Applying Machine Learning to Identify Human Written or AI Written

Trang Khong

Reason of This Project

- Chat GPT (Chat Generative Pre-trained Transformer) was released in November 2022.
- Chat GPT is being used for translation, conversational AI, coding, and education.
- Want explore the difference between ChatGPT’s response and humans' response.
- Need a model that can recognize whether the text is AI response or human response.

Goal of This Project

- Collected human answers and questions from Reddit by using PRAW (The Python Reddit API Wrapper)
- The AI answers were collected from Openai API by asking the same questions as were collected from Reddit to ChatGPT
- Combine those data together.
- Build a model that can predict whether the text is human answer or ai answer.

List of Six Subreddit Topic

- NoStupidQuestion
- Questions
- Askreddit
- MorbidQuestions
- TooAfraidToAsk
- AskScienceFiction

Features Use for Modeling

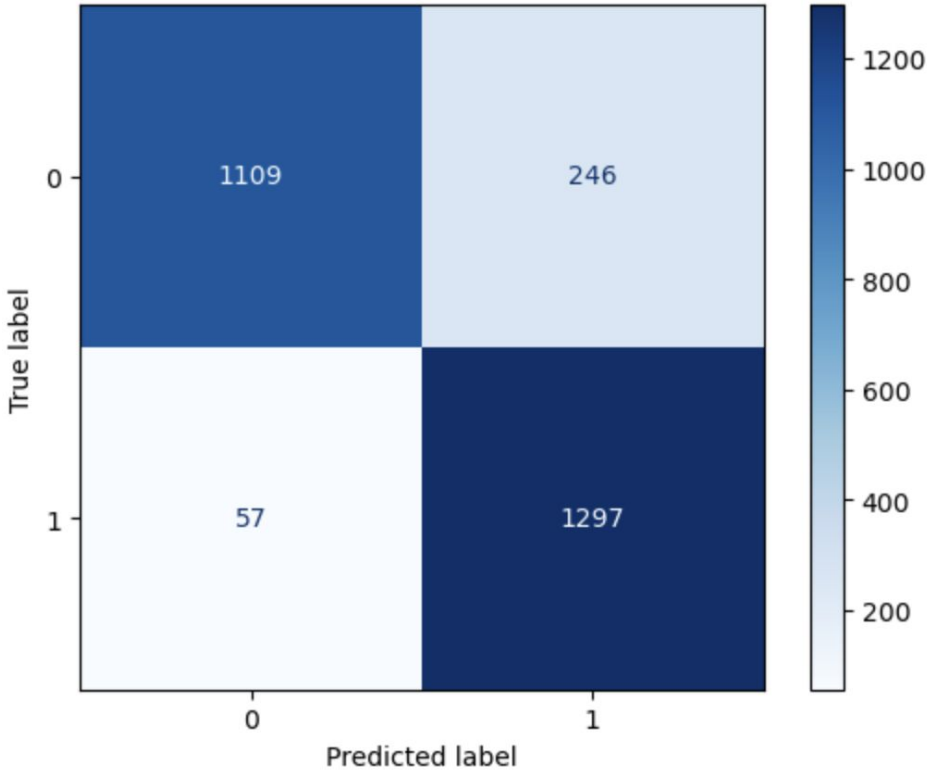
Feature	Type	Dataset	Description
answer	object	project3_answer	List of human answer and AI answer
result	int64	project3_answer	List of 1 and 0, 1 stand for human answer and 0 stand for AI answer

```
pipe_params = {
    'cvec__max_features' : [20000, 40000],
    'cvec__min_df' : [3, 5],
    'cvec__max_df' : [0.9, 0.95],
    'cvec__ngram_range' : [(1, 1),(1, 2)],
    'cvec__stop_words' : [None ,'english', nltk_stop]
}
```

Model 1: CountVectorizer & BernoulliNB

```
{'cvec__max_df': 0.9,
 'cvec__max_features': 40000,
 'cvec__min_df': 3,
 'cvec__ngram_range': (1, 2),
 'cvec__stop_words': None}
```

