NLP Appendix

Glossary

Term	Explanation
Preprocessing	Applying techniques to normalize text before feeding to a model
Stemming	Gets the base of a word by removing the last few characters (often a prefix or suffix). Sometimes can result in non-words but is computationally efficient and aims for approximation rather than accuracy
Lemmatization	Gets the base dictionary form of a word (lemma) by considering the part of speech. Able to distinguish between verbs and nouns (and others). Generally much more accurate but also more expensive to compute
Feature Extraction	Converting text to structured vectors to be used as a feature in a model
Bag of Words (BOW)	Feature extraction technique. Does not consider the order or structure of words (man eats apple = apple eats man). Each dimension in the vector is a unique word from all the words and the number represents the frequency. Simple but efficient
Term Frequency Inverse Document Frequency	Like BOW but additionally normalized by dividing frequencies with total words in a text. Additionally measures importance by taking the log of documents divided by documents with a certain term. Means that more frequent terms are less informative. While allows for keyword importance rather than treating them all the same, will have a similar problem where bank of river = financial bank
Word Embeddings	Vectors which allow words with similar meanings to have a similar numerical representation. This allows to capture word relationships that may appear (dogs are to cats as cats are to mice). Generally requires large datatsets to train on but pretrained vectors like Twitter 25 exist. Captures semantic relationships but can be expensive to train your own or read from large vectors
Twitter 25	Specific set of word embeddings trained using GLOVE. Each vector (word) has 25 dimensions (relationships). This is small compared to other embeddings and although faster to compute may not capture enough representation to be useful. Depending on what they are trained on also may not generalize well
Support Vector Machines	ML model that separates classes in a feature space by finding the right hyperplane (simple to think of as a line which becomes a plane in higher dimensions or more features) to get the best separation. Aims to maximise the margin between points and the plane to get the most separation (distance based). Effective with many dimensions but can require tuning

Random Forests	ML model using ensemble ie creates multiple decision trees. Randomly splits based on random features (hence the need for multiple trees). Then takes the most frequent prediction from the trees as the output. Can be accurate due to ensemble but in sparse feature space the trees may not have enough information to split on well and can become computationally expensive to tune
Multilayer Perceptron	Neural Network ML model that has an input layer, hidden layer (at least 1) and an output layer. All layers are connected through nodes and all nodes are connected. Non linear transformations occur in the hidden layers and before the output layer to learn non linear patterns. Done through back propagation. Weights for each node are calculated going forward in the network and then upon getting the output are passed backward to update the weights. This process continues until a certain threshold is met. Can be flexible and fit any kind of data but intensive to compute and tune and susceptible to not getting the right weights
DistilBERT	Transformer model that is smaller than BERT. Approximates BERTs learning boundaries. Faster but might lose complexity.
Transformer	Type of neural net that can handle sequence but processed simultaneously rather than sequentially. Significance of words is weighed across the whole sentence rather than in order. Can be extremely resource intensive to train but achieves top results
Unigram	A single word taken as a feature eg bird
Bigram	A pair of words taken as a feature eg birds fly
Trigram	Three words taken as a feature eg birds fly south
N gram	Extends beyond uni, bi and tri taking as many words together. Can lead to data sparsity problems and large dimensionality
Hyperparameter	Settings set before the model trains to control the behaviour of the model
Singular Value Decomposition (SVD)	Mathematical technique used to reduce a matrix down to important features. Can be computationally expensive and will not work in a live setting. But can be updated in the background and makes the models faster to train on given the dimensionality reduction
Part of Speech	Tags words as to if they are nouns, verbs etc

Intermediate Results and Notes

1. Finetuning models was left as personally this was done multiple times in other courses on the degree
    a. I wanted to focus on learning about the different kinds of text preprocessing in detail instead as that was new for me

2. Results for SVD 100 and 1000. 100 shows very bad results, likely a loss of a lot of information. 1000 is comparable to 2000 in the main report. These values were picked based on the paper mentioned in the report that also employed SVD

combined_shrunk_results_df

Average Accuracy	Pre-Processing	SVM	RF	MLP	Feature Count
0	X_lemma_bow_uni_shrunk	0.644642	0.585346	0.651208	100
0	X_lemma_bow_bi_shrunk	0.535119	0.545956	0.529493	100
0	X_lemma_bow_tri_shrunk	0.509276	0.501458	0.502709	100
0	X_lemma_tfidf_uni_shrunk	0.676843	0.595978	0.668298	100
0	X_lemma_tfidf_bi_shrunk	0.559296	0.542830	0.544290	100
0	X_lemma_tfidf_tri_shrunk	0.521467	0.506775	0.510318	100
0	X_stem_bow_uni_shrunk	0.656001	0.589517	0.660691	100
0	X_stem_bow_bi_shrunk	0.539601	0.536473	0.538455	100
0	X_stem_bow_tri_shrunk	0.510317	0.507504	0.502918	100
0	X_stem_tfidf_uni_shrunk	0.689765	0.601918	0.676217	100
0	X_stem_tfidf_bi_shrunk	0.552523	0.542415	0.539600	100
0	X_stem_tfidf_tri_shrunk	0.523759	0.505315	0.514171	100

