Krithik Duraisamy
LING 471

Assignment 5 Writeup

PART 1
Positive Datasets

Review Processing	20 Most Frequently used Words with Frequency in Training Set
Original POS reviews	the 148405 and 84271 a 79424 of 75339 to 65208 is 55356 in 45793 that 31942 I 28725 it 26981 this 25958 /><br 24617 as 23928 with 22031 was 21309 for 20867 but 16456 his 16199 The 15953 on 15385
Cleaned POS reviews	the 149984 and 86459 a 80098 of 75615 to 65926 is 56712 in 46950 it 37826 I 35603 that 34375 s 33601 this 27190 as 24495 with 22557 The 21780 was 21778 for 21390 film 20395 movie 18508 but 17102
Lowercased POS reviews	the 172492 and 89697 a 83247 of 76543 to 66681 is 57189 in 49970

	it	47291
	i	38235
	that	35575
	s	34031
	this	33183
	as	26153
	with	23207
	for	22310
	was	21900
	but	20818
	film	20661
	movie	18757
	his	17225
No stopwords POS reviews	film	20335
	movie	18161
	one	13177
	like	8879
	good	7552
	story	6672
	time	6352
	great	6262
	well	6258
	see	5838
	also	5536
	would	5351
	really	5308
	even	4935
	much	4617
	first	4434
	people	4419
	get	4242
	best	4220
	love	4138
Lemmatized POS reviews	film	24624
	movi	21666
	one	13682
	like	10258
	time	8309
	good	7670
	stori	7362
	see	7229
	charact	7056
	make	6935
	well	6554
	get	6437
	great	6434
	watch	6171
	love	5921
	also	5536
	show	5512
	would	5351
	realli	5308
	even	5088

