Artificial Intelligence Capstone Project I

Dataset:

Initially, my plan was to focus on image classification for clothing. However, a mere two days before the deadline, I experienced a sudden change of heart. I was struck by a more innovative and compelling idea: fake news detection. In today's digital landscape, the proliferation of fake news has reached alarming levels, sowing confusion, and misinformation among the masses.

To ensure the efficacy of my project, I opted to source my dataset from Politifact, a highly respected fact-checking website renowned for its thorough examination of news articles and statements from public figures. Leveraging web crawler techniques, I gathered data from Politifact. Besides Politifact, I also collected data from Twitter, because Twitter is a social media that full of fake news. Alongside my classmate, 陳沂亨, we manually annotated labels for each news title (0 for fake news title and 1 for real news title). The essence of my model lies in its ability to discern the authenticity of news titles, swiftly distinguishing between genuine information and falsified reports.

I prepared 4 kinds of datasets to do experiments. They are, biased dataset (20% real news title and 80% fake news title), unbiased dataset (45% real news title and 55% fake news title), unbiased large dataset (contain approximately 2500 news title), dataset in French (Just translate large dataset into French). And I also prepare two testing datasets, one is for English, and another is for French.

Algorithms:

I use two supervised learning algorithms and one unsupervised learning algorithm to process the dataset. The two supervised learning algorithms are Support Vector Machine (SVM) and Convolutional Neural Network (CNN), and the unsupervised learning algorithm, Latent Dirichlet Allocation, and why I used these methods because SVM can handle both linear and non-linear relationships between features and labels, making it suitable for capturing complex patterns in textual data; CNNs are adept at capturing local patterns or features within data. In the context of text classification, the words in a sentence can be considered as a sequence of data. And LDA can be used to extract features from the text data that capture the underlying structure of the documents. These features can then be used as input to other machine learning models for classification tasks.

Analysis:

- Supervised Learning:
  - Support Vector Machine:

    Accuracy: 0.80

    Precision: 0.70

    Recall : 0.86

    F1 Score: 0.81

    AUROC: 0.84

    Confusion Matrix: $\begin{bmatrix} 141 & 132 \\ 23 & 43 \end{bmatrix}$

  - Convolutional Neural Network:

    Accuracy: 0.88

    Precision: 0.90

    Recall : 0.817

    F1 Score: 0.829

    AUROC: 0.90

    Confusion Matrix: $\begin{bmatrix} 161 & 97 \\ 28 & 53 \end{bmatrix}$

  - Conclusion:

    In conclusion, both SVM and CNN models demonstrated strong performance in classifying fake news articles, with SVM achieving an accuracy of 0.8 and CNN achieving an accuracy of 0.88. The SVM model exhibited balanced precision, recall, and F1-scores above 0.77, along with a strong AUROC score of 0.84. CNN, on the other hand, demonstrated robust performance in precision and recall, with F1-scores above 0.8 and an AUROC score of 0.9. These results indicate that both SVM and CNN are effective in distinguishing between fake and real news articles, making them valuable tools for fake news detection tasks.

- Unsupervised Learning:

- Latent Dirichlet Allocation:
  Accuracy: 0.687

	Precision	Recall	F1 Score	Support
0	0.8	0.82	0.81	273
1	0.16	0.14	0.15	66
Accuracy	x	x	0.69	339
Macro Avg.	0.48	0.48	0.48	339
Weighted Avg.	0.67	0.69	0.68	339

- Conclusion:

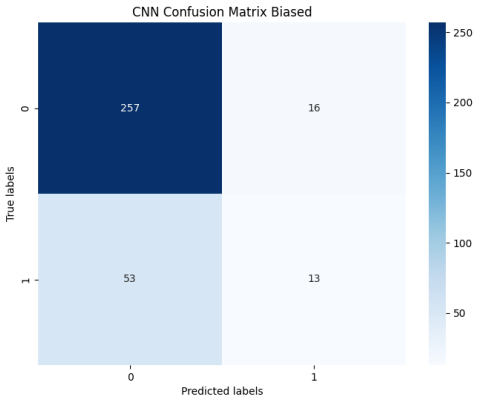
  In unsupervised learning, LDA showed moderate performance in classifying fake news. While LDA achieved an accuracy of 0.687, its precision, recall, and F1-scores were comparatively lower, indicating a less precise classification compared to supervised learning approaches. The model's classification report revealed challenges in effectively identifying fake news articles, with a precision of 0.16 and a recall of 0.14 for the minority class (fake news). Overall, LDA may not be as effective as supervised learning methods like SVM and CNN for fake news detection, suggesting the importance of labeled data for training robust classification models.

