Few-shot Classification of Disaster-related Tweets

Stanford CS224N Custom Project

Jubayer Ibn Hamid
Department of Computer Science
Stanford University
jubayer@stanford.edu

Jitendra Nath Pandey
Department of Computer Science
Stanford University
jnpandey@stanford.edu

Sheikh Rifayat Daiyan Srijon
Department of Computer Science
Stanford University
srijon@stanford.edu

Abstract

Social media is a very powerful tool in helping emergency aid centres and response operators in coordinating a response to a crisis. Platforms like Twitter allow information to travel fast making coordination with people at the scene is easier and, therefore, allowing response operators to attain higher situational awareness (Vieweg, 2012). However, lack of filtration methods on these platforms means that there remains possibilities of the spreading of false news. This skepticism has curtailed our ability to respond to crises in a timely manner. An AI-driven solution to this problem *needs* to be able to perform well even when it has been trained on a small labelled dataset. As (Chowdhury et al., 2020) discusses, most of the work in this domain has been with regards to classifying posts that have been written in English only and if one were to finetune a model for each disaster, the dataset would be even smaller. In this project, we analyse the performance of language models (both base models and those that incorporate few-shot learning) in classifying disaster-related Tweets as either true or false on few-shot datasets. Particularly we analyse the performance of base DistilBERT models with pretraining, with supervised contrastive learning that enhances the loss function to get better results with fewer training examples, and with Prototypical Neural Networks. We find that large language models like DistilBERT are good at few-shot learning of classification of disaster-related tweets even without incorporating few-shot learning techniques and show lower degradation of performance with shrinking of datasets. Our research reinforces the hypothesis from OpenAI (2020) that pretraining scaled-up language models on large corpuses of data improves task-agnostic performance using strong generalised, representation of language and that finetuning on noisy datasets worsens performance in few-shot learning. Our analysis of the results suggests that large pretrained language models perform very well at few-shot learning due to learning of strong representations of language make them task-agnostic few-shot learners. In particular, we find that, comparatively, fine-tuning can even *worsen* performance when noisy datasets damage the representational learning of these large language models.

1 Key Information to include

- Mentor: Swastika Dutta
- External Collaborators (if you have any): N/A
- Sharing project: N/A

2 Introduction

Disasters are high-pressure situations in which response operators need to act fast and deploy resources very efficiently. For that, it is crucial that they have access to key information that increases their situational awareness. Research suggests that social media can be a powerful tool in enabling that (Vieweg, 2012). However, on most social media platforms like Twitter there is a lack of filtration method specifically aimed at filtering posts regarding disasters on the basis of whether they are true or not. As such, social media posts are both high reward (if you respond correctly to posts that spread true information) and high risk (if you respond on the basis of *misinformation* spread by posts). It is imperative that we can filter disaster-related posts based on whether they are spreading misinformation or true information.

One of the biggest challenges in this area is the shortage of data. First of all, there is a shortage of data in multiple languages. This severely limits performance of models on non-English languages as seen in the results section of Chowdhury et al. (2020). Furthermore, even for English language, there are strong imbalances in data. For example, in the dataset CREDBANK (Mitra and Gilbert, 2015), the vast majority (> 95 percent) has been labelled as certainly accurate whereas *only one* has been labelled as certainly inaccurate - which suggests that data imbalances are very likely in the realm of disaster-tweet classification. In many cases, we want to finetune models with respect to each instantiation of disasters by using only data from that disaster respectively and in such cases insufficiency of data becomes an even more acute problem.

In this project, we implemented analysed the performance of three different models; a DistilBERT model with cross-entropy loss, a DistilBERT with supervised contrastive loss and a prototypical neural network. We trained these models on datasets of various sizes and evaluated them to analyse their performance in few-shot classification. We conclude that the DistilBERT and Prototypical Neural Network performs better (with some differences in precision versus recall) at few-shot classifications but, overall, baseline language models are good at learning from small datasets, which confirms the study (OpenAI, 2020).

3 Related Work

Few-shot classification aims to learn a classifier using a small number of labelled training examples. Several different approaches have been taken train a model to do this. Initialised-based methods tackles the problem by training models to be able to learn to finetune; some attempt to train to learn good model intialisations (Finn et al., 2017) whereas other models are trained to learn an optimiser (Ravi and Larochelle, 2017). In this paper, we explore distance-metric learning based methods; specifically we analyse the performance of Prototypical Neural Networks (Snell et al., 2017) which are models that learn to embed classes into a class-space and learn to embed each input into the class space. The class label is then found by measuring similarities via norms in that class-space.

Supervised contrastive loss in classification is a family of loss-functions that are widely used in natural-language processing problems (Beliz Gunel). In the case of binary classification, we work with a batch of training examples of size N : $\{x_i, y_i\}_{i=1,...N}$. Furthermore, let N_{y_i} be the total number of examples in the batch that have the same label as y_i. Let $y_{i,c}$ be the true label and $\hat{y}_{i,c}$ is the model output for the probability of the i-th example belonging to the class c.

