Gender classification by first name

Konstantin Lekomtsev
Data Science Career Track

Mentor: Nishant Gupta

Objective and Dataset overview

Objective: classify a person's first name as male or female.

Info about the dataset: 5000 first names. 2662 are male, 2365 are female.

	Name	Gender	LastLetter	LastTwoLetter	FirstLetter
0	ashutosh	0	h	sh	a
1	meghamala	1	a	la	m
2	sahib	0	b	ib	s
3	pragya	1	a	ya	p
4	kranti	1	i	ti	k

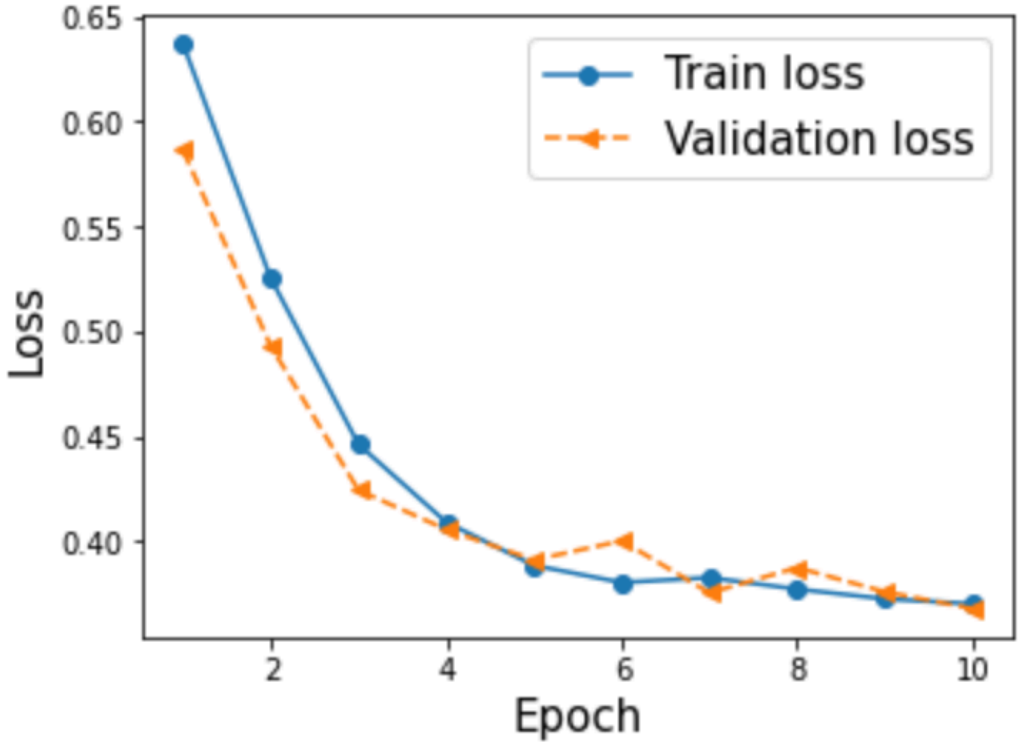
Approach

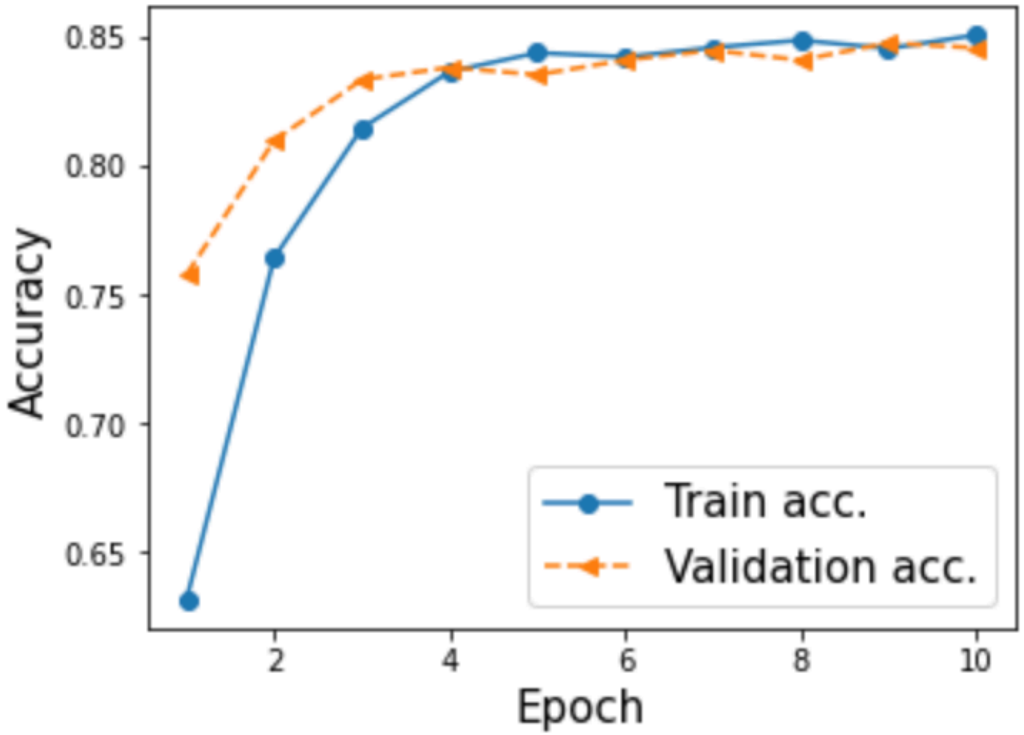
- Apply Naive Bayes classifier in the last letter, last two letters, and the first letter of the names.
- Use term-frequency times inverse document-frequency (tfidf) vectorization with 2-gram and 3-gram sequences to cerate features for SVM classifier and multilayer perceptron classifier.
- Use simple integer encoding and apply a recurrent neural network with one and multiple long short-term memory (LSTM) layers and dropout layers.

