Remy Zachwieja remyz
LING471 SP 2021 Assignment5

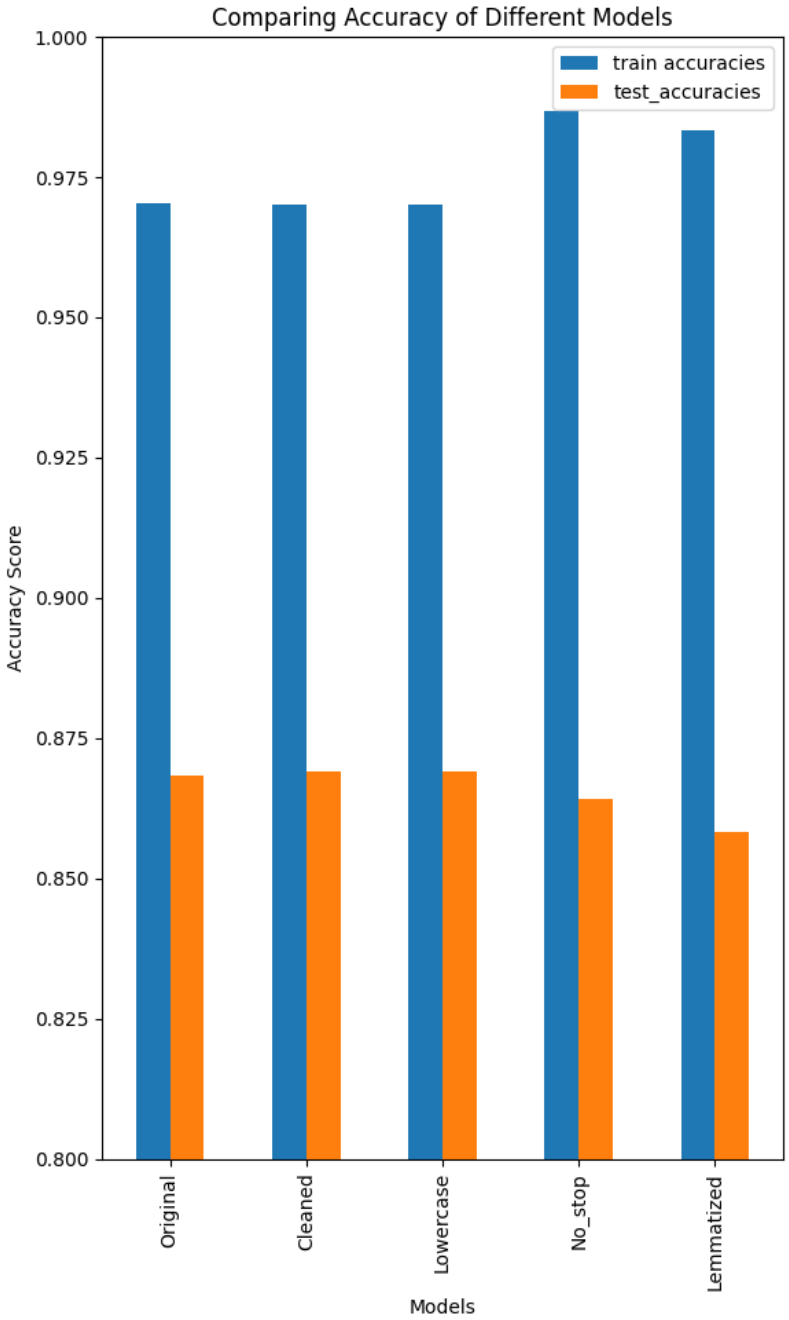
20 Most Common Words in Train Data (POSTIVE)

