

“Emotion” and “Tweet Eval’s Irony” Datasets

Data Splits

Using split configuration, we have the split size as below:

	Train	Validation	Test
Emotion	16000	2000	2000
Tweet Eval	2862	955	784

Models Performance

After fine tuning the models:

	Dataset	Training-accuracy	Validation-accuracy	Test-accuracy
Base BERT	Emotion	93.5%	92.2%	91.8%
DistilBERT	Emotion	99%	93.7%	92%
ALBERT	Emotion	96.5%	93.9%	92.2%
Base BERT	Tweet Eval	96%	70.7%	69.4%
DistilBERT	Tweet Eval	67.7%	59.4%	63.6%
ALBERT	Tweet Eval	88.8%	68.3%	66.8%

Comparing the results on test-set accuracy before and after fine-tuning:

	Before fine-tuning	After fine-tuning
Base BERT	10.6%	91.8%
DistilBERT	9.75%	92%
ALBERT	9.75%	92.2%
Base BERT	52.1%	69.4%
DistilBERT	52.2 %	63.6%
ALBERT	45.5%	66.8%

Confusion Matrix for validation set for tweet eval's irony task:

Base BERT: [[331 168]

[111 345]]

DistilBERT: [[320 179]

[209 247]]

ALBERT: [[352 147]

[156 300]]

Qualitative Analysis

Emotion dataset

	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5
text	i accept the medication until i dont feel too troubled by those i will never have the full benefices from them	i get up to refill my coffee and feel that pleasant and familiar ache it reminds me how much i miss the whole body conversations you can have when you're sitting on a good good horse	ive explained that he is very creative and loves to makes things and i feel that he is very smart and intelligent and he is lacking in some areas that i agree with	i was feeling sentimental	when i was walking around all alone at night
Real label	0	1	1	0	4
BERT Prediction	0	1	1	0	3
text	i began feeling shaky my heart was sort of skipping around i felt like someone who had been drinking coffee all day long	i have a feeling hes going to be way more successful than i am	i eat or sleep i cant get myself to feel the life loving energy i felt so easily before	i love those kiddos and yet am left feeling so helpless	im just feeling personally devastated that this happened at my college in the school im studying under
Real label	4	1	2	0	0
DistilBERT Prediction	4	1	2	4	0
text	i eat or sleep i cant get myself to feel the life loving energy i felt so easily before	i love those kiddos and yet am left feeling so helpless	im just feeling personally devastated that this happened at my college in the school im studying under	im still feeling a little shocked over yesterdays news that pope benedict xvi has decided to resign	i want to feel pretty or handsome or something
Real label	2	0	0	5	1
ALBERT Prediction	2	4	0	5	1

Tweet Eval dataset

	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5
text	I bet I'm the first to use that one	Chenua Achebe - when things fall apart	@user @user the upper deck definitely seemed like Philly South.	@user COD AW SERVERS #OFFLINE	My year is ending perfectly 😊
Real label	1	0	0	0	1
BERT Prediction	1	0	1	0	0
text	Women do some crazy ass shit just to be with a guy and the least they can do for her is break her heart! all but some	@user @user @user Probably folks like #FullMcIntosh who don't buy games. #gamergate	MORON, no one has ever said a ball is square #STRAWMAN @user @user	girls put so much effort in dressing nice for school..I'm over here in my oversized sweats and vans #maybetomorrow	Bruh. I just realized I'm wearing her shirt. The chick who dumped me. Lmfao
Real label	0	1	0	1	1
DistilBERT Prediction	0	1	0	1	1
text	I bet I'm the first to use that one	Chenua Achebe - when things fall apart	@user @user the upper deck definitely seemed like Philly South.	@user COD AW SERVERS #OFFLINE	My year is ending perfectly 😊
Real label	1	0	0	0	1
ALBERT Prediction	0	0	1	1	1

Experimentation Procedure Discussion

In the experimentation process for fine-tuning for both tasks, an extensive hyperparameter search was conducted to optimize model performance. For the project, we used the Ray library to tune the hyperparameters of our models to efficiently find the best hyperparameters by searching through a specified range for each parameter. We focused on four main hyperparameters: learning rate, batch size, number of epochs, and weight decay, setting a specific search range for each based on our available computational resources. Due to the limits of our GPU power, we had to work with a smaller portion of our data—only 1/10th of our training and testing sets by utilizing dataset sharding. It allowed a rapid iteration over numerous configurations with a manageable computation. Our goal was to maximize the accuracy of our models by finding the optimal combination of these hyperparameters.

The hyperparameter tuning was performed over 30 trials, each evaluating a different combination of parameters, thereby providing a robust comparison to determine the most effective configuration. However, we faced challenges due to the limited memory of our GPU, which restricted our ability to experiment with more tuning strategies, such as freezing certain layers of the models. In some cases, like when tuning the BERT model on an emotion dataset, the tuning process got terminated because of the limitation in GPU space. Despite these all, we adjusted the number of tuning trials based on the task and the computational resources we had, aiming to use the best hyperparameters we found to improve our models. This process varied in length depending on the specific task and the limitations we faced.