

# Predicting Initiating Events in Narrative Texts

By: Adrian Chavez-Loya

## Background

Story telling is a key component of interpersonal communication and the study of narrative ability in children can provide critical insights into their language development. Narrative sample analysis is a process in which an individual produces a narrative and then a Speech-Language Pathologist (or similar practitioner) analyzes the quality. One tool for measuring this quality is the Monitoring Indicators of Scholarly Language (MISL). It provides an objective measure of the macrostructure story elements (e.g. Characters, Setting, Initiating Event) as well as the microstructure or grammatical elements.

The process of scoring the macrostructure can be very time consuming though, which leads to less effective ongoing monitoring. This dataset provides the first publicly accessible data for attempting to automate scoring of the macrostructure via Machine Learning.

## Dataset:

`AutomatedNarrativeAnalysisMISLData.csv`

## Task

We will predict the Initiating Event ( **IE** ) label. The **IE** is scored as either 0, 1, 2, or 3 but for our purposes it is acceptable to predict this as either a continuous or categorical output. If you predict it as continuous, it is necessary to constrain the prediction in some way, therefore, categorical may be easier.

For predictor variables, we have two choices: either the raw text or the text features (or both, technically). The text features are every column **except** Char , Sett , IE , Plan , Act , and Con . Those 6 variables are the output scores but again we'll just be focusing on IE for now. Also, exclude the ID column.

Using cross-validation, we will explore the many different classification algorithms we discussed to find the model with the highest performance (I'll leave it to you to define performance).

```
In [ ]: import pandas as pd
from sklearn.model_selection import train_test_split, cross_val_s
from sklearn.preprocessing import StandardScaler
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.linear_model import LogisticRegression
from sklearn.neighbors import KNeighborsClassifier
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier, GradientBoos
from sklearn.svm import SVC
from sklearn.metrics import accuracy_score, f1_score

df = pd.read_csv('AutomatedNarrativeAnalysisMISLData.csv')
```

```
In [ ]: df.head()
```

Out [ ]:	ID	vecOfNarratives	Char	Sett	IE	Plan	Act	Con	ENP	DESPC	...	WRI
<b>0</b>	1	The mom wanted to go and pet the alien dog. B...	1	0	2	1	2	0	2	1	...	430.7
<b>1</b>	2	There was two little kids who were walking in...	1	1	3	0	2	3	3	1	...	399.8
<b>2</b>	3	these aliens came to earth. and this girl a...	1	1	1	0	1	0	2	1	...	462.7
<b>3</b>	4	One time this alien ship came down to earth. ...	1	1	1	1	1	0	2	1	...	425.2
<b>4</b>	5	Aliens came down from the planet they came fr...	0	0	2	0	2	0	1	1	...	452.6

5 rows × 117 columns

```
In [ ]: # Dropped unnecessary ID column
df.drop(columns=['ID'], inplace=True)
```

```
In [ ]: df.head()
```

Out [ ]:	vecOfNarratives	Char	Sett	IE	Plan	Act	Con	ENP	DESPC	DESSC	...
0	The mom wanted to go and pet the alien dog. B...	1	0	2	1	2	0	2	1	11	...
1	There was two little kids who were walking in...	1	1	3	0	2	3	3	1	29	...
2	these aliens came to earth. and this girl a...	1	1	1	0	1	0	2	1	7	...
3	One time this alien ship came down to earth. ...	1	1	1	1	1	0	2	1	5	...
4	Aliens came down from the planet they came fr...	0	0	2	0	2	0	1	1	6	...

