# **Tweets Disaster or Not ? EXPERIMENTS**

[**Link to the Kaggle Problem**](https://www.kaggle.com/c/nlp-getting-started)

**Data Split: 70/30**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **S.no** | **MODEL** | **PROCEDURE** | **F1 SCORE** | **LESSONS LEARNT** |
|  | **BERT** | 1. **Manual Preprocess + Bert Preprocess** 2. **Dropped missing values columns** | **0.77** | **Avoid Manual Preprocess** |
| **2.** | **BERT** | 1. **Bert Preprocess** 2. **Included Missing values columns** | **0.79** | **Bert preprocess > Manual**  **More Features = More Score** |
| **3.** | **BERT** | 1. **Bert Preprocess** 2. **Dropped missing values columns** | **0.78** | **Less Features = Less Score** |
| **4.** | **ELECTRA** | 1. **Electra Preprocess** 2. **Dropped missing values columns** 3. **Epoch = 1** | **0.74** | **Simpletransformers library needs only 2 columns as input** |
| **5.** | **ELECTRA** | 1. **Manual Preprocess was included** | **0.79** | **Manual cleaning increased F1 score to 0.79 from 0.74** |
| **6.** | **ELECTRA** | **Hyperparameter Tuning**  **epochs=2 , learning rate = 1e-5** | **0.79** | **True positive increased to 730/958 from 725/958** |
| **7.** | **BERT** | 1. **Did Feature Engineering** 2. **Included raw text & clean text** 3. **Calculated Predict\_proba** | **0.81** | **F1 score increased to 0.81 from 0.79 for 4 epoch** |

**SCIKIT-LEARN MODELS - COMMON PROCEDURE**

1. **Manual Preprocess**
2. **Countvectorizer**
3. **Encoding : Tf-idf Transformer**

|  |  |  |
| --- | --- | --- |
| **No** | **MODEL** | **F1 SCORE** |
| **1.** | **GradientBoostingClassifier** | **0.64** |
| **2.** | **MultinomialNB** | **0.72** |
| **3.** | **LogisticRegression** | **0.74** |
| **4.** | **SVM** | **0.74** |
| **5.** | **Nearest Centroid** | **0.73** |
| **6.** | **KNN** | **0.72** |

**Data Split: 50/25/25**

**GRADIENT BOOSTING CLASSIFIER**

**Encoding : Leave One Out Encoder**

|  |  |  |  |
| --- | --- | --- | --- |
| **NO** | **PROCEDURE** | **TRAIN F1 SCORE** | **VAL F1 SCORE** |
| **1.** | **Default parameters without manual preprocess** | **0.61** | **0.60** |
| **2.** | **Parameter Tuning** | **0.65** | **0.60** |
| **3.** | **Feature Engineering + Parameter Tuning** | **0.69** | **0.65** |

**Did some experiments on Parameter Tuning & chose the parameters with high val F1score**

**MULTINOMIAL NAIVE BAYES**

**Encoding : Leave One Out Encoder**

|  |  |  |  |
| --- | --- | --- | --- |
| **NO** | **PROCEDURE** | **TRAIN F1 SCORE** | **VAL F1 SCORE** |
| **1.** | **Default parameters without manual preprocess** | **0.48** | **0.50** |
| **2.** | **Feature Engineering** | **0.60** | **0.63** |

**NOTES:**

**DELETION**

1. **As MultiNomialNB has low F1 score, we will not use it in Level-1**
2. **As KNN has low F1 score, we will not use it in Level-1**
3. **As Nearest Centroid has no predict\_proba,we will not use it in Level-1**
4. **As XgBoost has low F1 score, we will not use it in Level-1**

**ADDITION**

**NEW ADDITION OF MODELS WITH HIGH F1 SCORE**

1. **ExtraTreesClassifier**
2. **CatBoost4**
3. **BernoulliNB = Naive\_Bayes**

**LEVEL-1 : ML MODELS**

**Trained on Set - 1 (50%)**

**Validated on Set - 2 (25%)**

|  |  |  |
| --- | --- | --- |
| **NO** | **MODEL** | **Set- 2: F1 SCORE** |
| **1.** | **Logistic Regression** | **0.74** |
| **2.** | **SVM** | **0.75** |
| **3.** | **GradientBoostingClassifier** | **0.71** |
| **4.** | **ExtraTreesClassifier** | **0.73** |
| **5.** | **CatBoost** | **0.72** |
| **6.** | **Xgboost** | **0.72** |

