Error Analysis

Team 11 - English

Tudor Pitulice (234803)

Noah Ivanisevic (235738)

Zakariae El Moumni (226324)

Introduction

During task 5, team 11 fine tuned numerous models for the task of sentiment classification. The team have developed in total a numeber of 6 models: Naive Bayes, Logistic Regression, Recurrent Neural Networks (or shortly RNNs), Long Short Term Memory models (LSTM) and 2 transformers. All presented difficulties in classifying correctly the emotion of the sentence, however most suitable for the task in question is the transformer. Thus, this document represents a brief explanation of the behaviour of the DistilBERT model.

DistilBERT

DistilBERT is the child of the parent model BERT, which are both great transformer models developed by Hugging Face and respectively Google in 2018. BERT was created for Natural Language Processing tasks such as: text classification, question answering, and more. DistilBERT on the other hand, is a smaller variant of BERT which uses the same concept of bidirectional processing of the text with a twist! DistilBERT has 40% parameters retaining 97% of the accuracy of BERT. So, BERT excels in accuracy but is resource-intensive, while DistilBERT offers efficiency with minimal trade-offs, making it suitable for practical deployments.

Our Implementation and Error Analysis

Tudor was responsible for the development of the transformer models on the duration of Block C. The start of the process started on week 4 and finished in week 7, which represented a significant challenge for the team. The morale of the team was low when the performance of the models were not satisfactory and did not meet the requirements of the clients in spite of the efforts the team put in. During iteration 2 of the DistilBERT model, the f1 score reached 0.28 weighted average with an accuracy of 41%.

The classification report presented below, shows a more holistic view of the performance of the model per class.

	Precision	Recall	F1-score
Anger	0.00	0.00	0.00
Disgust	0.00	0.00	0.00
Fear	0.00	0.00	0.00
Happiness	1.00	0.01	0.01
Neutral	0.39	0.99	0.56
Sadness	0.00	0.00	0.00
Surprise	0.78	0.13	0.22
Accuracy			0.41
Weighted Average	0.58	0.41	0.28

From the classification report, the models seems like it cannot identify emotion such as anger, disgust, fear or sadness. This factor is explained by the imbalanced distribution of emotion in the test set. The team tested the performance of the model on a dataset labelled by the company pipeline. Thus, the distribution is as followed:

Emotion	Number of appearances
Neutral	347
Surprise	299
Happiness	182
Fear	54
Anger	34
Sadness	24
Disgust	1

This explain a part of the behaviour of the model. The model predicts the majority of the 2 classes correctly: Neutral and Surprise which encompasses almost 68% of the test set. This skewed distribution of the test set is letting us know only the performance on these 2 classes.

Also, another important insight is that for the emotion „happiness”, the precision is equal to 1.00 while the recall is 0.01. That means that the model is very carefull on the prediction that it makes to be correct. It has predicted only 1% of the true positives as happiness, all of them being correct. The rest 99% of the sentences for „happiness” are being classified as another emotion. This insight, is telling us that either the model is too focused on as few false positives as possible with the cost of having a lot of false negatives for this class.

With the information available at this point, we know that the performance of the model is not satisfactory on any of the emotions. To further investigate the issue, the team have checked individual predictions that were wrong.

Text: Let's see if they figure it out. So it's to vote for someone...

True emotion: surprise

Predicted emotion: neutral

Text: Since half of you could not agree on someone to eliminate someone has to now randomly be eliminated....

True emotion: surprise

Predicted emotion: neutral

Text: More than half of us have to agree to drop one person...

True emotion: happiness

Predicted emotion: neutral

Text: Let's see if they can make a decision this time and the one thing that these contestants all have in...

True emotion: surprise

Predicted emotion: neutral

Text: And were eliminated in my global hit series on prime video beast games...

True emotion: happiness

Predicted emotion: neutral

This provides a clear insight into the underlying problem of the model. This is a small sample of the whole output, however, nonetheless useless. The output of the client pipeline is not the ground truth and presents flaws. Before jumping the conclusions, one important factor to keep in mind is that the pipeline of the output is classifying not on 7 emotions, but in 27 emotions. When manually checking the test set in Excel, the prediction seemed satisfactory with only around ~6% error. Thus, the problem lies in the mapping of the emotions in the pipeline. The emotion that are not intense, can be classified as neutral instead and by double checking the mapping we can boost the performance of the model.