The Importance of Large vs. Relevant Data in Transfer Learning: Classifying Sentiment on Specialized Online Forums

Patricia Degner
School of Information, UC Berkeley

ABSTRACT

Online, there are many specialized forums and websites for niche hobbies and interests. These communities often have vernacular unique to that community. Attempting sentiment analysis of these online communities using a generally-trained classifier, such as DistilBERT, can lead to ineffective models due to this specialized vocabulary [2]. One potential solution is to train a specialized BERT model for that domain, as is done with FinBERT. Another is to use transfer learning. However, a drawback of transfer learning is the necessity of labeled data, which can be difficult to acquire for niche groups.

This paper tunes DistilBERT using two datasets: one is large, but less relevant to the task. The other is smaller, but more relevant. This experiment seeks to determine which dataset is better for transfer learning using a specific online forum, a rock climbing website called Mountain Project, as an example.

1 INTRODUCTION

The goal of this project is to create a model that can label the sentiment of an online forum in a niche community using limited training data. Although this project is focused on the rock climbing community, the methods described here, if successful, could easily be used in other domains as well. This could be useful for the participants in a given community who want to know the best techniques and buy the best gear. It would also be useful for companies that supply products for that industry; it would be useful for these companies to know how users feel about their products, and which keywords participants are using when they make a positive or negative comment about the product.

2 RELATED WORK

The main approach to this type of sentiment analysis is transfer learning. The idea is to take a general language model, such as BERT or DistilBERT, and then fine tune the model on available labeled data [1]. This is done in hate speech detection [3], or aspect-target sentiment classification [4], and is the approach taken for this experiment.

There have been a few attempts to make the standard BERT model more specialized. Most notably, DistilBERT uses knowledge distillation to retain the most important parts of BERT, while reducing overall computational expense with little sacrifice in performance. Because DistilBERT is cheaper to run, it has been used in place of BERT in this paper to illustrate the effects of transfer learning. If one desires more accuracy at the expense

of computational complexity, then it would be trivial to change the model from DistilBERT to BERT and achieve slightly better results [5].

Another approach to specialization is to train a new BERT model altogether. This was done by Dogu Araci when he created FinBERT, a specialized BERT for the finance industry. This approach used a large, unlabeled dataset to train a BERT model, then further trained the model on a smaller, labeled dataset [2].

3 METHODS

3.1 DATA:

This experiment required three datasets, all scraped from the internet.

3.1.1 Forum Data

First, the target data was scraped from the Mountain Project forums called “Climbing Gear Discussion” and “Climbing Gear Reviews”. In total there are 177,172 climbing gear discussion posts, and 12,309 climbing gear review posts. Because each post can contain multiple sentiments about multiple items, the data needed to be split further.

In a simplistic attempt to split the posts roughly into sentences, each post was split into new lines every time there was a period followed by a space. Although this approach could be problematic for some applications such as text generation or part of speech tagging, it was assumed that the impact would be minimal for sentiment analysis. After splitting each post, there were 572,786 unlabeled forum post examples. I went through and manually labeled 4,050 posts for model training and evaluation.

3.1.2 Route Data

The second dataset is filled with climbing route descriptions, the average number of stars given to that route by the users, and the number of votes that the route has. A total of 116,700 routes were scraped. Rows with empty descriptions and non-English entries were removed.

Routes without many votes are going to have unreliable star ratings. The reasons for this are detailed in Appendix A. I have removed routes that have fewer than 10 votes on the star ratings. This cut my route dataset down to 31,022 examples.

Another potential problem with this dataset is that it is not directly related to the final task. This dataset describes climbing routes, but the goal is to describe climbing gear. There is some overlap; both routes and gear can be “solid”, “old-school”, or “bomber”. However, you wouldn’t describe gear as “exciting” and you wouldn’t describe a route as “useful”. The idea is that this dataset can help DistilBERT understand the meaning of climbing slang in general, but the concern for lack of relevance led me to scrape another dataset.

3.1.3 Trailspace Data

There is a website called Trailspace that allows users to write reviews about outdoor gear. These are often detailed descriptions with pro-con lists and an associated overall star rating. It seems to be the ideal labeled dataset for this experiment, but I was only able to scrape 1,099 examples.

