'negative': 'The premise and hypothesis are not related to each other in the context of the given text.', 'unknown': 'It is unclear if the premise and hypothesis are related to each other in the context of the given text.'}
Action	Action refers to the presence of an action taking place in the text, such as a physical action, an event, or a process.	Does the text contain an action?	['positive', 'negative', 'unknown']	{'positive': 'The text includes an action, such as a physical action, an event, or a process.', 'negative': 'The text does not contain any action or event.', 'unknown': 'The text does not have any information about an action or event.'}
Location	Location refers to the physical setting of the text, such as an interior or exterior space, a geographic location, or a specific place.	Does the text contain any information about the location?	['positive', 'negative', 'unknown']	{'positive': 'The text includes information about the location, such as a physical space, geographic location, or specific place.', 'negative': 'The text does not contain any information about the location.', 'unknown': 'The text does not have any information about the location nor is there any information about the location.'}
Object Mention	Object Mention refers to the presence of a physical object in the text, such as a person, animal, vehicle, tool, or other tangible item.	Does the text mention any physical objects?	['positive', 'negative', 'unknown']	{'positive': 'The text mentions a physical object, such as a person, animal, vehicle, tool, or other tangible item.', 'negative': 'The text does not mention any physical objects.', 'unknown': 'The text does not provide enough information to determine if any physical objects are mentioned.'}
Sentiment	Sentiment refers to the overall tone of the text, such as positive, negative, or neutral. It can be expressed through words, phrases, or expressions.	Does the text contain a positive, negative, or neutral sentiment?	['positive', 'negative', 'neutral', 'unknown']	{'positive': 'The text contains a positive sentiment, such as joy, enthusiasm, or encouragement.', 'negative': 'The text contains a negative sentiment, such as sadness, anger, or fear.', 'neutral': 'The text contains a neutral sentiment, such as a lack of emotion or a neutral opinion.', 'unknown': 'The text does not contain any sentiment or it is unclear what the sentiment is.'}

	Unnamed: 0	Contextual Relevance	Action	Location	Object Mention	Sentiment
0	Contextual Relevance	1.000000	0.393893	0.252520	0.250003	0.085678
1	Action	0.393893	1.000000	-0.021096	0.371259	0.132506
2	Location	0.252520	-0.021096	1.000000	0.151646	-0.029771
3	Object Mention	0.250003	0.371259	0.151646	1.000000	0.153675
4	Sentiment	0.085678	0.132506	-0.029771	0.153675	1.000000

	class	feature	coefficient	model
15	NaN	Contextual Relevance	0.704629	XGBoost
16	NaN	Action	0.047454	XGBoost
17	NaN	Location	0.036449	XGBoost
18	NaN	Object Mention	0.211467	XGBoost
19	NaN	Sentiment	0.000000	XGBoost

	model	train_accuracy	test_accuracy
0	Logistic Regression	0.58	0.435
1	XGBoost	0.62	0.455

Financial phrasebank results

	text	label	mentions financial performance	mentions financing arrangements	mentions market activity	mentions growth	mentions risk
71	The company did not distribute a dividend in 2005 .	1	-1	-1	-1	-1	-1
74	The offer of some 30 million shares aimed to raise more than x20ac 500 million US$ 640 million , was expected to be completed by Oct. 9 , Outokumpu said .	1	-1	0	-1	0	0
11	The chain posted sales of 298 million euros for full 2005 , a rise of 19.5 percent , year-on-year .	2	-1	-1	-1	1	0
30	Raisio 's bid to buy Glisten is a `` win-win '' deal for both companies , the chairman of the UK snacks firm told just-food today 10 February .	2	-1	0	-1	0	0
183	`` BasWare 's product sales grew strongly in the financial period , by 24 percent .	2	1	-1	0	1	1

