# Modeling Label Semantics for Predicting Motivation

**Steven Secreti**
Stony Brook University, Stony Brook, New York
SBU ID: 112911405
steven.secreti@stonybrook.edu

## Abstract

Understanding the motivation of an agent in a given context is a difficult task that requires implicit knowledge of interhuman dynamics. Understanding these dynamics is typically inherent for humans but is remarkably hard for machines. This problem can be mapped to the multi-label classification task of predicting human psychological needs and motives. In order to make machines better understand these implicit dynamics, it is necessary to utilize techniques that exploit the fact that the psychological needs and motives themselves have unique semantic meanings as opposed to treating these labels as anonymous classes to predict. I hypothesize that by acknowledging the semantics of need and motive labels, the attention of the model can be more accurately positioned to represent the input story in the most optimal way for the task at hand. Through empirical evaluations, I show that modeling label semantics does provide benefits in the needs and motives prediction tasks when the dataset used is limited to sentences in which annotators deemed an action as having occurred. Furthermore, I show that the model can experience a significant increase in accuracy when the number of samples used in each batch during training is increased.

## 1 Introduction

A task that is relatively easy for humans but much harder for machines is to explain the reasoning behind the action that an agent takes. A primary tool humans use to answer this question is intuition. More specifically, humans have the innate ability to detect and consider interhuman dynamics. What makes this task difficult for machines is that implicit interhuman dynamics are not explicitly stated. “Motive” is defined as “An emotion, desire, physiological need, or similar impulse that acts as an incitement to action”. While considering this definition, we can map the task of explaining why an agent took a particular action to the task of predicting the agent’s motive. This task has been previously defined on the ROCStories dataset and mapped to a classification task with the goal of predicting the motive or emotion of the particular characters in the story on a line-by-line basis (Rashkin et al., 2018). ”Motives” here are defined by two primary theories in psychology, Maslow’s Hierarchy of Needs and Reiss’ Motives (Maslow, 1943; Reiss, 2002). The overall goal of this project is to improve the accuracy of the multi-label prediction of an entity’s motive within short story contexts. This will be done using methods such as modeling label semantics, which has been shown to improve accuracy on the similar task of predicting the emotion of entities within short story contexts (Gaonkar et al., 2020). Improving the accuracy in a task like this is important because the ability to predict the motives of agents at a reasonable confidence level is the first step in further generating a specific explanation for the actions of an individual in more detail.

There have been a few broad approaches in attempting to solve this problem previously. As mentioned previously, (Rashkin et al., 2018) had the ROCStories dataset annotated for motivation and emotion across 15,000 stories. This resulted in 56,000 character-line pairs with an annotated motivation and 105,000 with an annotated emotion. The basic approach used by (Rashkin et al., 2018) was to provide as an input to the model the target character, story context sentences, and current sentence and as output receive the applicable labels, either emotion or motive depending on the task, to the character in the current sentence. Paul and Frank (2019) expand upon the approach that was implemented by Rashkin et al. (2018) by utilizing a commonsense knowledge resource to extract knowledge paths that are matched with context representations to more accurately predict the target character's emotion(s) or motive(s). Most recently, Gaonkar et al. (2020) have further enhanced the accuracy in predicting the emotions of target characters by realizing that the labels (emotions) themselves have semantics that, when considered in creating the representations for the input story, help the model to more accurately predict the target label.

The primary gap that is missing from the approach as implemented by Rashkin et al. (2018) is that the model treated the labels it was predicting as anonymous classes. This, as established in (Gaonkar et al., 2020), was a mistake. Furthermore, it is also shown in (Paul and Frank, 2019) that considering general commonsense knowledge provided great benefits for this prediction task.

The particular ideas that I am trying to address this problem are those which are mentioned above that were implemented by Gaonkar et al. (2020); the idea that considering the semantics of the labels themselves in training a model to perform a classification task will yield benefits and improve accuracy over the case that label semantics are not considered. Namely, I am trying to test whether or not expanding this methodology to the task of predicting a target character's motive in a short story context is as effective as the results observed in (Gaonkar et al., 2020) in the task of predicting a target character's emotion in a short story context.