3.2 GOAL:

The goal of this project is to label the sentiment of the Mountain Project forum posts. This experiment seeks to determine which dataset is better for transfer learning in this task: the large, but less relevant route dataset, or the smaller, more relevant gear dataset.

3.3 PLAN:

With untuned DistilBERT as a baseline, the plan is to compare the effects of tuning DistilBERT in three ways: 1. Using the larger dataset, 2. Using the smaller dataset, 3. Using the larger and smaller datasets combined.

On top of each DistilBERT is a small, identical neural network. This network is trained on 4050 labeled examples of the target task with a random seed set to 42 to prevent variation in the way the data is split. The lower DistilBERT layers were locked, meaning they were not re-trained by the forum data. By keeping the networks identical, the only variation between the models is the dataset (or lack thereof) on which DistilBERT was trained. This will allow me to conclude which dataset did the best at tuning DistilBERT for predicting forum posts, without introducing noise from different types of models or variation in data split. Because there are three categories (positive, negative, and neutral), categorical cross-entropy was used as a loss function.

3.4 BASELINE:

The baseline model is DistilBERT with a neural network on top that is trained on 4,050 labeled forum posts.

4 RESULTS AND DISCUSSION

4.1 MODEL TRAINING

Below, Table 1 compares the training, dev, and testing accuracy, and the loss of each model:

	Baseline	Route	Route/Gear	Gear
train_acc	0.952	0.935	0.952	0.940
dev_acc	0.807	0.800	0.793	0.762
test_acc	0.805	0.816	0.819	0.770
loss	0.672	0.635	0.704	0.703

Table 1: comparison of models

The baseline model fared well, with a test accuracy of 80.5%. To my surprise, the gear model performed worse than the untuned DistilBERT, with a test accuracy of 77.0%. Images 1 and 2 shed some light on why this is the case.

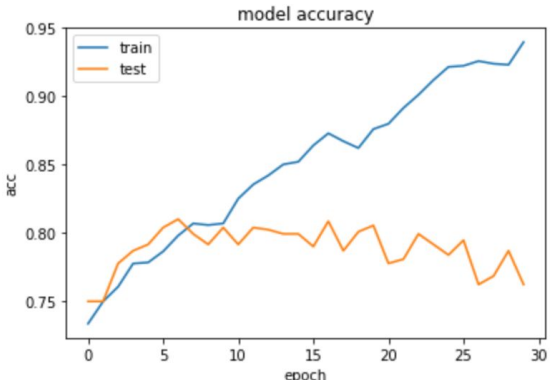


Image 1: Gear model accuracy of train and dev data by epoch

In image 1, it appears that the gear data does help the model at first, but then the model becomes overfit. Dev accuracy suffers and becomes more volatile.

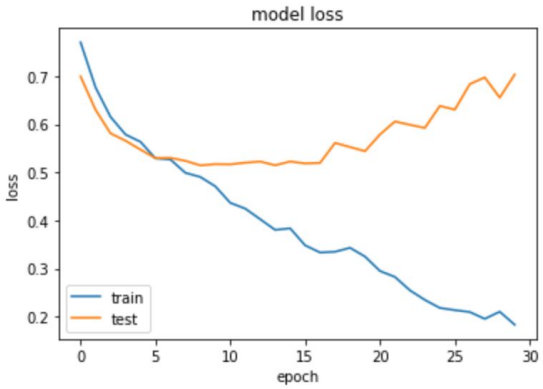


Image 2: Gear model loss of train and dev data by epoch

In image 2, it appears that the loss drops at first, then climbs as the model overfits. The graphs of all the other models are similar and can be seen in Appendix B image 1.

4.2 MODEL RETRAINING

To reduce overfitting and improve accuracy, I chose to retrain each model to a more optimal epoch. The baseline, route, route plus gear, and gear model were trained to 8, 10, 8, and 7 epochs, respectively. Table 2 shows the results of this retraining.

	Baseline	Route	Route/Gear	Gear
train_acc	0.968	0.978	0.97	0.966
dev_acc	0.801	0.812	0.801	0.806
test_acc	0.801	0.814	0.816	0.795
Δtest_acc	+.006	-.002	-.004	+.025
loss	0.74	0.748	0.8965	0.7

Table 2: Retrained model scores

As expected, the gear model improved test accuracy quite a bit, by 2.5%. The other models fared about the same, but loss was reduced in all. The route plus gear model performed the best, so I chose to explore this model further.

