

Group: always twenty-two

Members: Yufeng Song (yfsong), Lan Xu (lanxu)

Methodology: We started by recognizing this as a text-classification problem and auditing the splits: validation/test had a balanced human-vs-machine ratio, but the 310k training set didn't, so we down-sampled to (a) match that ratio and (b) be as large as our laptops could handle for a more representative sample. We set two baselines: a Dummy classifier (most-frequent vs. stratified; best F1 = 0.37 on the validation set) and an untuned TF-IDF (unigram) + Logistic Regression (F1 = 0.70). Then we tuned a full pipeline via grid search (TF-IDF analyzer/1–2 n-grams/min_df/max_df/sublinear-tf plus LR's solver/C/class_weight), checked ROC/threshold trade-offs, and picked an operating point at $p \geq 0.30$, giving F1 = 0.747 with best params: C=4.0, class_weight=balanced, solver=liblinear, analyzer=word, max_df=0.99, min_df=2, n-grams=(1,2), sublinear_tf=True. We also trained Linear SVC and Complement Naive Bayes; SVC edged out LR on validation (F1 = 0.764), while CNB lagged (F1 = 0.70) with weaker AUC and a very low optimal threshold (~0.05–0.10), so we skipped submitting CNB alone. Finally, we built a simple ensemble, averaging calibrated scores from tuned Linear SVC + tuned LR + CNB and re-optimizing the decision threshold, which achieved F1 = 0.792 on Kaggle and ranked #2 among all our submitted outcomes (**score = 0.77797**). Additionally, we explored lightweight nonlinear models to further improve performance while adhering to competition constraints that prohibited deep learning methods. We systematically evaluated three approaches: Random Forest (F1 = 0.7146, AUC = 0.7022), Gradient Boosting (F1 = 0.7246, AUC = 0.7424), and Multinomial Naive Bayes (F1 = 0.6782, AUC = 0.6648). Gradient Boosting emerged as the strongest performer among these lightweight nonlinear models, achieving the highest F1 score of 0.7246 with an optimal threshold at 0.15. We trained the final Gradient Boosting model on the full dataset and incorporated it into our ensemble alongside the previously optimized Linear SVC, Logistic Regression, and Complement Naive Bayes models. This enhanced four-model ensemble achieved an F1 score of 0.7922 on the validation set with a decision threshold of 0.25, which is our #1 submitted outcomes (**score = 0.77805**).