Now, suppose, $\Phi(\cdot) \in \mathbf{R}^d$ is the encoder that outputs the l_2 normalized final encoder hidden layer before the softmax projection. The overall loss is a weighted average of cross-entropy (CE) and the proposed supervised contrastive learning (SCL) loss, as denoted by $\mathcal{L}$:

$$\mathcal{L} = (1 - \lambda)\mathcal{L}_{CE} + \lambda\mathcal{L}_{SCL}$$

where λ is a hyperparameter, $\mathcal{L}_{CE}$ is the cross-entropy loss defined as

$$\mathcal{L}_{CE} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{2} y_{i,c} \cdot log\hat{y}_{i,c}$$

and the contrastive loss is

$$\mathcal{L}_{SCL} = -\sum_{i=1}^{N}\frac{1}{N_{y_i} - 1}\sum_{j=1}^{N} 1\{i \neq j\} \cdot 1\{y_i = y_j\} \log\frac{\exp\left(\Phi(x_i) \cdot \Phi(x_j)/\tau\right)}{\sum_{k=1\ i\neq k}^{N}\exp\left(\Phi(x_i) \cdot \Phi(x_k)/\tau\right)}$$

.

Our final implementation of the supervised contrastive loss function was inspired by the implementation in Khosla et al. (2020).

Prototypical Neural Networks (PNN) are a family of neural networks that aim to do few-shot classification of training examples across *unseen* classes. Suppose S_i be the set of all training examples which are in class i. PNN computes a M-dimension representation/embedding (called a *prototype*) for each class through the function $f_\phi : \mathbb{R}^D \to \mathbb{R}^M$ with parametres ϕ. Using this, the prototype for class i is defined as

$$c_i = \frac{1}{|S_i|} \sum_{(x_i, y_i) \in S_i} f_\phi(x_i).$$

Using this, we can compute the probability of each input being in class i using

$$p_\phi(y = i|x) = \frac{\exp(-d(f_\phi(x), c_i))}{\sum_k \exp(-d(f_\phi(x), c_k))}$$

where $d(x, y)$ is the Euclidean-norm of x and y.

4 Approach

The first model we trained is a **DistilBERT model**. We pretrained this model by masked-language modelling. Instead of pretraining on all kinds of tweets, we pretrained on disaster-related tweets only so that the model learns stronger representations for words that specifically appear in disaster-related tweets. We then finetuned the model by adding a softmax classifier. In this first model, we only used cross-entropy loss.

Next, we trained a DistilBERT model with supervised contrastive loss and then tuned the hyperparametres λ and τ. In order to ensure that the model is forced to capture similarities between examples in one class and contrasting them with examples in other classes, we ensure that $\lambda \geq 0.5$.

Lastly, we trained a Prototypical Neural Network using a learnable embedding matrix. We used a distilBERT model with a prototypical neural network head. Across all models, we used dropout to prevent overfitting on small datasets.

All these models have been pretrained and fine-tuned (HuggingFace) with 5 transformer blocks and incorporating dropout (to reduce overfitting) and layernorm. We also employed early stoppping of training when we saw insignificant improvement in performance to prevent overfitting which we think is an important concern when we train on few-shot datasets. The models were trained to be able to identify a tweet related to a disaster as either true (correct information) or false (misinformation). Note that we are *specifically* not training the models to idenity uninformative posts, we only want to identify them as either true or false.

5 Experiments

5.1 Data

We trained on three datasets - two from Kaggle (Kaggle) and the other being the CRISIS6 dataset (CrisisMMD). These are labelled datasets that have Tweets labelled as either true or false. We tested our algorithms on four different sizes of datasets - 5700 (full), 1000, 100 and 10 examples. Each of these four datasets were created by randomly sampling datapoints from the full dataset.

5.2 Evaluation method

We evaluate the performance of each model on each dataset using both accuracy and F1 score which combines the precision and recall scores of a model. To define the F1 score, define the following:

True Positives (TP): Number of samples correctly predicted as "positive."
False Positives (FP): Number of samples wrongly predicted as "positive."
True Negatives (TN): Number of samples correctly predicted as "negative."

False Negatives (FN): Number of samples wrongly predicted as "negative."
Then, the F1 score can be computed a $F1 = \frac{TP}{TP+\frac{1}{2}(FP+FN)}$.

