Movie Genre Prediction - Milestone 2

Group 2: Charles, Giffin, Jon G, Canxiu

Recap: Multi-Label Movie Genre Classification

Solve Multi-label classification problem by using problem transformation methods to transform the multi-label problem into a set of binary classification problems, which can then be handled using single-class classifiers.

genres
Drama|Adventure|Science Fiction
Family|Animation|Adventure|Comedy
Comedy|Animation|Family
Action|Adventure|Crime
Science Fiction|Fantasy|Action|Adventure

Horror: 0.02%
Romance: 0.02%
Adventure: 99.96%
Documentary: 0.0%

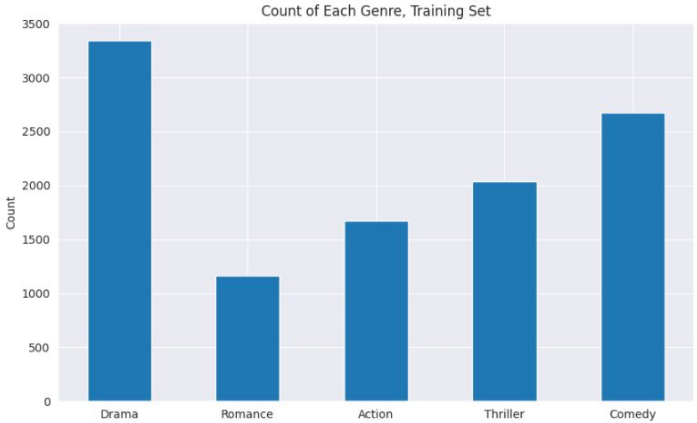
Feature engineering for consideration:

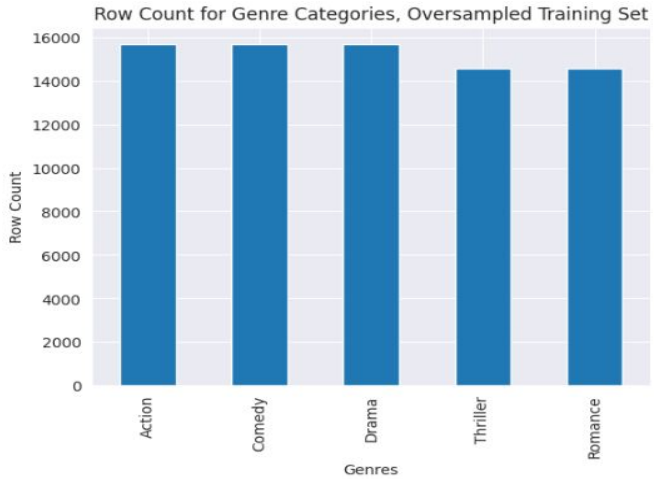
1. Counting methods (Bag of Words, Bag-of-ngram and TF-IDF)
2. Word Embedding model (Word2Vec, GloVe and fasttext)
3. Language model (BERT)
4. Topic model (LDA/LSI)
5. Document Clustering with Similarity Features

FE Methods Selection

Approach:

1. Handle imbalanced data:
    1. Convert multi-label target to single-column powerset
    2. Apply SMOTE oversampling on training data
    3. Convert single column powerset back to a multi-label target
2. Use One-vs-the-rest (OvR) multiclass strategy and logistic regression model to handle the multi-label problem
3. Choose accuracy, f1, precision score, recall score and roc auc score as performance measurements
4. Compare word2vec (shape: 500), doc2vec (shape: 100), GloVe (shape: 96) and TF-IDF (shape: 185,395) performances created by logistic regression classification





GloVe: (Global Vectors)

- Glove Vectorizes words through the use of both global statistics and local statistics in a corpus of documents. Global statistics looks at dependencies across the entire corpus of documents.

