

---

# Text Analysis: Identifying Comedy Movies

**Carlos Jaime**

The University of Texas at Arlington  
Carlos.Jaime@mavs.uta.edu

**Sai Sowmith Reddy Chintha**

The University of Texas at Arlington  
Saisowmithreddy.chintha@mavs.uta.edu

**Kashish Jain**

The University of Texas at Arlington  
Kashish.Jain@mavs.uta.edu

**Tatsat Pandey**

The University of Texas at Arlington  
Tatsat.Pandey@mavs.uta.edu

## **OUTLINE**

1. Abstract
2. Introduction
3. Data
4. Models
  - a. Logistic Regression
  - b. K-Nearest Neighbours
  - c. Stochastic Gradient Descent
  - d. Decision Trees
  - e. Random Forest
5. Conclusion
  - a. Logistic Regression
6. Appendix
  - a. Frequency distribution of the top 30 tokens

### **1.Abstract**

In this group project, we analysed text descriptions of movies and developed a predictive model that can predict whether or not a movie is a comedy. The descriptions/story were used to create textual features of the movie descriptions/story and multiple models were created in order to find the best performing model.

### **2.Introduction**

There is nothing more frustrating than not finding the right movie to watch. Understanding movie descriptions/story could give an insight of the type of genre the movie is. It can be a comedy, romance, horror or another type of film. Sorting through movies by their genre would make it easier for movie viewers to find the movie they want to watch. The goal of this group project is to predict whether or not a movie is a comedy based on the movie description/story.

### **3.Data**

The data used for this group project was provided by Dr. Mahyar S Vanghefi. This data consisted of 3 csv files found below: movie\_story\_student\_file.csv, movie\_story\_evaluation\_file.csv and movies.csv.

```

**movie_story_evaluation_file.csv**
movie_id story
0 122349 Growing up in the Mission district of San Fran...
1 122351 A soldier returns home from the Iraq war only ...
2 122361 Marco the Monkey works as a beach officer. But...
3 187901 When an honest cop, Vijay Kumar\'s family is r...
4 187903 Kathiresan aka Kaththi, a criminal, escapes fr...
...
3493 131062 In the middle of nowhere, 20 years after an ap...
3494 131064 After living for years as a struggling artist ...
3495 131066 Ronal is a young barbarian with low self-estee...
3496 131068 Ziege, H\xc3\xa4schen and Max have now moved t...
3497 131070 During their childhood, Hanna and Clarissa wer...

```

[3498 rows x 2 columns]

```

**movie_story_student_file.csvv**
movie_id story
0 131072 A girl who always tends to fall in love with t...
1 196609 Bigfoot has come to the town of Ellwood City, ...
2 131074 At an altitude of 18,000 feet, Alaska\'s Mount...
3 196611 In her first special since 2003, Ellen revisit...
4 196613 Mike and Sulley are back at Monsters Universit...
...
19995 56801 The iconic creatures from two of the scariest ...
19996 122337 When a bored-with-life English teacher meets a...
19997 187875 Herbert Blount is a crowdfunding contributor f...
19998 187873 REAL BOY is the coming-of-age story of Bennett...
19999 56805 Following a childhood tragedy, Dewey Cox follo...

```

[20000 rows x 2 columns]

```

**movies.csv**
movieId title \
0 27509 Carolina (2005)
1 27618 Sound of Thunder, A (2005)
2 27788 Jacket, The (2005)
3 27821 Interpreter, The (2005)
4 27839 Ring Two, The (2005)
...
23493 209051 Jeff Garlin: Our Man in Chicago (2019)
23494 209085 The Mistletoe Secret (2019)
23495 209133 The Riot and the Dance (2018)
23496 209157 We (2018)
23497 209163 Bad Poems (2018)

```

```

genres
0 Comedy|Romance
1 Action|Adventure|Drama|Sci-Fi|Thriller
2 Drama|Mystery|Sci-Fi|Thriller
3 Drama|Thriller
4 Drama|Horror|Mystery|Thriller
...
23493 (no genres listed)
23494 Romance
23495 (no genres listed)
23496 Drama
23497 Comedy|Drama