Algorithms comparison

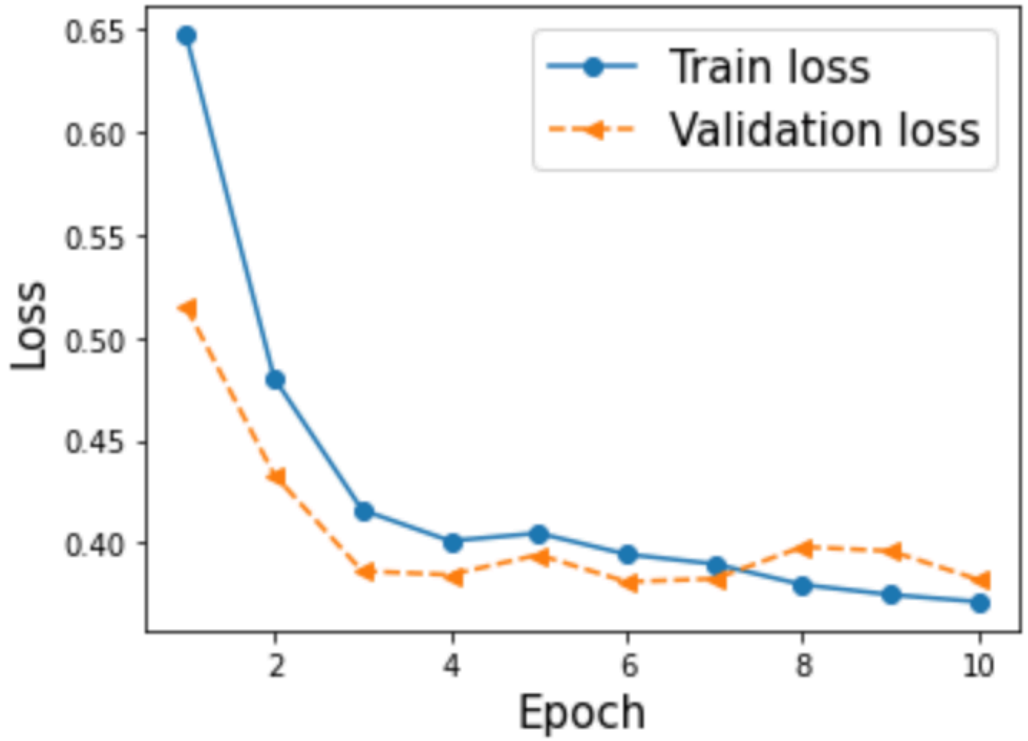
Model	AUC (test)
NB, last letter	0.85
NB, last two letters	0.88
NB, first letter	0.55
SVM, tfidf vect.	0.82
MLP, tfidf vect.	0.73
LSTM, integer vect.	0.9
multi LSTM, integer vect.	0.9