- In comparison, a vectorizer such as Word2Vec uses local statistics where dependencies are captured within small units of text such as individual phrases

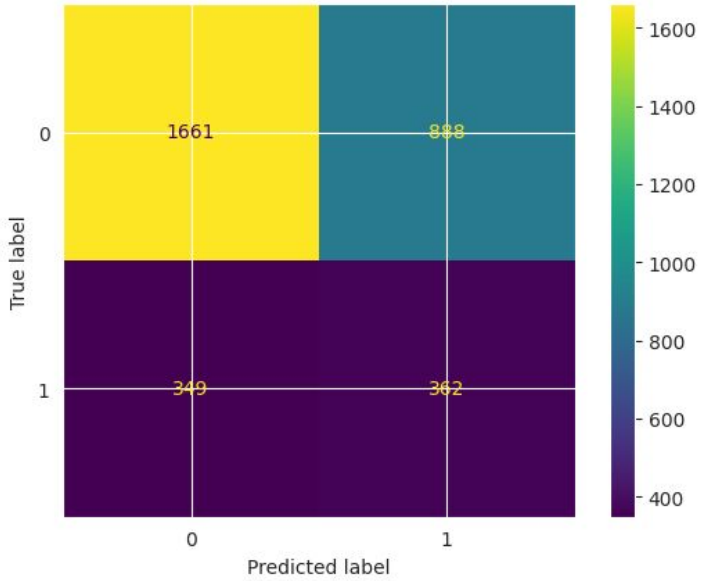
- Glove uses a co-occurrence matrix to see how many times a given word occurs with another given word across a corpus

Supporting article link:

https://towardsdatascience.com/light-on-math-ml-intuitive-guide-to-understanding-glove-embeddings-b13b4f19c010

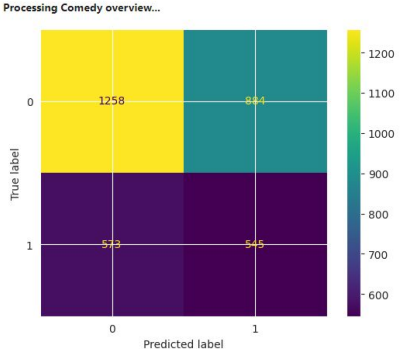
FE with GloVe modelling

Processing Action overview...



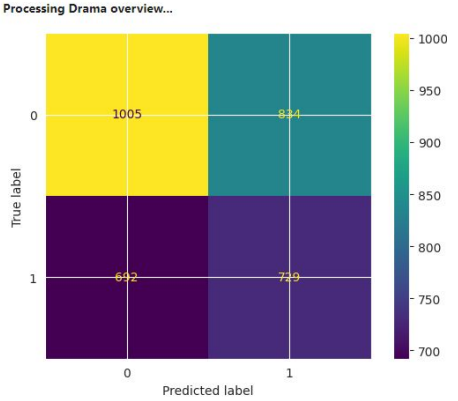

```
Test accuracy is 0.6205521472392638
Test f1 is 0.3691993880673127
Test precision score is 0.2896
Test recall score is 0.509142053445851
Test roc auc score is 0.5803850714463464
```

Processing Comedy overview...



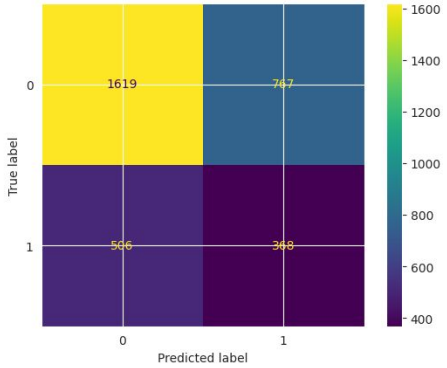

```
Test accuracy is 0.5530674846625767
Test f1 is 0.4279544562230075
Test precision score is 0.3813855843247026
Test recall score is 0.4874776386404293
Test roc auc score is 0.5373896129710083
```

Processing Drama overview...