Original		Cleaned		Lowercased		No Stopwords		Lemmatized	
the	148404	the	149984	the	172492	film	20335	film	24624
and	84265	and	86459	and	89697	movie	18160	movi	21666
a	79423	a	80098	a	83247	one	13177	one	13682
of	75339	of	75615	of	76543	like	8879	like	10258
to	65208	to	65926	to	66681	good	7552	time	8309
is	55356	is	56712	is	57189	story	6672	good	7670
in	45791	in	46950	in	49970	time	6352	stori	7362
that	31939	it	37826	it	47291	great	6262	see	7229
I	28723	I	35603	i	38235	well	6258	charact	7056
it	26976	that	34375	that	35575	see	5838	make	6935
this	25957	s	33601	s	34031	also	5536	well	6554
/><br	24618	this	27190	this	33183	would	5351	get	6437
as	23928	as	24495	as	26153	really	5308	great	6434
with	22030	with	22557	with	23207	even	4935	watch	6171
was	21308	The	21780	for	22309	much	4617	love	5921
for	20866	was	21778	was	21900	first	4434	also	5536
but	16452	for	21390	but	20818	people	4419	show	5512
his	16199	film	20395	film	20661	get	4242	would	5351
The	15943	movie	18508	movie	18757	best	4220	realli	5308
on	15385	but	17102	his	17225	love	4138	even	5088

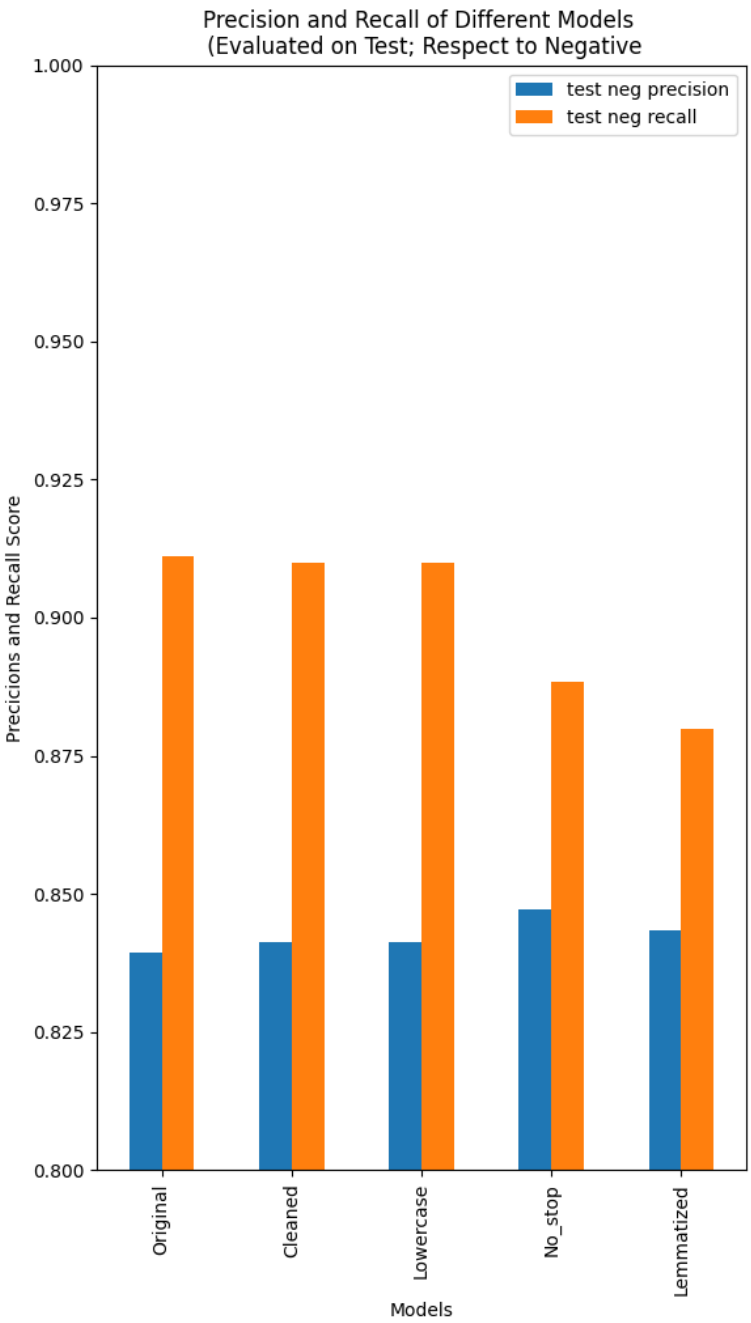
20 Most Common Words in Train Data (NEGATIVE)

