

Final Project Presentation

CS-UY 4563

2:00pm Section

May 2, 2022

Prof Linda M. Sellie

Chuangyang Jin, Alex Yan

Introduction

Dataset from Kaggle containing 10876 real tweets:

- 2 Classes (relating / not relating to a real disaster)
- 6851 Training Sample used
- 762 Testing Sample used

Performance measured by F1 score

- Imbalanced dataset (in real life dataset will be imbalanced)

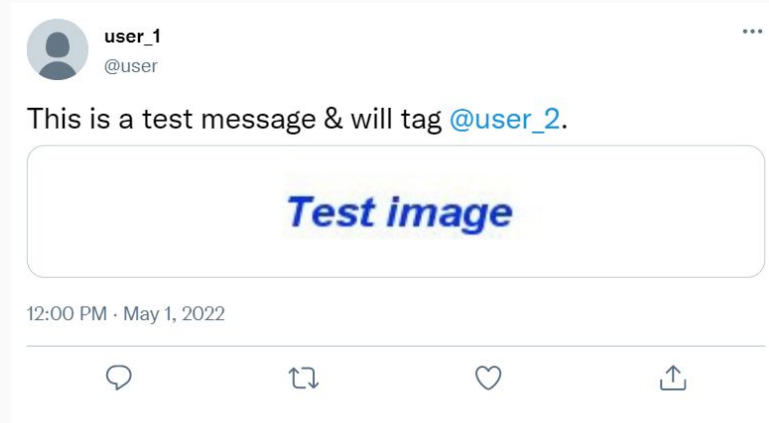
Preprocessing

Filter out meaningless message in the raw text.

&
@
https://

Common stop word filter applied at feature transformation.

Web View:



Raw Text:

This is a test message & will tag @user2 https://t.co/image_url.



Filtered Text:

This is a test message will tag URL.

Transforming Feature

Token Count:

Transforms the texts into a feature vector with the number of times each word appears in the collection of every tweet text.

Output: 14443 features

	the	red	dog	cat	eats	food
1. the red dog →	1	1	1	0	0	0
2. cat eats dog →	0	0	1	1	1	0
3. dog eats food →	0	0	1	0	1	1
4. red cat eats →	0	1	0	1	1	0

Transforming Feature

TF-IDF

(Term Frequency–Inverse Document Frequency)

Output: 14443 features

TFIDF

For a term i in document j :

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

$tf_{i,j}$ = number of occurrences of i in j
 df_i = number of documents containing i
 N = total number of documents

	blue	bright	can	see	shining	sky	sun	today
1	1	0	0	0	0	1	0	0
2	0	1	0	0	0	0	1	1
3	0	1	0	0	0	1	1	0
4	0	1	1	1	1	0	2	0



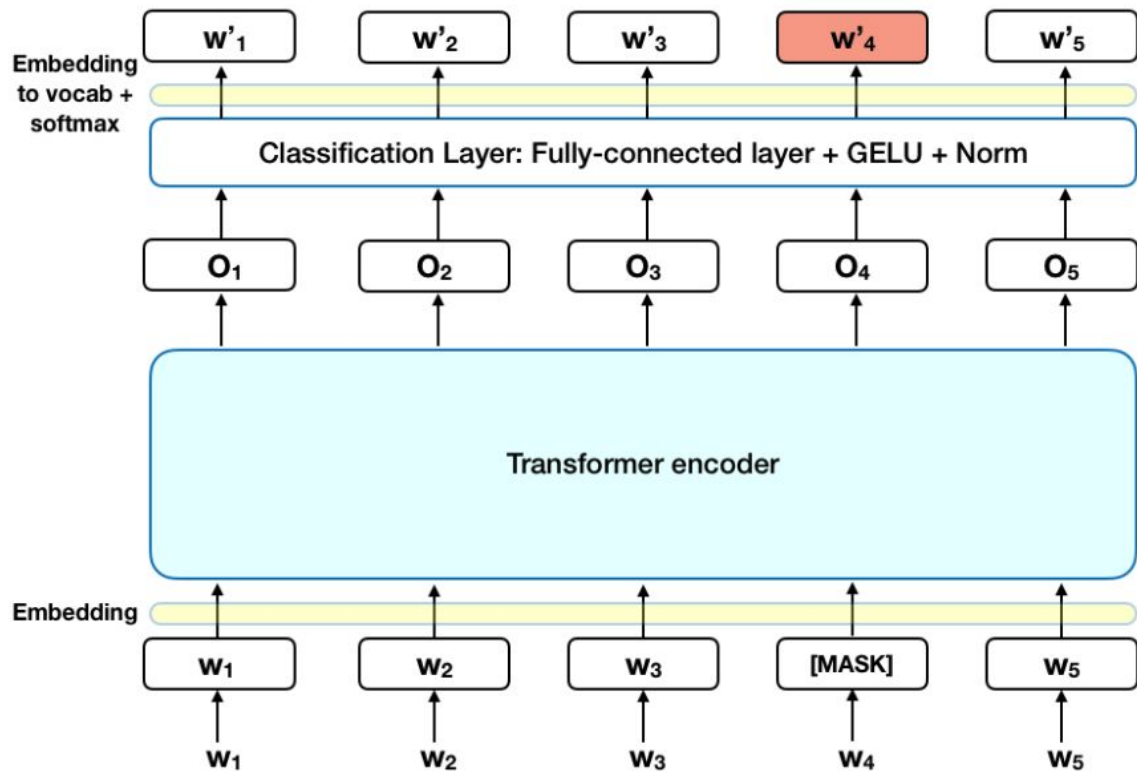
	blue	bright	can	see	shining	sky	sun	today
1	0.301	0	0	0	0	0.151	0	0
2	0	0.0417	0	0	0	0	0.0417	0.201
3	0	0.0417	0	0	0	0.100	0.0417	0
4	0	0.0209	0.100	0.100	0.100	0	0.0417	0