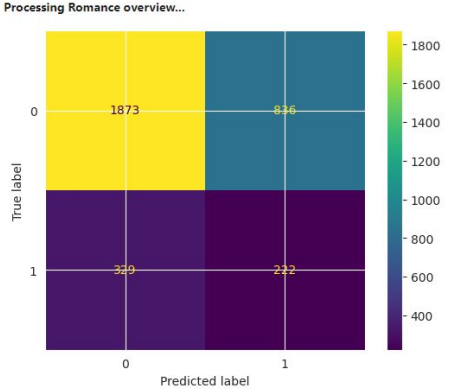

```
Test accuracy is 0.5319018404907976
Test f1 is 0.4886058981233244
Test precision score is 0.46641074856046066
Test recall score is 0.5130190007037297
```

Processing Thriller overview...




```
Test accuracy is 0.6095092024539878
Test f1 is 0.366351418616227
Test precision score is 0.32422907488986785
Test recall score is 0.42105263157894735
Test roc auc score is 0.5497970618079145
```

Processing Romance overview...




```
Test accuracy is 0.6426380368098159
Test f1 is 0.2759477936606588
Test precision score is 0.20982986767485823
Test recall score is 0.4029038112522686
Test roc auc score is 0.5471514257442591
```

TF-IDF Vectorizer
(Term Frequency-Inverse Document Frequency)

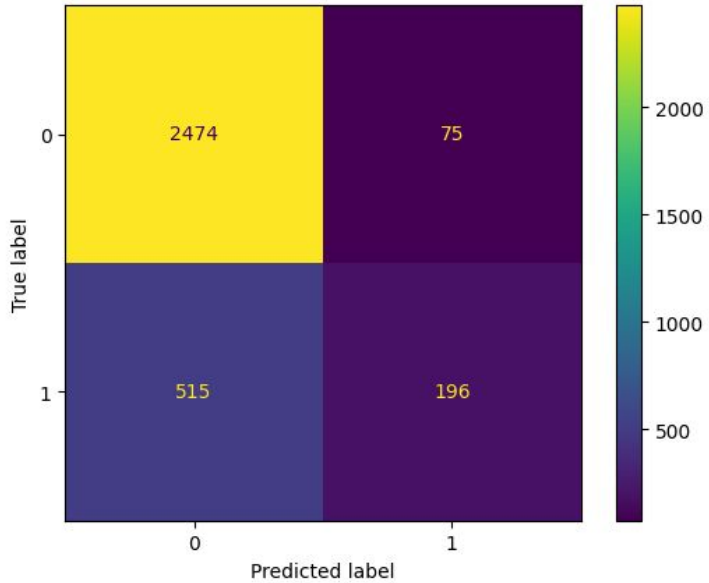
- Term Frequency: How often a term occurs in a document of text. The idea is that the more a term occurs, the more important that term is to the document
- Inverse Document Frequency: How rarely a term occurs across a a corpus of documents
- *IDF will assign higher importance to words that occur less often across the corpus of documents
- The TFIDF Formula is TF/IDF or TF * IDF

Supporting article link:

https://www.capitalone.com/tech/machine-learning/understanding-tf-idf/

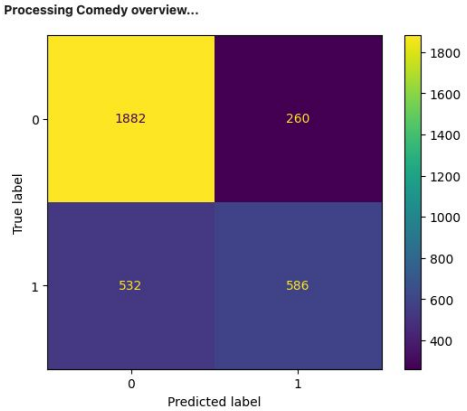
FE with TF-IDF modelling

Processing Action overview...



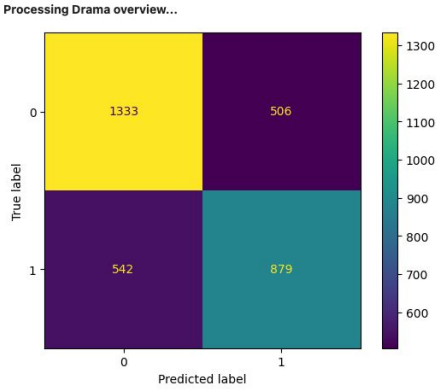

```
Test accuracy is 0.8190184049079755
Test f1 is 0.39918533604887985
Test precision score is 0.7232472324723247
Test recall score is 0.27566807313642755
Test roc auc score is 0.6231223849401243
```

Processing Comedy overview...