5.3 Experimental details

First, we report the hyperparametres for our models. Note that for all the models, we early stopped training if we saw no improvement in validation loss over 10 epochs in order to prevent overfitting on small datasets. For the DistilBERT model with cross-entropy loss, the hyperparametres (after tuning) are as follows:

Hyperparametre	*Value*
Learning rate	10^{-5}
Decay factor	0.8 (after every 10 epochs)
Number of training epochs	200
Dropout (for each transformer block)	0.1
Dropout (final)	0.5
Number of transformer blocks	5

For the DistilBERT model with supervised contrastive learning loss, the hyperparametres (after tuning) are as follows:

Hyperparametre	*Value*
Learning rate	10^{-5}
Decay factor	0.8 (after every 10 epochs)
Number of training epochs	200
Dropout (for each transformer block)	0.1
Dropout (final)	0.5
λ (weight of SCL loss function)	0.9
Number of transformer blocks	5

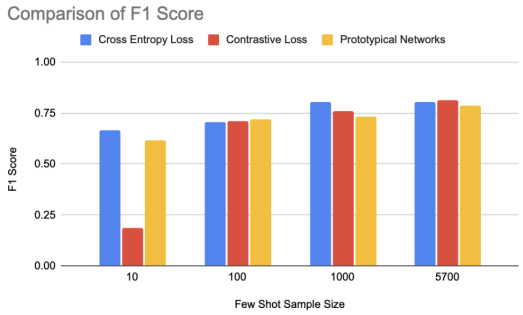
For the Prototypical Neural Networks model, the hyperparametres (after tuning) are as follows:

Hyperparametre	*Value*
Learning rate	$6 \cdot 10^{-5}$
Decay factor	0.8 (after every 10 epochs)
Number of training epochs	200
Dropout (for each transformer block)	0.1
Dropout (final)	0.5
λ (weight of SCL loss function)	0.9
Number of transformer blocks	5

As mentioned before, we used four different sizes of datasets. For the smaller ones, we randomly sampled x training examples where x is the size of the training set. We resampled if we saw significant imbalance in the dataset.

5.4 Results

First, we analyse the F1 scores attained by the three models on different datasets:

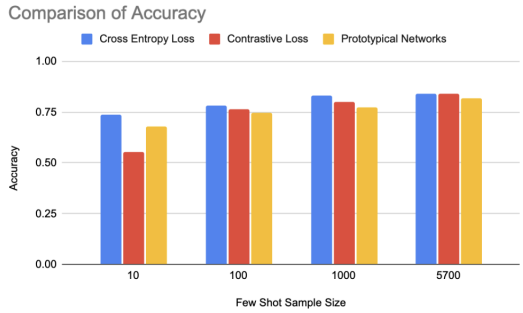


First, we observe that the performance of distilBERT with supervised contrastive learning (SCL) loss falls rapidly as the dataset gets to a size of 10. On the smallest dataset of 10 examples, SCL attains an F1 score of only 0.186 while the others gets > 0.6.

Before this, however, SCL performed as well as the other models; in fact, it performed better than distilBERT with cross-entropy loss on 100 training examples. With the *full* dataset, SCL performs better than all other models - SCL attains F1 score of 0.813 whereas distilBERT attains (a marginally less value of) 0.806 and prototypical NN attains 0.784.

It is clear that the performance of DistilBERT with cross-entropy loss and Prototypical Neural Networks are more stable with differences in sizes of dataset and have less variance in model performance.

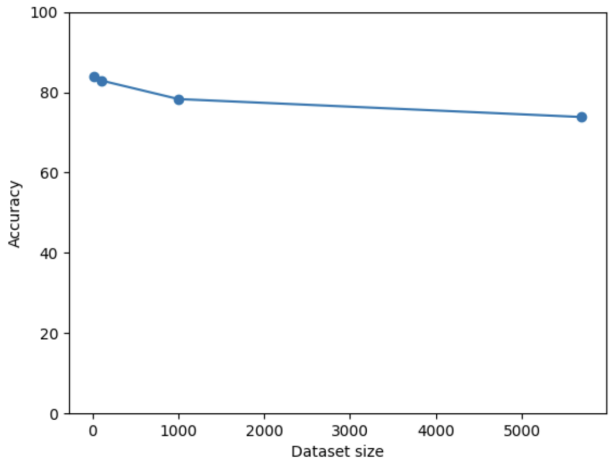
Next, we analyse the accuracy of the models.