To establish a new baseline, I have fine-tuned a pre-trained BERT uncased language model for the multi-label need(s) and motive(s) prediction tasks. Furthermore, I have tested the previously mentioned ideas by applying to the need and motive labels two of the methods that were applied to emotion labels described in the paper ”Modeling Label Semantics for Predicting Emotional Reactions” (Gaonkar et al., 2020). The first of the two methods is a label attention network. This label attention network is based on the Label-Embedding Attentive Network (LEAM) architecture which is used to produce label-focused representations of the input (Wang et al., 2018). LEAM is particularly used to compute attention scores between the label and the token representations of the input before the classification layer. The second method is using label sentences as additional input. The idea behind this is that by generating label sentences and passing them as input to the transformer-based model (in this case BERT), the transformer-based self-attention mechanism of the model can be exploited and focused more clearly on the most relevant aspects of the input.

As mentioned previously, the dataset used for this task is the ROCStories dataset that had been annotated by Rashkin et al. (2018) with character-line pairs and the motivations of those characters. Each character-line pair was annotated by 3 unique annotators that had undergone some training for the task. For the sake of this project and in accordance with the research performed in (Paul and Frank, 2019), the dataset was preprocessed to include only one of each character-line pair, maintaining labels that had at least 2 annotators in agreement. In other words, I performed majority voting on the data in order to compile an agreed-upon dataset. I will use the baseline micro-averaged precision, recall, and f1 scores of both of the motivation prediction tasks provided by Rashkin et al. (2018) in order to evaluate the effectiveness of my ideas. Furthermore, I will use my BERT baseline to determine if the label semantics techniques benefit the task when compared to a transformer-based language model. In addition to this, I will compare my results to the results achieved in (Paul and Frank, 2019) in order to determine which methodology is most effective for this particular task.

1. I implemented BERT for the tasks of predicting Maslow's Needs and Reiss' Motives and showed that fine-tuning BERT helped to improve accuracy over the previously established baseline in both tasks

2. Our evaluation shows that by increasing the training batch size from 8 samples to 16 samples while keeping all other variables consistent, we are able to increase accuracy, in the best case, by 11 F1 points.

3. Analysis shows that the label semantics techniques used improved performance on the task when using the dataset including only the sentences in which annotators marked an action as occurring and failed when using the entire dataset.

4. Based on my work and by analyzing the work of Paul and Frank (2019), I theorize that label semantics techniques would provide the best benefit to this task when performed on the dataset including only those samples in

which actions occurred and in conjunction with a commonsense knowledge resource.

## 2 Motivation Inference

The basic input for the task at hand is a character-line pair within a short story context. More specifically, the model receives as input the target sentence, in which there exist one or more characters, followed by a separator, and then context sentences which include all the sentences that preceded the target sentence in the short story, followed by a separator and then the target character within the target sentence to perform the motivation inference on. An example input is as follows: ”Tim entered his school’s annual costume contest.//No Context//Tim”. In this case, the target sentence is ”Tim entered his school’s annual costume contest”. Since this is the first sentence in the short story, ”No Context” was provided in the position for the context sentences in the input. Furthermore, the character being targeted here is ”Tim”. For the Maslow’s Needs prediction task, the true output label for this instance is [”Esteem”]. For the Reiss’ Motives prediction task, the true output label for this instance is [”Competition”]. Here, it is clear that the key challenge in determining the correct output for the given input sentence is the fact that coming up with the labels ”Esteem” and ”Competition” require an understanding of commonsense human knowledge. In this particular example, it is evident that the underlying knowledge that a contest is an inherent form of competition is necessary to then understand the motivation of Tim in entering this event. Furthermore, connections can be made with commonsense human knowledge to associate Tim’s school costume contest with his inherent self-esteem. More explicitly, most humans understand that within a school there is a prominent force of social dynamics between peers. This concept can guide the average human and has guided the annotators in selecting the label ”Esteem” as the reason behind Tim joining the costume contest. This task of extrapolating ideas and themes from the information given is relatively innate for humans and, as mentioned previously, is much harder for machines. Therefore, the requirement of this reasoning in determining the correct output label is the main challenge that must be addressed in order to solve this problem most effectively. State-of-the-art solutions for this task selectively filter multi-hop relation paths from a commonsense knowledge resource to interpret the input and most effectively predict the motives of the target characters in this task (Paul and Frank, 2019).