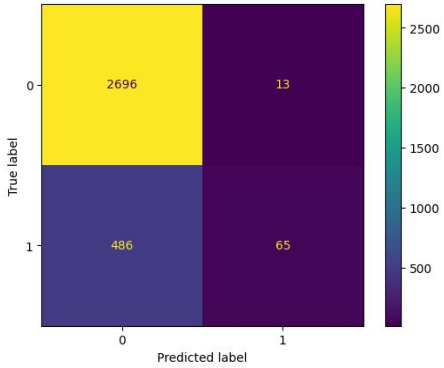

```
Test accuracy is 0.7570552147239263
Test f1 is 0.5967413441955193
Test precision score is 0.6926713947990544
Test recall score is 0.5241502683363148
Test roc auc score is 0.7013841911242732
```

Processing Drama overview...



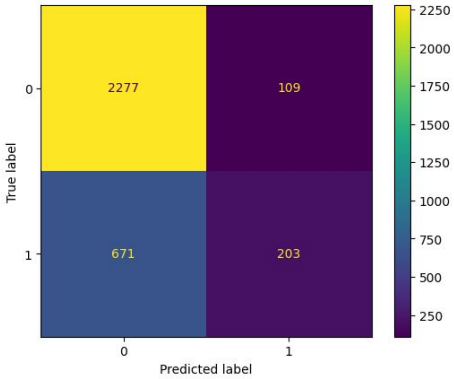

```
Test accuracy is 0.6785276073619632
Test f1 is 0.6265146115466856
Test precision score is 0.6346570397111914
Test recall score is 0.6185784658691063
Test roc auc score is 0.671714464038414
```

Processing Romance overview...



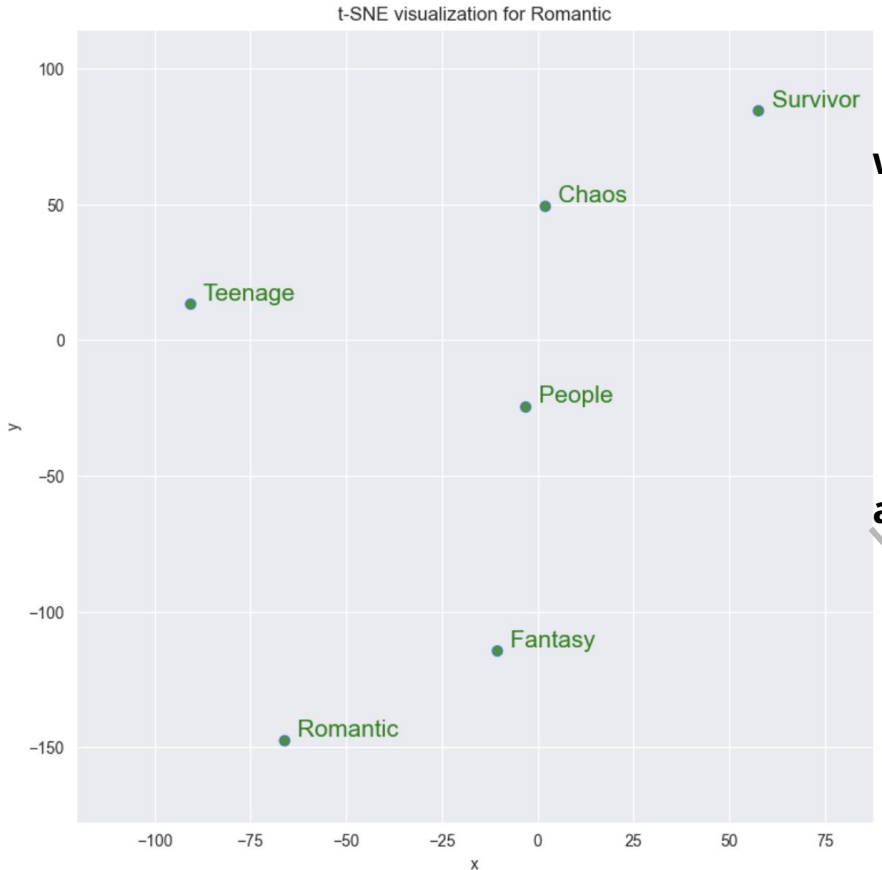

```
Test accuracy is 0.8469325153374233
Test f1 is 0.20667726550079493
Test precision score is 0.8333333333333334
Test recall score is 0.11796733212341198
Test roc auc score is 0.5565842568555524
```