combined_shrunk_results_df

Average Accuracy	Pre-Processing	SVM	RF	MLP	Feature Count
0	X_lemma_bow_uni_shrunk	0.728741	0.560754	0.724676	1000
0	X_lemma_bow_bi_shrunk	0.572010	0.540955	0.565443	1000
0	X_lemma_bow_tri_shrunk	0.509796	0.507089	0.502188	1000
0	X_lemma_tfidf_uni_shrunk	0.743122	0.566382	0.732699	1000
0	X_lemma_tfidf_bi_shrunk	0.593476	0.537410	0.572324	1000
0	X_lemma_tfidf_tri_shrunk	0.513650	0.501771	0.509899	1000
0	X_stem_bow_uni_shrunk	0.734889	0.553773	0.730305	1000
0	X_stem_bow_bi_shrunk	0.571696	0.540956	0.567736	1000
0	X_stem_bow_tri_shrunk	0.512611	0.508442	0.504690	1000
0	X_stem_tfidf_uni_shrunk	0.747812	0.572218	0.741976	1000
0	X_stem_tfidf_bi_shrunk	0.596601	0.541790	0.575134	1000
0	X_stem_tfidf_tri_shrunk	0.514485	0.518862	0.504375	1000

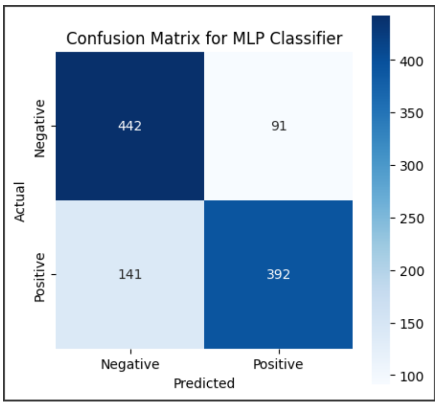
3. 1 iteration model tuning
   a. Tuning did not increase the score
   b. Grid search or random search could have been used over different hyperparameters to improve results

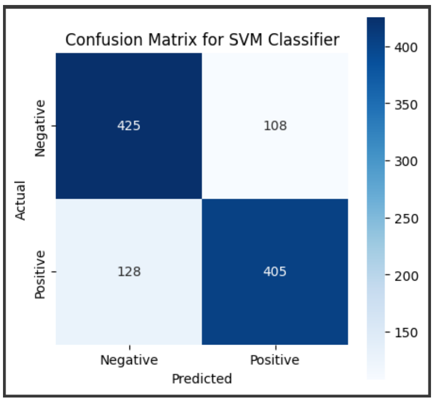
```
svm_classifier_complex = SVC(kernel='rbf',
                    C=10,
                    gamma='scale')

rf_classifier_complex = RandomForestClassifier(n_estimators=100,
                                               max_depth=None,
                                               min_samples_split=2,
                                               random_state=49)

mlp_classifier_complex = MLPClassifier(hidden_layer_sizes=(100, 50, 25),
                                       activation='relu',
                                       solver='adam',
                                       alpha=0.001,
                                       learning_rate_init=0.001,
                                       max_iter=200,
                                       early_stopping=True,
                                       n_iter_no_change=5,
                                       tol=0.001,
                                       random_state=15)
```

results_df_complex_stem_tfidf_uni_shrunk

Average Accuracy	Pre-Processing	SVM Complex	RF Complex	MLP Complex	Feature Count
0	X_stem_tfidf_uni_shrunk	0.74729	0.625051	0.738224	2000

4. Test confusion matrices





5. Models had to be kept simple as the computation was running to about 40 minutes per preprocessing combination on during my training experiments. Kept simple to save time but acknowledged as a limitation
6. Base model training scores using just unigram BOW – for third baseline

	Pre-Processing	SVM	RF	MLP	Feature Count
Average Accuracy					
0	X_bow_uni	0.740932	0.670488	0.752082	18683

7. Error analysis dataframes. Only showing 2 and the first 10 rows.
    a. Testing notebook has all the frames created and can be run easily if needed
        i. Label 0 SVM 1 MLP 1
        ii. Label 0 SVM 0 MLP 1
        iii. Label 0 SVM 1 MLP 0
        iv. Label 0 SVM 0 MLP 0
        v. Label 1 SVM 1 MLP 1
        vi. Label 1 SVM 0 MLP 0
        vii. Label 1 SVM 1 MLP 0
        viii. Label 1 SVM 0 MLP 1