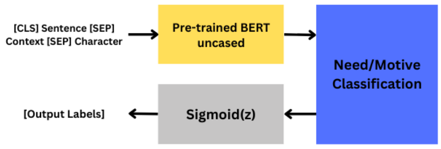
[CLS] Sentence [SEP] Context [SEP] Character
Pre-trained BERT uncased
Need/Motive Classification
Sigmoid(z)
[Output Labels]

Figure 1: BERT Baseline System

### 2.1 Baseline Model(s)

The baseline system that I used to perform the motivation inference was inspired by Gaonkar et al. (2020). Specifically, I adapted the new baseline system that they had established: a pre-trained, uncased BERT language model with a task-specific linear classification layer on the end of it. The baseline system was trained using a Binary Cross Entropy loss function. The classification layer on the end of the BERT model was set to output a logits vector with the dimensions of the number of labels in the task at hand (5 for the Maslow’s Needs inference task and 19 for the Reiss’ Motives prediction task). The logits were passed to a sigmoid activation function and then thresholded at a value of 0.5. That is, logits that were greater than or equal to 0.5 were set to 1 and logits less than 0.5 were set to 0. 1 here indicated that the model predicted the sample to have the label that corresponded with the position in the output vector in which the 1 was present. Figure 1 illustrates the key components of the baseline system implemented for this research.

### 2.2 The Issues

There are a few issues with the baseline model that have come to light in performing this research. The first key issue is the same as the one that was previously mentioned; the baseline model treats the labels as anonymous classes to learn and disregards the fact that the labels themselves have contextual semantic meaning that can be used in the process of matching them to the correct inputs. As previously stated, this complex problem requires a reasonable understanding of commonsense human dynamics as well as a regard for the contextual implications of the labels themselves. Since the labels themselves are words that are used as an umbrella representation for some underlying idea or theme, the baseline system treating them as anonymous classes logically seems to provide a disadvantage to the model's performance in the motivation inference tasks. In addition to this, one primary issue I experienced with the baseline system, in general, was the lack of data available for this task. As this task was considered more complex by Rashkin et al. (2018), it was more expensive to annotate the short stories with motivation labels and thus there was substantially less data available for these tasks than there was for the emotion inference task.

## 3 Label Semantics using Embeddings

The premise behind the ideas implemented in this project is that performance can be improved by modeling the semantics of labels in the motivation inference tasks, that have labels with intrinsic semantic meaning. The following ideas in particular are inspired by Gaonkar et al. (2020) in the application of these techniques to the emotion inference task on the parallel dataset. At a high level, the strategy is to model label semantics by creating embeddings that model the underlying semantics of the label name and then represent each label with its respective embedding. This can easily be done by realizing that the labels for both the Maslow's Needs task and the Reiss' motives task correspond to actual words. Since this is the case, the label embeddings can be initialized with their corresponding word embeddings. The label embeddings can be used in an attempt to improve the accuracy of the baseline BERT system in the following ways.

### 3.1 Label Embedding Attention Network

Label embeddings primarily can be used to guide the attention of the BERT encoder to produce representations of the input that focus primarily on maintaining information related to either the Maslow's Needs labels or the Reiss' Motives labels. For the sake of this project, and in following suit with the work performed by Gaonkar et al. (2020) in the emotion inference task, I have adopted the Label-Embedding Attentive Network (LEAM) architecture in order to generate the described label-focused representations (Wang et al., 2018). The main function of the LEAM architecture is to determine attention scores between the labels and the input token representations that are to be classified. These attention scores are used to appropriately weight each token such that the contribution of each token is based on the relevance of that token in the context of the output labels themselves. Since this work was derived from the work done by Gaonkar et al. (2020), I use the same system in terms of using LEAM to compute an attention matrix that is computed over the hidden states produced by the encoder and the label embeddings. For this project, the encoder used is the BERT features for each token in the text, denoted by $B_t$, and each of the label sentences, denoted by $J$. To compute the final representations (weighted combination of the contextual representations of the input), the attention matrix is used with the compatibility matrix $H$, as computed in (Wang et al., 2018) and similarly computed in (Gaonkar et al., 2020). The output of this computation is the final representation, $y$, and is used in the final classification layer.