Processing Thriller overview...




```
Test accuracy is 0.7607361963190185
Test f1 is 0.342327150084317
Test precision score is 0.6506410256410257
Test recall score is 0.2322654462242563
Test roc auc score is 0.5932911472529496
```

Word2Vec and Doc2Vec



Movie Overview:
'romantic fantasy movie people last survivor ancient line goddess worshiper sell art shop teenage daughter rhea fall love poetically inclined boy begin develop magical power throw everything chaos'

word2vec

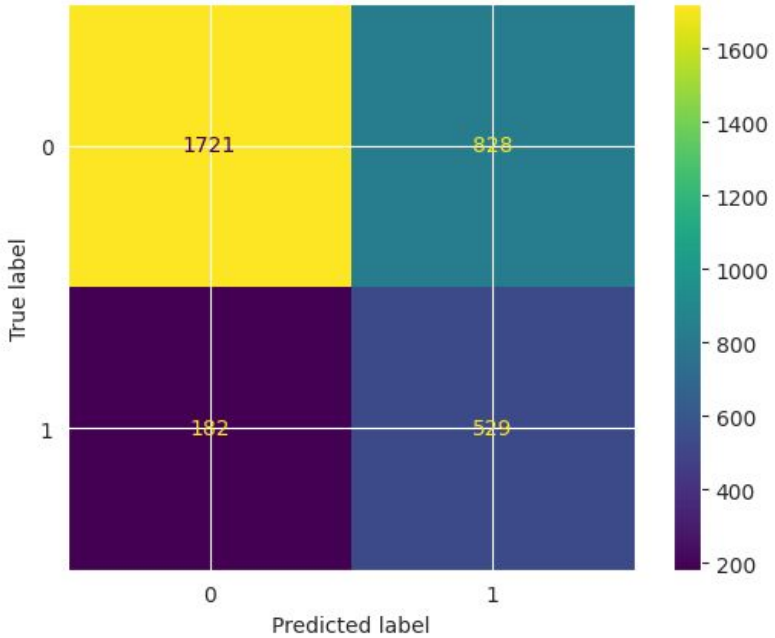
doc2vec

average

```
array([ 0.15376482,  0.18092218,  0.11861369,  0.06002235, -0.14580119,
       -0.2147511 ,  0.06483621,  0.36061509,  0.14044209, -0.008824  ,
       -0.05754338,  0.10503522,  0.11005998,  0.02758399,  0.08955757,
       -0.2138584 , -0.17213251, -0.12228273, -0.01229672,  0.04875104,
        0.0501123 , -0.06083884,  0.14568402, -0.09636139,  0.15324458,
        0.10478802,  0.06202611,  0.01406976, -0.28569955, -0.00701886,
        0.12390759,  0.04840128, -0.0555181 , -0.02675169,  0.0895873 ,
        0.10506309, -0.04137709, -0.04168905, -0.08121119, -0.19932454,
       -0.08275044,  0.05585648, -0.160104  ,  0.0392762 , -0.14579296,
       -0.17716806, -0.12774382,  0.05987273, -0.03823655, -0.003711  ,
       -0.02019309, -0.04265463,  0.03332094, -0.26431747,  0.09969765,
       -0.11019437,  0.06895421, -0.00653785, -0.0171375 ,  0.0910629 ,
        0.08043635, -0.01361404, -0.03029563,  0.04031828, -0.10234138,
        0.1048274 , -0.05981804,  0.0773802 ,  0.08275738,  0.03052277,
       -0.13086363,  0.03658277,  0.00764878, -0.07038752,  0.12078158,
        0.2045487 , -0.0618625 ,  0.00429446,  0.14007089,  0.12574692,
       -0.05450244,  0.05354711, -0.11334193,  0.13482648, -0.32191707,
        0.15772745, -0.06555435,  0.11681744,  0.18442382,  0.10433377,
        0.02870045,  0.04180757, -0.11611081,  0.07104141,  0.12312881,
```

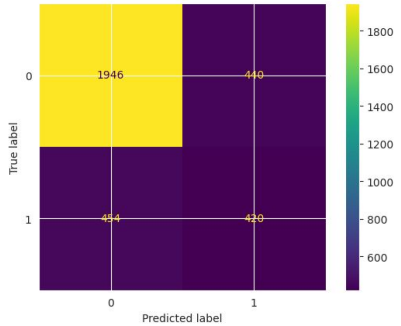
FE with word2vec modelling

Processing Action overview...



