IDENTIFICATION OF SPAM MESSAGES

HSLU DEEP LEARNING BOOTCAMP

AGNES LEI

AMELIA BUDIHARTO

FATHIMA ABDUL RAHMAN

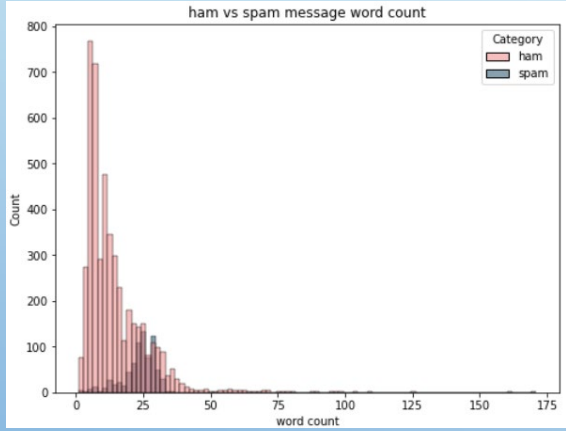
HOSSAM FAHMY

OVERVIEW

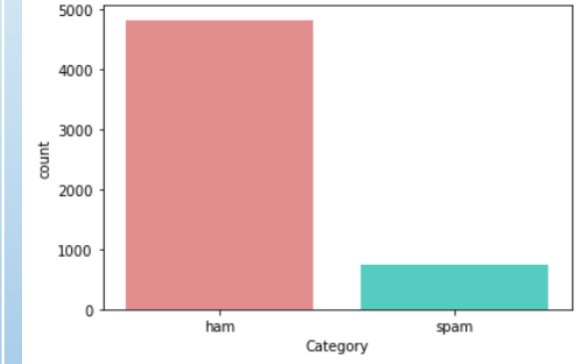
- EXPLANATORY DATA ANALYSIS
- DEEP NEURAL NETWORKS MODELS
- MODEL EVALUATION + ‘BEST’ MODEL
- POSSIBLE IMPROVEMENT

DATASET EDA

Word count



Imbalanced Data



Foreign Characters

```
df_eda[df_eda['len_foreign_words'
```

	Category	foreign_words
22	ham	[ü]
35	ham	[ü]
125	ham	[Ü, ü]
140	ham	[ü]
206	ham	[Ü]
...	...	...
5354	ham	[ü, ü, ü, ü]
5400	ham	[ü‌ud]
5407	ham	[ü, ü]
5473	ham	[ü]
5568	ham	[ü]

Repeated Messages

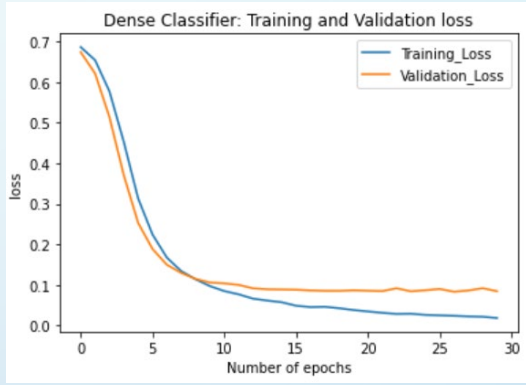
label	ham	spam
count	4825	747
unique	4516	653
top	Sorry, I'll call later	Please call our customer service representativ...
freq	30	4