Concept Name	Concept Description	Concept Question	Possible Responses	Response Guide
mentions financial performance	mentions financial performance refers to the discussion or mention of financial performance, such as net sales, operating profit, or net interest income, within a given text, conversation, or communication.	Does the text mention any financial performance metrics?	['positive', 'negative', 'unknown']	{'positive': 'The text includes words or phrases related to financial performance, such as net sales, operating profit, or net interest income.', 'negative': 'The text does not mention any financial performance metrics.', 'unknown': 'The text does not have any information about financial performance.'}
mentions financing arrangements	mentions financing arrangements refers to the discussion or mention of financing instruments, such as loans, bonds, or equity, within a given text, conversation, or communication.	Does the text mention any financing arrangements?	['positive', 'negative', 'unknown']	{'positive': 'The text includes words or phrases related to financing arrangements, such as loans, bonds, or equity.', 'negative': 'The text does not mention any financing arrangements.', 'unknown': 'The text does not have any information about financing arrangements.'}
mentions market activity	mentions market activity refers to the discussion or mention of market activity, such as stock prices, trading volumes, or market trends, within a given text, conversation, or communication.	Does the text mention market activity?	['positive', 'negative', 'unknown']	{'positive': 'The text includes words or phrases related to market activity, such as stock prices, trading volumes, or market trends.', 'negative': 'The text does not mention market activity, or any words related to market activity.', 'unknown': 'The text does not have any information about market activity.'}
mentions growth	mentions growth refers to the discussion or mention of growth, such as increased sales, increased profits, or increased market share, within a given text, conversation, or communication.	Does the text mention any growth?	['positive', 'negative', 'unknown']	{'positive': 'The text includes words or phrases related to growth, such as increased sales, increased profits, or increased market share.', 'negative': 'The text does not mention any growth, or any words related to growth.', 'unknown': 'The text does not have any information about growth.'}
mentions risk	mentions risk refers to the discussion or mention of potential risks, such as market volatility, economic downturns, or political instability, within a given text, conversation, or communication.	Does the text mention any potential risks?	['positive', 'negative', 'unknown']	{'positive': 'The text mentions potential risks, such as market volatility, economic downturns, or political instability.', 'negative': 'The text does not mention any potential risks.', 'unknown': 'The text does not have any information about potential risks.'}

	Unnamed: 0	mentions financial performance	mentions financing arrangements	mentions market activity	mentions growth	mentions risk
0	mentions financial performance	1.000000	-0.062319	0.100913	0.421768	0.087488
1	mentions financing arrangements	-0.062319	1.000000	0.171679	0.185088	0.363805
2	mentions market activity	0.100913	0.171679	1.000000	0.274713	0.249711
3	mentions growth	0.421768	0.185088	0.274713	1.000000	0.256134
4	mentions risk	0.087488	0.363805	0.249711	0.256134	1.000000

	class	feature	coefficient	model
15	NaN	mentions financial performance	0.821041	XGBoost
16	NaN	mentions financing arrangements	0.006004	XGBoost
17	NaN	mentions market activity	0.003620	XGBoost
18	NaN	mentions growth	0.163507	XGBoost
19	NaN	mentions risk	0.005828	XGBoost

	model	train_accuracy	test_accuracy
0	Logistic Regression	0.77	0.775
1	XGBoost	0.77	0.720

Rotten tomatoes results

	text	label	Humor	Romantic Tone	Cinematic Quality	Moral Message	Character Development
50	it's one of those baseball pictures where the hero is stoic , the wife is patient , the kids are as cute as all get-out and the odds against success are long enough to intimidate , but short enough to make a dream seem possible .	0	0	-1	0	0	1
138	solaris " is a shapeless inconsequential move relying on the viewer to do most of the work .	0	0	-1	0	-1	0
100	it is one more celluloid testimonial to the cruelties experienced by southern blacks as distilled through a caucasian perspective .	0	-1	-1	0	-1	0
147	it's super- violent , super-serious and super-stupid .	0	0	-1	0	-1	-1
178	a smart , sweet and playful romantic comedy .	1	1	1	0	0	0

