Classification of Questions

Petr Lorenc 2017
FIT CTU Prague
12.5.2017

Why ?

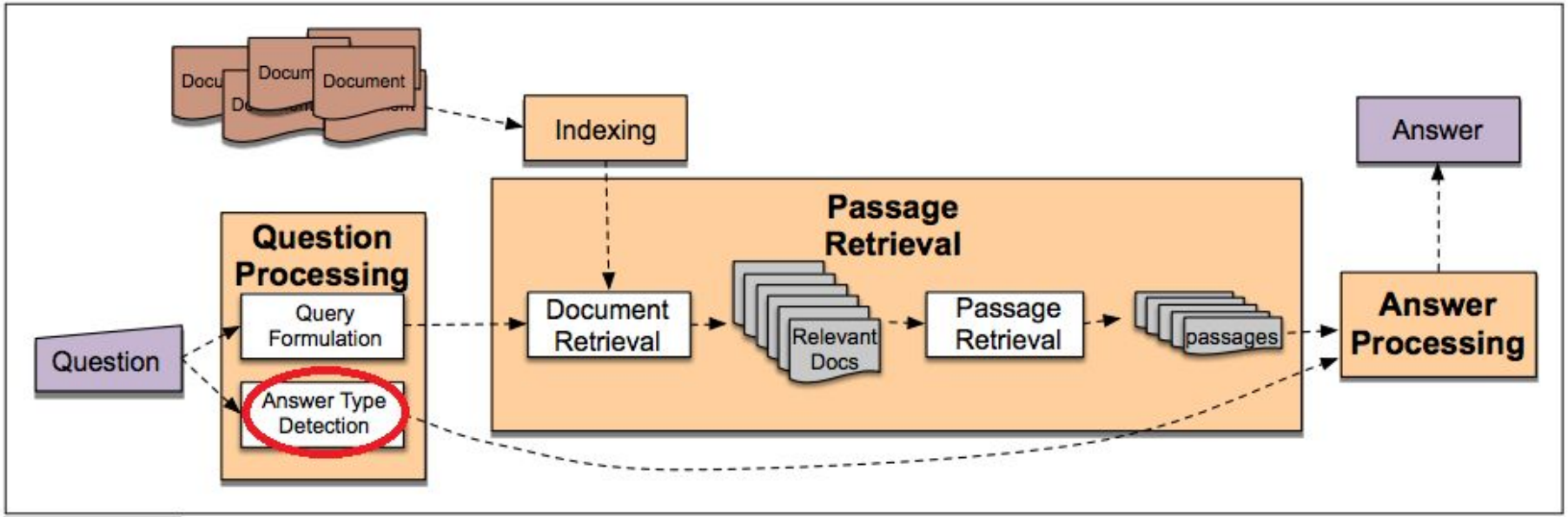
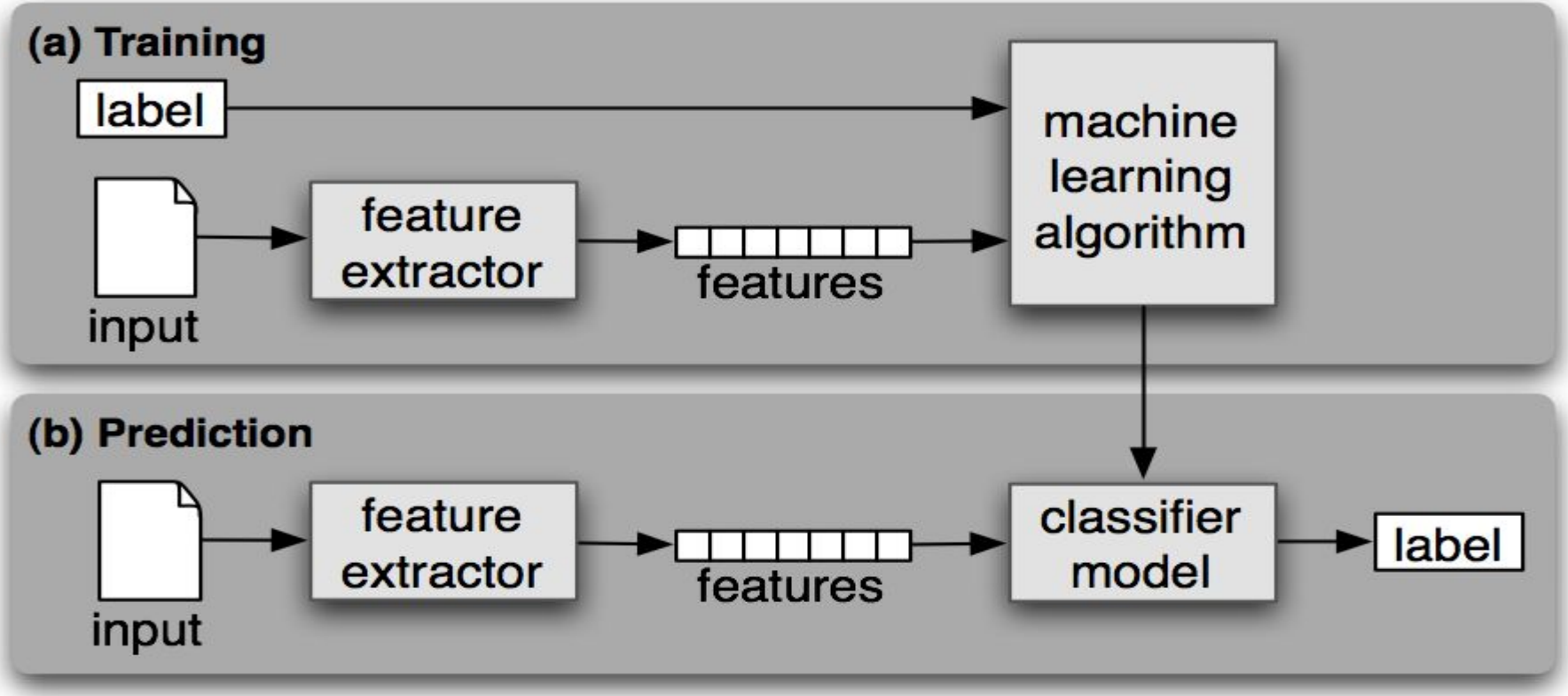