5 rows × 116 columns

```
In [ ]: # Extracted features/target 'IE'
X = df.drop(columns=['Char', 'Sett', 'IE', 'Plan', 'Act', 'Con',
y = df['IE']
```

```
In [ ]: # Tf-Idf Vextorizer
tfidf = TfidfVectorizer(max_features=1000)
text_features = tfidf.fit_transform(df['vecOfNarratives']).toarray

# Combine text features with other features
X_text = pd.DataFrame(text_features, columns=[f'text_feat_{i}' for i in range(text_features.shape[1])])
X = pd.concat([X.reset_index(drop=True), X_text], axis=1)

# Converted columns to numeric values, dropped others
X = X.apply(pd.to_numeric, errors='coerce')
X.dropna(axis=1, inplace=True)

# Split data into training and test sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Standardized data with standard scaler
scaler = StandardScaler()
X_train = scaler.fit_transform(X_train)
X_test = scaler.transform(X_test)
```

```
# Chose models to use for cross-validation
models = {
    'Logistic Regression': LogisticRegression(max_iter=1000),
    'KNN': KNeighborsClassifier(),
    'Decision Tree': DecisionTreeClassifier(),
    'Random Forest': RandomForestClassifier(),
    'Gradient Boosting': GradientBoostingClassifier(),
    'AdaBoost': AdaBoostClassifier(),
    'SVM': SVC()
}
```

```
In [ ]: # Cross-validation and model evaluation
results = {}
for model_name, model in models.items():
    scores = cross_val_score(model, X_train, y_train, cv=5, scoring='accuracy')
    results[model_name] = scores.mean()
```

```
In [ ]: # Results of Model Performance
print("Model Performance (Accuracy):")
for model_name, score in results.items():
    print(f"{model_name}: {score:.4f}")

# Trained the best model on the entire training set and evaluate
best_model_name = max(results, key=results.get)
best_model = models[best_model_name]
best_model.fit(X_train, y_train)
y_pred = best_model.predict(X_test)
test_accuracy = accuracy_score(y_test, y_pred) # Evaluation for test set
test_f1 = f1_score(y_test, y_pred, average='weighted')

print(f"\nBest Model: {best_model_name}")
print(f"Test Accuracy: {test_accuracy:.4f}")
print(f"Test F1 Score: {test_f1:.4f}")
```

```
Model Performance (Accuracy):
Logistic Regression: 0.5285
KNN: 0.4077
Decision Tree: 0.3805
Random Forest: 0.5557
Gradient Boosting: 0.5255
AdaBoost: 0.4106
SVM: 0.5014
```

```
Best Model: Random Forest
Test Accuracy: 0.5060
Test F1 Score: 0.4672
```

- The Random Forest model showed the highest cross-validation accuracy, but its performance on the test set was slightly lower,

- indicating a potential overfitting issue.
- Logistic Regression and Gradient Boosting also showed competitive cross-validation scores, suggesting they might be worth further exploration or tuning.
- Further tuning of hyperparameters, such as adjusting the number of trees in Random Forest or the regularization parameter in Logistic Regression, could potentially improve performance on the test set.
- Feature engineering, such as experimenting with different text features or additional domain-specific features, could also enhance model performance.

## Hyperparameter Tuning

```
In [ ]: from sklearn.metrics import accuracy_score, f1_score

# Define the parameter grid
param_grid = {
    'n_estimators': [50, 100, 200],
    'max_depth': [None, 10, 20, 30],
    'min_samples_split': [2, 5, 10],
    'min_samples_leaf': [1, 2, 4],
    'max_features': ['auto', 'sqrt', 'log2']
}

# Initialize the Random Forest model and Grid Search
rf_model = RandomForestClassifier(random_state=42)
grid_search = GridSearchCV(estimator=rf_model, param_grid=param_g

# Fit GridSearchCV
grid_search.fit(X_train, y_train)

print("Best Parameters:", grid_search.best_params_) # Best parame
print("Best Cross-Validation Accuracy:", grid_search.best_score_)
best_rf_model = grid_search.best_estimator_ # Best Model
y_pred = best_rf_model.predict(X_test)
test_accuracy = accuracy_score(y_test, y_pred)
test_f1 = f1_score(y_test, y_pred, average='weighted')
print(f"Test Accuracy: {test_accuracy:.4f}") # Test accuraxy/F1 S
print(f"Test F1 Score: {test_f1:.4f}")
```

[illegible]

[illegible]



