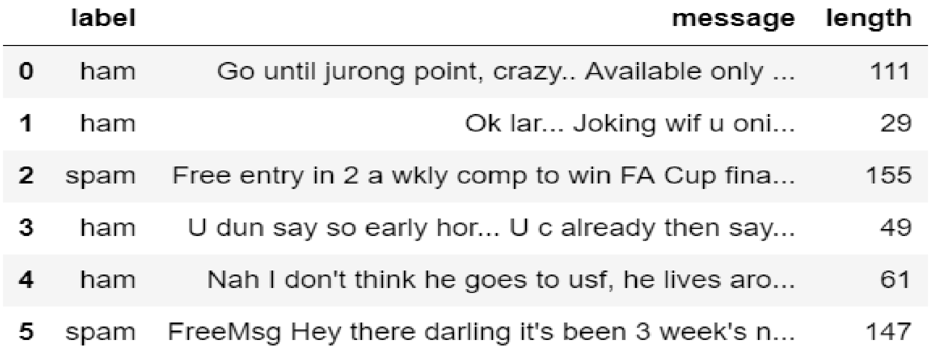
Name: Rohan Patel
SFSU ID: 917583698

SMS Spam Detection – Project Update

1. Current Status

So far, I have done some basic data analysis and feature engineering on the dataset. I downloaded the SMS dataset from the UCI Machine Learning Repository. I used python's pandas library to convert the dataset into a pandas dataframe. Initially the dataframe had 2 columns: label and message. The 'label' column indicates if the message is spam or non-spam (ham). The 'message' column is simply the text message itself. Next, I added a new column to this dataframe called 'length'. The length column stores the length of the message. Figure 1 shows the head of this dataframe.



	label	message	length
0	ham	Go until jurong point, crazy.. Available only ...	111
1	ham	Ok lar... Joking wif u oni...	29
2	spam	Free entry in 2 a wkly comp to win FA Cup fina...	155
3	ham	U dun say so early hor... U c already then say...	49
4	ham	Nah I don't think he goes to usf, he lives aro...	61
5	spam	FreeMsg Hey there darling it's been 3 week's n...	147

Figure 1.

In addition to this, I have also done some data visualization using python's matplotlib library. I created a histogram of the 'length' column to see how message length is distributed, and if there are any interesting patterns in this distribution. Figure 2 shows the distribution of all the messages. It is interesting to see there are 2 peaks in this distribution, indicating that the spam and ham messages might have different distributions. Figure 3 gives more information about how spam and ham messages distributed.

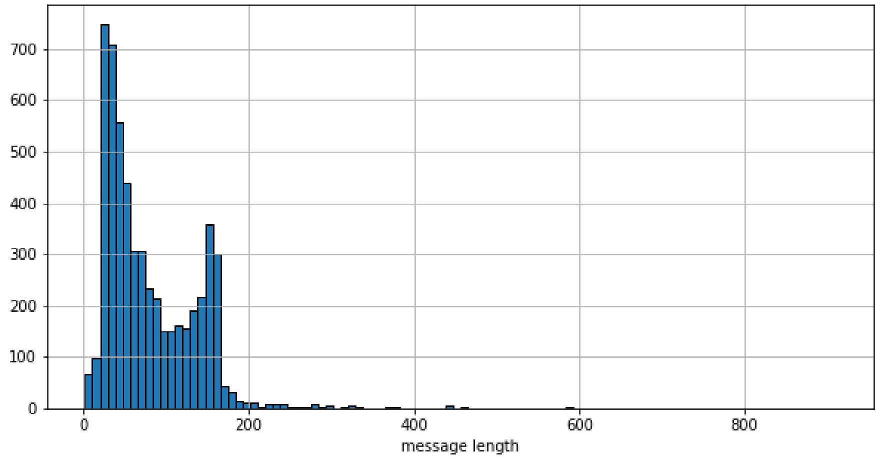


Figure 2.

Name: Rohan Patel
SFSU ID: 917583698

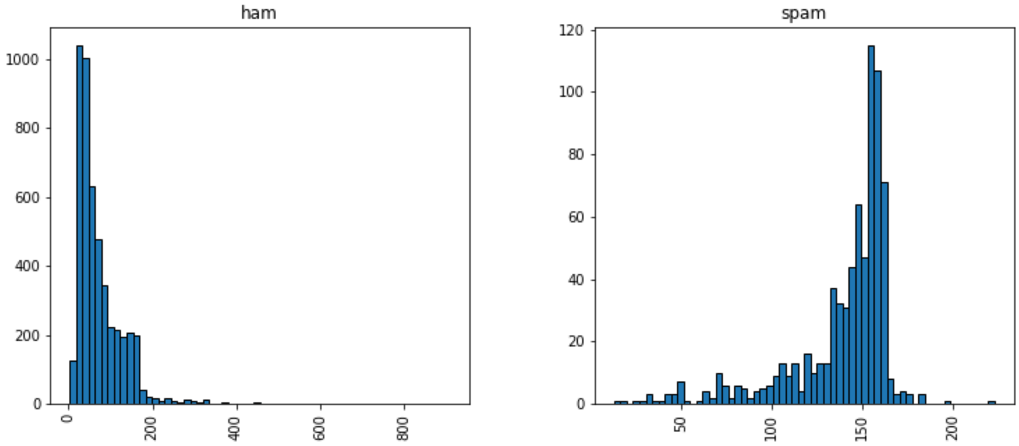


Figure 3.

2. Challenges

The only challenge I have faced so far is that I haven’t been able to find any good SMS datasets besides the one from the UCI repository that I’m currently using. I did find another dataset, but the only problem is that this dataset only contains spam messages and no ham messages. As a result, there is no way for me to test if my final approach will have any bias or not. I plan on meeting the professor this week during her office hours to address this issue.

3. Next Steps

For my next project update, I will be working on text pre-processing. I will be removing stopwords and punctuations from the messages and lower all of the text. I am also considering stemming of the words, but I’m not absolutely certain about this right now.