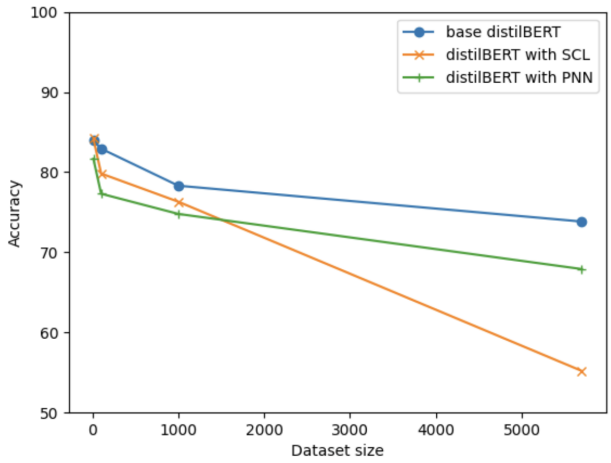
We note that the distilBERT with cross-entropy loss is superior to all models on all sizes of data in terms of accuracy. Once again, we note that the performance of SCL drops significantly when we make the dataset a size of 10; on size 10, it attains an accuracy of only 0.552 while on datasets ≥ 100, it attained accuracy ≥ 76. Before that, it attained accuracy approximately the same as DistilBERT with cross-entropy and (marginally) outperforming prototypical neural network.

6 Analysis

Firstly, we observe that the base distilBERT model with cross-entropy loss performs notably well at few-shot learning, which reaffirms analysis in OpenAI (2020).



Note that the accuracy attained by the distilBERT model changes very little as the dataset size is changed. This coincides with the hypothesis that scaling up language models greatly improves task-agnostic, few-shot performance. The DistilBERT model is a large language model with 66 million parametres and has been trained on 3.3 billion words. In fact, using supervised contrastive learning loss and prototypical neural head on top of the distilBERT model seemed to either make no improvement in performance or sometimes *worsen* the performance.



We believe this is because large language models attain strong representations of language in general and that Tweets are not out-of-distribution enough to hurt their performance in this particular domain.

Furthermore, we notice that there is a significant change in performance as the dataset is shrunk from size 100 to 10, compared to other shrinks. This suggests that there is significant noise in the dataset which "cancels out" only when the dataset is extremely large. For small datasets, when finetuned with this small, noisy dataset, model performance sees large variance. Notably, the performance degrade when shrinking from 5700 to 100 training examples is not as large as it is when shrinking from 100 to 10. Our hypothesis here is that a training set of 100 examples is still fairly strongly representative of the distribution of Tweets related to disasters but 10 examples is harmful for robustness.

Our hypothesis is further strengthened by observing the performance of prototypical neural networls. Note that for binary classification the prototype for each class will be

$$c_i = \frac{1}{|S_i|} \sum_{(x_i, y_i) \in S_i} f_\phi(x_i).$$

If the dataset is extremely noisy, this unweighted mean embedding of the two classes will vary a lot. This is because if the embeddings are of 768 dimensions, then the variance

$$|I^T \Sigma I|$$

where I is the identity and Σ is the covariance matrix of f_ϕ is going to be very large. This means its performance in a *different* noisy dataset of small size will be bad. In comparison, DistilBERT without the prototypical neural head will not stray from the original representations much when it employs cross-entropy loss on the small noisy datasets.

7 Conclusion

In conclusion, we find that base large language models without incorporating few-shot learning loss functions are relatively better at classification of disaster-related Tweets than those that do employ them. Further work could be done to explore other large language models on few-shot learning datasets. Furthermore, due to limitation of GPU capacity, we could not try some other interesting experiments but further work should explore deploying pretrained large language models without *any* finetuning at few-shot learning by just adding classifier heads on top of them.

References

Alexis Conneau Ves Stoyanov Beliz Gunel, Jingfei Du. Supervise contrastive learning for pre-trained language model fine-tuning. In *ICLR 2021*.

Jishnu Ray Chowdhury, Cornelia Caragea, and Doina Caragea. 2020. Cross-lingual disaster-related multi-label tweet classification with manifold mixup. In *Association for Computational Linguistics 2020*.

CrisisMMD. Crisismmd: Multimodal crisis dataset.

Chelsea Finn, Sergey Levine, and Peter Abbeel. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. *Proceedings of the 34th International Conference on Machine Learning, Sydney, Australia, PMLR 70, 2017*.

HuggingFace. Fine-tuning a masked language model. *https://huggingface.co/course/chapter7/3?fw=tf*.

Kaggle. Disaster tweets dataset. *https://www.kaggle.com/datasets/vstepanenko/disaster-tweets*.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *arXiv preprint arXiv:2004.11362*.

Tanushree Mitra and Eric Gilbert. 2015. Credbank: A large-scale social media corpus with associated credibility annotations. *Vol. 9 No. 1 (2015): Ninth International AAAI Conference on Web and Social Media*.

OpenAI. 2020. Language models are few-shot learners.

Sachin Ravi and Hugo Larochelle. 2017. Optimisation as a model for few-shot learning. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Jake Snell, Kevin Swersky, and Richard S. Zemel. 2017. Prototypical networks for few-shot learning. *CoRR*, abs/1703.05175.

Sarah Elizabeth Vieweg. 2012. Situational awareness in mass emergency: A behavioral and linguistic analysis of microblogged communications. In *PhD. Thesis, University of Colorado at Boulder.*