```

[23498 rows x 3 columns]

The movies.csv dataset was used to create a new column to contain the target feature (genres) for movie\_story\_student\_file.csv and movie\_story\_evaluation.csv. This was done by joining the data on movieid/movie\_id. Next, this new column was turned into a binary format and identified 1 for comedy and 0 for non-comedy. The final dataset for movie\_story\_student\_file.csv and movie\_story\_evaluation.csv are shown below:

- Movie\_story\_student\_file.csv

	movieid	story	genres
0	131072	A girl who always tends to fall in love with t...	1
1	196609	Bigfoot has come to the town of Ellwood City, ...	1
2	131074	At an altitude of 18,000 feet, Alaska's Mount...	0
3	196611	In her first special since 2003, Ellen revisi...	1
4	196613	Mike and Sulley are back at Monsters Universi...	1
...	...	...	...
18881	56801	The iconic creatures from two of the scariest ...	0
18882	122337	When a bored-with-life English teacher meets a...	0
18883	187875	Herbert Blount is a crowdfunding contributor f...	0
18884	187873	REAL BOY is the coming-of-age story of Bennett...	0
18885	56805	Following a childhood tragedy, Dewey Cox follo...	1

18886 rows × 3 columns

- Movie\_story\_evaluation.csv

	movie_id	story	genres	ComedyGenre2
0	122349	Growing up in the Mission district of San Fran...	Drama	0
1	122351	A soldier returns home from the Iraq war only ...	Horror Thriller	0
2	122361	Marco the Monkey works as a beach officer. But...	Animation Children Comedy	1
3	187901	When an honest cop, Vijay Kumar's family is r...	Action Romance	0
4	187903	Kathiresan aka Kaththi, a criminal, escapes fr...	Action Drama Romance	0
...	...	...	...	...
3493	131062	In the middle of nowhere, 20 years after an ap...	Drama Fantasy Sci-Fi	0
3494	131064	After living for years as a struggling artist ...	Comedy	1
3495	131066	Ronal is a young barbarian with low self-estee...	Adventure Animation Fantasy	0
3496	131068	Ziege, Hixc3xa4schen and Max have now moved t...	Comedy	1
3497	131070	During their childhood, Hanna and Clarissa wer...	Drama Mystery Thriller	0

3498 rows × 4 columns

Later in the modeling, Movie\_story\_student\_file.csv is used to train models and find the best performing model and Movie\_story\_evaluation.csv is used to test the best performing model created with Movie\_story\_student\_file.csv.

In order to turn the unstructured data into structured, word embedding was used. We decided to take this approach because we noticed we had over 100,000 features when we used the count vectorizer method from sklearn. Also, Word embedding seems to be a better approach with dealing with high dimensional data. Regarding word embedding, GLoVe 300d was used as our pre trained word embedding method.

## 4.The Model

We chose to use logistic regression, decision tree, random forest and Stochastic Gradient Descent for binary classification of comedy movies. The data was split 70% training and 30% test using train\_test\_split from sklearn.

### 4.1 Logistic regression

As stated earlier, we used word embedding and GLoVe 300d was used as our pre trained word embedding method to reduce the dimension to 300 features. This modeling approach implements probability of a document belonging to 1 or 0.

$$y = \frac{1}{1 + e^{-z}}, z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_m X_m$$

$$prob(y^{(i)} = 1|X^{(i)}) = \frac{1}{1 + e^{-z^{(i)}}}, z^{(i)} = \beta_0 + \beta_1 X_1^{(i)} + \beta_2 X_2^{(i)} + \dots + \beta_m X_m^{(i)}$$

GridsearchCV was used for cross validation to find the best  $\lambda$  and 'C' value to find the best inverse of regularization strength for our logistic regression model.