	Concept Name	Concept Description	Concept Question	Possible Responses	Response Guide
0	Humor	Humor refers to the presence of comedic elements in a text, such as jokes, puns, or satire.	Does the text contain any comedic elements?	['positive', 'negative', 'unknown']	{'positive': 'The text includes comedic elements, such as jokes, puns, or satire.', 'negative': 'The text does not contain any comedic elements.', 'unknown': 'The text does not provide enough information to determine if it contains comedic elements.'}
1	Romantic Tone	Romantic Tone refers to the presence of romantic elements in a text, such as expressions of love, affection, or admiration.	Does the text contain any romantic elements?	['positive', 'negative', 'unknown']	{'positive': 'The text includes expressions of love, affection, or admiration.', 'negative': 'The text does not contain any romantic elements.', 'unknown': 'The text does not provide enough information to determine if it contains any romantic elements.'}
2	Cinematic Quality	Cinematic Quality refers to the degree to which a movie is well-crafted, with good production values, and an engaging story.	Does the text mention the cinematic quality of the movie in a positive way?	['positive', 'negative', 'unknown']	{'positive': 'The text explicitly mentions or implies that the movie has good cinematic quality, such as good production values, engaging story, or well-crafted scenes.', 'negative': 'The text explicitly mentions or implies that the movie has poor cinematic quality, such as bad production values, unengaging story, or poorly crafted scenes.', 'unknown': 'The text does not mention anything about the cinematic quality nor is there any information about the cinematic quality.'}
3	Moral Message	Moral Message refers to the presence of a moral lesson or message in the text, such as a lesson about right and wrong, justice, or truth.	Does the text contain a moral message?	['positive', 'negative', 'unknown']	{'positive': 'The text includes a moral message, such as a lesson about right and wrong, justice, or truth.', 'negative': 'The text does not contain a moral message.', 'unknown': 'The text does not have any information about a moral message.'}
4	Character Development	Character Development refers to the degree to which characters in the text are fleshed out, with distinct personalities, motivations, and arcs.	Does the text contain evidence of character development?	['positive', 'negative', 'unknown']	{'positive': 'The text contains evidence of characters with distinct personalities, motivations, and arcs.', 'negative': 'The text does not contain evidence of characters with distinct personalities, motivations, and arcs.', 'unknown': 'The text does not have any information about character development.'}

	Unnamed: 0	Humor	Romantic Tone	Cinematic Quality	Moral Message	Character Development
0	Humor	1.000000	0.198142	0.453003	0.386847	0.379987
1	Romantic Tone	0.198142	1.000000	0.268649	0.334081	0.235793
2	Cinematic Quality	0.453003	0.268649	1.000000	0.459624	0.569837
3	Moral Message	0.386847	0.334081	0.459624	1.000000	0.538088
4	Character Development	0.379987	0.235793	0.569837	0.538088	1.000000

	class	feature	coefficient	model
10	NaN	Humor	0.004007	XGBoost
11	NaN	Romantic Tone	0.191974	XGBoost
12	NaN	Cinematic Quality	0.033759	XGBoost
13	NaN	Moral Message	0.663066	XGBoost
14	NaN	Character Development	0.107194	XGBoost

	model	train_accuracy	test_accuracy
0	Logistic Regression	0.84	0.830
1	XGBoost	0.84	0.825

Yahoo answers results

	label	text	Mentions Immigration	Mentions Solutions	Mentions Emotion	Mentions Specific Details	Mentions Respect
0		Why do so many answers given refer to websites. I am sure that people have probably checked out sites before a [TITLE] On some occassions you will not find a solution to your problem on the web, that feels right for you. With thousands of people using Answers the Human Factor can hit the nail on the Head. So come on people help others with your Knowledge, Ideas, Solutions properly, not give a web site just to get points, [CONTENT] Websites are good for facts and factual information especially accurate information. Sometimes the answer could be too long or involved, so websites can be useful. Not everyone knows how to look for specific websites. However not every question can be answered by facts. Sometimes it`s opinion, personal experience, or just fun. Both are useful depending on the question. [ANSWER]	-1	0	1	0	0
0		your feelings about the sins in the Bible? [TITLE] After 243 questions, i'm starting to get the way most christians interpret the bible. A sin is a sin is a sin. There is no 'in between', a sin is a sin. Because it's God's word. Ok, i begin to get that part now. (i am kinda slow in these matters, sorry)\n\nI was wondering, seriously, do you ever have a feeling about all this? Do you ever wish something God tells you to be wrong, would in fact be not all that wrong? Specially the trivial items, like wearing clothes of two different fabrics, homosexuality, eating shelfish?\n\nWouldn't life be much easier if those things were never mentioned in the Bible??\n\nDon't get me wrong, i understand that is your feeling that God's word should always be concidered as the full Truth, i just wonder how it makes you feel. [CONTENT] Things have to be taken in context. I agree that the wearing of mixed fabrics and refraining from shellfish are not about morality, they are about sanctity - God set ...	-1	-1	-1	0	-1
3		Essay contest? [TITLE] For school we are having this essay contest, its 200 dollars for the winner. I know this seems lame but i need the money to buy my mom a nice mothers day gift (my dad is kinda an a** and her mothers day always sucks) so the topic is "If i could travel anywhere I would go..." \n\nso help? any suggestions. your supposed to think outside of the box [CONTENT] back in time. hit all the major events that happened. see what people really had to live through compared to now. [ANSWER]	-1	0	1	0	0
2		How do I get in great shape by June??? Help!? [TITLE] I'm 35. I used to have a great body. I've gained some weight, although I carry it well and look nice in clothes....summer time is coming and I would look crazy on the beach! :-/ It doesn't seem to be as easy as it used to. I need good advice on how to take it off and keep it off. Any suggestions? [CONTENT] I used to weigh 311 lbs. I had tried everything, had my stomach stapled, teeth wired, and worked with a personal trainer- I never lost more than 10lbs. A doctor that I know told me about this product on this website www.realsauna.com and I tried it! After only 2 months of using the suit with no diet, I LOST 33 LBS. I have been using the REALSAUNA suit now for 2 years, 3 times a week for 30min, and I am now down to 155 lbs! My idea weight! Try the product, it will really be worth every penny! I promise! [ANSWER]	-1	0	1	1	0
2		does doing gym excersise increases blood pressure, what is the normal bp(120/80) is it true? [TITLE] [CONTENT] 120/80 is quoted as the 'normal' BP but there is more a healthy range (which changes according to the latest trend). \n\nWhen u do excercise ur BP rises because your heart is trying to get more blood to you tissues and also trying to get rid of the heat produced (it's why u go all red). But over time your resting blood pressure does decrease. But this might be a combination of losing weight and becoming fitter (which leads to a lower heart rate). \n\nSo the answer is really yes it does, while your doing the exercise and a bit after. [ANSWER]	-1	0	-1	0	-1