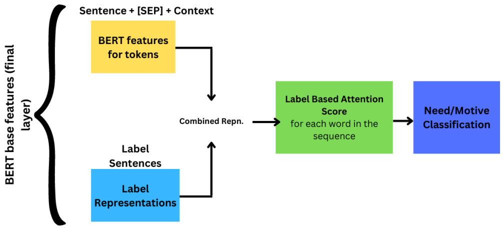
BERT base features (final layer)
Sentence + [SEP] + Context
BERT features for tokens
Label Sentences
Label Representations
Combined Repn.
Label Based Attention Score for each word in the sequence
Need/Motive Classification

Figure 2: Label-Embedding Attentive Network using BERT Features.

$$H = (J^T B_t) \oslash \hat{H} \tag{1}$$

Figure 2 illustrates the key components of the baseline system coupled with the LEAM architecture for this task.

### 3.2 Label Sentences

The second approach used by this paper was also inspired by Gaonkar et al. (2020). This idea is based on the premise of using contextual embeddings produced from a transformer-based model like BERT instead of learning the label embeddings from scratch. This idea is primarily aimed at exploiting the self-attention mechanism in BERT by including label semantics within the input itself. This is accomplished by generating and including within the input to the model, task-specific label sentences, denoted by $L_s$, of the form ”[character] needs [label]” for the Maslow’s Needs motivation inference task and of the form ”[character] is motivated by [label]” for the Reiss’ Motives motivation inference task. For each instance in the Maslow’s Needs inference task, I add five sentences covering all of the needs labels. Similarly for each instance in the Reiss’ Motives inference task, I add nineteen sentences covering all of the motives labels. The main idea here is that the self-attention mechanism of the model will automatically learn to consider and focus on the contextual semantics of these label sentences when generating the representation for the input text.

### 3.3 Implementation Details

The primary implementation for this project was based on the implementation done by Gaonkar et al. (2020). I began by starting with the original codebase used for that research and attempted to replicate the experiments that were performed for the emotion inference task. Unfortunately, the associated codebase was missing several files and modules that were required to run the experiments. Based on the files that were present, I was able to extrapolate many of the modules that were missing and created my own implementations. More details about the codebase itself and the files I created and modified can be found in the README of the codebase which is linked in section 4.6. As far as the actual implementation for the motivation inference task, I was able to map the code used in (Gaonkar et al., 2020) to the two tasks performed in this project. First, I preprocessed my data to match the same format as the data used by Gaonkar et al. (2020). Then, I modified the pre-trained BERT model to expect the labels from each of the respective motivation inference tasks. Similarly for the label semantics techniques, I was able to map the architecture that was already in place to the Maslow’s Needs and Reiss’ Motives tasks. I created a new driver file for each experiment that I conducted and details on how to run each of these as well as what changes were made to the original codebase can similarly be found in the README file at the root directory of the codebase.

## 4 Evaluation

The purpose of my evaluation is to determine if the applied label semantics techniques within this project are effective in increasing the performance of the model at the specified tasks in which predicting the labels requires an understanding of the inherent contextual semantics of the labels themselves. A general method of evaluating a system for this task requires using the same dataset on the system and comparing the performance of the system to the benchmarks established by Rashkin et al. (2018) and the state-of-the-art results established by Paul and Frank (2019). A basic approach to evaluating whether or not the system is good is to determine if the system improves upon the accuracy of the benchmarks or if the system establishes a new state-of-the-art. There are a few simple questions to ask in order to evaluate the system implemented by this project: ”Can modeling label semantics help improve performance on the task?”, ”When does modeling label semantics help improve performance on the task?”, and ”When does modeling label semantics hurt model performance?”. The answers to these questions and a specific description of my implementation are detailed below.