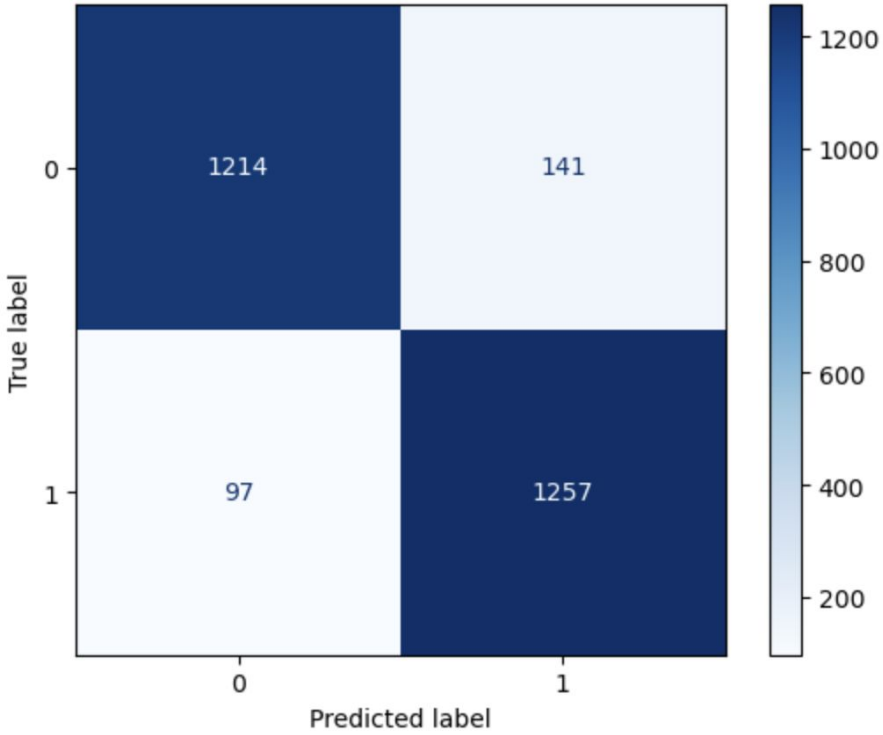
```
accuracy: 0.8881506090808416
recall: 0.9579025110782866
precision: 0.8405703175631886
f1 score: 0.8954090438384534
```



Model 2: CountVectorizer & LogisticRegression

```
{'cvec__max_df': 0.9,
 'cvec__max_features': 20000,
 'cvec__min_df': 3,
 'cvec__ngram_range': (1, 2),
 'cvec__stop_words': None}
```

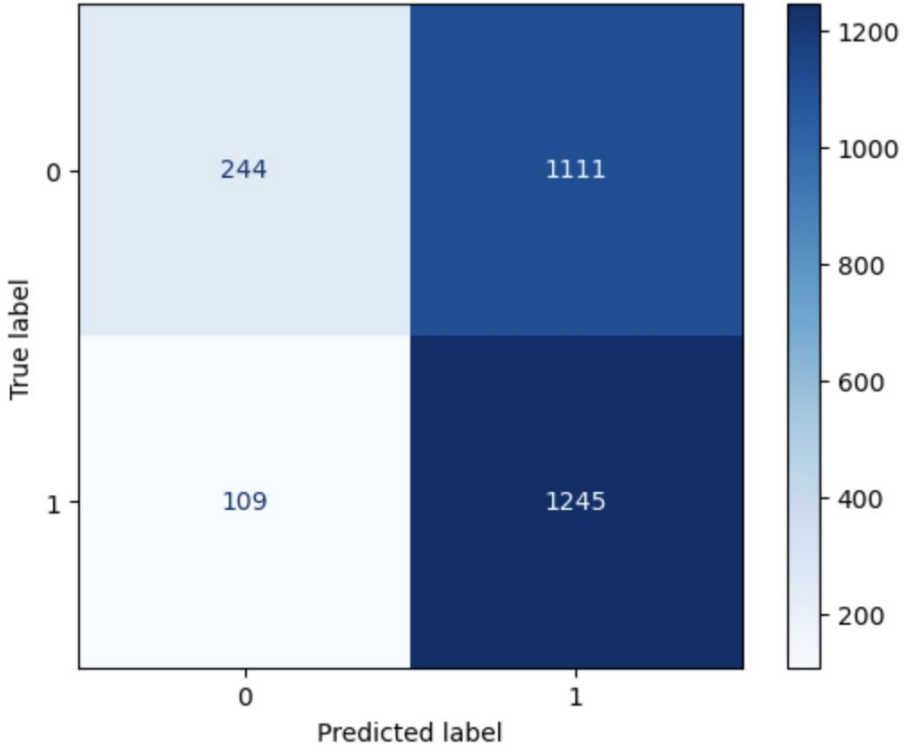
```
accuracy: 0.9121447028423773
recall: 0.9283604135893648
precision: 0.8991416309012875
f1 score: 0.9135174418604651
```