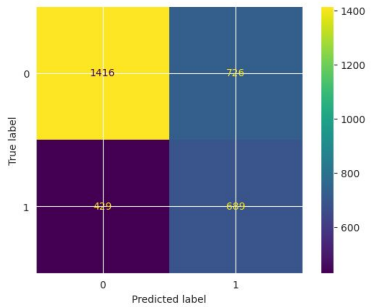

```
Test accuracy is 0.6901840490797546
Test f1 is 0.511605415860735
Test precision score is 0.3898305084745763
Test recall score is 0.7440225035161744
Test roc auc score is 0.7095946177839797
```

Processing Thriller overview...



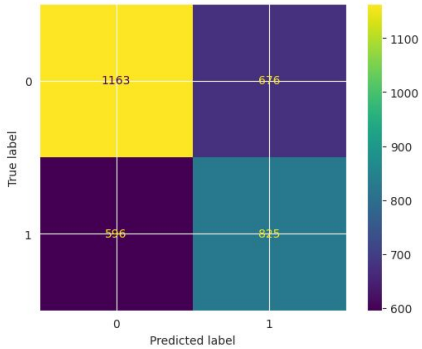

```
Test accuracy is 0.7257668711656442
Test f1 is 0.4844290657439446
Test precision score is 0.4883720930232558
Test recall score is 0.4805491990846682
Test roc auc score is 0.6480700731383106
```

Processing Comedy overview...



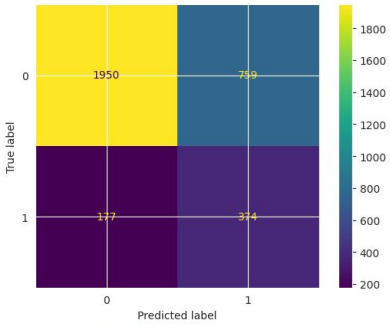

```
Test accuracy is 0.6457055214723927
Test f1 is 0.5440189498618239
Test precision score is 0.4869257950530035
Test recall score is 0.6162790697674418
Test roc auc score is 0.638671747768875
```

Processing Drama overview...




```
Test accuracy is 0.6098159509202454
Test f1 is 0.5646817248459959
Test precision score is 0.5496335776149234
Test recall score is 0.5805770584095707
Test roc auc score is 0.6064929881498643
```

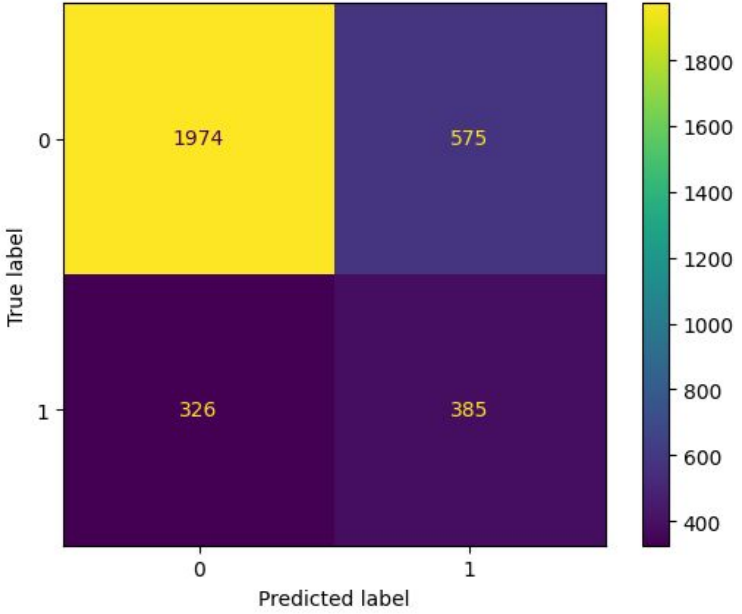
Processing Romance overview...




```
Test accuracy is 0.7128834355828221
Test f1 is 0.4441805225653206
Test precision score is 0.3300970873786408
Test recall score is 0.6787658802177858
Test roc auc score is 0.6992943465319272
```