Concept Name	Concept Description	Concept Question	Possible Responses	Response Guide
Mentions Immigration	Mentions immigration refers to any discussion or mention of immigration-related subjects, concepts, or policies within a given text, conversation, or communication.	Does the text contain any words related to immigration?	['positive', 'negative']	{'positive': 'The text includes words or phrases related to immigration, such as specific policies, immigration trends, or immigration-related events.', 'negative': 'The text is does not mention immigration, or any words related to immigration'}
Mentions Solutions	Mentions solutions refers to any discussion or mention of potential solutions to the problem presented in the text, conversation, or communication.	Does the text contain any discussion or mention of potential solutions?	['positive', 'negative', 'unknown']	{'positive': 'The text includes words or phrases related to potential solutions, such as specific actions, strategies, or approaches.', 'negative': 'The text does not mention any potential solutions.', 'unknown': 'The text does not provide enough information to determine if the text contains any discussion or mention of potential solutions.'}
Mentions Emotion	Mentions emotion refers to any discussion or mention of emotions within a given text, conversation, or communication.	Does the text contain any words related to emotion?	['positive', 'negative']	{'positive': 'The text includes words or phrases related to emotion, such as joy, sadness, anger, or fear.', 'negative': 'The text does not mention emotion, or any words related to emotion.'}
Mentions Specific Details	Mentions specific details refers to the discussion or mention of specific facts, figures, or details within a given text, conversation, or communication.	Does the text contain any specific details?	['positive', 'negative', 'unknown']	{'positive': 'The text includes specific facts, figures, or details.', 'negative': 'The text does not mention any specific facts, figures, or details.', 'unknown': 'The text does not have enough information to determine if it contains specific details.'}
Mentions Respect	Mentions respect refers to any discussion or mention of respect, courtesy, or politeness within a given text, conversation, or communication.	Does the text contain any words or phrases related to respect, courtesy, or politeness?	['positive', 'negative', 'unknown']	{'positive': 'The text includes words or phrases related to respect, courtesy, or politeness.', 'negative': 'The text does not mention respect, courtesy, or politeness.', 'unknown': 'The text does not have any information about respect, courtesy, or politeness.'}