4.3 MODEL ANALYSIS

The route and gear model provided a test accuracy of 81.6%. This begs the question: what is happening in the other 18.4%?

4.3.1 String length

An initial look at string lengths does not show that the lengths of the mismatched strings are terribly different from the lengths of the whole dataset (see Appendix B image 2 for charts).

4.3.2 Word counts

Next, I looked at the counts of words that were mislabeled compared to the counts of words correctly labeled, both with and without stop words. In each, the counts looked similar except for two words: “cam” and “hex”. These each refer to a type of climbing gear, and I think they are mislabeled for different reasons.

Hexes are “old-school” gear that work, but are outdated. Therefore, people don’t really buy them anymore and there are a lot of mixed feelings in the forums about whether or not they are still useful. This may have confused the model when it comes to classifying sentiment when the subject is hexes.

When there is a big sale on gear, people often post about it in the forums. As I was labeling data, if the post was simply “25% off on cams at website.com”, I labeled it as neutral. Cams are expensive, they go on sale frequently, and climbers need a lot of them, so sales on cams are posted frequently; all of these postings were labeled as neutral. There is also a lot of debate about the best kind of cams, which

can lead to sentences with multiple sentiments, causing the label to come out as neutral. Additionally, people talk about cams in a neutral way when the recommended gear for a specific climb. I think that these things led my model to believe that cams are almost always neutral. This is supported by image 3.

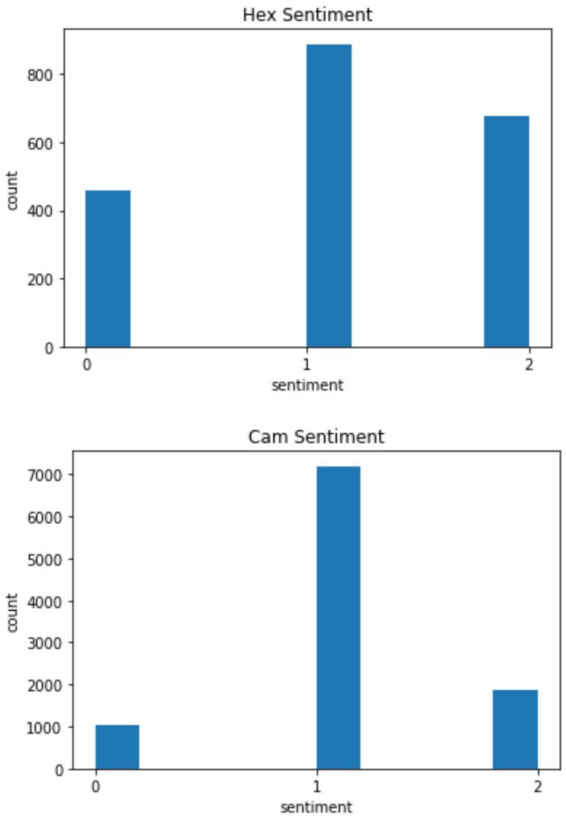


Image 3: Sentiment about hexes is quite controversial. Sentiment about cams are more often listed as neutral. Notice the difference in the number of examples; cams are far more popular than hexes.

4.3.3 Perfect bad examples

In sum, my model confuses sentiment in the following cases:

1. When there is a sale mentioned: *"25% off Black Diamond cams, the best on the market"* true: 2, label: 1
2. When the post is not directly related to climbing: *"The republican party does not support our use of public land"* true: 0, label: 1 (I suspect this is because my model is not trained for it)
3. When parallel cracks are mentioned: *"Cams are good for parallel cracks"* true: 2, label: 0 (I suspect this is because cams are the only kind of gear that work well in parallel cracks. Most posts say things like "tricams are good unless it is a parallel crack.")
4. When hexes are mentioned: *"Hexes are fine for your first rack."* true: 2, label: 0

5 CONCLUSION

This experiment attempts to determine if a large, less relevant dataset is better than a smaller, more relevant dataset for analyzing sentiment on a niche online forum. My model that had the most training examples performed the best. However, it is unclear whether this is due to the additional examples being more relevant, or if it is simply due to more examples.