C:[1, 0.75, 0.65, 0.5, 0.25, 0.1]

Best 'C' was 1

In sample accuracy: 79.485

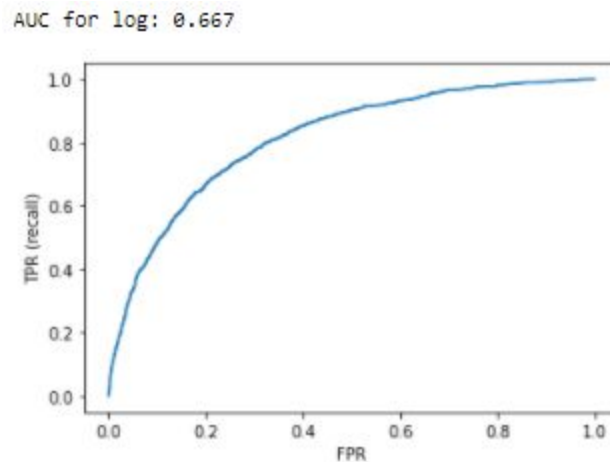
Out of sample accuracy: 79.1666

```
Confusion Matrix:
[[4110  310]
 [ 940  640]]
Classification Report:
              precision    recall  f1-score   support

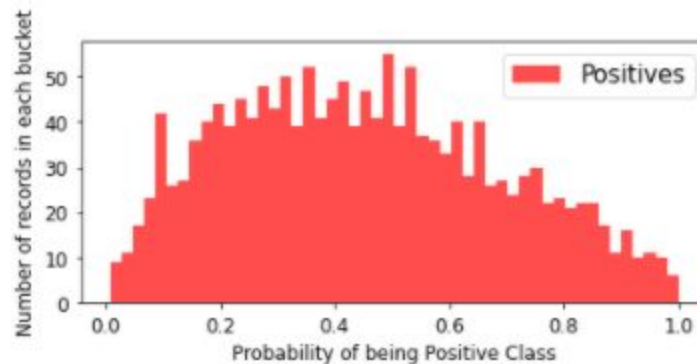
     0       0.81         0.93         0.87         4420
     1       0.67         0.41         0.51         1580

 accuracy          0.79         6000
 macro avg         0.74         0.67         0.69         6000
 weighted avg         0.78         0.79         0.77         6000
```

This model predicted more non comedy cases correctly than correct comedy cases. Also, it produced less type 1 errors than type 2 errors. Also, we will compare precision and recall with the next model later.



Looking at the AUC curve above, we can see how accurate it predicts 1 as 1 and 0 as 0. This auc is not very pleasing.



The above graph displays the probability of a document as a comedy. Probabilities that were greater than 0.5 will result as comedy and anything less, will result as a non comedy. The distribution is good.

## 4.2 K-Nearest Neighbours

We have used GridSearch method to determine the best number of neighbours for our model. And It turns out to be 19 in our case.

N\_neighbors: range(1 to 20]

N\_splits: 5

```
Best Parameter: {'n_neighbors': 19}
Best Cross Validation Score: 0.7641428571428571
```

The in-sample accuracy is 78.97857142857143

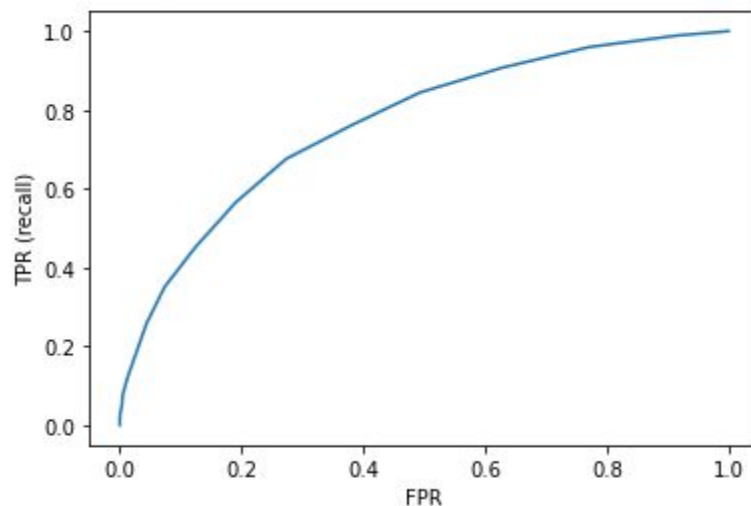
The out-of-sample accuracy is 77.23333333333333

```
Confusion Matrix
[[4226  194]
 [1172  408]]
Classification Report
```