	Unnamed: 0	Mentions Immigration	Mentions Solutions	Mentions Emotion	Mentions Specific Details	Mentions Respect
0	Mentions Immigration	1.000000	0.054999	0.189627	0.117391	0.149503
1	Mentions Solutions	0.054999	1.000000	0.166090	0.259527	0.518854
2	Mentions Emotion	0.189627	0.166090	1.000000	0.061526	0.230999
3	Mentions Specific Details	0.117391	0.259527	0.061526	1.000000	0.401496
4	Mentions Respect	0.149503	0.518854	0.230999	0.401496	1.000000

	class	feature	coefficient	model
50	NaN	Mentions Immigration	0.414348	XGBoost
51	NaN	Mentions Solutions	0.122475	XGBoost
52	NaN	Mentions Emotion	0.147878	XGBoost
53	NaN	Mentions Specific Details	0.165418	XGBoost
54	NaN	Mentions Respect	0.149882	XGBoost

	model	train_accuracy	test_accuracy
0	Logistic Regression	0.26	0.14
1	XGBoost	0.36	0.09

Amazon reviews results

text	label	mentions value for money	mentions customer service	mentions product quality	mentions product features	mentions satisfaction
The playback was disrupted and very poor. All other connections to other sites were fine. Very hard to watch thisWith such poor playback from Amazon.	0	0	0	0	-1	-1
I have a 3 year old daughter, whom is now into all the Princess'. She watched Little Mermaid one time, and after the enchantment of the fins had her hooked to wanting to watch it every day. When Little Mermaid II came on Disney channel one night, it changed everything for her. She loves Melody, Ariel's daughter. No other Disney Princess movie represents a princess that has a mother. They are all passed on, or non existent. Melody and Ariel has a great relationship, like any young adolescence child and her mother would. I think this movie is a positive role model for children, and yet there is nothing to buy with Melody represented, which is a shame.	1	0	0	0	-1	0
The package arrived on time and great condition. The dvd itself was amazing. I was laughing the entire hour the show was going for. Dimitri Martin is probably my favorite comedian.	1	0	0	0	-1	1
This book is a fantasy and dangerous. Most of the items on this book and all of Buscaglia books are fantasy. Yes, Leo kind of acknowledge that it is wishful thinking what he is talking about but that statement is meant for you to let down your guard. And if what he said is able to pass into your subconscious mind, it can be very dangerous. As his book has many trojan virus that if pass into your subconscious mind can wreck havoc. Just as a computer virus can wreck havoc to your computer system, so can Buscaglia mental virus within his book. I will give you an example of a virus in his book. One of the book point was "Learn to trust again" and "Love trust". If that one point was able to by pass your conscious mind defense and into your subconscious mind, it can wreck havoc to your human computer (Mind). And there are many more virus within his book. Be wary of Buscaglia book readers.	0	0	0	0	-1	0
I run a small imaging workgroup. This scanner seemed worth a try for scanning odd formats: textiles & orginial art.This is about the 20th scanner that I have installed over the years--the only one that would not function on either platform.Clever design: too bad that the hardware is so buggy.	0	0	0	-1	-1	-1

Concept Name	Concept Description	Concept Question	Possible Responses	Response Guide
mentions value for money	mentions value for money refers to the discussion or mention of the cost-effectiveness of a product or service relative to its quality or performance.	Does the review mention value for money?	['positive', 'negative', 'unknown']	{'positive': 'The review explicitly mentions or implies that the product or service provides good value for money, such as being cost-effective, high-quality, or worth the price.', 'negative': 'The review explicitly mentions or implies that the product or service does not provide good value for money, such as being expensive, low-quality, or not worth the price.', 'unknown': 'The review does not mention anything about the value for money nor is there any information about the cost-effectiveness of the product or service.'}
mentions customer service	mentions customer service refers to the discussion or mention of the customer service related to the product or service, such as the response time, helpfulness, and overall customer experience.	Does the text mention customer service in a positive or negative way?	['positive', 'negative', 'unknown']	{'positive': 'The text mentions customer service in a positive way, such as fast response time, helpfulness, or overall good customer experience.', 'negative': 'The text mentions customer service in a negative way, such as slow response time, unhelpfulness, or overall bad customer experience.', 'unknown': 'The text does not mention customer service nor is there any information about customer service.'}
mentions product quality	mentions product quality refers to the discussion or mention of the quality of the product or service, such as the materials used, design, manufacturing techniques, and attention to detail.	Does the text mention the quality of the product in a positive or negative way?	['positive', 'negative', 'unknown']	{'positive': 'The text explicitly mentions or implies the product has good quality, such as well-made, sturdy, or long-lasting construction.', 'negative': 'The text explicitly mentions or implies the product has poor quality, such as being prone to defects or wearing out quickly.', 'unknown': 'The text does not mention anything about the quality nor is there any information about the quality.'}
mentions product features	mentions product features refers to the discussion or mention of the specific features of the product or service, such as its capabilities, functions, or components.	Does the text mention any specific features of the product or service?	['positive', 'negative', 'unknown']	{'positive': 'The text mentions specific features of the product or service, such as its capabilities, functions, or components.', 'negative': 'The text does not mention any specific features of the product or service.', 'unknown': 'The text does not have any information about the product or service features.'}
mentions satisfaction	mentions satisfaction refers to the discussion or mention of a positive or negative experience with the product or service.	Does the text mention satisfaction or dissatisfaction with the product or service?	['positive', 'negative', 'unknown']	{'positive': 'The text explicitly mentions or implies satisfaction with the product or service.', 'negative': 'The text explicitly mentions or implies dissatisfaction with the product or service.', 'unknown': 'The text does not mention anything about satisfaction or dissatisfaction with the product or service.'}