Original		Cleaned		Lowercased		No Stopwords		Lemmatized	
the	138587	the	140120	the	162488	movie	23773	movi	27772
a	75661	a	76294	a	79022	film	18580	film	22201
and	68373	and	70490	and	74337	one	12671	one	13102
of	67629	to	68017	to	68904	like	11089	like	12170
to	67356	of	67987	of	68747	even	7629	make	8194
is	47869	is	49440	is	50006	good	7306	bad	7838
in	39779	I	41460	it	47670	bad	7244	even	7735
I	32791	in	40831	i	44253	would	6989	get	7585
that	32613	it	38993	in	43586	really	6087	time	7446
this	31102	that	35872	this	38947	time	5971	good	7405
it	27425	this	33226	that	37535	see	5360	charact	7081
/><br	26319	s	31589	s	31984	story	5134	watch	7036
was	25387	was	26027	was	26259	much	5003	would	6989
for	20193	movie	24140	movie	24536	get	4980	see	6491

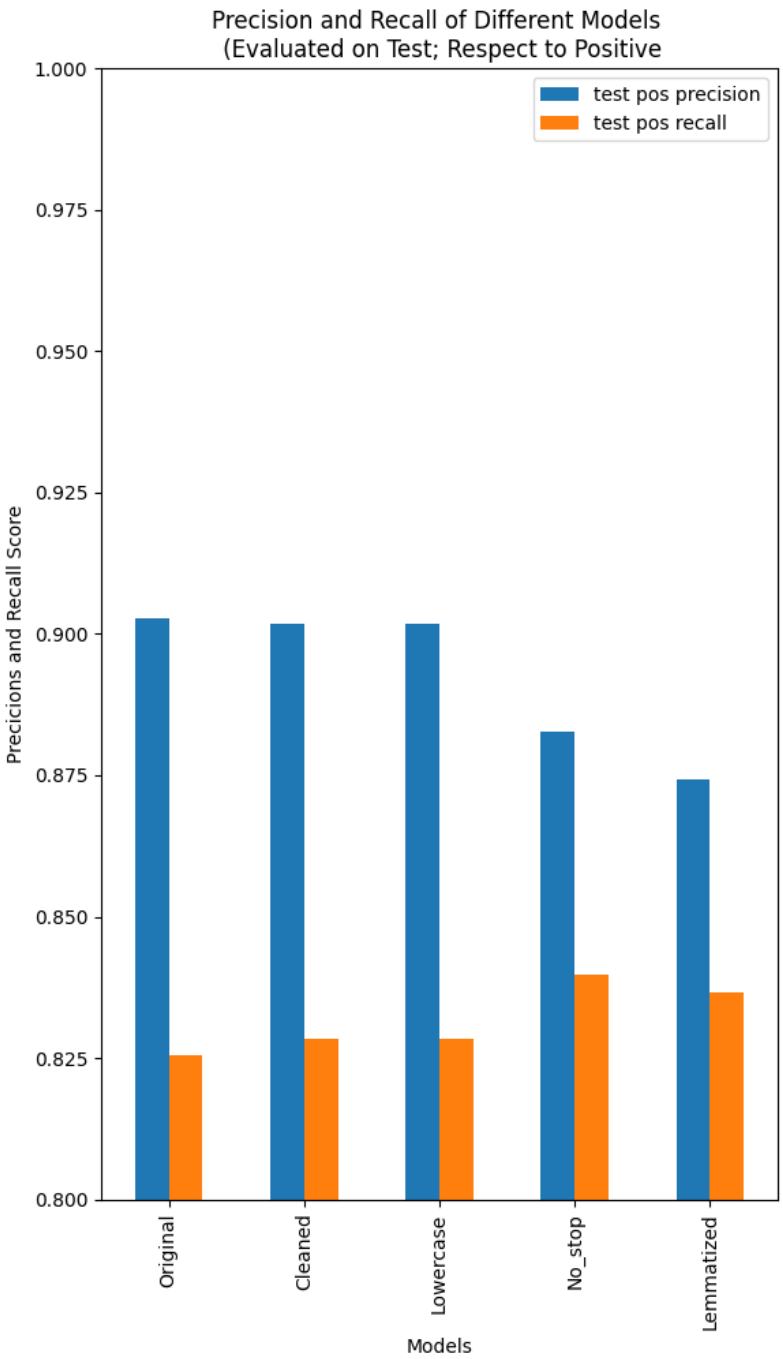
with	19684	The	21662	for	21836	people	4768	realli	6087
as	18575	for	20765	but	21759	make	4694	look	5804
but	17325	with	20336	with	20820	could	4638	stori	5616
movie	17118	t	20160	t	20549	made	4478	scene	5566
The	16338	as	19078	as	20486	plot	4095	act	5270
on	15379	film	18778	film	18964	well	4080	much	5004



