Brendan Magdamo, Brendan Manning, Stephanie Scheerer

Projects in Data Science

Professor Belorkar

12 April, 2021

3.4 - Performance Report

Performance Metric: F1 Score

Definition, Computation

F1 score is defined as the harmonic mean of precision and recall. The F1 score can be calculated as $score = 2\frac{precision \cdot recall}{precision + recall}$. Precision is defined as $P = \frac{T_p}{T_p + F_p}$ and recall is defined as $R = \frac{T_p}{T_p + F_n}$.

A convenient method for computing the F1 score can be found in the `sklearn.metrics.f1_score` package. A typical invocation might look like `f1_score(y_true, y_pred)` where `y_true` is a 1D array containing the "ground truth" labels and `y_pred` is a 1D array containing the predicted labels returned by the model. Further configuration can be done with `average` property in the case of multiclass/multilabel classification.[1]

Motivation for Use

[1] This is outside the scope of our project since the model simply returns a 1 (disaster) or 0 (no disaster).

The Kaggle competition is utilizing F1 score for evaluation which is the primary reason why we selected it for our performance metric but it is also a standard method for evaluating machine learning models as it provides a holistic view of model performance. Accuracy is the simplest evaluation metric and depending on the situation can be a good choice. However, in the case of building a model with training and test data, accuracy only provides a partial understanding of model performance as it does not account for the difference between true and false positives. This can pose a problem as it can lead to overfitting which is why F1 is a better choice for model evaluation. F1 score accounts for the tradeoff between precision and recall which provides a more detailed view of performance than simple accuracy. Furthermore, the calculation for F1 is relatively straightforward compared to other methods for evaluation such as Mean Squared Error or Cross Entropy Loss that require a stronger understanding of mathematics. For F1 score, someone without a technical background can intuitively grasp the concept of a basic 2x2 confusion matrix of true and false positives and negatives and how that translates into the formula for F1 making it a good choice for our selected performance metric.

Evolution of F1 Score

Figure 1 below shows the evolution of our F1 score during the different phases of our project. Here I will give a brief description of what had been accomplished during each of these phases.

Baseline - Our baseline model included the vectorization of words using the TF-IDF method. During preprocessing, we removed emojis, stop words, links, and punctuation. The F1 score achieved was 0.74663.

DistilBERT with Optuna - In this phase, we tried using a DistilBERT model, which runs faster than a standard BERT model. We used Optuna to do hyperparameter optimization. The F1 score achieved was 0.773.

Simple Transformers with RoBERTa - Here, we changed the model that we were using to RoBERTa, which is more accurate than DistilBERT. We used the Simple Transformers package to do our training. Stop words and emojis were kept in at this stage. The F1 score achieved was 0.81857.

Back Translation - As a form of data augmentation, we used back translation which is the process of translating the text to a different language and then translating it back to the original language. We translated to-and-from French for these Tweets. The F1 score achieved was 0.82592.

Abbreviation and non-ASCII Replacement - At this stage we altered our preprocessing steps; we had noticed that some emojis had not rendered correctly and were being displayed as non-ASCII characters. In order to account for this, we removed all of these non-ASCII characters that may have been confusing the model. We also replaced abbreviations with their true meanings at this stage. The F1 score achieved was 0.83512.

Simple Transformers with Optuna - During the final stage of modeling, we used Optuna to optimize the hyperparameters for the Simple Transformer model. The F1 score achieved was 0.83818.

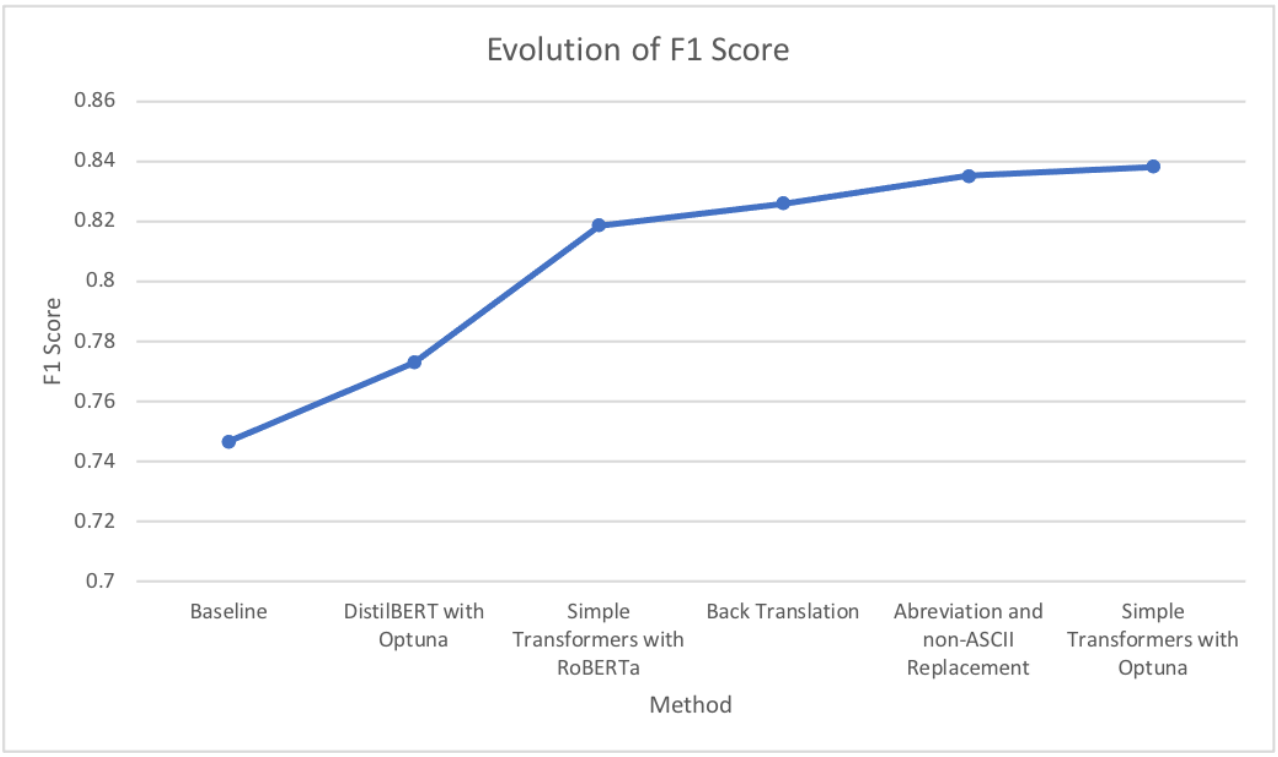

Figure 1: Evolution of F1 scores over different phases of our project.