Figure 28.2 IR-based factoid question answering has three stages: question processing, passage retrieval, and answer processing.

How?


(a) Training
label
input
feature extractor
features
machine learning algorithm
(b) Prediction
input
feature extractor
features
classifier model
label

Representation of data

1. Tf-Idf
2. Bi-grams + Tf-Idf
3. Vector representation (with/without scale normalization)
    a. Word2Vec
    b. Glove
    c. Doc2Vec
    d. FastText

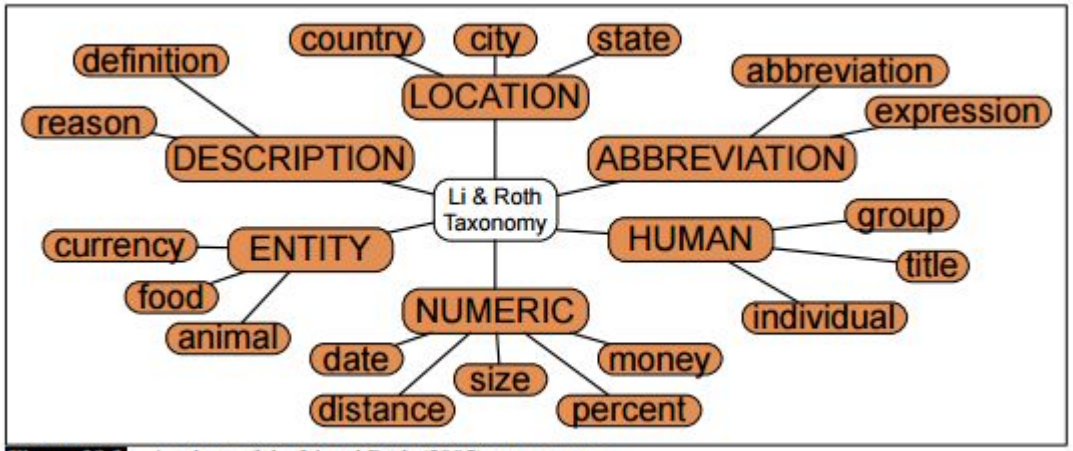
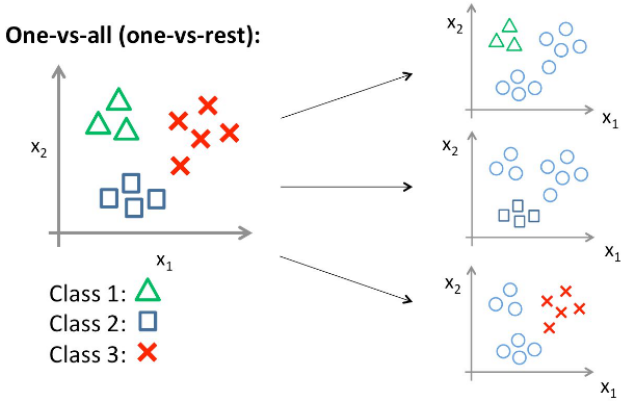
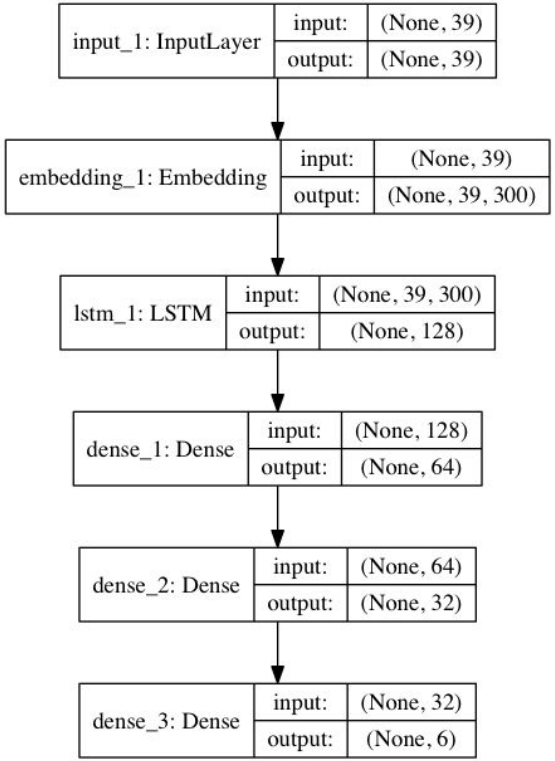


Figure 28.3 A subset of the Li and Roth (2005) answer types.