Model 3: CountVectorizer & KNeighborsClassifier

```
{'cvec__max_df': 0.9,
 'cvec__max_features': 20000,
 'cvec__min_df': 5,
 'cvec__ngram_range': (1, 1),
 'cvec__stop_words': None}
```

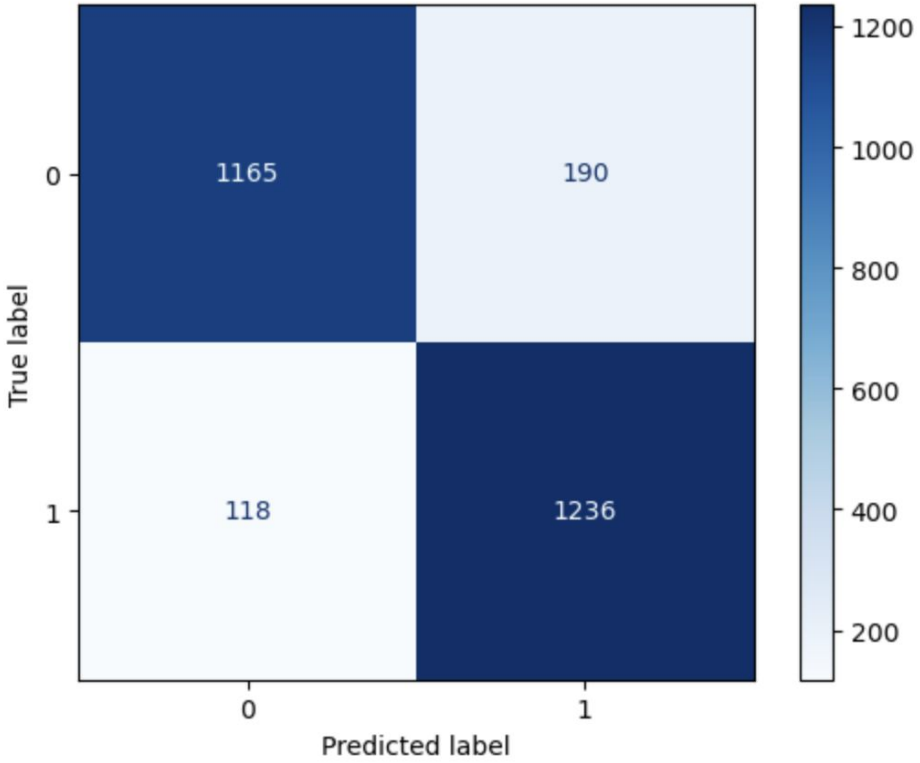
```
accuracy: 0.5496493170911776
recall: 0.9194977843426884
precision: 0.5284380305602716
f1 score: 0.671159029649595 7
```



Model 4 : CountVectorizer & MultinomialNB

```
{'cvec__max_df': 0.9,
 'cvec__max_features': 40000,
 'cvec__min_df': 3,
 'cvec__ngram_range': (1, 2),
 'cvec__stop_words': None}
```