### 4.1 Dataset Details

The first important aspect of this system to describe is the details of the dataset. As mentioned previously, the dataset used for this task is the ROCStories dataset which has been annotated by Rashkin et al. (2018) with Maslow’s Needs labels and Reiss’ Motives labels by 3 individual annotators for character-line pairs in short story contexts. I preprocessed this original dataset to include only one of each character-line pair maintaining labels that had been selected by at least two annotators. That is, there had to be some consensus or majority agreement for the label applied to a character in a character-line pair to persist into the dataset used within this project. The annotators also distinguished between the character-line pairs in which an action was said to have occurred and those lines in which no action had occurred. Due to the nature of this task in explaining the motive behind a character, which is typically regarded as the reasoning behind a character taking a specific action, I processed the data further to create a second dataset maintaining only those character-line pairs in which an action had occurred. The first dataset consisting of all character-line pairs with motivation annotations was comprised of 36,640 samples. The second dataset consisting of only character-line pairs in which an action occurred

| **Maslow Models** | P | R | F1 |
|---|---|---|---|
| Rashkin et al. (2018) | 29.30 | 44.18 | 35.23 |
| Paul and Frank (2019) | 57.90 | 66.07 | 61.72 |
| **Reiss Models** | P | R | F1 |
| Rashkin et al. (2018) | 21.38 | 28.70 | 24.51 |
| Paul and Frank (2019) | 31.74 | 43.51 | 36.70 |

Table 1: Baseline Results on ROCStories with Maslow and Reiss motivation labels

was comprised of 26,683 samples. The samples within these datasets were separated into training and testing sets in accordance with the previous separation established by Rashkin et al. (2018). Furthermore, unique datasets were created for either task: one consisting of the data labeled with Maslow's Needs and the other consisting of the data labeled with Reiss' Motives.

### 4.2 Evaluation Measures

The primary method of evaluation for the sake of this project will be comparing the results I've obtained through running my experiments to the results established previously by Rashkin et al. (2018). More specifically, the micro-averaged F1 scores of running the testing set through the models I've trained will be used as the primary metric of comparison between my experiments and the baseline established by Rashkin et al. (2018).

### 4.3 Baselines

Table 1 showcases the baseline results that the models created by this project will be compared to. Only the best results are recorded from both Rashkin et al. (2018) and Paul and Frank (2019). The primary baseline here are the results referenced in (Rashkin et al., 2018). These results were obtained for the Maslow's Needs task using a CNN and for the Reiss' Motives task using an LSTM. The results for both tasks from Paul and Frank and recorded for the sake of parallel research and as a reference to the state-of-the-art in this field.

### 4.4 Results

Table 2 showcases the results of the models created for this project that were trained and evaluated on the dataset consisting of all character-line pairs. It can be seen here that the standard pretrained BERT model performed best out of all the experiments run for both the Maslow's Needs and Reiss' Motives prediction tasks. The pre-trained BERT model performed better at these tasks than the baseline score established in (Rashkin et al., 2018). The baseline BERT model scored approximately 0.6 F1 points higher for the Maslow Needs prediction task than the best model created by Rashkin et al. (2018). Similarly and more significantly, the baseline BERT model scored approximately 6 F1 points higher for the Reiss' motives prediction task than the best model created by Rashkin et al. (2018). Unfortunately, the label semantics techniques did not outperform the standalone pre-trained BERT model. An analysis of why this may have occurred can be found in more detail below in section 4.5.

| **Maslow Models** | P | R | F1 |
|---|---|---|---|
| BERT Pretrained | 32.38 | 40.27 | 35.83 |
| BERT + LEAM | 64.43 | 19.06 | 29.42 |
| BERT + Label Sentences | 67.99 | 16.55 | 26.62 |
| **Reiss Models** | P | R | F1 |
| BERT Pretrained | 35.02 | 27.23 | 30.64 |
| BERT + LEAM | 66.41 | 15.45 | 27.27 |
| BERT + Label Sentences | 64.42 | 11.58 | 29.40 |

Table 2: Results from BERT baseline implemented by this project as well as experiments using label semantics techniques on the dataset consisting of all character-line pairs.

| **Maslow Models** | P | R | F1 |
|---|---|---|---|
| BERT Pretrained | 58.26 | 27.48 | 37.34 |
| BERT + LEAM | 56.38 | 32.78 | 41.46 |
| BERT + Label Sentences | 58.05 | 35.82 | 44.30 |
| **Reiss Models** | P | R | F1 |
| BERT Pretrained | 59.27 | 16.91 | 26.31 |
| BERT + LEAM | 52.29 | 18.48 | 27.31 |
| BERT + Label Sentences | 57.75 | 13.79 | 22.26 |

Table 3: Results from BERT baseline implemented by this project as well as experiments using label semantics techniques on the dataset consisting of only the character-line pairs in which an action occurred.