I have additional route data (the routes with 9 or fewer votes). Although I suspect that this data may be less reliable, it could be useful in follow-up experiments. I could train models on only route data of different sizes, up to the 116,700 examples that I scraped, then compare. This would tell me if the additional accuracy was solely due to more data, or if the specificity of the small gear dataset helped.

Although it cannot be concluded that inclusion of more a smaller but more

relevant labeled data improves the model, it can be concluded that more data is indeed better than less, even if the larger dataset is a little bit less relevant. This is evidenced by the comparison of the gear only and route only models. However, the relevance of the gear dataset may or may not have improved the final model; further experimentation is needed to make that conclusion.

APPENDIX A

Why might a route with fewer ratings provide unreliable data?

First of all, routes with fewer ratings are more prone to outliers. Route star ratings are purely subjective, and the enjoyment of the route can depend on factors such as strength or height of the climber, how the weather was when the person tried it, how previously climbed routes compare to this one, etc. On any route, one person may think it is terrible, where another loved it. Therefore, the average star rating becomes more reliable as more people vote on it.

Second, the first ascent party will often inflate the rating of a route because routes are expensive, in time and money, to put up. Each bolt on a rock climb costs $5-$10, so a standard sport climb with 10-12 bolts costs around $100 out of the personal pocket of the climber who established it. Each bolt can take fifteen minutes to an hour to place, for a total of three to ten hours per route. A trad climb with no bolts can still be a big investment; a trad climber will be taking on greater risk, since they will not know exactly what gear they will need to safely complete the climb. It also takes time to find a potential new route, and then clear the route of loose rocks, dirt, and other debris. There can be legal problems getting permission to put up rock climbs on private or public land. All of these issues make route establishment a very time-consuming process. The reward for the first ascent party is their name in the guidebook and on Mountain Project, and they get to name the route. However, many first ascent parties want others to climb the routes they put so much effort into. So, the first ascent party will sometimes inflate the star rating on Mountain Project to entice more climbers to try it.

On the other hand, sometimes climbers in a small local community do not want their small crag to be swarmed with crowds drawn by lots of high-star climbs. In this case, the climbers may mark good routes as being very bad, to scare away masses of climbers.

Finally, there are many “closed projects” listed on Mountain Project. This means that the climb has been found, permission to climb it obtained, and it has been cleaned and bolted (if there are bolts on it), but the climb is very technically difficult and the person who put in the work to establish it has not yet been able to climb it without falling. It is common courtesy to allow the establisher to climb and name the route first, so these climbs will not have any ascents and will have unreliable star ratings.

APPENDIX B

DistilBERT only:

acc = 80.4%

Route only:

