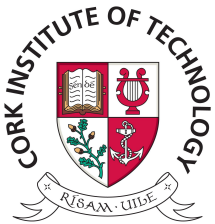


Joao Tiago Viegas

R00157699

Higher Diploma

in Science

in

Data Science and Analytics

December 2018

A study on sentiment classification of news headlines

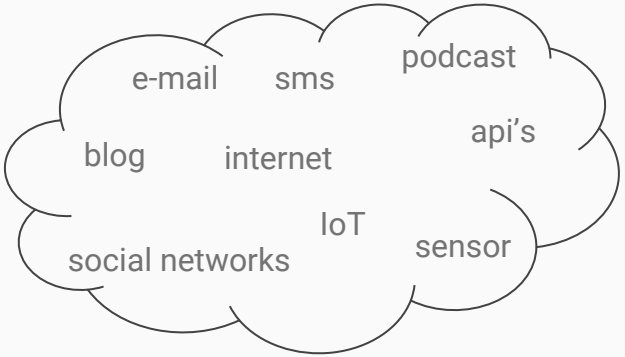
using deep learning

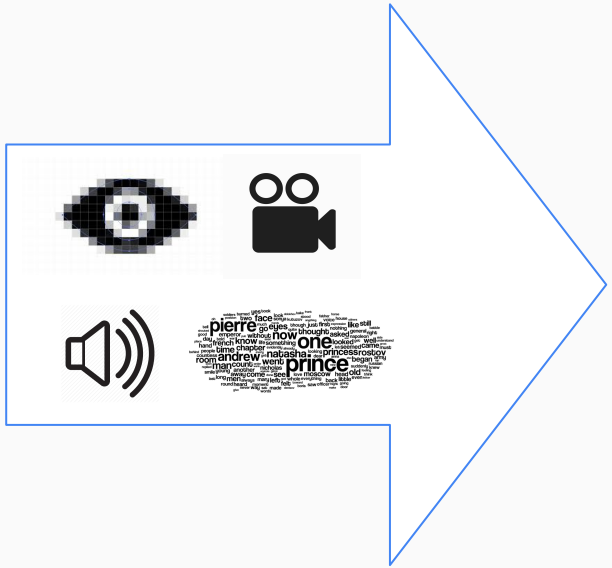
Intro

- motivation and dataset
- methodology
- findings
- conclusion

motivation and dataset

motivation





product feedback

<u>sentiment analysis</u>

fraud detection

automatic translation

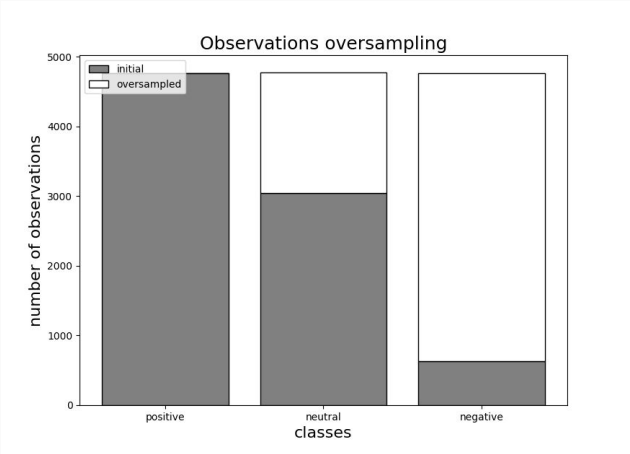
recommendation systems

security surveillance

dataset - sentiment analysis

141	to cut 5,000 jobs in U.S.--sources	−1
142	Devon IT, and VMware to Host	0
143	Guangzhou Metro Corporation Works With to Modernize Rapid Transit in China	1

- news headlines subset from reuters;
- dates range: 08.01 2007 -> 02.10.2018;
- 9385 observations;
- 10-fold cross validation sample:
  - Training set: 8447 observations;
  - negative: 632, neutral: 3044, positive: 4771;