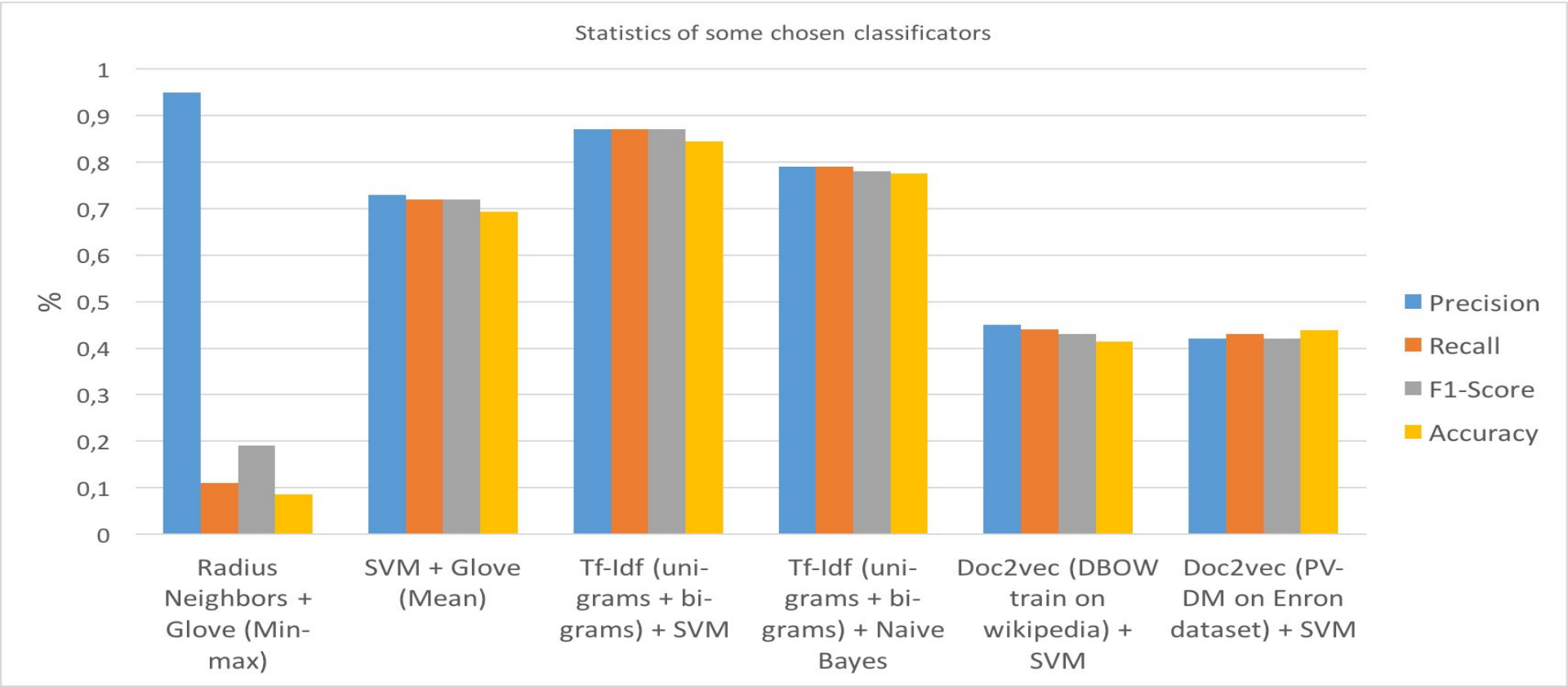
Algorithms

1. Naive Bayes
2. **SVM (+ bigrams) - 83%**
3. kNN
4. Neural network
   a. **LSTM - 89 %**
   b. CNN
   c. LSTM + CNN

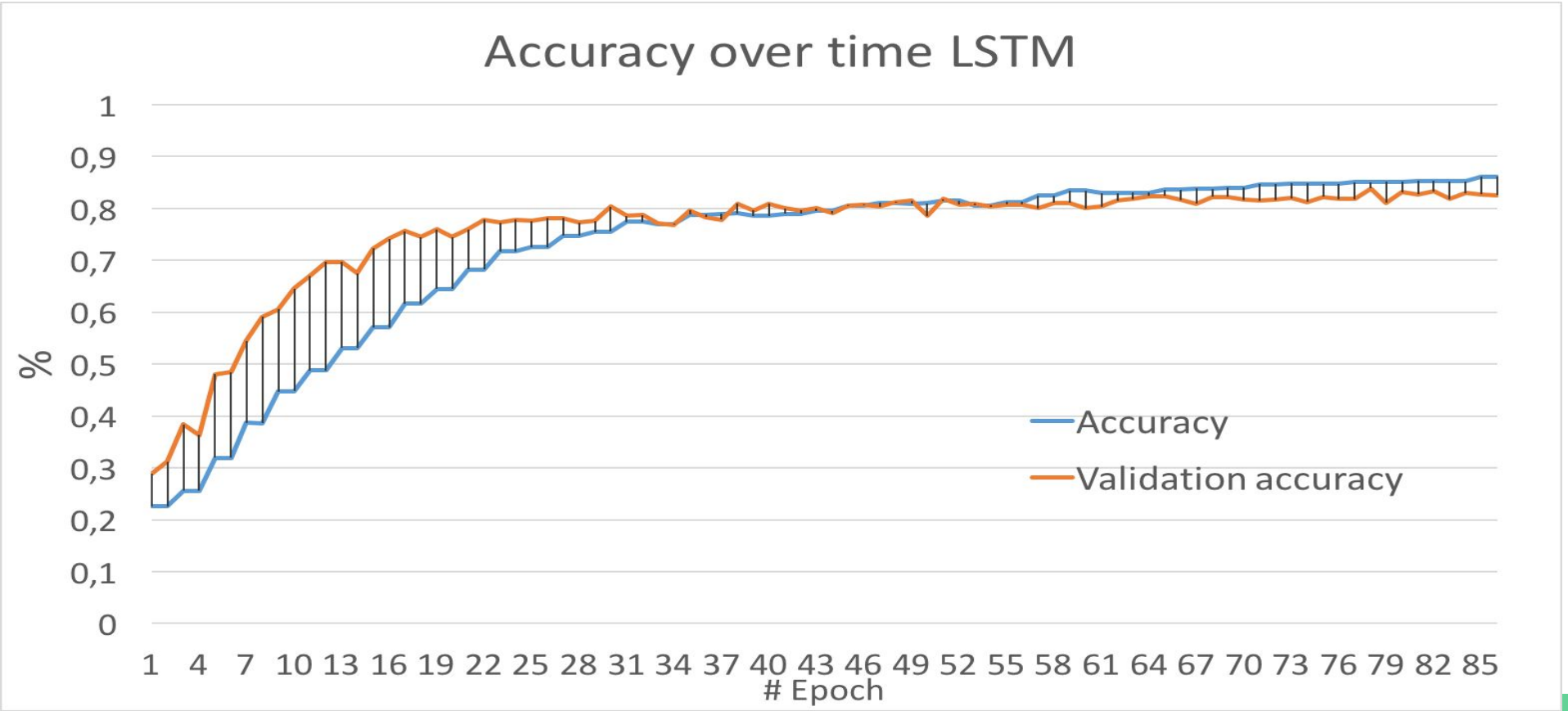




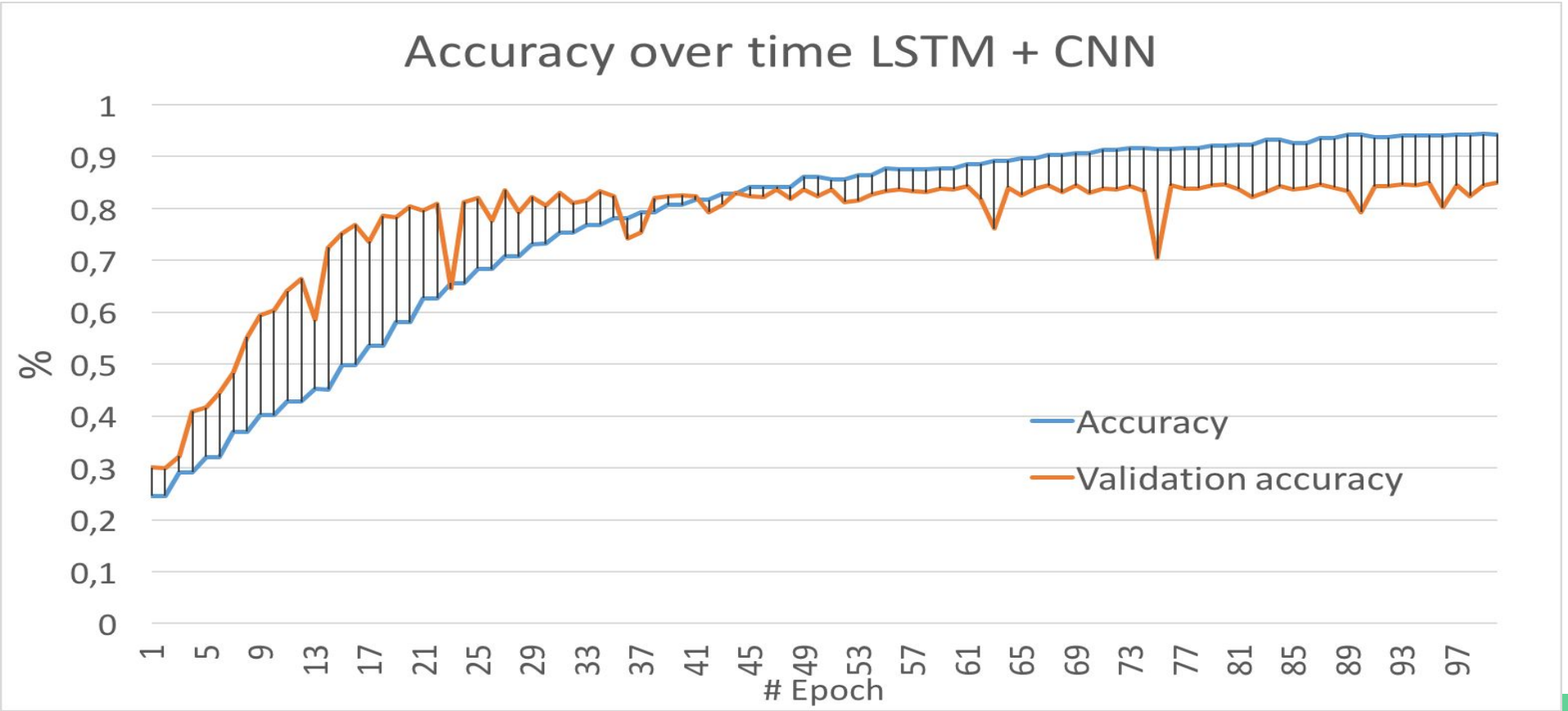
Results


Statistics of some chosen classificators
%
1
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
Radius Neighbors + Glove (Min-max)
SVM + Glove (Mean)
Tf-Idf (uni-grams + bi-grams) + SVM
Tf-Idf (uni-grams + bi-grams) + Naive Bayes
Doc2vec (DBOW train on wikipedia) + SVM
Doc2vec (PV-DM on Enron dataset) + SVM
Precision
Recall
F1-Score
Accuracy

Results


Accuracy over time LSTM
1
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
%
1 4 7 10 13 16 19 22 25 28 31 34 37 40 43 46 49 52 55 58 61 64 67 70 73 76 79 82 85
Epoch
Accuracy
Validation accuracy

Results


Accuracy over time LSTM + CNN
1
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
%
1
5
9
13
17
21
25
29
33
37
41
45
49
53
57
61
65
69
73
77
81
85
89
93
97
Epoch
Accuracy
Validation accuracy

Results

- FastText is library from Facebook, which has build-in function to classify text (based on train data), it also provide similar functionality like word2vec

	# Unique Words	Precision	Recall	Time to train 5 Epoch
Train on **Quora** Test dataset + Retrain on **Base** dataset	136 665	**0.862**	**0.862**	7 min
Train on **Quora** Train dataset + Retrain on **Base** dataset	57 165	0.85	0.85	94 s
Train on **Base** dataset	7 111	0.818	0.818	8 s

Reference

https://web.stanford.edu/ jurafsky/slp3/28.pdf

Frameworks

Code

Available at: https://github.com/petrLorenc/EmailReply

Report at: https://github.com/petrLorenc/EmailReply/blob/master/report.pdf