Capstone 3 - Project Report

Introduction

WIth the ever expanding news sharing outlets including various apps and emailed newsletters that are not reachable for all the public, there is always going to be an undercurrent of fake news that is dangerous to the world. While parts of the problem are hard to mitigate, this project is an effort to explore the learning models using real and fake news and find ways to separate real news from fake news.

Modelling Methods and Results

The data set is developed by University of Victoria and found to be of good quality. It is prelabeled. The fake news dataset had a few tweets and the real news dataset specifies the label of the fact checking source, Reuters as a confirmation of the cases that it was indeed provided by Reuters.Using those clues the chosen subset is in the categories of US and US political news.It is cleaned and prepared for modelling using a 70/30 train/test split respectively. Few supervised classification methods are applied. Those methods are chosen as their performance on binary classification of text data is considered better and recommended.
The classifiers attempted are - Naive Bayes, Logistic Regression, SGDC (Stochastic Gradient Descent Classifier), SVM (Support Vector Machine classifier). Features are generated using a TF-IDF (unigram) transformer. The results are p[resented below:

1. **Model: Multinomial Naive Bayes**

   Score: 0.8539922730199614

   Classification Report:

	precision	recall	f1-score	support
0	0.98	0.65	0.78	2516
1	0.81	0.99	0.89	3696
accuracy			0.85	6212
macro avg	0.89	0.82	0.84	6212
weighted avg	0.88	0.85	0.85	6212

   Confusion_matrix:
   ```
   [[1644  872]
    [  35 3661]]
   ```

2. **Model: Logistic Regression**

   Score: 0.9634578235672892

   Classification Report:

	precision	recall	f1-score	support
0	0.95	0.96	0.95	2516
1	0.97	0.97	0.97	3696
accuracy			0.96	6212
macro avg	0.96	0.96	0.96	6212
weighted avg	0.96	0.96	0.96	6212

   Confusion_matrix:
   ```
   [[2403  113]
    [ 114 3582]]
   ```

3. **Model: Logistic Regression with SGDC**

Score: 0.8539922730199614

Classification Report:

	precision	recall	f1-score	support
0	0.95	0.96	0.95	2516
1	0.97	0.97	0.97	3696
accuracy			0.96	6212
macro avg	0.96	0.96	0.96	6212
weighted avg	0.96	0.96	0.96	6212

Confusion_matrix:

[[32 2484]
[0 3696]]

4. **Model: SVM (Support Vector Machine)**

Score: 0.9608821635544108

Classification Report:

	precision	recall	f1-score	support
0	0.95	0.95	0.95	2516
1	0.97	0.97	0.97	3696
accuracy			0.96	6212
macro avg	0.96	0.96	0.96	6212
weighted avg	0.96	0.96	0.96	6212

Confusion_matrix:

[[2401 115]
[128 3568]]

Conclusions

The chosen problem of fake news detection is highly nuanced; it is both an NLP problem and finding an efficient algorithm. The models also need to be continuously retrained as the nature of fakeness gets better and comes closer to real news. Though the accuracy scores of NB is not better than the other models, both NB an dSGDC performed better in classifying the fake news better. Perhaps an ensemble model would do better. In the future, deep learning on a larger dataset, testing on a never-before-seen dataset would provide more insights to make the current work even better.