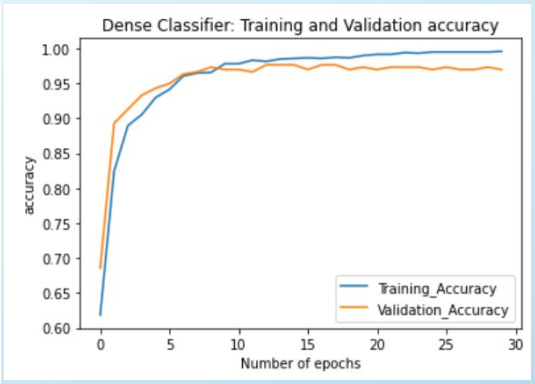
DEEP NEURAL NETWORK MODEL

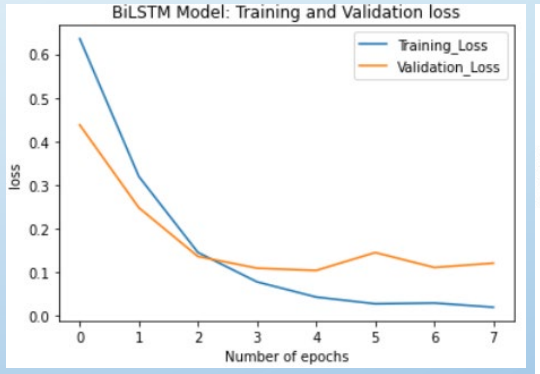
Model	Pre-processing	Model Architecture *from tensorflow.keras.layers*	Hyperparameters
Dense	Tokenization, Sequencing, Padding, Resampling, Decoding	Sequential + GlobalAveragePooling1D + Dense + Dropout + Dense	vocab_size = 500 embeding_dim = 16 drop_value = 0.2 n_dense = 24
LSTM		Sequential + Embedding + LSTM + Dense	Learning rate = 0.001 Epochs = 100 (with early stopping – stopped at 8) Batch size = 16 Verbose=2 activation= sigmoid
Bidirectional LSTM		Functional +Embedding +bidirectional LSTM + Dense	Epochs = 30 (with early stopping at 5) Recurrent dropout = 0.2 Verbose = 2 Activation = ReLU and sigmoid
Bidirectional GRU		Sequential + Embedding + Bidirectional GRU + GRU + Dense	Optimizer = Adam Learning rate = 0.01 Dropout = 0.5 Recurrent dropout = 0.3 Activation = relu and sigmoid Epoch = 100 Batch size = 64
GRU + Glove Embedding		Sequential + Glove Embedding + GRU + Dense	

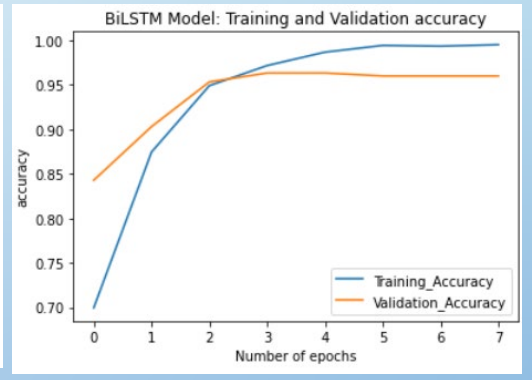
MODEL EVALUATION

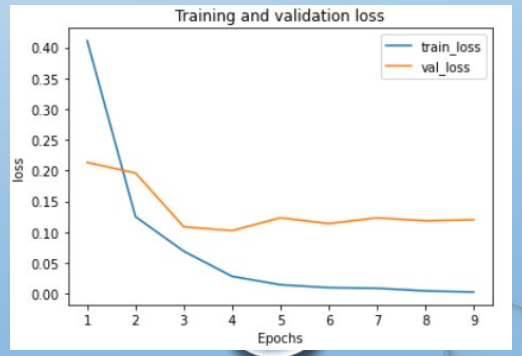
Model	Accuracy	Loss
Dense	97.32%	8.46%
LSTM	97.65%	7.95%
Bidirectional LSTM	96.65%	13.62%
Bidirectional GRU	*Imbalance Data* 97.87% *Balance Data* 92.2%	*Imbalance Data* 15.9% *Balance Data* 31.4%
GRU + Glove Embedding	*Imbalance Data* 97.38% *Balance Data* 92.61%	*Imbalance Data* 8.07% *Balance Data* 30.8%