[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]



```
[CV] END max_depth=10, n_estimators=100; total time= 1.2s
[CV] END max_depth=20, max_features=log2, min_samples_leaf=1, min
_samples_split=10, n_estimators=200; total time= 1.9s
[CV] END max_depth=20, max_features=log2, min_samples_leaf=2, min
_samples_split=2, n_estimators=200; total time= 1.9s
[CV] END max_depth=20, max_features=log2, min_samples_leaf=2, min
_samples_split=5, n_estimators=100; total time= 0.9s
[CV] END max_depth=20, max_features=log2, min_samples_leaf=2, min
_samples_split=10, n_estimators=50; total time= 0.4s
[CV] END max_depth=20, max_features=log2, min_samples_leaf=2, min
_samples_split=10, n_estimators=50; total time= 0.3s
[CV] END max_depth=20, max_features=log2, min_samples_leaf=2, min
_samples_split=10, n_estimators=100; total time= 0.7s
[CV] END max_depth=20, max_features=log2, min_samples_leaf=4, min
_samples_split=2, n_estimators=50; total time= 0.3s
[CV] END max_depth=20, max_features=log2, min_samples_leaf=4, min
_samples_split=2, n_estimators=50; total time= 0.3s
[CV] END max_depth=20, max_features=log2, min_samples_leaf=4, min
_samples_split=2, n_estimators=100; total time= 0.8s
[CV] END max_depth=20, max_features=log2, min_samples_leaf=4, min
_samples_split=2, n_estimators=200; total time= 1.1s
[CV] END max_depth=20, max_features=log2, min_samples_leaf=4, min
_samples_split=5, n_estimators=200; total time= 1.0s
[CV] END max_depth=20, max_features=log2, min_samples_leaf=4, min
_samples_split=10, n_estimators=200; total time= 1.0s
[CV] END max_depth=30, max_features=sqrt, min_samples_leaf=1, min
_samples_split=5, n_estimators=50; total time= 0.5s
[CV] END max_depth=30, max_features=sqrt, min_samples_leaf=1, min
_samples_split=5, n_estimators=50; total time= 0.3s
[CV] END max_depth=30, max_features=sqrt, min_samples_leaf=1, min
_samples_split=5, n_estimators=100; total time= 0.6s
[CV] END max_depth=30, max_features=sqrt, min_samples_leaf=1, min
_samples_split=5, n_estimators=100; total time= 0.6s
[CV] END max_depth=30, max_features=sqrt, min_samples_leaf=1, min
_samples_split=5, n_estimators=100; total time= 0.6s
[CV] END max_depth=30, max_features=sqrt, min_samples_leaf=1, min
_samples_split=5, n_estimators=100; total time= 0.6s
[CV] END max_depth=30, max_features=sqrt, min_samples_leaf=1, min
_samples_split=5, n_estimators=100; total time= 0.5s
[CV] END max_depth=30, max_features=sqrt, min_samples_leaf=1, min
_samples_split=5, n_estimators=200; total time= 1.0s
[CV] END max_depth=30, max_features=sqrt, min_samples_leaf=2, min
_samples_split=10, n_estimators=100; total time= 0.4s
[CV] END max_depth=20, max_features=sqrt, min_samples_leaf=4, min
_samples_split=10, n_estimators=100; total time= 0.8s
[CV] END max_depth=20, max_features=sqrt, min_samples_leaf=4, min
_samples_split=10, n_estimators=200; total time= 1.6s
[CV] END max_depth=20, max_features=log2, min_samples_leaf=1, min
_samples_split=2, n_estimators=100; total time= 1.2s
[CV] END max depth=20, max features=log2, min samples leaf=1, min
```

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]



[illegible]

[illegible]