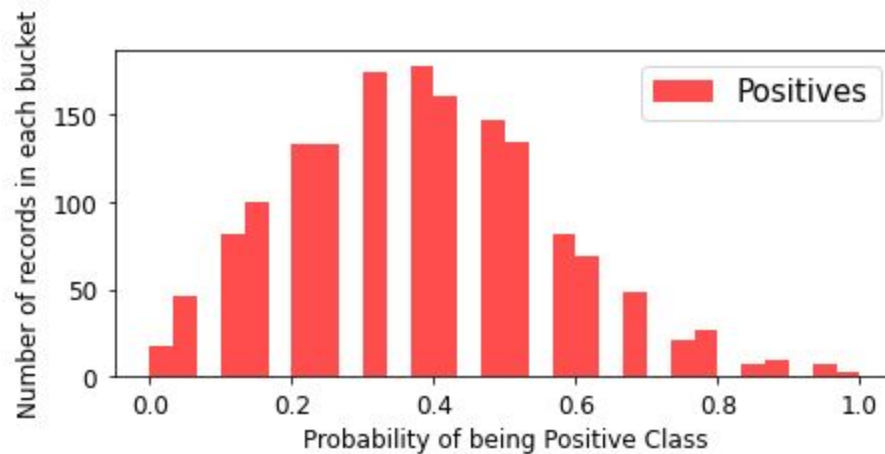
	precision	recall	f1-score	support
0	0.78	0.96	0.86	4420
1	0.68	0.26	0.37	1580
accuracy			0.77	6000
macro avg	0.73	0.61	0.62	6000
weighted avg	0.76	0.77	0.73	6000

This model's accuracy is lower to logistic regression, where the variance is low and the bias is reasonable; so, logistic regression is performing better than this model.

AUC for log: 0.607







### 4.3 Stochastic Gradient Descent

We decided to use a classic model from text classification and natural language. This model handles sparse data very well.

```
Confusion Matrix
[[4325  95]
 [1277 303]]
Classification Report
      precision    recall  f1-score   support

     0       0.77      0.98      0.86      4420
     1       0.76      0.19      0.31      1580

 accuracy      0.77
 macro avg     0.77      0.59      0.58
weighted avg     0.77      0.77      0.72
```

In sample accuracy: 77.4

Out of sample accuracy: 77.1

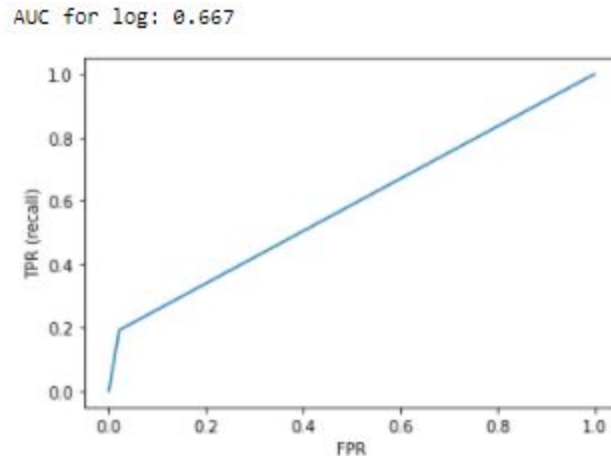
This model's accuracy is similar to logistic regression, where the variance is low and the bias is reasonable; but, logistic regression is still performing better than this model. Although this model does not commit as many type 1 error as the logistic regression, it lacks the accuracy when a comedy is a comedy and it predicted it as a comedy correctly.

Looking at the recall here, it is fairly low.

Recall is calculated using the formula below:

$$\text{Recall} = \text{TruePositives} / (\text{TruePositives} + \text{FalseNegatives})$$

Here, we can assume this model is producing too many false negatives and it is causing the recall to drop. Unlike, the logistic model.



Looking at the AUC curve above, we can see how accurate it predicts 1 as 1 and 0 as 0. This auc is not very pleasing.