Transforming Feature

BERT

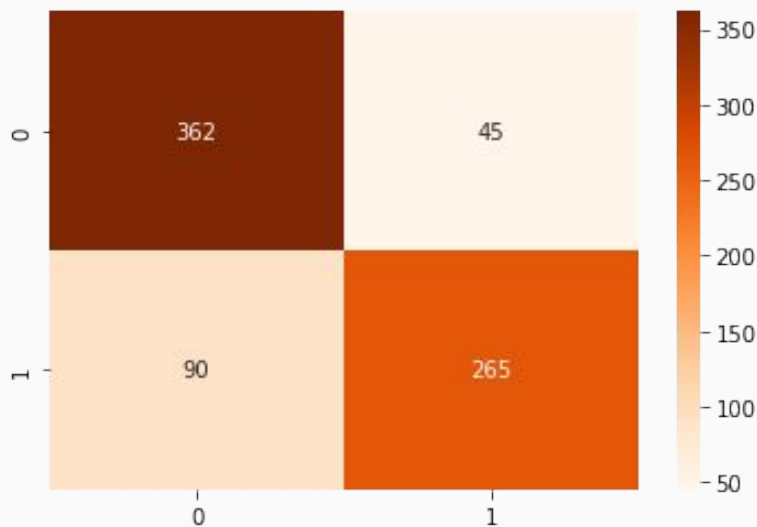
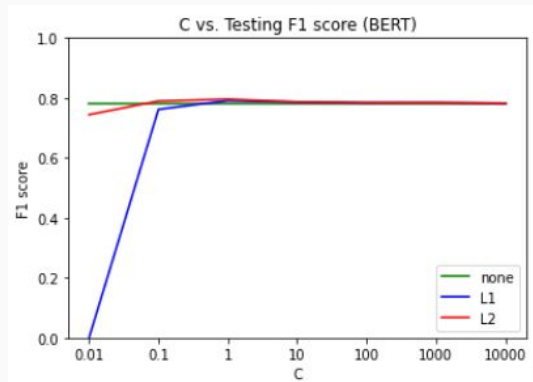
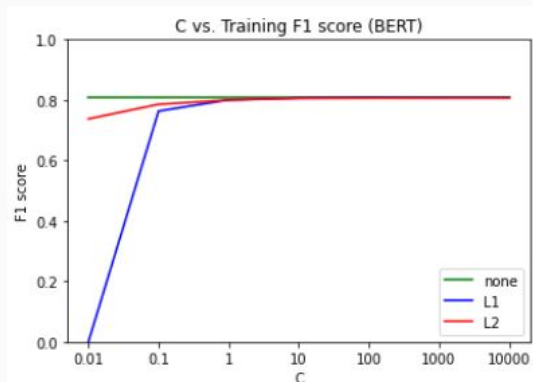
(Bidirectional Encoder Representations from Transformers)