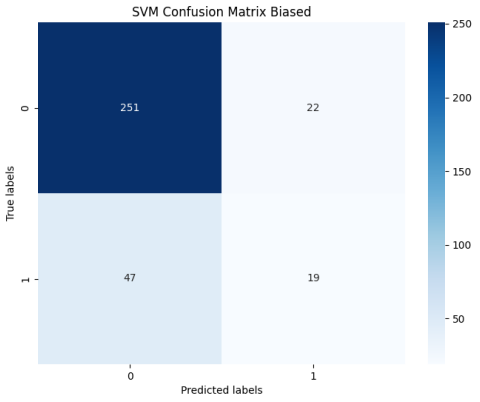
Experiments:

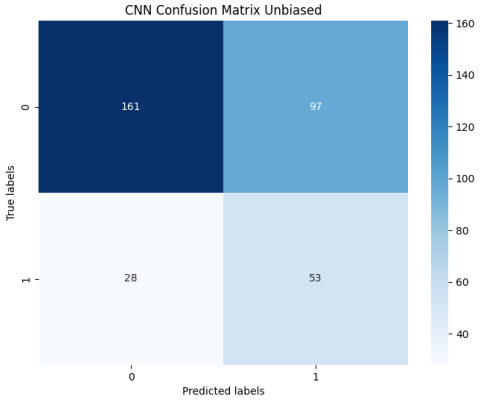
I conducted various experiments on my dataset, which involved biased and unbiased approaches, as well as working with larger datasets and exploring different languages, and I measure it in accuracy.

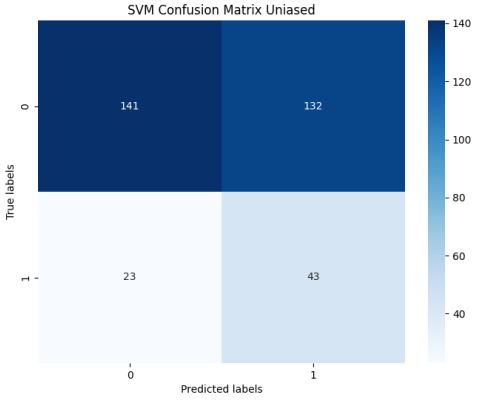
	Biased	Unbiased	Large	French
CNN	0.9	0.92	0.97	0.8
SVM	0.82	0.85	0.9	0.75
LDA	0.28	0.68	0.77	0.42

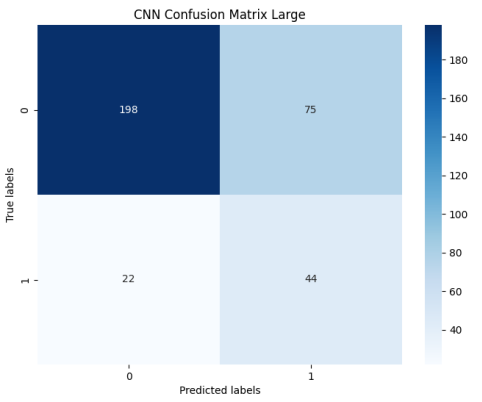
And the confusion matrices are shown below.