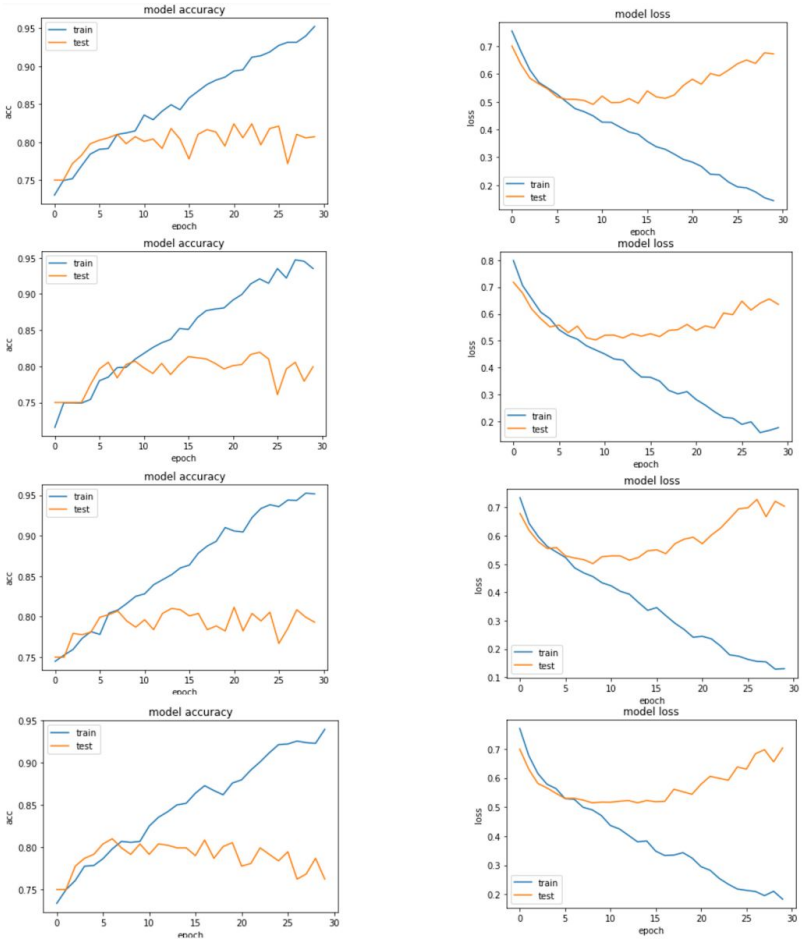
acc = 81.6%

Route and gear:

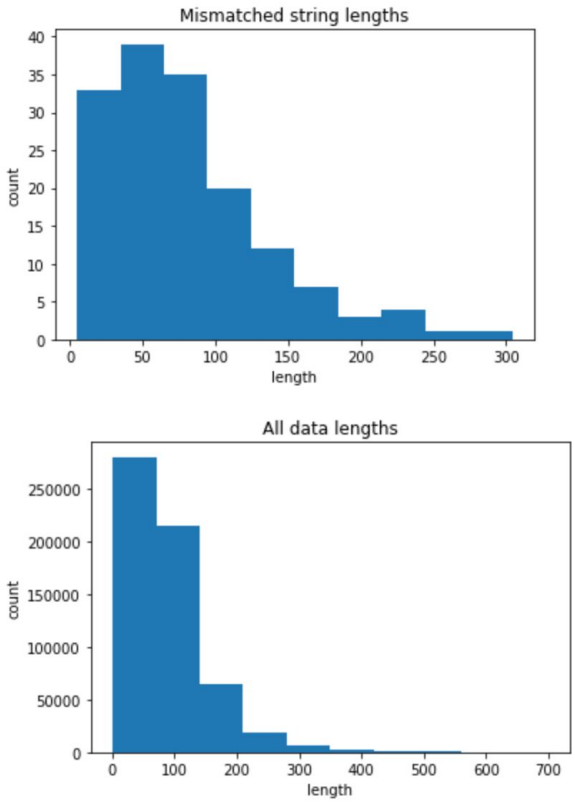
acc = 81.9%

Gear only:

acc = 77.0%



Appendix image 1: This image depicts the charts of the first round of model training. "Acc" refers to test accuracy. The "test" in the graphs refers to the dev data.



Appendix image 2: Length of mismatched strings compared to all strings.

BIBLIOGRAPHY:

[1] Dong, Xin, and Gerard De Melo. “A Helping Hand: Transfer Learning for Deep Sentiment Analysis.” *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, doi:10.18653/v1/p18-1235.

[2] Liu, Zhuang, et al. “FinBERT: A Pre-Trained Financial Language Representation Model for Financial Text Mining.” *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, 2020, doi:10.24963/ijcai.2020/622.

[3] Mozafari, Marzieh, et al. “A BERT-Based Transfer Learning Approach for Hate Speech Detection in Online Social Media.” *Complex Networks and Their Applications VIII Studies in Computational Intelligence*, 2019, pp. 928–940., doi:10.1007/978-3-030-36687-2_77.

[4] Rietzler, Alexander, et al. “Adapt or Get Left Behind:Domain Adaptation through BERT Language Model Finetuning For Aspect-Target Sentiment Classification.” *DeepOpinion.ai*, 2019, doi:https://arxiv.org/pdf/1908.11860.pdf .

[5] Sanh, Victor, et al. “DistilBERT, a Distilled Version of BERT: Smaller,Faster, Cheaper and Lighter.” *HuggingFace*, 1 Mar. 2020, doi:https://arxiv.org/pdf/1910.01108.pdf.

[6] Tetsuya Nasukawa and Jeonghee Yi. 2003. Sentiment analysis: capturing favorability using natural language processing. In *Proceedings of the 2nd international conference on Knowledge capture* (*K-CAP ’03*). Association for Computing Machinery, New York, NY, USA, 70–77. DOI:https://doi.org/10.1145/945645.945658