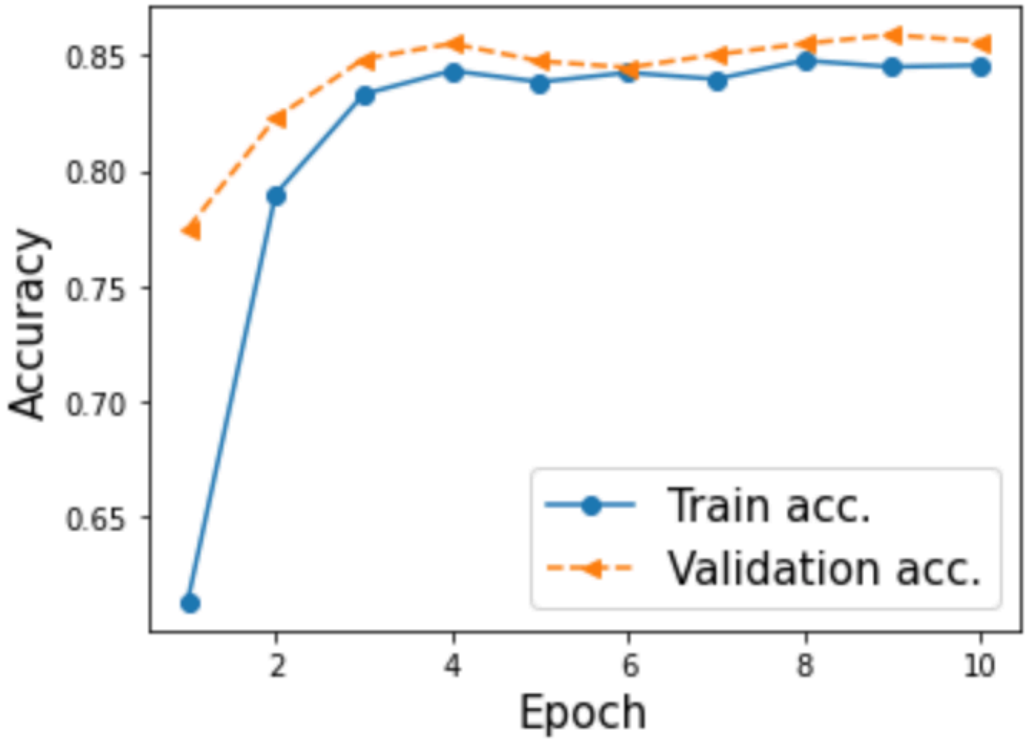
RNN with one bidirectional LSTM - training





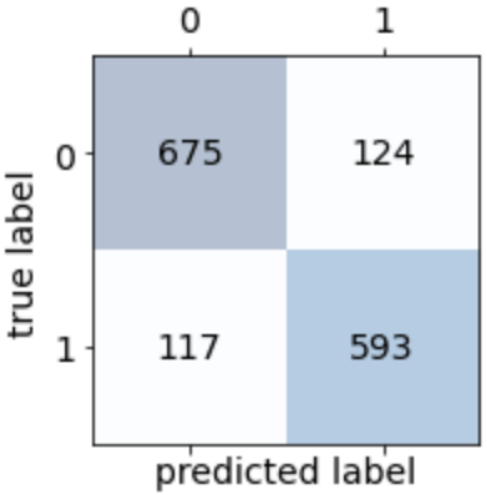
RNN with two bidirectional LSTMs and dropout layers - training





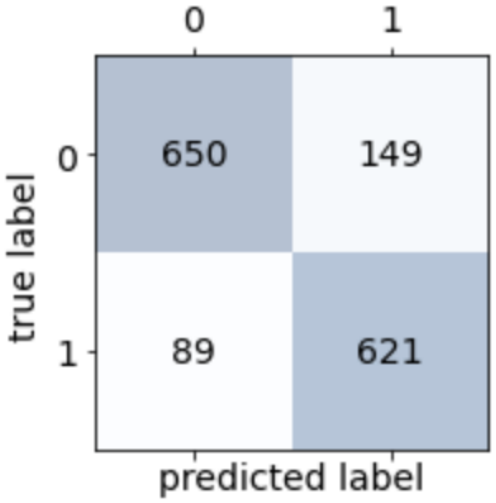
RNNs - comparison

One LSTM:



	precision	recall	f1-score
male	0.85	0.84	0.85
female	0.83	0.84	0.83
accuracy			0.84

Two LSTMs with dropout layers:



	precision	recall	f1-score
male	0.88	0.81	0.85
female	0.81	0.87	0.84
accuracy			0.84

Conclusions

- The top performers were the RNNs with one and two bidirectional LSTMs and dropout layers.

- Naive Bayes classifier on the last two letters was the second best, with significantly shorter computation time.