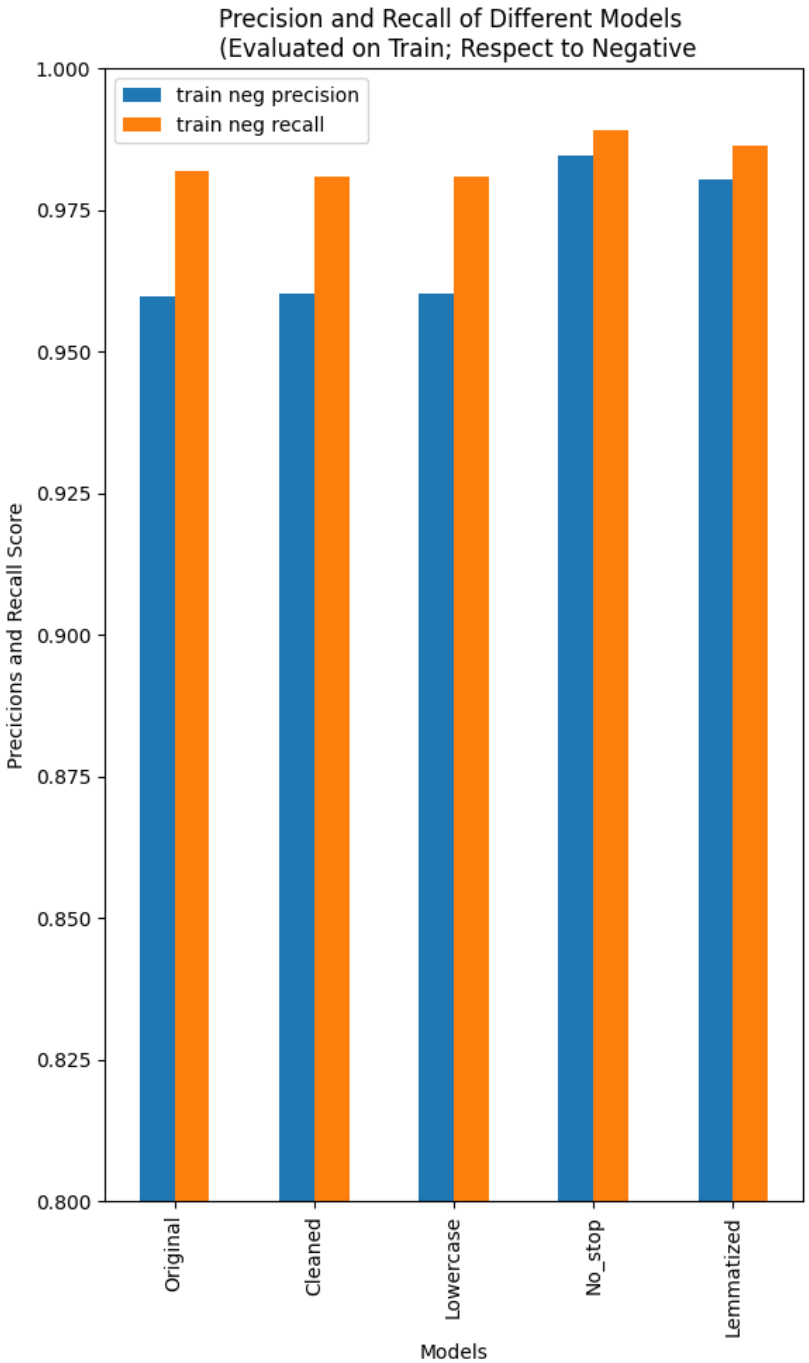
This graph plots the **accuracy** of all the models evaluated on **both** the **test** and **train** sets