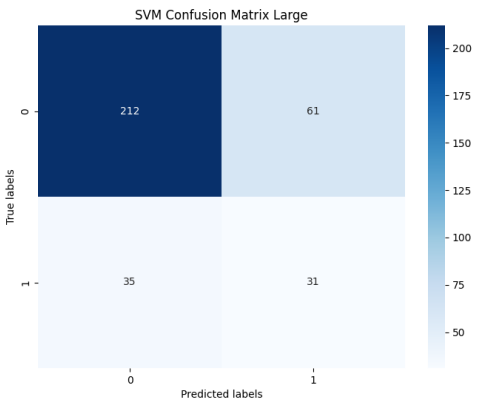


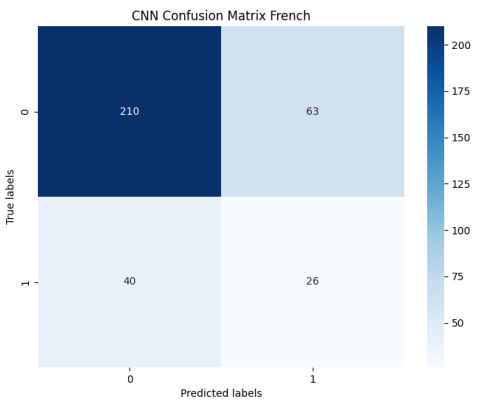


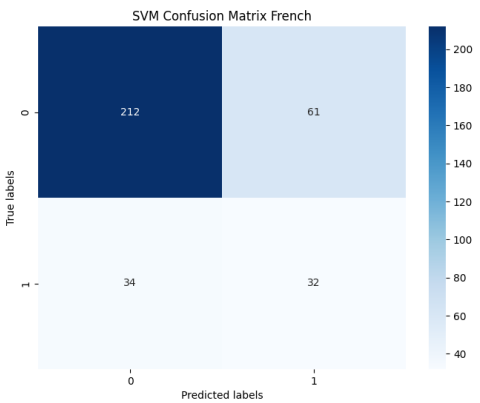












Based on the experiments conducted on various datasets with different biases, sizes, and languages, the following insights can be derived:

- Unbiased datasets consistently outperform biased ones across all models.
- Larger datasets generally lead to better performance.
- French-language datasets tend to perform lower compared to English-language ones.
- CNN outperforms SVM and LDA across all scenarios, indicating its effectiveness in fake news detection tasks.

These observations suggest that unbiased, larger English-language datasets are preferable for training CNN-based models for fake news detection.

Discussion:

- Expectations vs. Results:
  Generally aligned with expectations, especially regarding unbiased and larger datasets yielding better performance. French-language datasets performed comparatively worse, possibly due to vocabulary differences and dataset size limitations.

- Factors Affecting Performance:
  Dataset characteristics, including bias, size, and language, significantly impact model performance. Biased datasets hinder generalization, while language differences affect vocabulary and linguistic nuances, influencing model accuracy.

- Experiments with More Time:
  Explore advanced preprocessing techniques, additional features, advanced architectures (e.g., RNNs, transformers), ensemble learning, and domain adaptation methods, and explore different languages to identify which ones yield the best performance, considering linguistic characteristics and dataset availability.

- Learnings and Remaining Questions:
  Dataset quality is critical for model performance in fake news detection. Language-specific preprocessing is crucial. Remaining questions include investigating more sophisticated feature representations and developing interpretable models for deeper insights.

Dataset Documents:

Link to dataset: Datasets

Part of large dataset:

large

Title	Label
The U.S. Code states that a person convicted of treason would be “taken by the posse to the nearest busy intersection and at high noon hung by the neck until dead.”	0
President Joe Biden was “rushed to hospital unexpectedly.”	0
“We’re having the largest classes of correctional officers we’ve ever had before.”	1
Texas wildfires are a “deliberate” attack on U.S. food supply.	0
The Texas wildfires were started by directed energy weapons.	0
Texas petition to outlaw “aerosolized spraying” is evidence of chemtrails.	0
There are less than two weeks until the deadline to register to vote in Wisconsin.	0
Setting on Apple’s new Journal app “lets anyone near you know your FULL NAME and EXACTLY where you’re geo-located.”	0
Michael Moore is supporting Trump in the 2024 election.	0
If you went “anywhere in the world,” you could get a prescription filled for 40% to 60% less than it costs in the U.S.	1
“Your legal name is in fact a corporation. This is why you always see your name written in ALL CAPS.”	0
In New York City “local elections, illegal immigrants can vote.”	0
The Olympic boxing committee “decided this year that men can box against women in the Olympics.”	0
“Chuck Schumer caught on hot mic” plotting to “blame REPUBLICANS AND MAGA” for reform delays.	0
President Joe Biden’s comments about Texas homes are evidence that homes of a certain color are spared from wildfires.	0
Poland’s foreign minister said the U.S. is “dysfunctional and unreliable” because it has not sent more aid to Ukraine.	0
“7.2 million illegals entered the U.S. under Biden administration,” showing that “the ‘Great Replacement’ is not a theory, it’s a reality.	0
Apple iOS 17.4: iMessage Gets Post-Quantum Encryption in New Update	1
The Morning After: Nintendo’s next console may not arrive until 2025	1
Stephen Dubner is bullish on the podcast industry	1
A top investor in global brands from Airbnb to Snap says we need swift ‘guardrails’ to get countries on the same page about AI	1
RIP Apple Car. This Is Why It Died	1
A New Headset Aims to Treat Alzheimer’s With Light and Sound	1
Nvidia sued over copyright issues in AI platform	1
Own an actual Call of Duty gold bar to celebrate its 20th anniversary, but you’ll need to be quick	1
MWC 2024: all the phones, wearables, and gadgets announced in Barcelona	1
Donald Trump called his wife "Mercedes" instead of Melania.	0
Solar flares caused AT&T’s cellphone outage.	0
Taylor Swift said she “will leave the United States if Donald Trump becomes President in 2024.”	0
“North Carolina has the longest voting period in the country … and we have the most ways of voting.”	1

Part of French dataset:

French

Title	Label
Le Code américain indique qu'une personne reconnue coupable de trahison serait «prise par le groupe jusqu'à l'intersection occupée la plus proche et à midi accroché au cou jusqu'à mort».	0
Le président Joe Biden a été «transporté d'urgence à l'hôpital de façon inattendue».	0
«Nous avons les plus grandes classes d'agents correctionnels que nous ayons jamais eues auparavant.»	1
Les incendies de forêt du Texas sont une attaque «délibérée» contre l'approvisionnement alimentaire aux États-Unis.	0
Les incendies de forêt du Texas ont été lancés par des armes énergétiques dirigées.	0
La pétition du Texas pour interdire la «pulvérisation aérosolisée» est la preuve de Chemtrails.	0
Il y a moins de deux semaines avant la date limite pour s'inscrire pour voter dans le Wisconsin.	0
Définir la nouvelle application de journal d'Apple "Permet à quiconque près de chez vous connaissez votre nom complet et exactement où vous êtes géo-localisé."	0
Michael Moore soutient Trump lors des élections de 2024.	0
Si vous alliez «n'importe où dans le monde», vous pourriez obtenir une ordonnance remplie de 40% à 60% de moins qu'elle coûte aux États-Unis aux États-Unis	1
«Votre nom légal est en fait une société.C'est pourquoi vous voyez toujours votre nom écrit dans toutes les plafonds.	0
À New York, «élections locales, les immigrants illégaux peuvent voter».	0
Le comité de boxe olympique «a décidé cette année que les hommes pouvaient se bloquer contre les femmes aux Jeux olympiques».	0
«Chuck Schumer a pris le micro chaud» comploter pour «blâmer les républicains et Maga» pour des retards de réforme.	0
Les commentaires du président Joe Biden sur les maisons du Texas sont la preuve que les maisons d'une certaine couleur sont épargnées des incendies de forêt.	0
Le ministre polonais des Affaires étrangères a déclaré que les États-Unis étaient "dysfonctionnels et peu fiables" car il n'a pas envoyé plus d'aide à l'Ukraine.	0
"7,2 millions d'illégaux sont entrés aux États-Unis sous l'administration Biden", montrant que "le" grand remplacement "n'est pas une théorie, c'est une réalité.	0
Le poulet et les vaches ont des niveaux élevés d'œstrogènes, ce qui conduit les hommes qui les mangent à chauve, développent des seins et ne peuvent pas cultiver la masse musculaire.	0
«L'une de mes responsabilités en tant que secrétaire d'État est les actes de contre-insignes adoptés par l'Assemblée législative et signé par le gouverneur Evers.»	0
«Les Rothschild étaient à l'origine de la Réserve fédérale, et ils contrôlent désormais le système financier mondial.»	0
Cinquante pour cent des votes pour Joe Biden lors des élections primaires présidentielles du Michigan étaient envoyées par la poste, non vérifiées, pas de signature, pas de pièce d'identité.	0
Ce n'est pas l'ancien président Donald Trump qui a promu le vaccin Covid-19, mais un «sosie».	0
Le gouvernement a fait plus pour arrêter la distribution de l'ivermectine et de l'hydroxychloroquine que pour arrêter la distribution du fentanyl.	0
La Loi sur les puces et les sciences de 2022 «a attiré 640 milliards de dollars d'investissements des entreprises privées».	0
La photo montre l'incendie de Smokehouse Creek au Texas.	0
"L'étude révèle que la fausse viande de Bill Gates provoque des" cancers turbo "chez l'homme."	0
Donald Trump a appelé sa femme "Mercedes" au lieu de Melania.	0
Les poussées solaires ont provoqué une panne de téléphone cellulaire AT&T.	0
Taylor Swift a déclaré qu'elle "quitterait les États-Unis si Donald Trump devient président en 2024."	0
«La Caroline du Nord a la plus longue période de vote du pays… et nous avons la plupart des façons de voter.»	1