df_label0_svm1_mlp1

	text	preprocessed_text	text_preprocessed_stem	label	svm_predicted	mlp_predicted
544	a film without surprise geared toward maximum comfort and familiarity .	film without surprise geared toward maximum comfort familiarity	film without surpris gear toward maximum comfort familiar	0	1	1
545	fessenden continues to do interesting work , and it would be nice to see what he could make with a decent budget . but the problem with wendigo , for all its effective moments , isn't really one of resources .	fessenden continues interesting work would nice see could make decent budget problem wendigo effective moments really one resources	fessenden continu interest work would nice see could make decent budget problem wendigo effect moment realli one resourc	0	1	1
546	spirit is a visual treat , and it takes chances that are bold by studio standards , but it lacks a strong narrative .	spirit visual treat takes chances bold studio standards lacks strong narrative	spirit visual treat take chanc bold studio standard lack strong narr	0	1	1
548	this is a children's film in the truest sense . it's packed with adventure and a worthwhile environmental message , so it's great for the kids . parents , on the other hand , will be ahead of the plot at all times , and there isn't enough clever innuendo to fil	children film truest sense packed adventure worthwhile environmental message great kids parents hand ahead plot times enough clever innuendo fil	children film truest sens pack adventur worthwhil environment messag great kid parent hand ahead plot time enough clever innuendo fil	0	1	1
554	what ensues are much blood-splattering , mass drug-induced bowel evacuations , and none-too-funny commentary on the cultural distinctions between americans and brits .	ensues much bloodsplattering mass druginduced bowel evacuations nonetoofunny commentary cultural distinctions americans brits	ensu much bloodsplatt mass druginduc bowel evacu nonetoofunni commentari cultur distinct american brit	0	1	1
559	sandra bullock and hugh grant make a great team , but this predictable romantic comedy should get a pink slip .	sandra bullock hugh grant make great team predictable romantic comedy get pink slip	sandra bullock hugh grant make great team predict romant comedi get pink slip	0	1	1
580	a wannabe comedy of manners about a brainy prep-school kid with a mrs . robinson complex founders on its own preciousness -- and squanders its beautiful women .	wannabe comedy manners brainy prepschool kid mrs robinson complex founders preciousness squanders beautiful women	wannab comedi manner braini prepschool kid mr robinson complex founder precious squander beauti women	0	1	1
587	the connected stories of breitbart and hanussen are actually fascinating , but the filmmaking in invincible is such that the movie does not do them justice .	connected stories breitbart hanussen actually fascinating filmmaking invincible movie justice	connect stori breitbart hanussen actual fascin filmmak invinc movi justic	0	1	1
588	a depressingly retrograde , 'post-feminist' romantic comedy that takes an astonishingly condescending attitude toward women .	depressingly retrograde postfeminist romantic comedy takes astonishingly condescending attitude toward women	depressingli retrograd postfeminist romant comedi take astonishingli condescend attitud toward women	0	1	1
589	return to never land is much more p . c . than the original version (no more racist portraits of indians , for instance) , but the excitement is missing .	return never land much p c original version racist portraits indians instance excitement missing	return never land much p c origin version racist portrait indian instanc excit miss	0	1	1

df_label1_svm0_mlp0

	text	preprocessed_text	text_preprocessed_stem	label	svm_predicted	mlp_predicted
4	red dragon " never cuts corners .	red dragon never cuts corners	red dragon never cut corner	1	0	0
7	weighty and ponderous but every bit as filling as the treat of the title .	weighty ponderous every bit filling treat title	weighti ponder everi bit fill treat titl	1	0	0
9	generates an enormous feeling of empathy for its characters .	generates enormous feeling empathy characters	gener enorm feel empathi charact	1	0	0
12	mostly , [goldbacher] just lets her complicated characters be unruly , confusing and , through it all , human .	mostly goldbacher let us complicated characters unruly confusing human	mostli goldbach let us complic charact unruli confus human	1	0	0
18	as it turns out , you can go home again .	turns go home	turn go home	1	0	0
19	you've already seen city by the sea under a variety of titles , but it's worth yet another visit .	already seen city sea variety titles worth yet another visit	alreadi seen citi sea varieti titl worth yet anoth visit	1	0	0
22	grown-up quibbles are beside the point here . the little girls understand , and mccracken knows that's all that matters .	grownup quibbles beside point little girls understand mccracken knows matters	grownup quibbl besid point littl girl understand mccracken know matter	1	0	0
28	devotees of star trek ii : the wrath of khan will feel a nagging sense of deja vu , and the grandeur of the best next generation episodes is lacking .	devotees star trek ii wrath khan feel nagging sense deja vu grandeur best next generation episodes lacking	devote star trek ii wrath khan feel nag sens deja vu grandeur best next gener episod lack	1	0	0
30	what's so striking about jolie's performance is that she never lets her character become a caricature -- not even with that radioactive hair .	striking jolie performance never let us character become caricature even radioactive hair	strike joli perform never let us charact becom caricatur even radioact hair	1	0	0
31	the main story . . . is compelling enough , but it's difficult to shrug off the annoyance of that chatty fish .	main story compelling enough difficult shrug annoyance chatty fish	main stori compel enough difficult shrug annoy chatti fish	1	0	0