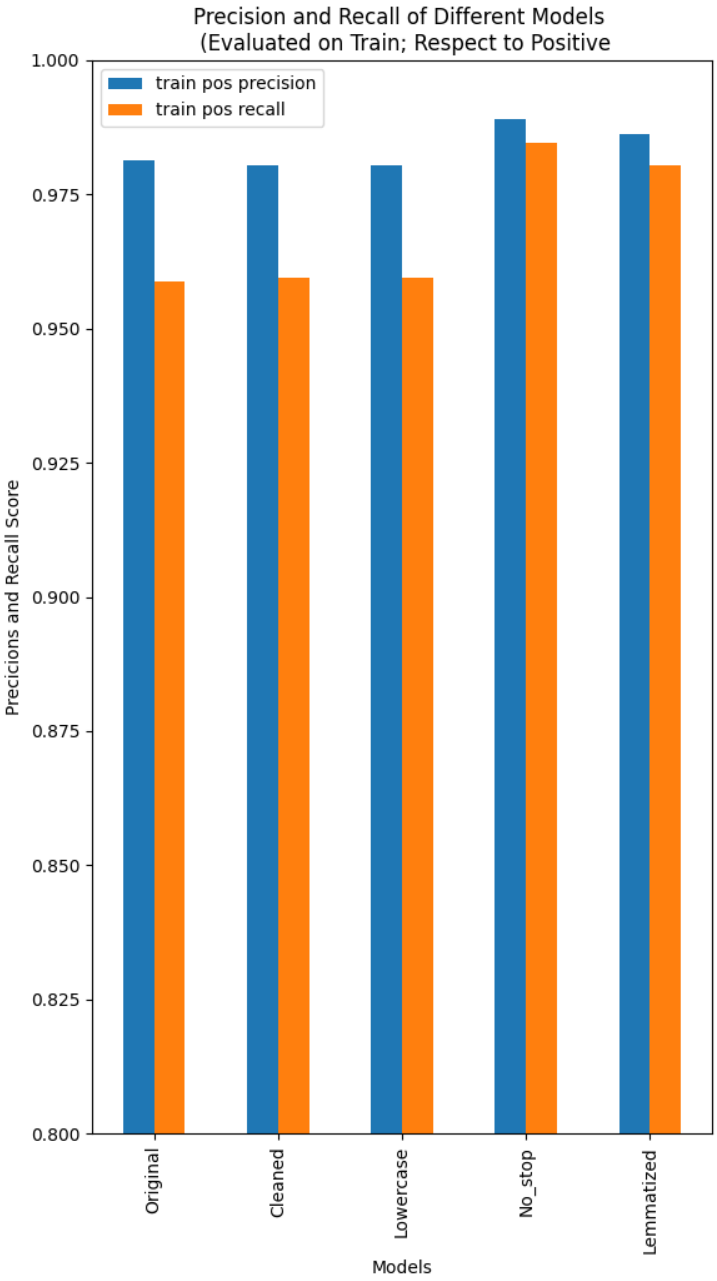
This graph plots the **precision** and **recall** for all models when evaluated on the **test** set, with respect to **negative**.



This graph plots the **precision** and **recall** for all models when evaluated on the **test** set, with respect to **positive**.



This graph plots the **precision** and **recall** for all models when evaluated on the **train** set, with respect to **negative**.



This graph plots the **precision** and **recall** for all models when evaluated on the **train** set, with respect to **positive**.

1. Being a morphologically simple language, I've noticed that the lemmatizer tends to take off and remove suffixes on a lot of words that don't have any suffix on them. This may limit the reliability for embeddings for those specific words if they are incorrectly transformed into other words.

2. Given that the IMDB dataset is so large, small things like /><br tag or other tags are present in the text. You might not think this would create a large difference in the model, but with such a large dataset we can see how much the cleaning model changes our numbers. Especially looking at some of the most commonly occurring words, there are large differences in values of a few words there.

3. Looking through some of the reviews, most of them are written fairly casually, and proper capitalization is sometimes overlooked, things like 'I' or other words not being capitalized changes values for certain words. This may not have a huge effect, and in the graphs it is apparent that the cleaned and lowercase models are usually quite similar, but it does create slight variation.

4. Although the Linguistic operations didn't give us results that were incredibly amazing, it was still interesting to see how small changes on how the words are represented as tokens and such changes the values