\_samples\_split=5, n\_estimators=100; total time= 0.8s  
[CV] END max\_depth=20, max\_features=sqrt, min\_samples\_leaf=4, min  
\_samples\_split=5, n\_estimators=200; total time= 1.6s  
[CV] END max\_depth=20, max\_features=sqrt, min\_samples\_leaf=4, min  
\_samples\_split=10, n\_estimators=200; total time= 1.7s  
[CV] END max\_depth=20, max\_features=log2, min\_samples\_leaf=1, min  
\_samples\_split=2, n\_estimators=200; total time= 2.1s  
[CV] END max\_depth=20, max\_features=log2, min\_samples\_leaf=1, min  
\_samples\_split=5, n\_estimators=200; total time= 2.3s  
[CV] END max\_depth=20, max\_features=log2, min\_samples\_leaf=1, min  
\_samples\_split=10, n\_estimators=100; total time= 1.0s  
[CV] END max\_depth=20, max\_features=log2, min\_samples\_leaf=2, min  
\_samples\_split=2, n\_estimators=50; total time= 0.5s  
[CV] END max\_depth=20, max\_features=log2, min\_samples\_leaf=2, min  
\_samples\_split=2, n\_estimators=100; total time= 1.1s  
[CV] END max\_depth=20, max\_features=log2, min\_samples\_leaf=2, min  
\_samples\_split=2, n\_estimators=200; total time= 1.8s  
[CV] END max\_depth=20, max\_features=log2, min\_samples\_leaf=2, min  
\_samples\_split=5, n\_estimators=200; total time= 1.6s  
[CV] END max\_depth=20, max\_features=log2, min\_samples\_leaf=2, min  
\_samples\_split=10, n\_estimators=100; total time= 0.6s  
[CV] END max\_depth=20, max\_features=log2, min\_samples\_leaf=4, min  
\_samples\_split=2, n\_estimators=50; total time= 0.3s  
[CV] END max\_depth=20, max\_features=log2, min\_samples\_leaf=4, min  
\_samples\_split=2, n\_estimators=100; total time= 0.8s  
[CV] END max\_depth=20, max\_features=log2, min\_samples\_leaf=4, min  
\_samples\_split=2, n\_estimators=200; total time= 1.1s  
[CV] END max\_depth=20, max\_features=log2, min\_samples\_leaf=4, min  
\_samples\_split=5, n\_estimators=100; total time= 0.5s  
[CV] END max\_depth=20, max\_features=log2, min\_samples\_leaf=4, min  
\_samples\_split=10, n\_estimators=50; total time= 0.3s  
[CV] END max\_depth=20, max\_features=log2, min\_samples\_leaf=4, min  
\_samples\_split=10, n\_estimators=100; total time= 0.5s  
[CV] END max\_depth=20, max\_features=log2, min\_samples\_leaf=4, min  
\_samples\_split=10, n\_estimators=200; total time= 1.1s  
[CV] END max\_depth=30, max\_features=sqrt, min\_samples\_leaf=1, min  
\_samples\_split=10, n\_estimators=200; total time= 1.3s  
[CV] END max\_depth=30, max\_features=sqrt, min\_samples\_leaf=1, min  
\_samples\_split=10, n\_estimators=200; total time= 1.2s  
[CV] END max\_depth=30, max\_features=sqrt, min\_samples\_leaf=1, min  
\_samples\_split=10, n\_estimators=200; total time= 1.0s  
[CV] END max\_depth=30, max\_features=sqrt, min\_samples\_leaf=2, min  
\_samples\_split=2, n\_estimators=50; total time= 0.2s  
[CV] END max\_depth=30, max\_features=sqrt, min\_samples\_leaf=2, min  
\_samples\_split=2, n\_estimators=50; total time= 0.2s  
[CV] END max\_depth=30, max\_features=sqrt, min\_samples\_leaf=2, min  
\_samples\_split=2, n\_estimators=50; total time= 0.2s  
[CV] END max\_depth=30, max\_features=sqrt, min\_samples\_leaf=2, min  
\_samples\_split=2, n\_estimators=50; total time= 0.2s  
[CV] END max\_depth=30, max\_features=sqrt, min\_samples\_leaf=2, min

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]



[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]



[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]



[illegible]