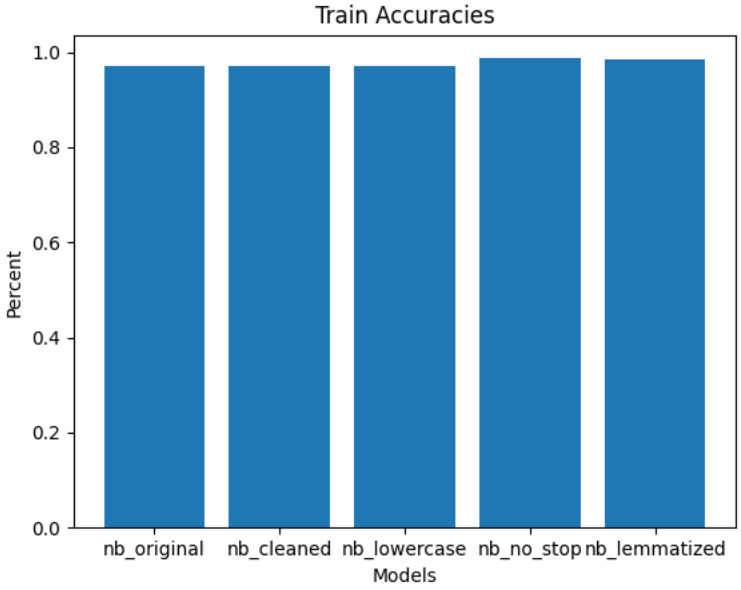
Negative Dataset

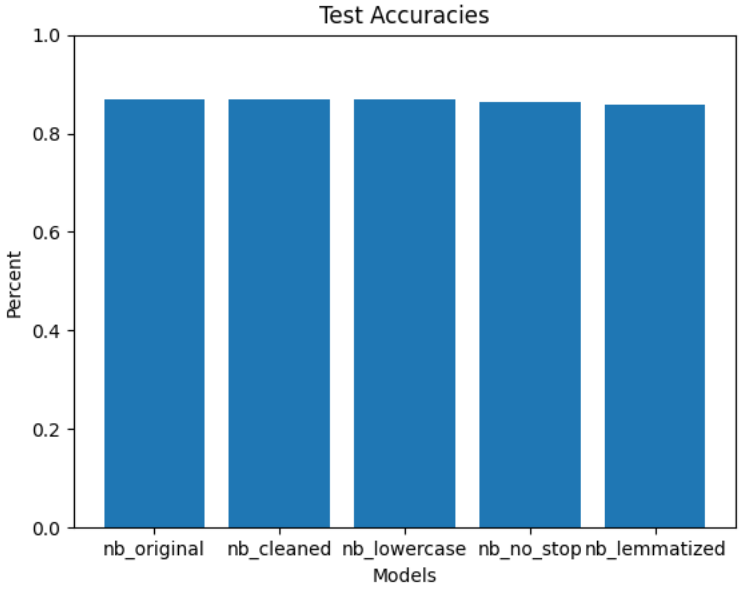
Review Processing	20 Most Frequently Used Words with Frequency in Training Set
Original NEG reviews	the 138593 a 75664 and 68380 of 67631 to 67356 is 47870 in 39781 I 32795 that 32615 this 31102 it 27433 /> 26319 was 25387 for 20195 with 19686 as 18576 but 17328 movie 17119 The 16347 on 15380
Cleaned NEG reviews	the 140120 a 76294 and 70490 to 68017 of 67987 is 49440 I 41460 in 40831 it 38993 that 35872 this 33226 s 31589 was 26027 movie 24140 The 21662 for 20765 with 20336 t 20160 as 19078 film 18778
Lowercased NEG reviews	the 162488 a 79022 and 74337 to 68904 of 68747 is 50006 it 47670 i 44253 in 43586 this 38947 that 37535 s 31984 was 26259 movie 24536

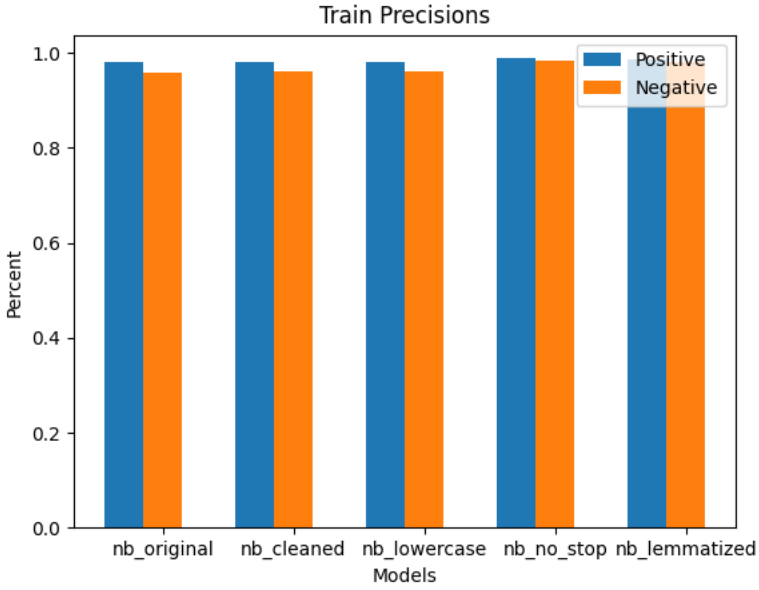
	for 21836 but 21759 with 20820 t 20549 as 20486 film 18964
No stopwords NEG reviews	movie 23773 film 18580 one 12671 like 11089 even 7629 good 7306 bad 7244 would 6989 really 6087 time 5971 see 5360 story 5133 much 5003 get 4980 people 4768 make 4694 could 4638 made 4478 plot 4095 well 4080
Lemmatized NEG reviews	movi 27772 film 22201 one 13102 like 12170 make 8194 bad 7838 even 7735 get 7585 time 7446 good 7405 charact 7081 watch 7036 would 6989 see 6491 realli 6087 look 5804 stori 5616 scene 5566 act 5270 much 5004