	Unnamed: 0	mentions value for money	mentions customer service	mentions product quality	mentions product features	mentions satisfaction
0	mentions value for money	1.000000	0.422563	0.735975	0.386207	0.630423
1	mentions customer service	0.422563	1.000000	0.497286	0.535970	0.442153
2	mentions product quality	0.735975	0.497286	1.000000	0.508502	0.651962
3	mentions product features	0.386207	0.535970	0.508502	1.000000	0.480454
4	mentions satisfaction	0.630423	0.442153	0.651962	0.480454	1.000000

	class	feature	coefficient	model
10	NaN	mentions value for money	0.237086	XGBoost
11	NaN	mentions customer service	0.000000	XGBoost
12	NaN	mentions product quality	0.000000	XGBoost
13	NaN	mentions product features	0.049580	XGBoost
14	NaN	mentions satisfaction	0.713335	XGBoost

	model	train_accuracy	test_accuracy
0	Logistic Regression	0.85	0.89
1	XGBoost	0.85	0.89

Concept generation prompt

Concept Feature Engineering Task

Below we are given a text dataset with accompanying labels. Our task is to identify a feature in the text that could be associated with the label. These labels are generated by a black box machine learning algorithm to score a given text on some dimension. Our task is to deconstruct and identify what high-level concepts the algorithm is using.

To do this, we will examine a sample of texts that have been scored by the algorithm. We will then define a potential concept that the model could be using. Below are some examples of concepts in json format. Each full concept definition comes with a concept name, description, question, response set, and response guide. The concept description provides an intuitive overview of the concept. The concept question is our tool for measuring the concept, this will be graded by a human annotator. The possible responses list the possible responses to the question and the response guide provides information on what each rating means.

1. A possible concept for toxicity detection
{
“Concept Name”: “explicit language”,
“Concept Description”: ""explicit language" refers to the use of words, phrases, or expressions that are offensive, vulgar, or inappropriate for general audiences. This may include profanity, obscenities, slurs, sexually explicit or lewd language, and derogatory or discriminatory terms targeted at specific groups or individuals.",
“Concept Question”: “Does the text contain explicit or vulgar language?”,
“Possible Responses”: ["positive", "negative"],
“Response Guide”: {
"positive": "The text contains explicit language, such as profanity, obscenities, slurs, or derogatory terms targeted at specific groups or individuals.",
"negative": "The text is free from explicit language and is appropriate for general audiences.”
}
} ###

2. A possible concept for product reviews
{
“Concept Name”: “good build quality”,
“Concept Description”: "build quality refers to the craftsmanship, durability, and overall construction of a product. It encompasses aspects such as materials used, design, manufacturing techniques, and attention to detail. A product with good build quality is typically considered to be well-made, sturdy, and long-lasting, while a product with poor build quality may be prone to defects or wear out quickly.",
“Concept Question”: “Does the review mention the build quality in a positive way?”,
“Possible Responses”: ["positive", "negative", "unknown"],
“Response Guide”: {
"positive": “The review explicitly mentions or implies the product has good build quality, such as well-made, sturdy, or long-lasting construction.”,
"negative": “The review explicitly mentions or implies the product has poor build quality, such as being prone to defects or wearing out quickly.”,
"unknown": “The review does not mention anything about the build quality nor is there any information about the build quality.”
}
}###