Table 3 showcases the results of the models created for this project that were trained and evaluated on the dataset consisting of only the character-line pairs in which an action occurred. For both tasks, the evaluation of the baseline BERT model showed that the model performed worse on the task than the BERT model when all of the character-line pair samples were given. However, when the label semantics techniques were applied to this dataset, a substantial increase was seen in the accuracy for the Maslow's Needs task. Similarly, a minor improvement in accuracy was seen in the Reiss' Motives task when the LEAM architecture was used. The techniques tested through the project were able to establish an improvement in accuracy over the benchmark established by Rashkin et al. (2018). Despite this, these techniques were not able to establish a new state-of-the-art and were still less effective than the techniques used in (Paul and Frank, 2019). For reference, all of these models were only able to be trained for 5 epochs due to limited computational resources. Furthermore, all of the Maslow's Needs models reported in tables 2 and 3 were trained with a training batch size of 16 samples. Unfortunately due to a lack of computational resources, the Reiss' Motives task was only trained with a training batch size of 8 samples.

### 4.5 Analysis

1. The label semantics techniques were only effective at improving upon the baseline established by BERT on the dataset containing only the character-line pairs in which an action occurred. I believe that this is due to a skewed underlying dataset. The dataset consisting of all character-line pairs contained an abundance of samples in which there were no labels at all for the given character-line pair. This is primarily due to an artifact of stories. Many stories contain sentences that are just meant to provide context on a given situation for the reader to understand what is going on. Take for example the sentence ”A huge thunderstorm blew through town.”. These sentences often do not have any actions occurring and thus induce more noise for the model handle while trying to accomplish the tasks at hand. I believe that the label semantics techniques were able to improve upon the baseline established by BERT on these tasks while utilizing the dataset containing only those character-line pairs in which an action occurred due to the fact that with this dataset the model was able to deal with much more relevant information with less noise. When limiting the character-line pairs to only those in which an action occurred, the samples left for the model to consider had more action-centric information that the attention mechanisms could learn to focus on to produce input representations that best captured the notable information related to the contextual semantics of the task-specific labels.

2. One potential reason why the BERT baseline model, as well as the applied label semantics techniques, did not have as significant of an improvement on the motivation inference tasks as they did on the emotion inference tasks as seen in (Gaonkar et al., 2020) is likely that the motivation inference tasks are more complex than the emotion prediction task. Furthermore, the motivation labels are less frequently applied to characters in the short stories than the emotion labels. This can be shown by analyzing the rate at which characters are labeled with emotions and comparing it to the rate at which the characters are labeled with motives from either theory. Generally speaking, the characters were more frequently labeled with an emotion. On average throughout the ROCStories dataset annotated with emotions and used by Gaonkar et al. (2020), there was approximately a 3-to-1 ratio of labeled as not having the emotion to labeled as having the emotion averaged across all emotion labels. Performing this analysis on the Maslow task data revealed an approximate 10-to-1 ratio of character labeled with not having a need to character labeled with having a need averaged across all needs. Even more significantly, this analysis of the Reiss task data revealed an approximate 51-to-1 ratio of character labeled with not having a motive to character labeled with having a motive averaged across all motive labels. Clearly, there were significantly more emotion labels present than there were need or motive labels present. This is not to say that this isn't a natural artifact in the differences in occurrences between emotions, needs, and motives. Rather, the argument I am making is that this analysis is evident of the motivation inference task being more complex and less explicit than the emotion inference task.