Output: 384 features



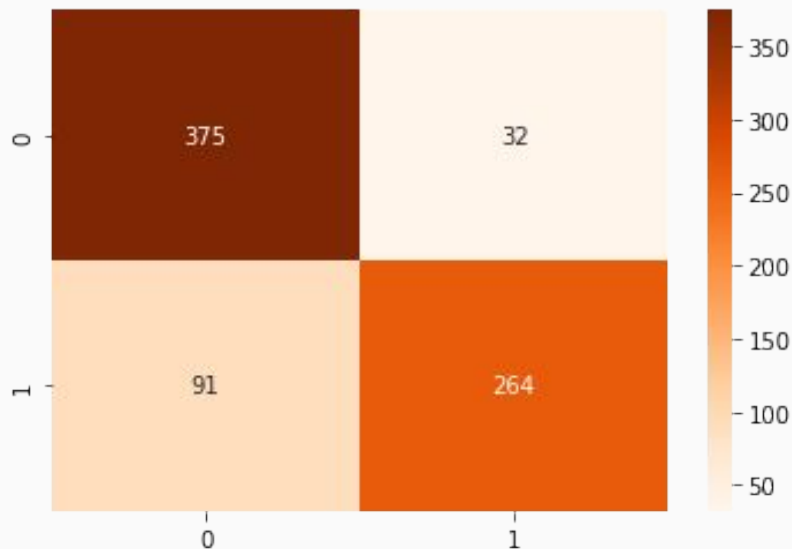
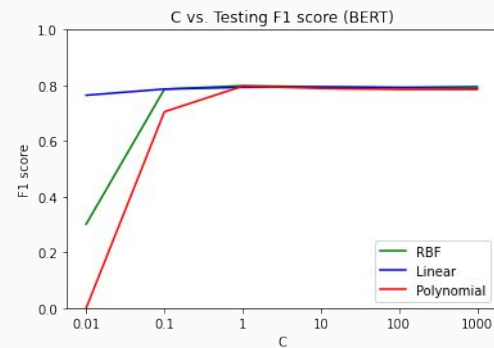
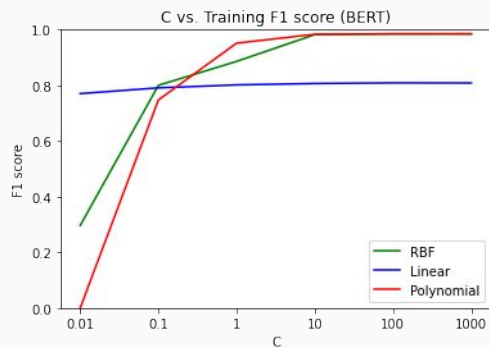
Logistic Regression

- 63 + 1 different hyperparameters
 - 3 Feature Transformation
 - L1, L2 Regularization
 - Different C values
- Trials reveal ideal C value is near 1
 - Best test F1 score of 0.796992.
 - Achieved at C = 0.8 with L2 Regularization and BERT
- Not overfitting too much.
 - Not a significant difference between training & testing F1 score



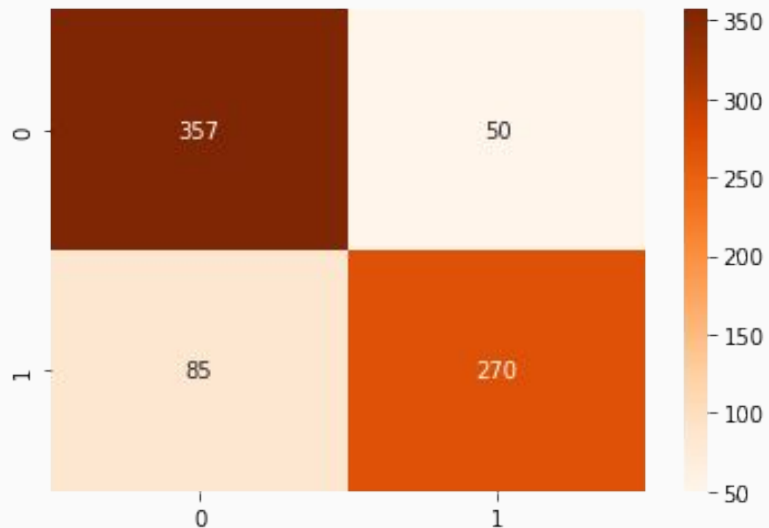
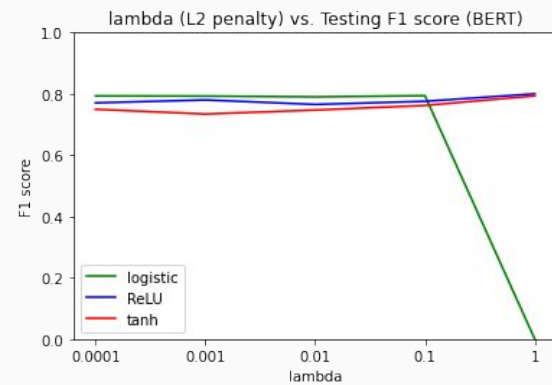
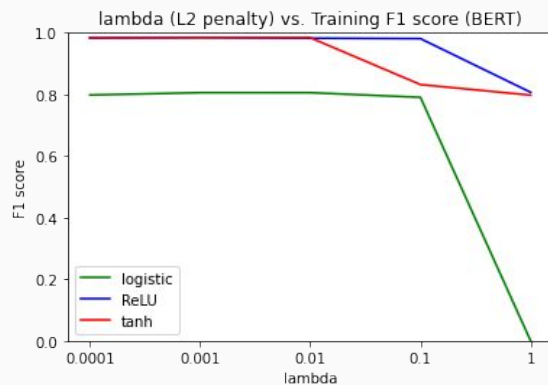
SVM