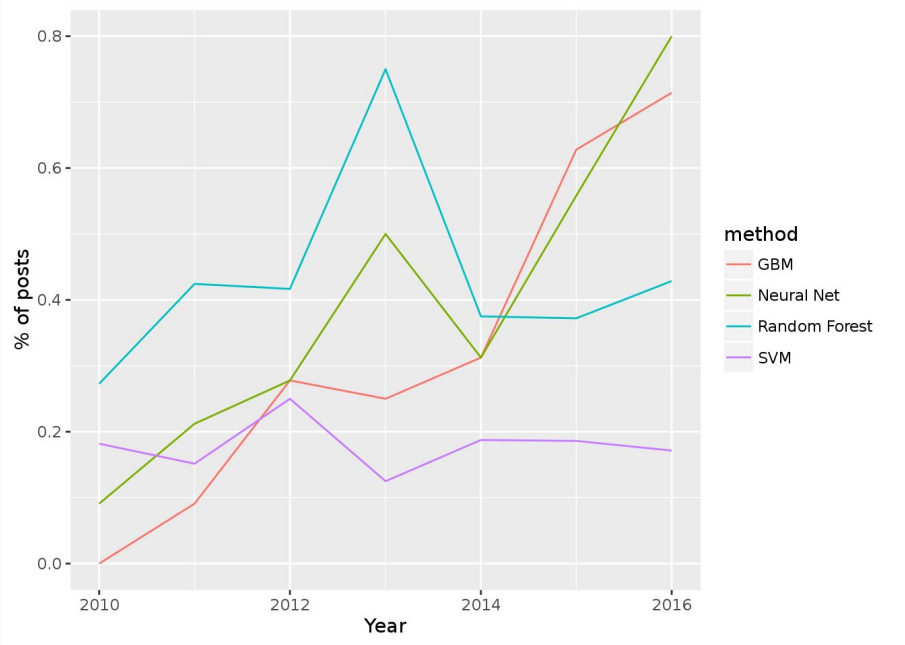


methodology

Baseline - Naïve Bayes

$$P(c|x) = \frac{P(x \mid c) * P(c)}{P(x)}$$

Deep learning - neural networks?

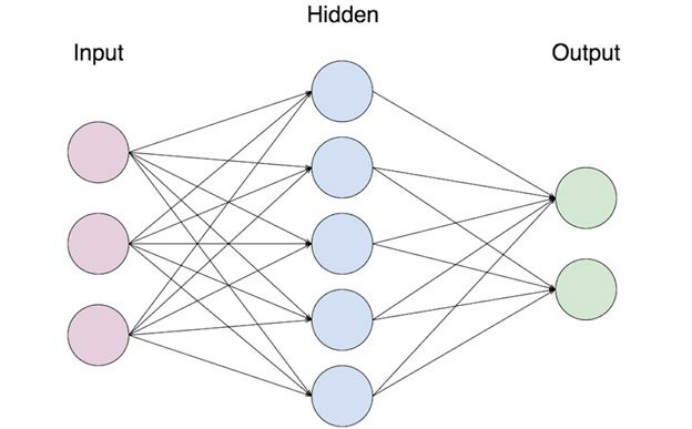


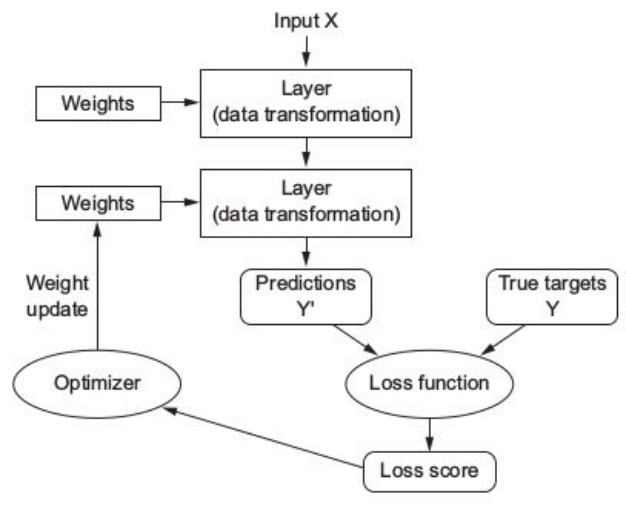
Neural networks algorithms adoption has been rising:

"...neural networks and gradient boosting machines...so far in 2016, they've been appearing in >70% of winners posts."

