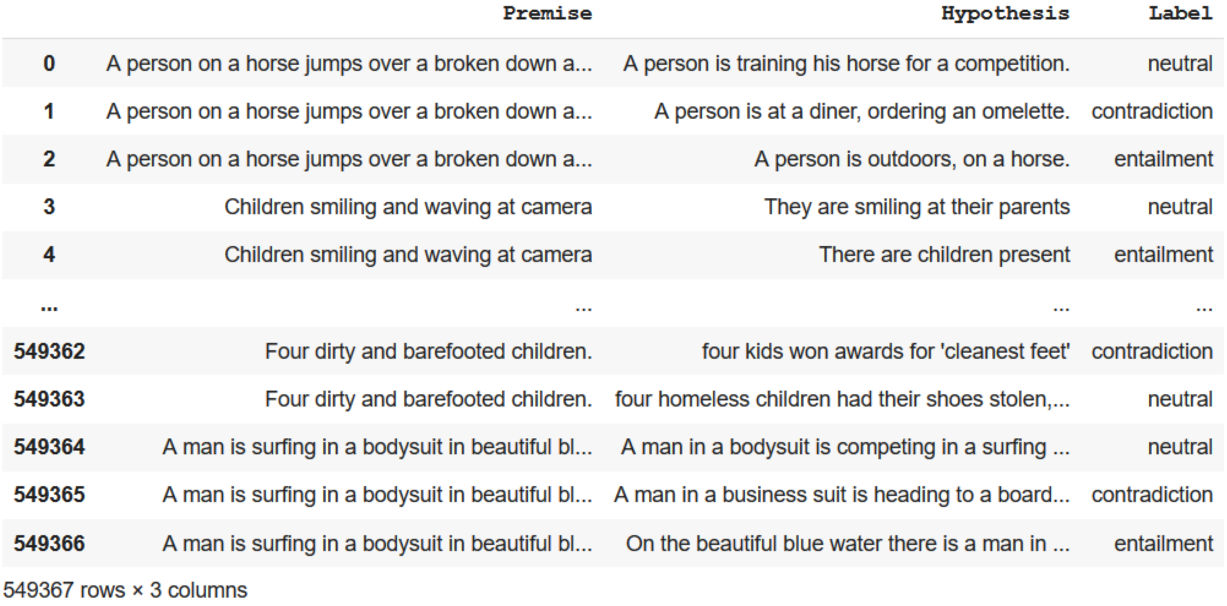
Siddharth Anil
5th year UG

DeepLearning Assignment 3 NLP

Dataset Cleaning: Here we are using SNLI dataset. Before vectorizing we need to clean the text data. We did the following:

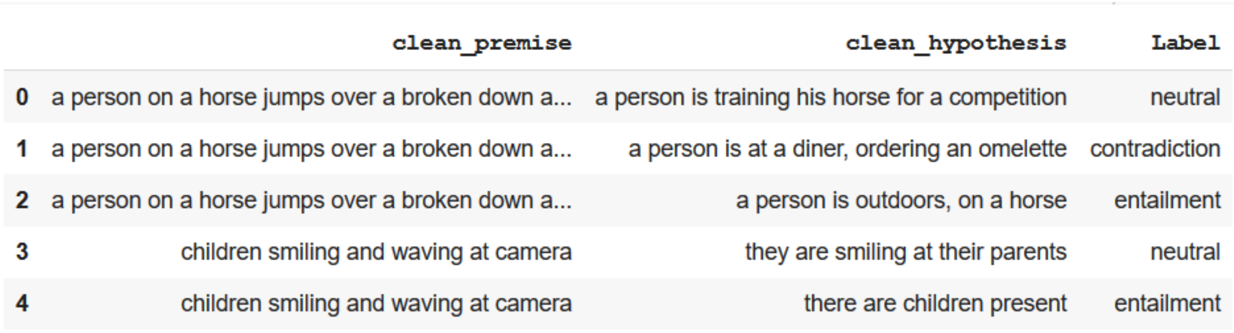
1. Convert to smaller case {*str.lower()*}
2. Remove Punctuation (like period and commas){using *str.replace*()}
3. Remove Special Characters and Numbers (which won't affect the overall meaning of hypothesis and premise) { using **re** module}
4. We also removed **stop words** which are the common generic words found in English language.(using the option in TFIDF vectorizes)



	Premise	Hypothesis	Label
0	A person on a horse jumps over a broken down a...	A person is training his horse for a competition.	neutral
1	A person on a horse jumps over a broken down a...	A person is at a diner, ordering an omelette.	contradiction
2	A person on a horse jumps over a broken down a...	A person is outdoors, on a horse.	entailment
3	Children smiling and waving at camera	They are smiling at their parents	neutral
4	Children smiling and waving at camera	There are children present	entailment
...	...	...	...
549362	Four dirty and barefooted children.	four kids won awards for 'cleanest feet'	contradiction
549363	Four dirty and barefooted children.	four homeless children had their shoes stolen,...	neutral
549364	A man is surfing in a bodysuit in beautiful bl...	A man in a bodysuit is competing in a surfing ...	neutral
549365	A man is surfing in a bodysuit in beautiful bl...	A man in a business suit is heading to a board...	contradiction
549366	A man is surfing in a bodysuit in beautiful bl...	On the beautiful blue water there is a man in ...	entailment

549367 rows × 3 columns

Fig1.SNLI Data Set before cleaning



	clean_premise	clean_hypothesis	Label
0	a person on a horse jumps over a broken down a...	a person is training his horse for a competition	neutral
1	a person on a horse jumps over a broken down a...	a person is at a diner, ordering an omelette	contradiction
2	a person on a horse jumps over a broken down a...	a person is outdoors, on a horse	entailment
3	children smiling and waving at camera	they are smiling at their parents	neutral
4	children smiling and waving at camera	there are children present	entailment

Fig1.SNLI Dataset after cleaning

Siddharth Anil
5th year UG

TF-IDF Vectorization : Here I used inbuilt tfidf vectorizer in sklearn module. It generated **33241 dimension** vector space. Since it was space consuming, I used sparse representation of our vectorized data.

Logistic Regression:

I gave the labels numbers as follows

```
'entailment':0,
'neutral':1,
'contradiction':2
'-':3
```

Passed the sparse vector matrix to logistic regression function as X and above numbers as Y.

It was not converging with default settings. I had to change the iteration number to 2000 from 100.

Results

With the above model, I got an accuracy of 63%. This is much better than random guessing but far from what we humans can achieve. This can be due to bad vectorization and also may be the sentences are not linearly separable . Vectorization can be dealt with Deep Learning based embeddings and instead of logistic regression we can use RNN based networks.