Part of testing dataset

testing

Title	Label
President Biden confirm(ed) everyone will receive their $1k gas checks this week!!	0
When Joe Biden “walked into this administration … 20 million people were on unemployment insurance benefits.”	0
Image shows Paul Pelosi, bruised, in a booking mugshot.	0
“Not a single person in the crowd on January 6 was found to be carrying a firearm. Not one.”	0
There are more females on Facebook than alive in the world.	0
“U.S. military at the White House arresting Congress.”	0
“Military take over on May 11 confirmed!”	0
Americans aren’t required to show IDs to vote.	0
“Donald Trump authorized up to 20,000 National Guard troops to protect the Capitol” before Jan. 6, 2021, but was "rejected" by Nancy Pelosi and Chuck Schumer.	0
“Bennie Thompson actively cheer-led riots in the ’90s.”	0
“A quarter of the entire acreage in the country that is under hemp production is here in North Carolina.”	0
“U.S. military arrests Michael Sussmann.”	0
The official pride flag was altered to include Ukrainian colors.	0
A “public health warning” was recently issued for fluoride toothpaste.	0
Gavin Newsom reportedly intervened at the request of Nancy Pelosi and directly ordered the California Highway Patrol to drop all charges” against Paul Pelosi for his DUI arrest.	0
Joe Biden approved a “new card” that “gives free health insurance to Americans” who are 25 and older.	0
Nearly 60% of all student loan debt is held by the rich and upper-middle class,” so forgiveness would give the wealthy a "financial windfall" but not really help low-income people.	1
“Bennie Thompson objected to the 2004 Presidential election.”	1
The Jan. 6, 2021, attack on the U.S. Capitol “was a dust-up.”	0
Baby boomers didn’t have autism, seizures, allergies and other ailments when they were kids.	0
Ray Liotta died because of the COVID-19 vaccine.	0
Says Barack Obama announced Joe Biden’s death.	0
“There are fewer Iowans working today than when Gov. Reynolds took office.”	1
Joe Biden resigned and “Trump is the new president.”	0
On building an NFL stadium in Virginia.	0
White supremacists are going to be “shooting up all Walmarts and will kill Blacks and Mexicans” in San Bernardino, California.	0
“USDA is predicting egg prices will be $12 a dozen by fall 2022.”	0
“Public schools are now as segregated by race and class as they were in the 1960s.”	1

References:

1. https://www.politifact.com
2. https://twitter.com
3. https://nytimes.com
4. https://newsapi.org
5. https://mehtaplustutoring-mlbootcamp20.github.io/Real_vs_Fake_News/
6. https://www.sciencedirect.com/science/article/pii/S2665917424000047
7. https://www.scitepress.org/PublishedPapers/2021/105620/105620.pdf
8. https://github.com/Cartus/Automated-Fact-Checking-Resources
9. https://github.com/ejupialked/fake-news-detection
10. https://www.sciencedirect.com/science/article/pii/S2405959521001375
11. https://arxiv.org/pdf/1806.00749.pdf
12. https://towardsdatascience.com/latent-dirichlet-allocation-lda-9d1cd064ffa2
13. https://medium.com/@corymaklin/latent-dirichlet-allocation-dfcea0b1fddc

Appendix:

CNN:

```python
import pandas as pd
import numpy as np
from sklearn.model_selection import cross_val_predict, cross_val_score, KFold
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score, roc_auc_score, confusion_matrix
import tensorflow as tf
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import Embedding, Conv1D, GlobalMaxPooling1D, Dense
from tensorflow.keras.preprocessing.text import Tokenizer
from tensorflow.keras.preprocessing.sequence import pad_sequences
import matplotlib.pyplot as plt
import seaborn as sns

# Load training dataset from CSV file
train_data = pd.read_csv('large.csv')

# Load testing dataset from CSV file
test_data = pd.read_csv('testing.csv')

# Combine train and test data to fit tokenizer on full dataset
combined_data = pd.concat([train_data['Title'], test_data['Title']], ignore_index=True)

# Tokenize text data and convert to sequences
tokenizer = Tokenizer()
tokenizer.fit_on_texts(combined_data)
X_train_sequences = tokenizer.texts_to_sequences(train_data['Title'])
X_test_sequences = tokenizer.texts_to_sequences(test_data['Title'])

# Pad sequences to ensure uniform length
max_sequence_length = max([len(seq) for seq in X_train_sequences + X_test_sequences])
X_train_padded = pad_sequences(X_train_sequences, maxlen=max_sequence_length)
X_test_padded = pad_sequences(X_test_sequences, maxlen=max_sequence_length)

# Define model architecture
model = Sequential()
model.add(Embedding(input_dim=len(tokenizer.word_index) + 1, output_dim=100, input_shape=(max_sequence_length,)))
model.add(Conv1D(128, 5, activation='relu'))
model.add(GlobalMaxPooling1D())
model.add(Dense(1, activation='sigmoid'))

# Compile model
model.compile(loss='binary_crossentropy', optimizer='adam', metrics=['accuracy'])

# Perform cross-validation on training data
kf = KFold(n_splits=5, shuffle=True, random_state=42)
accuracy_scores = []
precision_scores = []
recall_scores = []
f1_scores = []
auc_scores = []
conf_matrices = []

for train_index, val_index in kf.split(X_train_padded):
    X_train, X_val = X_train_padded[train_index], X_train_padded[val_index]
    y_train, y_val = train_data['Label'].iloc[train_index], train_data['Label'].iloc[val_index]

    # Train model
    model.fit(X_train, y_train, epochs=5, batch_size=64, verbose=0)

    # Predict Labels for validation data
    y_pred = (model.predict(X_val) > 0.5).astype(int)

    # Calculate evaluation metrics
    accuracy_scores.append(accuracy_score(y_val, y_pred))
    precision_scores.append(precision_score(y_val, y_pred))
    recall_scores.append(recall_score(y_val, y_pred))
    f1_scores.append(f1_score(y_val, y_pred))
    auc_scores.append(roc_auc_score(y_val, y_pred))
    conf_matrices.append(confusion_matrix(y_val, y_pred))

# Print average evaluation metrics from cross-validation
print("Average Accuracy:", np.mean(accuracy_scores))
print("Average Precision:", np.mean(precision_scores))
print("Average Recall:", np.mean(recall_scores))
print("Average F1 Score:", np.mean(f1_scores))
print("Average AUROC Score:", np.mean(auc_scores))

# Train model on full training data
history = model.fit(X_train_padded, train_data['Label'], epochs=5, batch_size=64, verbose=0)

# Predict Labels for testing data
y_test_pred = (model.predict(X_test_padded) > 0.5).astype(int)

test_conf_matrix = confusion_matrix(test_data['Label'], y_test_pred)

print("Confusion Matrix:")
print(test_conf_matrix)

plt.figure(figsize=(8, 6))
sns.heatmap(test_conf_matrix, annot=True, cmap='Blues', fmt='g')
plt.xlabel('Predicted labels')
plt.ylabel('True labels')
plt.title('CNN Confusion Matrix Large')
plt.show()
```