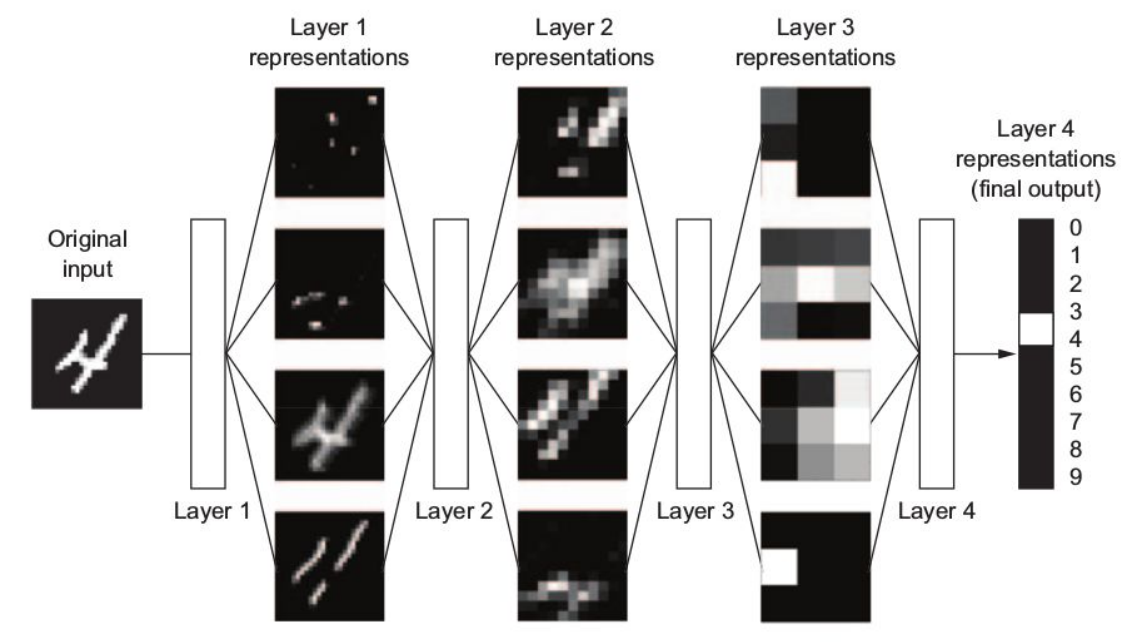
Anthony Goldbloom, Kaggle CEO

Deep Learning - Artificial Neural Networks

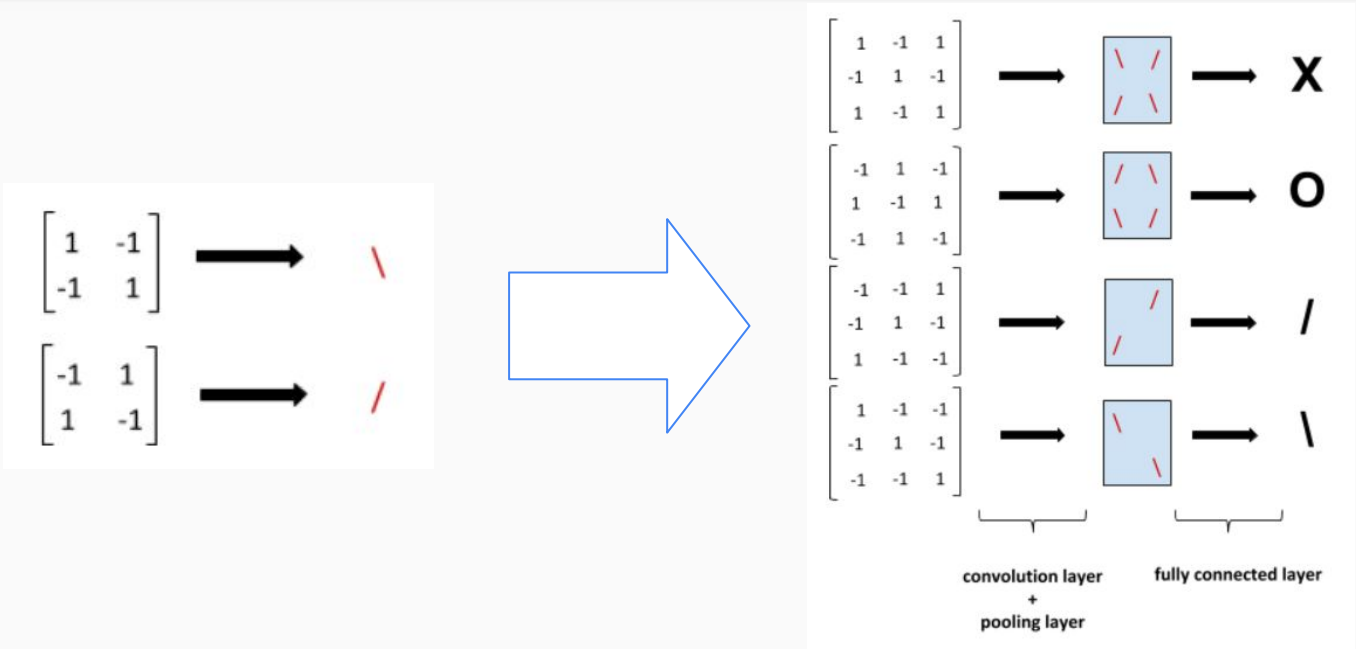




Deep Learning - Artificial Neural Networks



Deep Learning - Convolutional Neural Networks (CNN)



findings

neural network models

1 hidden layer

layers

```
Dense(32, activation='relu')
Dense(32, activation='relu')
Dense(3, activation='softmax')
```

learning

```
optimizer: rmsprop
loss: categorical_crossentropy
metrics: accuracy, recall, precision, f1_score
```