**IMPORTANT CORRECTION MADE IN THE UPCOMING MODELS**

**BEFORE: I had used f1\_score(average=’macro’) before.**

**AFTER : Now I changed f1\_score(average=’binary’)**

**After changing the F1 score ,**

**I added BernoulliNB model as it gave a good score &**

**Dropped the Xgboost model**

**BEFORE:**

**Part-1:**

* **Train: Set-1**
* **Val:Set-2**

**Part-2:**

* **Train: Set-2**
* **Val:Set-3**

**AFTER :**

**Part-1:**

* **Train: Set-1**
* **Val:Set-2**

**Part-2:**

* **Train: Set-1**
* **Val:Set-3**

**LEVEL-1 : INDIVIDUAL MODELS**

**TRAINING : PART-1**

**Files:**

1. [**Part-1 of Level-1 ML Models**](https://colab.research.google.com/drive/1mkjOh7tTe6cMfoHyN84BnVKu3veVQ0Lt?usp=sharing)
2. **BERT 5025.ipynb**

**Dataset :**

1. **Trained on Set - 1 (50%)**
2. **Validated on Set - 2 (25%)**

**Preprocess :**

1. **Leave One Out Encoder + Manual Cleaning + Feature Engineering for ML Scikit-Learn models**
2. **Bert preprocessing for Bert Model**

**Saved Set-1, Set-2 & Set-3 Data as csv files in order to use the same data to train Bert & other models**

**Improved Hyperparameter Tuning & increased the F1 scores.**

**Saved a new proba file : ml\_probabilities.csv in drive for future use in ensemble**

|  |  |  |
| --- | --- | --- |
| **NO** | **MODEL** | **Set- 2: F1 SCORE** |
| **1.** | **Logistic Regression** | **0.69** |
| **2.** | **SVM** | **0.69** |
| **3.** | **ExtraTreesClassifier** | **0.68** |
| **4.** | **GradientBoostingClassifier** | **0.67** |
| **5.** | **CatBoost** | **0.67** |
| **6.** | **BernoulliNB** | **0.67** |
| **7.** | **Bert** | **0.82** |

**TRAINING : PART - 2**

**Files:**

1. [**Part 2 - Level 1 ML Models Improved**](https://colab.research.google.com/drive/1BKMpl9Bb9j2c1qq4o9Xx07IAHXzTDcAd?usp=sharing)
2. **BERT 2525.ipynb**

**Dataset :**

1. **Trained on Set - 1 (50%)**
2. **Validated on Set - 3 (25%)**

**Preprocess :**

1. **Leave One Out Encoder + Manual Cleaning + Feature Engineering for ML Scikit-Learn models**
2. **Bert preprocessing for Bert Model**

|  |  |  |
| --- | --- | --- |
| **NO** | **MODEL** | **Set-3: F1 SCORE** |
| **1.** | **Logistic Regression** | **0.70** |
| **2.** | **SVM** | **0.70** |
| **3.** | **ExtraTreesClassifier** | **0.68** |
| **4.** | **GradientBoostingClassifier** | **0.68** |
| **5.** | **CatBoost** | **0.68** |
| **6.** | **BernoulliNB** | **0.69** |
| **7.** | **Bert** | **0.79** |

**LEVEL-2 : ENSEMBLE - STACKING**

**File: Ensemble 1.ipynb**

1. **Used predict\_probas of all level-1 models to train Level-2 model (logistic regression)**
2. **Adjusted the threshold to 0.2 for identifying more positives (1)**

**BEFORE THRESHOLD ADJUSTMENT:**

1. **Overall F1 Score: 0.73**
2. **F1 Score for 1: 0.68**
3. **True Positives: 529**

**AFTER THRESHOLD ADJUSTMENT: (0.2)**

1. **Overall F1 Score: 0.69**
2. **F1 Score for 1: 0.71**
3. **True Positives: 707**