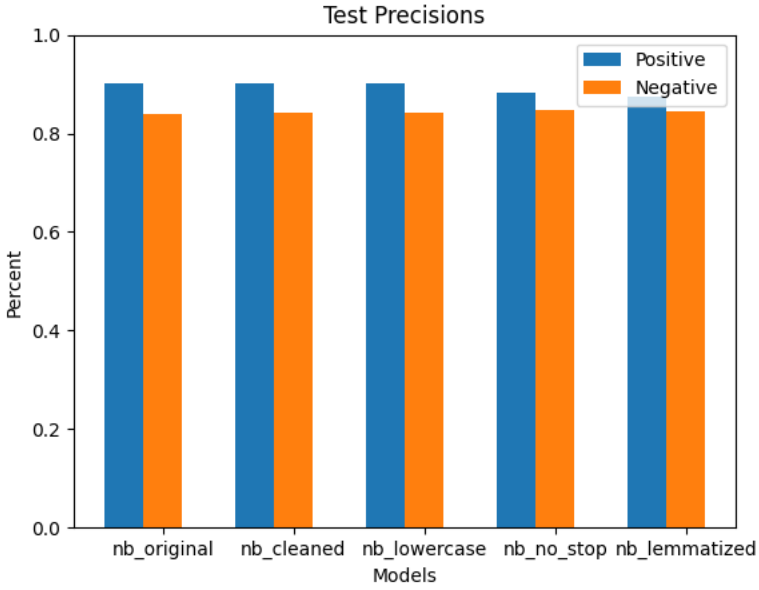
The lemmatizer did a fairly good job with much of the top 20 words matching with other rankings. Few of the words on the lists are not real English words, however, this should not significantly affect the output. Since the lemmatizer removes prefixes and suffixes, the corresponding list should have fewer words. “Movies” and “Movie” are placed together which should increase the accuracy.

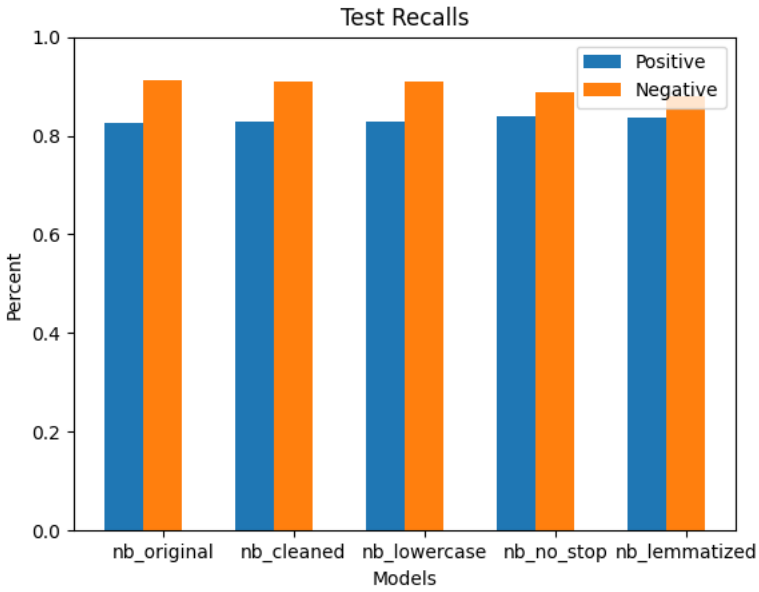
PART 2


Train Accuracies
1.0
0.8
0.6
0.4
0.2
0.0
Percent
nb_original
nb_cleaned
nb_lowercase
nb_no_stop
nb_lemmatized
Models


Test Accuracies
1.0
0.8
0.6
0.4
0.2
0.0
Percent
nb_original
nb_cleaned
nb_lowercase
nb_no_stop
nb_lemmatized
Models


Train Precisions
Positive
Negative
Percent
1.0
0.8
0.6
0.4
0.2
0.0
nb_original
nb_cleaned
nb_lowercase
nb_no_stop
nb_lemmatized
Models


Test Precisions
Positive
Negative
Percent
1.0
0.8
0.6
0.4
0.2
0.0
nb_original
nb_cleaned
nb_lowercase
nb_no_stop
nb_lemmatized
Models



The graphs indicate that the accuracy across all Naïve Bayes models were roughly the same on the test data sets at around 80%. Precision and recall for all models also were around 80%. Precision for the positive movie reviews were slightly higher than the negatives. Recall for the negative revies were slightly higher than the positives across all models. The difference between precision and recall may have to do with people tending to write harsher negative review than they would for positive review. The selection bias may attribute to the differences with harsher negative review than praising positive ones.