1 hidden layer with Glove pre-trained embeddings

layers

```
Embedding(...)
Flatten()
Dense(1024, activation='PReLU')
Dense(3, activation='softmax')
```

learning

```
optimizer: rmsprop
loss: categorical_crossentropy
metrics: accuracy, recall, precision, f1_score
```

2 hidden layers with word embeddings

```
Embedding(...)
Flatten()
Dense(32, activation='relu')
Dense(32, activation='relu')
Dense(3, activation='softmax')
```

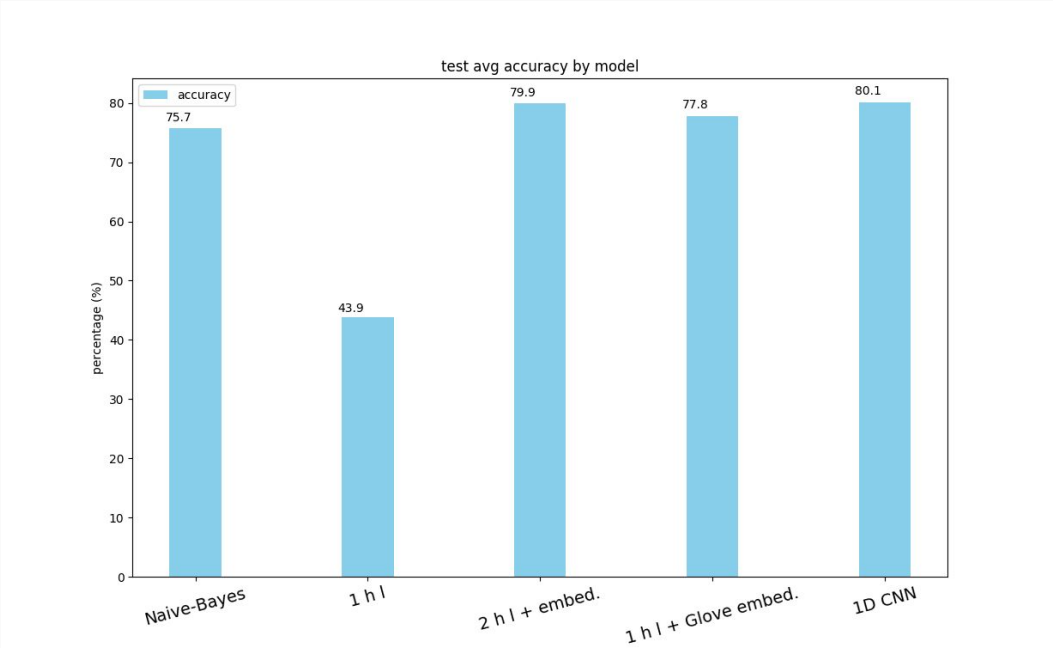
```
optimizer: rmsprop
loss: categorical_crossentropy
metrics: accuracy, recall, precision, f1_score
```

One dimensional Convolutional Neural Network

```
Embedding(...)
Conv1D(8, 24, activation='relu')
GlobalMaxPooling1D()
Dense(3, activation='softmax')
```

```
optimizer: rmsprop
loss: categorical_crossentropy
metrics: accuracy, recall, precision, f1_score
```

findings - models accuracy



conclusion

main findings

- Simple 1D CNN achieved better accuracy;
- 2 layers neural network similar to simple 1D CNN;
- GloVe pre-trained word embeddings vector not reflecting domain knowledge;
- Naïve Bayes not far away;

future work

- include full news article content;
- bigger dataset;
- further parameter testing;
- further topology experimentation;

conclusion

model	accuracy (avg +/- std)
Naïve Bayes	75.70% (+/- 1.58%)
1 hidden layer	43.93% (+/- 5.18%)
2 hidden layers with word embeddings	79.88% (+/- 1.30%)
1 hidden layer with Glove pre-trained embeddings	77.78% (+/- 1.90%)
one dimensional CNN	80.12% (+/- 0.85%)

code available @ https://github.com/jtviegas/studies/tree/master/cit/deep_learning/code/analysis

Thank You!