- 54 + 1 different hyperparameters
 - 3 Feature Transformation
 - Radial Basis Function / Linear / Polynomial (deg 3)
 - Different C values
- Trials reveal ideal C value is near 1
 - Best test F1 score of 0.811060.
 - Achieved at C = 1.7 with RBF and BERT
- Slight overfitting (more than Logistic Regression)
 - Around 10% difference in training & testing F1 score



Neural Network

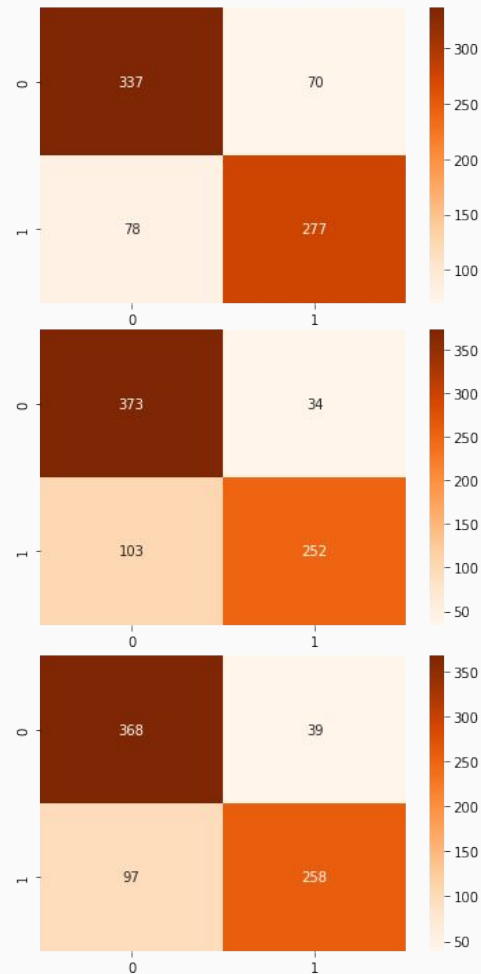
- 15 + 1 different hyperparameters
 - Only applied on BERT (otherwise too time consuming)
 - Different activation function: logistic, ReLU, tanh
 - Different C values
- Trials reveal ideal C value is near 1
 - Best test F1 score of 0.8.
 - Achieved at C = 1 with ReLU
- Less overfitting, roughly same as logistic regression.



Additional Methods

Naive Bayes
Random Forest
Gradient Boosting

- Try Gaussian Naive Bayes, Multinomial Naive Bayes, Complement Naive Bayes, Bernoulli Naive Bayes and Categorical Naive Bayes
- Vary the number of estimators (or trees) in the Random Forest, Gradient Boosting



Ensemble Model

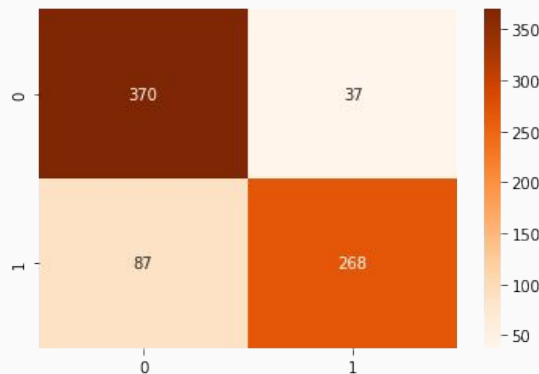
single model
may not work well



take advantage of
different models

First attempt: majority vote of three best models
→ fail to improve accuracy

Second attempt: rely mainly on the best model,
refer to the second, third, and fourth best
models as well
→ improve accuracy








Conclusion

Best result is produced by combining SVM, logistic regression, neural networks, and gradient boosting, achieving F1 score of 0.812121 (kaggle test set F1 score 0.82286 ranking 157/896)

Every model exhibited roughly the same F1 score using the best hyperparameter, probably because the dataset provided was not adequate for the training to reach a higher F1 score.

Model	LogReg	SVM	Neural Networks	Naive Bayes	Random Forest	Gradient Boosting	Ensemble Model
F1 Score	0.796992	0.811060	0.800001	0.789174	0.786834	0.791411	0.812121

153	ps40		0.82378	8	1mo
154	Leong Ivan		0.82378	4	1mo
155	Yuri	 	0.82316	2	2mo
156	Jaaack Wang		0.82286	10	1mo
157	Alex & Chuanyang		0.82286	2	1h