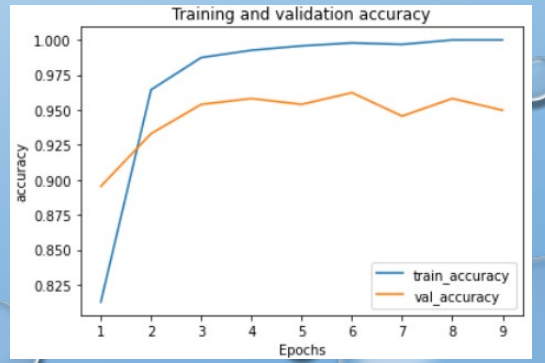










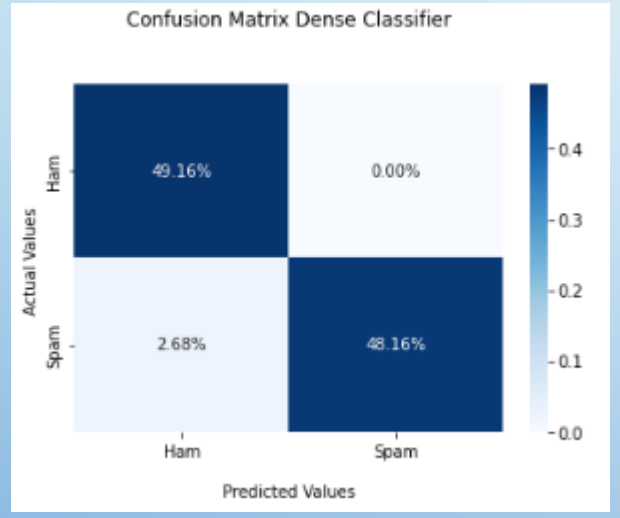


POSSIBLE IMPROVEMENT

- AUTOMATED HYPERPARAMETER TUNING (AUTOKERAS, AUTOML)
- EXTEND EDA AND PREPROCESSING (E.G., STANDARDIZE LANGUAGE)
- EXPLORE MORE SAMPLING APPROACHES (WEIGHT BALANCING, UPSAMPLING, SMOTE, OVERALL SAMPLE)
- EXPERIMENTING WITH DIFFERENT EMBEDDINGS (BERT, GLOVE, GPT2)

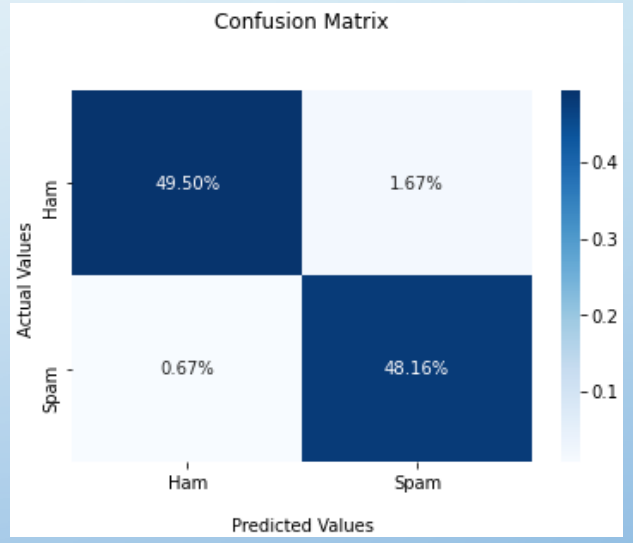
APPENDIX

Dense Model



	precision	recall	f1-score	support
0	0.49	1.00	0.66	147
1	0.00	0.00	0.00	152
accuracy			0.49	299
macro avg	0.25	0.50	0.33	299
weighted avg	0.24	0.49	0.32	299

LSTM Model



	precision	recall	f1-score	support
0	0.53	1.00	0.69	158
1	0.00	0.00	0.00	141
accuracy			0.53	299
macro avg	0.26	0.50	0.35	299
weighted avg	0.28	0.53	0.37	299