[illegible]

```
[CV] END max_depth=10, max_features=log2, min_samples_leaf=4, min
_samples_split=10, n_estimators=50; total time= 0.2s
[CV] END max_depth=20, max_features=sqrt, min_samples_leaf=2, min
_samples_split=5, n_estimators=100; total time= 0.6s
[CV] END max_depth=20, max_features=sqrt, min_samples_leaf=2, min
_samples_split=5, n_estimators=100; total time= 0.5s
[CV] END max_depth=20, max_features=sqrt, min_samples_leaf=2, min
_samples_split=5, n_estimators=100; total time= 0.6s
[CV] END max_depth=20, max_features=sqrt, min_samples_leaf=2, min
_samples_split=5, n_estimators=200; total time= 1.3s
[CV] END max_depth=20, max_features=sqrt, min_samples_leaf=4, min
_samples_split=2, n_estimators=50; total time= 0.3s
[CV] END max_depth=20, max_features=sqrt, min_samples_leaf=4, min
_samples_split=2, n_estimators=100; total time= 0.6s
[CV] END max_depth=20, max_features=sqrt, min_samples_leaf=4, min
_samples_split=2, n_estimators=200; total time= 1.3s
[CV] END max_depth=20, max_features=sqrt, min_samples_leaf=4, min
_samples_split=5, n_estimators=100; total time= 0.8s
[CV] END max_depth=20, max_features=sqrt, min_samples_leaf=4, min
_samples_split=10, n_estimators=50; total time= 0.4s
[CV] END max_depth=20, max_features=sqrt, min_samples_leaf=4, min
_samples_split=10, n_estimators=50; total time= 0.4s
[CV] END max_depth=20, max_features=sqrt, min_samples_leaf=4, min
_samples_split=10, n_estimators=100; total time= 0.9s
[CV] END max_depth=20, max_features=sqrt, min_samples_leaf=4, min
_samples_split=10, n_estimators=200; total time= 1.7s
[CV] END max_depth=20, max_features=log2, min_samples_leaf=1, min
_samples_split=2, n_estimators=200; total time= 2.1s
[CV] END max_depth=20, max_features=log2, min_samples_leaf=1, min
_samples_split=5, n_estimators=200; total time= 2.3s
[CV] END max_depth=20, max_features=log2, min_samples_leaf=1, min
_samples_split=10, n_estimators=200; total time= 2.1s
[CV] END max_depth=20, max_features=log2, min_samples_leaf=2, min
_samples_split=2, n_estimators=100; total time= 0.9s
[CV] END max_depth=20, max_features=log2, min_samples_leaf=2, min
_samples_split=5, n_estimators=50; total time= 0.5s
[CV] END max_depth=20, max_features=log2, min_samples_leaf=2, min
_samples_split=5, n_estimators=50; total time= 0.5s
[CV] END max_depth=20, max_features=log2, min_samples_leaf=2, min
_samples_split=5, n_estimators=100; total time= 0.9s
[CV] END max_depth=20, max_features=log2, min_samples_leaf=2, min
_samples_split=10, n_estimators=50; total time= 0.4s
[CV] END max_depth=20, max_features=log2, min_samples_leaf=2, min
_samples_split=10, n_estimators=100; total time= 0.7s
[CV] END max_depth=20, max_features=log2, min_samples_leaf=2, min
_samples_split=10, n_estimators=200; total time= 1.3s
[CV] END max_depth=20, max_features=log2, min_samples_leaf=4, min
_samples_split=2, n_estimators=100; total time= 0.7s
[CV] END max_depth=20, max_features=log2, min_samples_leaf=4, min
```

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]



[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]



[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

```
/Users/adrianchavezloya/anaconda3/lib/python3.11/site-packages/sklearn/model_selection/_validation.py:425: FitFailedWarning:
540 fits failed out of a total of 1620.
The score on these train-test partitions for these parameters will
be set to nan.
If these failures are not expected, you can try to debug them by
setting error_score='raise'.
```

Below are more details about the failures:

-----

238 fits failed with the following error:

Traceback (most recent call last):

```
File "/Users/adrianchavezloya/anaconda3/lib/python3.11/site-packages/sklearn/model_selection/_validation.py", line 732, in _fit_and_score
```

```
    estimator.fit(X_train, y_train, **fit_params)
```

```
File "/Users/adrianchavezloya/anaconda3/lib/python3.11/site-packages/sklearn/base.py", line 1144, in wrapper
```

```
    estimator._validate_params()
```

```
File "/Users/adrianchavezloya/anaconda3/lib/python3.11/site-packages/sklearn/base.py", line 637, in _validate_params
```

```
    validate_parameter_constraints(
```

```
File "/Users/adrianchavezloya/anaconda3/lib/python3.11/site-packages/sklearn/utils/_param_validation.py", line 95, in validate_parameter_constraints
```

```
    raise InvalidParameterError(
```

```
sklearn.utils._param_validation.InvalidParameterError: The 'max_features' parameter of RandomForestClassifier must be an int in the range [1, inf), a float in the range (0.0, 1.0], a str among {'sqrt', 'log2'} or None. Got 'auto' instead.
```