#### 4.4 Decision Tree

Using sklearn, we developed a decision tree model. This model used the gini index for node purity.

$$G(k) = \sum_{i=1}^m p(i) \times (1 - p(i))$$

Using decision trees was not better than logistic regression. This model was able to classify comedy movies 74% correct. It is predicting more non comedy movies when they were a comedy than logistic regression was.

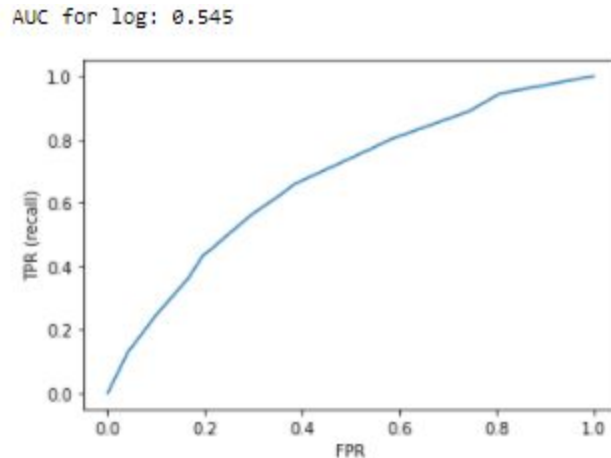
```
Confusion Matrix
[[4229 191]
 [1368 212]]
Classification report
      precision    recall  f1-score   support

     0       0.76      0.96      0.84       4420
     1       0.53      0.13      0.21       1580

 accuracy      0.74      6000
 macro avg     0.64      0.55      0.53      6000
 weighted avg   0.70      0.74      0.68      6000
```

In sample accuracy: 74.914

Out of sample accuracy: 74.016



Looking at the AUC curve above, we can see how accurate it predicts 1 as 1 and 0 as 0. This auc is not very pleasing.

#### 4.5 Random Forest

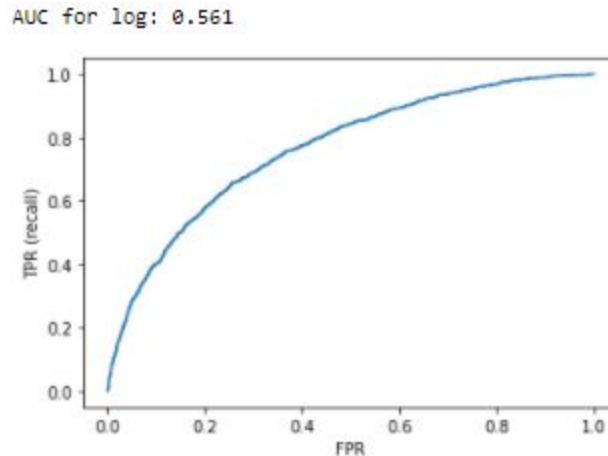
Naturally, Decision tree models suffer from overfitting. In order to avoid this, we decided on a random forest model next.

```
Confusion Matrix
[[4336  84]
 [1356 224]]
Classification report
      precision    recall  f1-score   support

     0       0.76      0.98      0.86      4420
     1       0.73      0.14      0.24      1580

 accuracy      0.76
 macro avg      0.74      0.56      0.55
weighted avg      0.75      0.76      0.69
```

Here, we can see a similar confusion matrix like Stochastic Gradient Descent, where it is producing a lot of type 2 errors. This is something that the logistic model does not do. Also, we can see that the recall is also suffering. This is similar to Stochastic Gradient Descent



Looking at the AUC curve above, we can see how accurate it predicts 1 as 1 and 0 as 0. This auc is not very pleasing.

In sample accuracy: 99.5  
Out of sample accuracy: 76.0

This model has very high variance compared to the bias. The in sample is almost 100% and one could possibly assume that this model is suffering from overfitting if more analysis is done with this model.

## 5. Conclusion

The goal of this group project was to predict whether or not a movie is a comedy based on the movie description/story. And, to use the movie\_story\_student\_file.csv dataset to train different models in order to find the best model for classifying comedy movies; and then using the Movie\_story\_evaluation.csv data set to further test our best model.