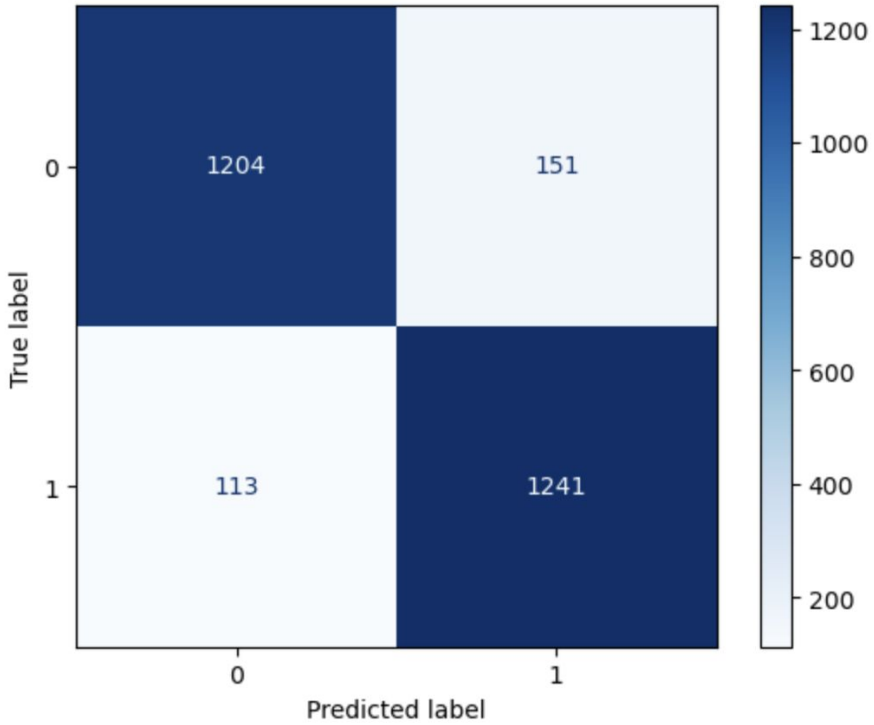
```
accuracy: 0.8863049095607235
recall: 0.912850812407681
precision: 0.8667601683029453
f1 score: 0.889208633093525
```



Model 5: TfidfVectorizer & LogisticRegression

```
{'logr__penalty': 'l2',
 'tvec__max_features': 20000,
 'tvec__ngram_range': (1, 2),
 'tvec__stop_words': None}
```

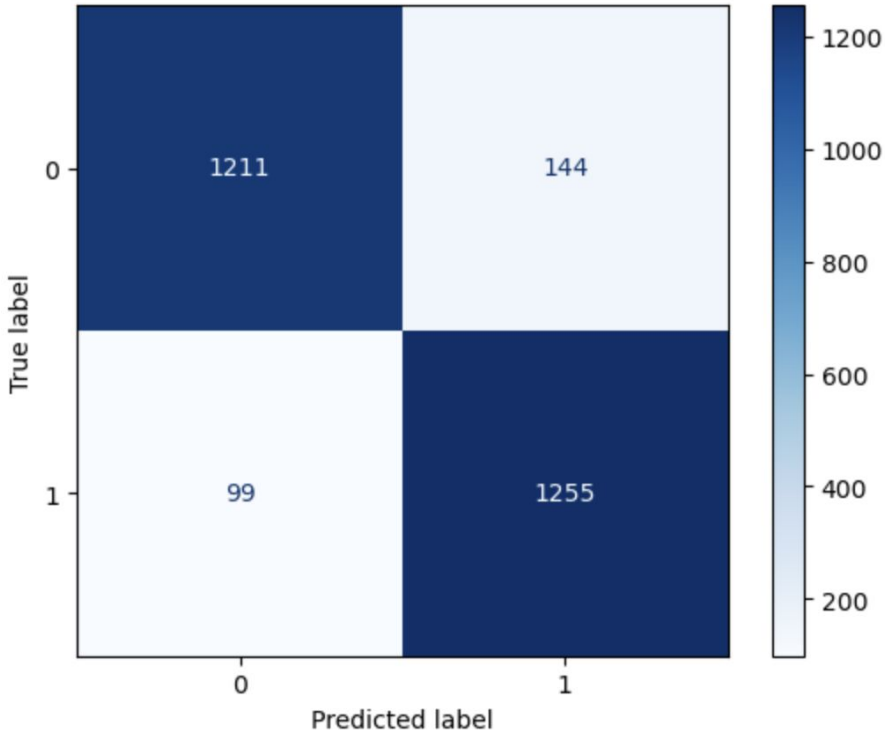
```
accuracy: 0.902547065337763
recall: 0.9165435745937962
precision: 0.8915229885057471
f1 score: 0.9038601602330663
```



Model 6: CountVectorizer and LogisticRegression (l1)

```
{'cvec__max_df': 0.85,
 'cvec__max_features': 40000,
 'cvec__min_df': 2,
 'cvec__ngram_range': (1, 2),
 'cvec__stop_words': None,
 'logr__penalty': 'l1'}
```