-----

302 fits failed with the following error:

Traceback (most recent call last):

```
File "/Users/adrianchavezloya/anaconda3/lib/python3.11/site-packages/sklearn/model_selection/_validation.py", line 732, in _fit_and_score
```

```
    estimator.fit(X_train, y_train, **fit_params)
```

```
File "/Users/adrianchavezloya/anaconda3/lib/python3.11/site-packages/sklearn/base.py", line 1144, in wrapper
```

```
    estimator._validate_params()
```

```
File "/Users/adrianchavezloya/anaconda3/lib/python3.11/site-packages/sklearn/base.py", line 637, in _validate_params
```

```
    validate_parameter_constraints(
```

```
File "/Users/adrianchavezloya/anaconda3/lib/python3.11/site-packages/sklearn/utils/_param_validation.py", line 95, in validate_parameter_constraints
```

```
    raise InvalidParameterError(
```



```
sklearn.utils._param_validation.InvalidParameterError: The 'max_f
eatures' parameter of RandomForestClassifier must be an int in th
e range [1, inf), a float in the range (0.0, 1.0], a str among
{'log2', 'sqrt'} or None. Got 'auto' instead.
```

warnings.warn(some_fits_failed_message, FitFailedWarning)						
/Users/adrianchavezloya/anaconda3/lib/python3.11/site-packages/sklearn/model_selection/_search.py:976: UserWarning: One or more of the test scores are non-finite: [						
nan	nan	nan		nan	nan	nan
	nan	nan	nan	nan	nan	na
n	nan	nan	nan	nan	nan	na
n	nan	nan	nan	nan	nan	na
n	nan	nan	nan	nan	nan	na
	nan	nan	nan	0.55576662	0.56173677	0.5464948
0.56770692	0.53749435	0.56467662	0.57376753	0.5737223	0.5767978	
3	0.55576662	0.55567616	0.56472185	0.58588874	0.57684306	0.5677521
5	0.56173677	0.55264586	0.55870647	0.54061511	0.55870647	0.5526458
6	0.54061511	0.55870647	0.55264586	0.53147897	0.55563094	0.5647218
5	0.5587517	0.57073722	0.57078245	0.53165988	0.53763003	0.5647218
5	0.54654003	0.55567616	0.55866124	0.56182723	0.55879692	0.5646766
2	0.54966079	0.57073722	0.5737223	0.55269109	0.55264586	0.5586612
4	0.54061511	0.54061511	0.54970602	0.54061511	0.54061511	0.5497060
2	0.54667571	0.54966079	0.55572139	nan	nan	na
n	nan	nan	nan	nan	nan	na
n	nan	nan	nan	nan	nan	na
n	nan	nan	nan	nan	nan	na
n	nan	nan	nan	nan	nan	na
n	nan	nan	nan	nan	nan	na
	0.561782	0.55870647	0.54658526	0.55273632	0.54364541	0.5677069
2	0.5737223	0.56463139	0.57073722	0.56788783	0.55866124	0.5707824
5	0.56472185	0.57381275	0.57385798	0.54961556	0.57078245	0.5708276
8	0.55554048	0.55264586	0.56169154	0.55554048	0.55264586	0.5616915