3. One aspect that helped significantly was tuning the hyperparameter of the train batch size from 8 samples to 16 samples. Increasing the train batch size improved the accuracy in

| **Maslow TrainBatch=8** | P | R | F1 |
|---|---|---|---|
| BERT Pretrained | 61.39 | 21.16 | 31.47 |
| BERT + LEAM | 60.42 | 25.83 | 36.19 |
| BERT + Label Sentences | 62.68 | 22.46 | 33.07 |
| **Maslow TrainBatch=16** | P | R | F1 |
| BERT Pretrained | 58.26 | 27.48 | 37.34 |
| BERT + LEAM | 56.39 | 32.78 | 41.46 |
| BERT + Label Sentences | 58.05 | 35.82 | 44.30 |

Table 4: The difference in accuracy scores on the Maslow Needs task when trained with a batch size of 8 vs a batch size of 16 on the dataset including only the character-line pairs in which an action occurred

some tasks with all other variables kept consistent by up to 11 F1 points. I think that this is due to the fact that during training, the model was able to see more examples at a time and thus was exposed to more labels present within a single batch. Thus, I believe the model was able to adjust its learnable parameters more effectively over the larger batch. Unfortunately due to a lack of necessary computational resources, I was only able to experiment with setting the train batch size to 16 for the Maslow's Needs task. I theorize, however, that a similar increase in accuracy would be seen in the Reiss' Motives task if trained with a larger train batch size as well. Table 4 shows a comparison between the scores when training with a batch size of 8 against training with a batch size of 16 on the dataset containing only character-line pairs in which an action occurred.

### 4.6 Code

The code for this project can be found at the following link: Google Drive. The README.md file at the root of the repository contains the source of the original codebase, the list of files that were modified for this project, a list of instructions on how to train and test the models described in this project, and a list of major requirements that are needed to run the experiments in this project.

## 5 Conclusions

In conclusion, there were a number of key takeaways that I learned from doing this project. Generally speaking, the label semantics techniques used in this project were able to improve accuracy in the motivation inference tasks over the best benchmarks established in (Rashkin et al., 2018). Furthermore, I was able to show that the label semantics techniques were able to improve upon the BERT baseline established by this project under certain conditions. Those conditions included using a larger training batch size increasing from 8 to 16 samples, as well as limiting the data used to train the model to the character-line pairs that were marked as having an action occur. Although this reduced the number of data samples available to the model by about 10,000, the benefits from this data distribution were substantial and well worth it. Although the performance on the multi-label Maslow's Needs and Reiss' Motives prediction tasks was able to be improved by modeling label semantics, a new state-of-the-art was not established due to the findings of Paul and Frank (2019) on these tasks while utilizing a commonsense knowledge resource. Moving forward in continuing the work this project was focused on, it would be interesting to consider the results of modeling label semantics in conjunction with the commonsense knowledge base techniques that were implemented in (Paul and Frank, 2019). I theorize that by combining these two techniques, a new state-of-the-art would be established on this task. Furthermore, with additional computational resources and further analysis, potentially applying proper threshold identification, for example, I believe that the performance of the models created in this project could be improved further. Overall, through this project, I have shown a substantial increase ( 9 micro-averaged f1 points) in one of the motivation inference tasks over the benchmark established by Rashkin et al. (2018). I believe that utilizing these label semantics techniques could help improve a variety of NLP tasks in which classification labels are words with their own contextual semantic meaning and not just anonymous classes.

## 6 References

## References

Radhika Gaonkar, Heeyoung Kwon, Mohaddeseh Bastan, Niranjan Balasubramanian, and Nathanael Chambers. 2020. Modeling label semantics for predicting emotional reactions.

A. H. Maslow. 1943. A theory of human motivation. *Psychological Review*, 50(4):370–396.

Debjit Paul and Anette Frank. 2019. Ranking and selecting multi-hop knowledge paths to better predict human needs.

Hannah Rashkin, Antoine Bosselut, Maarten Sap, Kevin Knight, and Yejin Choi. 2018. Modeling naive psychology of characters in simple commonsense stories.

Steven Reiss. 2002. Who am i?: 16 basic desires that motivate our actions define our personalities.

Guoyin Wang, Chunyuan Li, Wenlin Wang, Yizhe Zhang, Dinghan Shen, Xinyuan Zhang, Ricardo Henao, and Lawrence Carin. 2018. Joint embedding of words and labels for text classification. *CoRR*, abs/1805.04174.