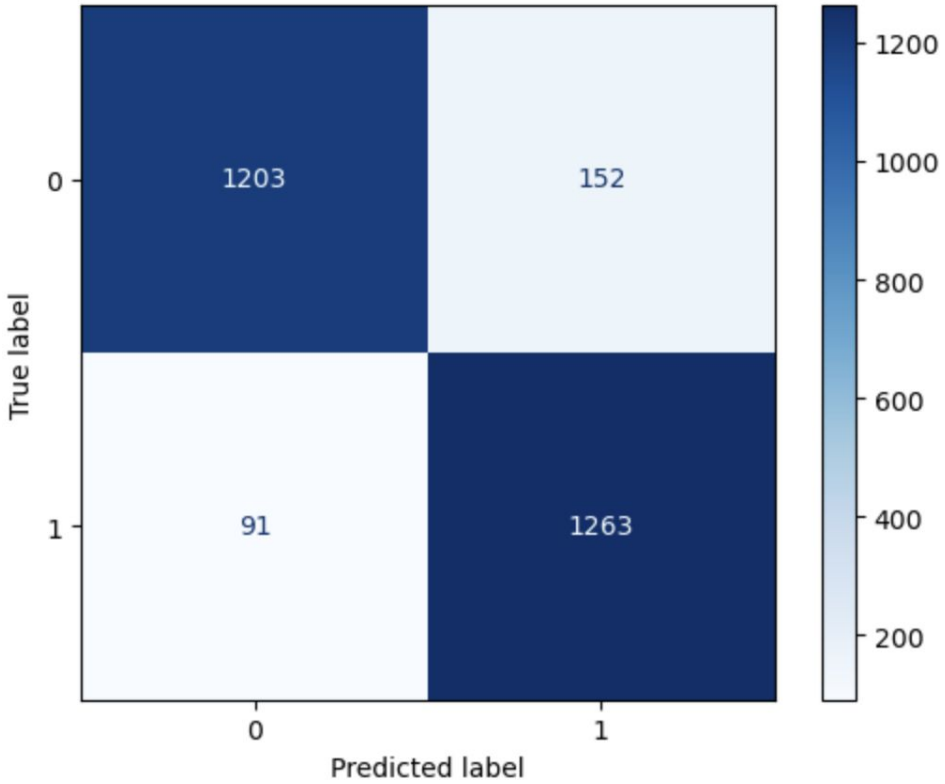
```
accuracy: 0.9102990033222591
recall: 0.9268833087149188
precision: 0.8970693352394568
f1 score: 0.9117326552851436
```



Model 7: CountVectorizer and AdaBoostClassifier

```
{'abc__n_estimators': 200,
 'cvec__max_df': 0.85,
 'cvec__max_features': 20000,
 'cvec__min_df': 3,
 'cvec__ngram_range': (1, 2),
 'cvec__stop_words': None}
```

```
accuracy: 0.9102990033222591
recall: 0.9327917282127031
precision: 0.8925795053003533
f1 score: 0.9122426868905742
```



The Result Summary of 7 Model

	accuracy	recall	precision	f1_score
model 1 (cvec & bnb)	0.888151	0.957903	0.840570	0.895409
model 2 (cvec & logr(l2))	0.912145	0.928360	0.899142	0.913517
model 3 (cvec & knn)	0.549649	0.919498	0.528438	0.671159
model 4 (cvec & mnb)	0.886305	0.912851	0.866760	0.889209
model 5 (tfidf & logr(l2))	0.902547	0.916544	0.891523	0.903860
model 6 (cvec & logr(l1))	0.910299	0.926883	0.897069	0.911733
model 7 (cvec & abc)	0.910299	0.932792	0.892580	0.912243

Conclusion and Recommendations

- LogisticRegression gave a higher accuracy score than KNeighborsClassifier, BernoulliNB and MultinomialNB.
- Model 2 has the highest accuracy score and F1 score but model 2 has the difference between test score and train score highest. Model 2 is overfit.
- Model 5, model 6 and model 7 were built to improve the overfit and accuracy score of model 2. The accuracy score of those models are lower than model 2 but the difference between train score and test score did improve.
- Model 2, model 6 and model 7 can be used to identify whether the text is human written or ai written.
- In the future, apply different classification and different params to find a higher accuracy score.

Reference

https://research.aimultiple.com/chatgpt-use-cases/#textual-applications