4	0.52243329	0.54663048	0.56169154	0.52849389	0.5345545	0.5798733
6	0.54070556	0.55581185	0.56779738	0.55572139	0.55572139	0.5677973
8	0.55278155	0.55870647	0.57073722	0.53464496	0.54658526	0.5646766
2	0.54056988	0.55572139	0.55866124	0.54364541	0.55273632	0.5526910
9	0.54364541	0.55273632	0.55269109	0.5375848	0.55273632	0.5526910
9						
n	nan	nan	nan	nan	nan	na
n	nan	nan	nan	nan	nan	na
n	nan	nan	nan	nan	nan	na
n	nan	nan	nan	nan	nan	na
n	nan	nan	nan	0.55879692	0.55870647	0.5525554
3	0.57073722	0.54052465	0.54654003	0.56770692	0.5737223	0.5798281
5	0.55273632	0.55567616	0.56472185	0.58588874	0.57381275	0.5707824
6	0.55870647	0.55264586	0.55870647	0.54061511	0.55870647	0.5526458
5	0.54061511	0.55870647	0.55264586	0.53147897	0.55563094	0.5647218
5	0.55572139	0.55866124	0.55567616	0.53469019	0.54364541	0.5647218
1	0.55251018	0.57078245	0.55866124	0.55581185	0.55879692	0.5616463
4	0.54663048	0.57376753	0.5737223	0.55572139	0.54966079	0.5586612
2	0.54061511	0.54061511	0.54970602	0.54061511	0.54061511	0.5497060
n	0.54667571	0.54966079	0.55572139	nan	nan	na
n	nan	nan	nan	nan	nan	na
n	nan	nan	nan	nan	nan	na
n	nan	nan	nan	nan	nan	na
n	nan	nan	nan	nan	nan	na
2	0.55576662	0.56173677	0.5464948	0.56770692	0.53749435	0.5646766
5	0.57376753	0.5737223	0.57679783	0.55576662	0.55567616	0.5647218

```

0.58588874 0.57684306 0.56775215 0.56173677 0.55264586 0.5587064
7
0.54061511 0.55870647 0.55264586 0.54061511 0.55870647 0.5526458
6
0.53147897 0.55563094 0.56472185 0.55879692 0.56770692 0.5707824
5
0.53165988 0.5375848 0.56472185 0.5464948 0.55567616 0.5586612
4
0.56182723 0.55879692 0.56467662 0.54966079 0.57073722 0.5737223
0.55269109 0.55264586 0.55866124 0.54061511 0.54061511 0.5497060
2
0.54061511 0.54061511 0.54970602 0.54667571 0.54966079 0.5557213
9]

```

```
warnings.warn(
```

```

Best Parameters: {'max_depth': None, 'max_features': 'sqrt', 'min
_samples_leaf': 2, 'min_samples_split': 5, 'n_estimators': 50}
Best Cross-Validation Accuracy: 0.585888738127544
Test Accuracy: 0.4819
Test F1 Score: 0.4300

```

## Model Comparison

```

In [ ]: from sklearn.linear_model import LogisticRegression
from sklearn.ensemble import GradientBoostingClassifier

# Initialized Logistic Regression model/Gradient Boosting model
logreg_model = LogisticRegression(random_state=42, max_iter=1000)
gb_model = GradientBoostingClassifier(random_state=42)

# Trained LR Model/GB Model
logreg_model.fit(X_train, y_train)
gb_model.fit(X_train, y_train)
y_pred_logreg = logreg_model.predict(X_test) # Predictions on tes
y_pred_gb = gb_model.predict(X_test)
test_accuracy_logreg = accuracy_score(y_test, y_pred_logreg)
test_f1_logreg = f1_score(y_test, y_pred_logreg, average='weighte
print("Logistic Regression Performance:") # Log. model regression
print(f"Test Accuracy: {test_accuracy_logreg:.4f}")
print(f"Test F1 Score: {test_f1_logreg:.4f}")
print()

test_accuracy_gb = accuracy_score(y_test, y_pred_gb)
test_f1_gb = f1_score(y_test, y_pred_gb, average='weighted')
print("Gradient Boosting Performance:") #Gradient Boosting result
print(f"Test Accuracy: {test_accuracy_gb:.4f}")
print(f"Test F1 Score: {test_f1_gb:.4f}")

```

Logistic Regression Performance:

Test Accuracy: 0.5542

Test F1 Score: 0.5273

Gradient Boosting Performance:

Test Accuracy: 0.5422

Test F1 Score: 0.5329

## Analysis of Model Comparison:

- Both Logistic Regression and Gradient Boosting models show comparable performance metrics.
- Logistic Regression slightly outperforms Gradient Boosting in terms of accuracy (55.42% vs. 54.22%), but Gradient Boosting has a slightly higher F1 score (53.29% vs. 52.73%).
- These results suggest that both models are reasonable choices, and the preference might depend on whether higher precision or recall is more critical for this application.

```
In [ ]: # Got feature importances from the best Random Forest model
feature_importances = best_rf_model.feature_importances_

# Created a DataFrame to display feature importances
feature_importances_df = pd.DataFrame({
    'Feature': X.columns,
    'Importance': feature_importances
})

# Sorted features by importance
feature_importances_df = feature_importances_df.sort_values(by='I

# Top 10 most important features
print("Top 10 Most Important Features:")
print(feature_importances_df.head(10))
```

### Top 10 Most Important Features:

	Feature	Importance
3	DESWC	0.027832
13	PCNARp	0.015797
46	LSAGN	0.014257
69	SYNLE	0.013994
9	DESWLsyd	0.012545
20	PCDCz	0.012459
79	DRPP	0.012344
2	DESSC	0.011905
66	SMCAUSlsa	0.011265
48	LDTTRc	0.010828

## Analysis of Feature Importance

- **DESWC (Descriptive Word Count):**

Measures the richness of descriptive language in the narrative. Higher values indicate more vivid and detailed descriptions, which improve the model's ability to identify well-developed Initiating Events.

- **PCNARp (Narrative Progression Percentage):**

Quantifies how effectively the narrative progresses through its key stages. It captures coherence and temporal flow, allowing the model to recognize structured storytelling patterns.

- **LSAGN (Lexical Sophistication Aggregate Norm):**

Reflects the overall sophistication and precision of vocabulary. Higher lexical sophistication contributes to clearer distinctions between simple and complex Initiating Events.

- **SYNLE (Syntactic Lexical Complexity):**

Represents the syntactic variety and structural complexity of sentences. It enhances the model's understanding of linguistic organization and grammatical depth.

- **DESWLsyd (Descriptive Word Lexical Diversity):**

Measures the diversity of descriptive word usage, indicating how varied the vocabulary is when describing scenes or events. High diversity supports richer narrative representation.

- **PCDCz (Narrative Cohesion Coefficient):**

Captures the level of narrative cohesion and connectivity between

sentences or ideas. Strong cohesion improves interpretability and event continuity.

- **DRPP (Discourse Referential Pronoun Proportion):**

Represents the frequency and usage of pronouns and discourse markers that maintain narrative continuity. Effective discourse use signals coherent event progression.

- **DESSC (Descriptive Semantic Coherence):**

Assesses the semantic consistency and alignment of descriptive language. Higher coherence indicates a more unified and meaningful event description.

- **SMCAUSIsa (Semantic Causality via LSA):**

Measures causal relationships within the narrative using latent semantic analysis. High causal linkage strength indicates well-connected and logically sequenced events.

- **LDTTRc (Lexical Diversity Type-Token Ratio - Corrected):**

Quantifies the lexical variety of the text while correcting for length effects. Greater lexical diversity reflects linguistic richness and improves differentiation between narrative quality levels.

## Summary and Conclusions

### Model Performance

This project applied machine learning to narrative data to predict the Initiating Event (IE) label. Several algorithms were tested, and the Random Forest model achieved the best performance, with a cross-validated accuracy of approximately 58.6%.

### Feature Importance

The model identified key linguistic and narrative features that most strongly influenced Initiating Event prediction:

- DESWC

- PCNARp
- LSAGN
- SYNLE
- DESWLsyd
- PCDCz
- DRPP
- DESSC
- SMCAUSlsa
- LDTTRc

**These features represent elements of descriptive richness, narrative coherence, syntactic complexity, and semantic structure—factors critical to identifying well-developed Initiating Events.**

## Conclusion

The Random Forest model provides a moderate but meaningful level of predictive accuracy for identifying Initiating Events in narrative text. Feature analysis highlights how linguistic complexity and coherence measures contribute to model performance. Continued experimentation with advanced NLP methods could further enhance accuracy and interpretability!