SVM:

```python
import pandas as pd
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.svm import SVC
from sklearn.metrics import confusion_matrix, roc_curve, roc_auc_score
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.pipeline import Pipeline
from sklearn.model_selection import cross_val_predict, cross_val_score


train_data = pd.read_csv('large.csv')
test_data = pd.read_csv('testing.csv')

X_train = train_data['Title']
y_train = train_data['Label']

X_test = test_data['Title']
y_test = test_data['Label']

# Define a pipeline for SVM classifier
svm_classifier = Pipeline([
    ('tfidf', TfidfVectorizer()),  # Convert text to TF-IDF features
    ('svm', SVC(kernel='linear', probability=True))  # Support Vector Machine classifier
])

# Use predict_proba for probability estimates
y_pred_cv = cross_val_predict(svm_classifier, X_train, y_train, cv=5, method='predict_proba')[:, 1]

# Calculate evaluation metrics using cross-validation
accuracy_cv = cross_val_score(svm_classifier, X_train, y_train, cv=5, scoring='accuracy').mean()
precision_cv = cross_val_score(svm_classifier, X_train, y_train, cv=5, scoring='precision').mean()
recall_cv = cross_val_score(svm_classifier, X_train, y_train, cv=5, scoring='recall').mean()
f1_cv = cross_val_score(svm_classifier, X_train, y_train, cv=5, scoring='f1').mean()
roc_auc_cv = cross_val_score(svm_classifier, X_train, y_train, cv=5, scoring='roc_auc').mean()

print("Cross-Validation Metrics:")
print("Accuracy:", accuracy_cv)
print("Precision:", precision_cv)
print("Recall:", recall_cv)
print("F1 Score:", f1_cv)
print('ROU-AUC:', roc_auc_cv)

# Fit the classifier on the full training data
svm_classifier.fit(X_train, y_train)

# Use predict_proba for probability estimates
y_pred_probs = svm_classifier.predict_proba(X_test)[:, 1]

fpr, tpr, thresholds = roc_curve(y_test, y_pred_probs)
auc_score = roc_auc_score(y_test, y_pred_probs)

conf_matrix = confusion_matrix(y_test, y_pred_probs.round())
print(conf_matrix)

# Plot confusion matrix
plt.figure(figsize=(8, 6))
sns.heatmap(conf_matrix, annot=True, cmap='Blues', fmt='g')
plt.xlabel('Predicted labels')
plt.ylabel('True labels')
plt.title('SVM Confusion Matrix Large')
plt.show()

```

LDA:

```python
import pandas as pd
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.decomposition import LatentDirichletAllocation
from sklearn.metrics import accuracy_score, classification_report

train_data = pd.read_csv('training_bias_french.csv')
test_data = pd.read_csv('testing_french.csv')

X_train = train_data['Title']
y_train = train_data['Label']

X_test = test_data['Title']
y_test = test_data['Label']

# Preprocess the text data
vectorizer = CountVectorizer()
X_train_counts = vectorizer.fit_transform(X_train)
X_test_counts = vectorizer.transform(X_test)

# Train the LDA model
lda = LatentDirichletAllocation(n_components=2, random_state=42)
X_train_lda = lda.fit_transform(X_train_counts)
X_test_lda = lda.transform(X_test_counts)

# Evaluate the model
# Predict the labels for the testing data
y_pred = [0 if topic[0] > topic[1] else 1 for topic in X_test_lda]

# Compute accuracy
accuracy = accuracy_score(y_test, y_pred)
print("Accuracy:", accuracy)

# Generate classification report
print("Classification Report:")
print(classification_report(y_test, y_pred))
```