Sentiment Analysis of Book Genres

- Shahrukh Sohail

Project Overview

Our project focuses on sentiment analysis. More specifically, researching and implementing machine learning models to predict book genre based on its synopsis .

Data Source

```
df=pd.read_csv('https://raw.githubusercontent.com/srkcheema/GenreAnalysis/main/data.csv')
df.head()
```

	Unnamed: 0	title	rating	name	num_ratings	num_reviews	num_followers	synopsis	genre
0	0	Sapiens: A Brief History of Humankind	4.39	Yuval Noah Harari	8,06,229	46,149	30.5k	100,000 years ago, at least six human species ...	history
1	1	Guns, Germs, and Steel: The Fates of Human Soc...	4.04	Jared Diamond	3,67,056	12,879	6,538	"Diamond has written a book of remarkable scop...	history
2	2	A People's History of the United States	4.07	Howard Zinn	2,24,620	6,509	2,354	In the book, Zinn presented a different side o...	history
3	3	The Devil in the White City: Murder, Magic, an...	3.99	Erik Larson	6,13,157	36,644	64.2k	Author Erik Larson imbues the incredible event...	history
4	4	The Diary of a Young Girl	4.18	Anne Frank	33,13,033	35,591	4,621	Discovered in the attic in which she spent the...	history

Process

1st

Data Exploration and Cleaning

Explore any unusable values within the dataset and transform or clean the data so that it does not yield inaccuracies

2nd

Data Processing

Format the data so that it is usable by the Machine Learning Models

3rd

Model Building and Training

Split the data into a train/test set and then train the different models using the train set

4th

Model Validation

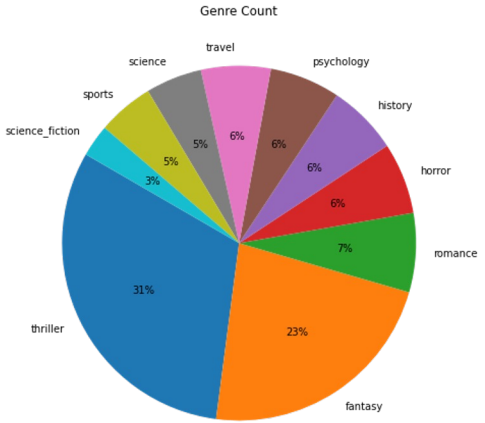
Use the test set to validate the accuracy metrics of the different model to check which model yielded the best result on our dataset

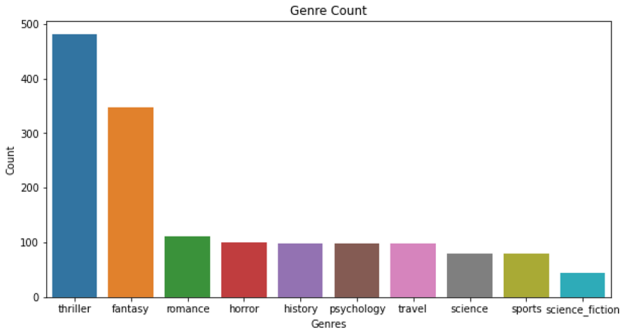
Data Processing

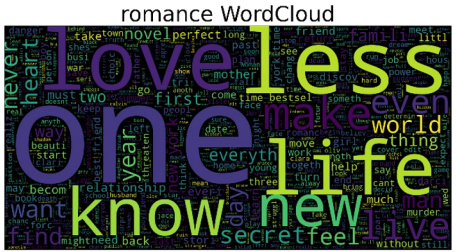
- Remove Unnamed column and converted data type to integer for reviews, ratings and followers column
- Modify the text in the synopsis column which includes:
  - Convert all the text to lowercase
  - Remove any text in square brackets, links present, punctuations, next line characters and words containing numbers
  - Remove stopwords
  - Perform stemming on the words

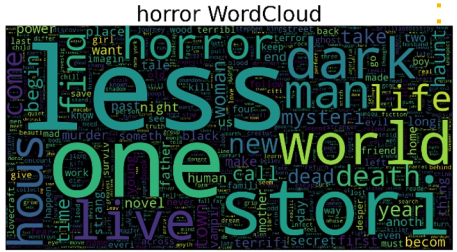
Data Visuals

We looked into created various data visuals to get any insight on the data that we might have missed









Model Training

- Performed Labelled encoding on genre column
- Vectorized Synopsis Text
- Split the data 80/20 to train on the different models
- Define a generic function to train the models on

Label Encoding

```
#Encoding the genre column
encoder=LabelEncoder()
df['genre'] = encoder.fit_transform(df['genre'])
```

Text Vectorization

```
# Splitting data into features(X) and targets(y)
x = df['synopsis']
y = df['genre'].values
```

```
# Splitting the data into train and test sets to use in models
X_train,X_test,y_train,y_test = train_test_split(x, y, test_size= 0.2, random_state= 42, stratify=y)
```

```
tfidf=TfidfVectorizer()
X_train = tfidf.fit_transform(X_train).toarray()
X_test = tfidf.transform(X_test).toarray()
```

Model Training

```
def train_model(model):
    model.fit(X_train,y_train)
    pred = model.predict(X_test)
    probability = model.predict_proba(X_test)
    accuracy = round(accuracy_score(y_test,pred),3)
    precision = round(precision_score(y_test,pred,average='weighted'),3)
    recall = round(recall_score(y_test,pred,average='weighted'),3)

    print('Accuracy: ', accuracy)
    print('Precision: ', precision)
    print('Recall: ', recall)

    fig, ax = plt.subplots(1, 2, figsize = (25,  8))
    ax1 = plot_confusion_matrix(y_test, pred, ax= ax[0], cmap= 'cool')
    ax2 = plot_roc(y_test, probability, ax= ax[1], plot_macro= False, plot_micro= False, cmap= 'summer')
```

ML Models Used

- Bernoulli Naive Bayes
- Categorical Naive Bayes
- Logistic Regression
- Random Forest

Model Validation

Random Forest Model had the greatest prediction accuracy of 61.4%, followed by Logistic Regression with an accuracy score of 60.1%.

Table 3: Prediction Results

Model	Accuracy
Bernoulli Naive Bayes	45.1%
Categorical Naive Bayes	31.2%
Logistic Regression	60.1%
Random Forest	61.4%

Future Scope

Experiment with changing number of genres

Increasing the scope of our dataset to other regions and languages of the world

Thanks!