3. A possible concept for topic analysis
{
“Concept Name”: “About Tech”,
“Concept Description”: "talks about technology refers to the discussion or mention of technology-related subjects, concepts, or advancements within a given text, conversation, or communication.",
“Concept Question”: “Does the text contain any words related to tech?”,
“Possible Responses”: ["positive", "negative"],
“Response Guide”: {
"positive": “The text includes words or phrases related to technology, such as specific devices, technological trends, innovations, or industry players.”,

"negative": “The text is does not mention technology, or any words related to technology”
}
}###

4. A possible concept for personality measurement
{
“Concept Name”: “Extraversion”,
“Concept Description”: "Extraversion represents the degree to which an individual is outgoing, sociable, and assertive, as opposed to being introverted, reserved, or shy.",
“Concept Question”: “Does the following profile description indicate that a person is extroverted?”,
“Possible Responses”: ["positive", "negative", "unknown"],
“Response Guide”: {
"positive": “The profile description includes characteristics or behaviors associated with extraversion, such as outgoing, sociable, assertive, or enthusiastic.”,
"negative": “The profile description includes characteristics or behaviors associated with introversion or the absence of extraversion traits, such as reserved, shy, or solitary.”,
"unknown": “The profile description does not have any information about this person's extraversion”
}
}###

In the task, we will generate concepts for the poem_sentiment dataset

Below is an explanation of the dataset and the labels therein:

Poem sentiment analysis. This dataset contains verses of poems with their sentiment labels. The goal is to predict the sentiment of a verse based on its text.

{'0': 'negative', '1': 'positive', '2': 'no impact', '3': 'mixed'}

Some additional pointers to keep in mind are the following:
1. In this exercise, we will restrict ourselves to making positive/negative/unknown questions
2. We want to make our concept as applicable to many items in the dataset as possible, so try to avoid being too specific.
3. Do not create "uninformative concepts" where it simply tests for the existence of a concept. For example "mentions sentiment" is bad because a yes or no answer does not tell us anything about the underlying quality. Comparatively, "mentions good sentiment" is good because a yes answer lets us know that the quality is good.

We want to create additional concepts to understand how the model is labelling examples.

To do this, we will look at some example texts along with their labels.
---
text:if the pure and holy angels
rating:1
text:"o lord, that didst smother mankind in thy flood,
rating:0
text:howled through the dark, like sounds from hell.
rating:0
text:and so,
rating:2
text:passing to lap thy waters, crushed the flower
rating:0
text:"does he mean himself, i wonder?
rating:2
text:in the shadow of the shores; as dead leaves wake,
rating:0
text:the sower scatters broad his seed,
rating:2
text:brightly expressive as the twins of leda,
rating:1
text:a spirit, neither here nor there,
rating:2
text:but half the secret told,

rating:2
text:the pain when it did live,
rating:0
text:gay little heart!
rating:1
text:ambrosial odours and ambrosial flowers,
rating:2
text:on their tracks his eyes were fastened,
rating:2
text:a story of the days of old,
rating:2
text:of night, and all things now retir'd to rest
rating:2
text:the soul with sweetness, and like an adept
rating:1
text:three lives, three strides, three foot-prints in the sand;
rating:2
text:tis that one told us it was life. 'for not
rating:2

As a reminder we already have the following concepts:
1. emotion:emotion refers to the expression of feelings or sentiments within a text. This may include words, phrases, or expressions that indicate a range of emotions from joy, sorrow, fear, anger, or other sentiments.
2. Religious References:Religious references refer to the use of words, phrases, or expressions that are related to religious beliefs or practices. This may include references to gods, deities, religious figures, or other religious concepts.

Keeping in mind the pointers above, below is an example of another additional positive/negative/unknown concept that we can add that is distinct from the current set of concepts.
Note the strict adherence to the json format
Definition: {

“Concept Name”: “Personification”,
“Concept Description”: "Personification refers to the use of words, phrases, or expressions that assign human characteristics to non-human objects or concepts. This may include giving objects or concepts human emotions, behaviors, or qualities.",
“Concept Question”: “Does the text contain any personification?”,
“Possible Responses”: ["positive", "negative", "unknown"],
“Response Guide”: {
"positive": “The text includes words or phrases that personify non-human objects or concepts, such as giving human emotions, behaviors, or qualities to them.”,
"negative": “The text does not contain any personification.”,
"unknown": “The text does not contain enough information to determine if there is personification.”
}
}