Logistic regression performed the best out of all of the models we trained. At first, we were worried that our initial logistic regression model was suffering from underfitting. This was because we found our accuracy to have a large amount of bias and very low variance but after using this new data set, the accuracy came very close to what it previously was predicting and no issues were found.

```

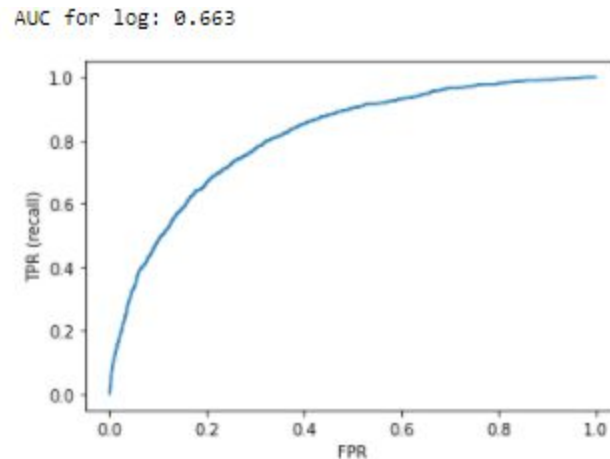
Confusion Matrix
[[2445 178]
 [ 530 345]]
Classification report

```

	precision	recall	f1-score	support
0	0.82	0.93	0.87	2623
1	0.66	0.39	0.49	875
accuracy			0.80	3498
macro avg	0.74	0.66	0.68	3498
weighted avg	0.78	0.80	0.78	3498

This model performed the when in terms of the confusion matrix, precision, recal and even f1-score. By looking at the concussion matrix, it did fairly okay and avoided too many type 1 and type 2 error compared to the rest of the models.

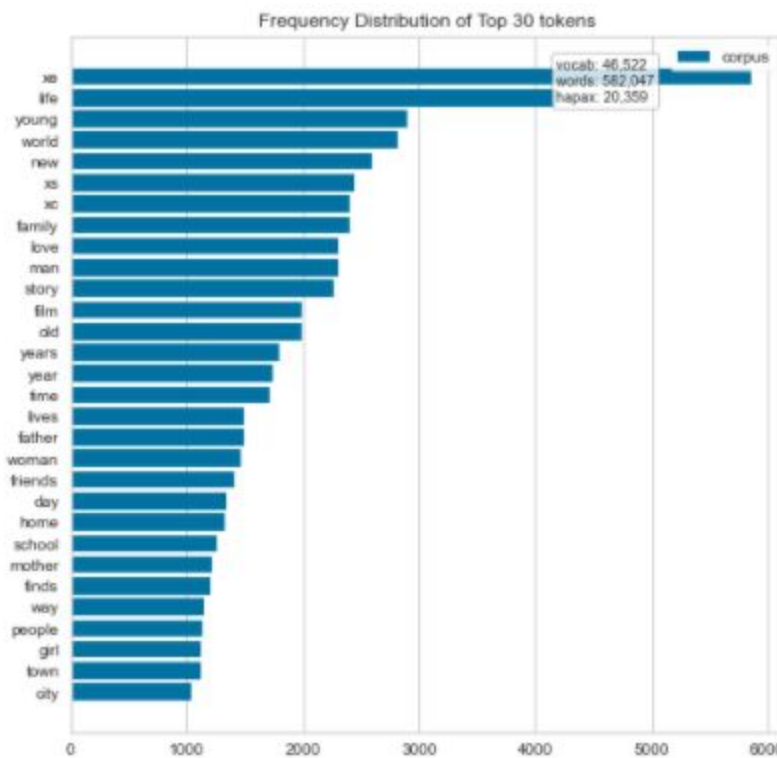
Evaluation data accuracy: 79.759



Also, as we saw from the rest of the models, the auc is fairly low and not very pleasing.

## 6. Appendix

### 6.1 Frequency Distribution of the Top 30 Tokens



Using data from the Movie\_story\_student\_file.csv data set before splitting, here is the frequency distribution of the top 30 tokens.