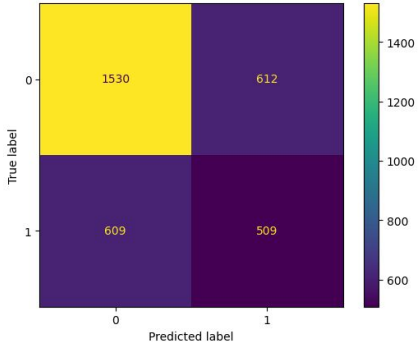
FE with doc2vec modelling

Processing Action overview...



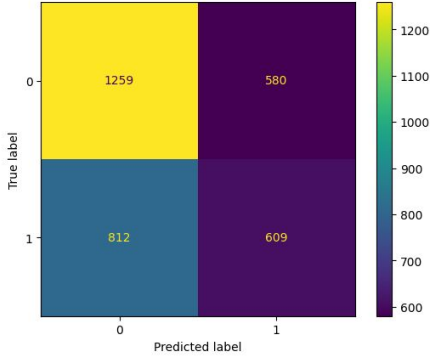

```
Test accuracy is 0.7236196319018405
Test f1 is 0.4608019150209456
Test precision score is 0.4010416666666667
Test recall score is 0.5414908579465542
Test roc auc score is 0.6579560998245914
```

Processing Comedy overview...



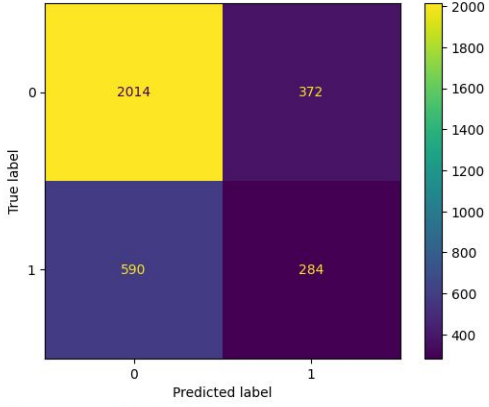

```
Test accuracy is 0.6254601226993866
Test f1 is 0.4546672621706119
Test precision score is 0.45405887600356826
Test recall score is 0.4552772808586762
Test roc auc score is 0.5847814975721953
```

Processing Drama overview...



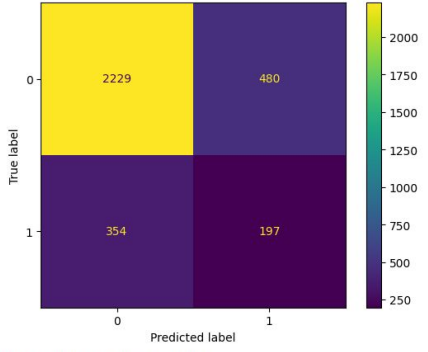

```
Test accuracy is 0.5730061349693252
Test f1 is 0.4666666666666667
Test precision score is 0.5121951219512195
Test recall score is 0.42857142857142855
Test roc auc score is 0.5565913151557523
```

Processing Thriller overview...




```
Test accuracy is 0.7049079754601227
Test f1 is 0.3712418300653595
Test precision score is 0.4329268292682927
Test recall score is 0.32494279176201374
Test roc auc score is 0.5845166599212416
```

Processing Romance overview...




```
Test accuracy is 0.7441717791411043
Test f1 is 0.3208469055374593
Test precision score is 0.2909896026587886
Test recall score is 0.35753176043557167
Test roc auc score is 0.5901723032521158

CPU times: user 1.67 s, sys: 267 ms, total: 1.94 s
Wall time: 12.2 s
```

Next Steps & Questions

1. Considering and researching on recommendation system
2. Use five labels or transform them to Label Powerset as one signal target variable
3. Comparing Classification Model performance
4. Pick Performance Metrics which better measure Multi-label problem