CS230 Winter 2018

Topic Analysis of Digital Media with Deep Learning

Nathan Zhao, Jiahui Wang, Xianling Long

Introduction

Differentiating content in news journalism can reveal insights as to what organizations tend to write about over time as well as elucidate how topics evolve over time. We apply deep learning to characterize topics of articles.

Data

We preprocess the documents by tokenization, dropping stop words, normalization (case-folding), stemming and lemmatization.

Table 1. Corpus Description metadata

Corpus	# Articles	length	tokens	# Labels
Guardian	10,000	675	102,000	27
Guardian Deployment	150,000	841	--	unlabeled
Economist	95,000	740	395,000	38

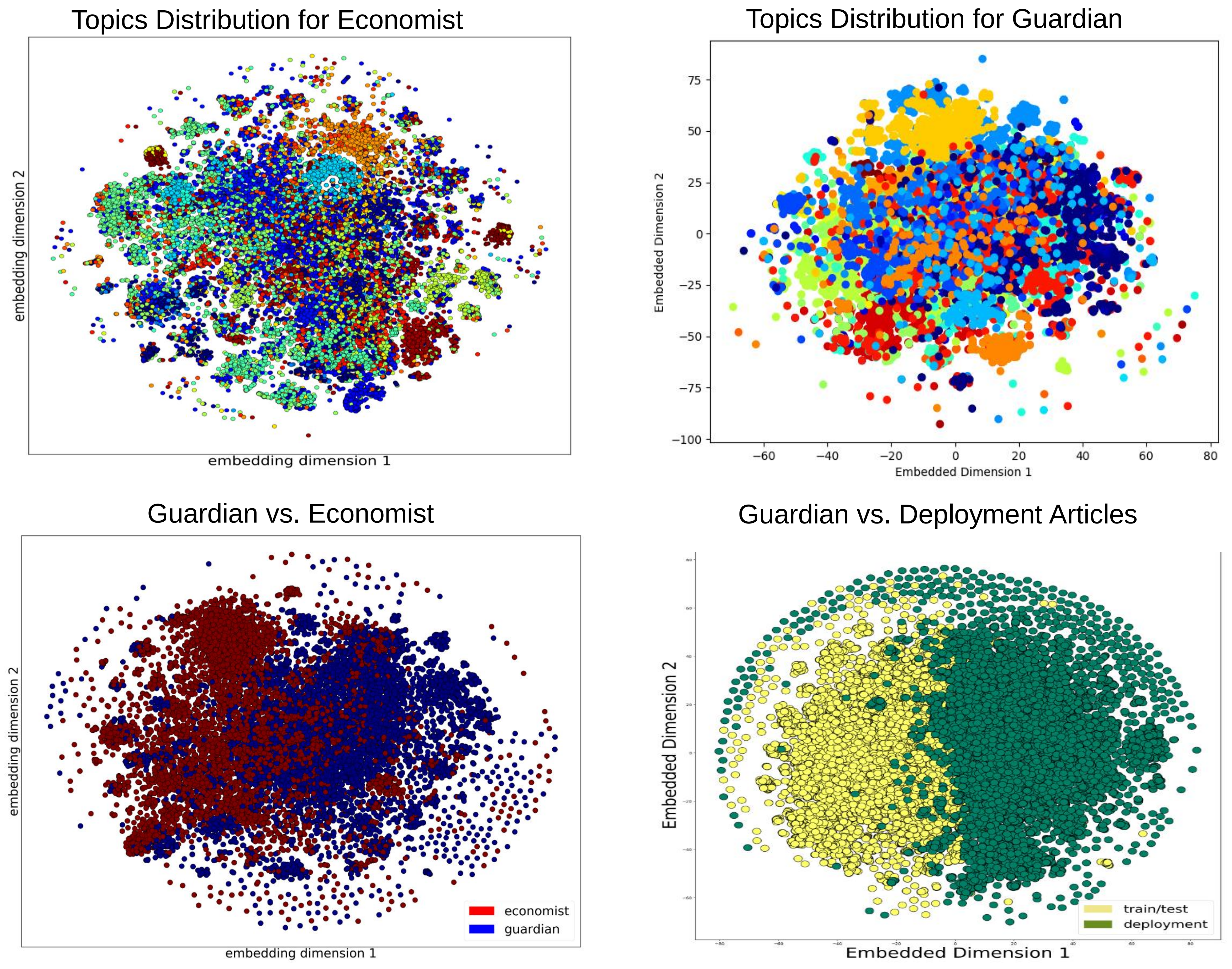
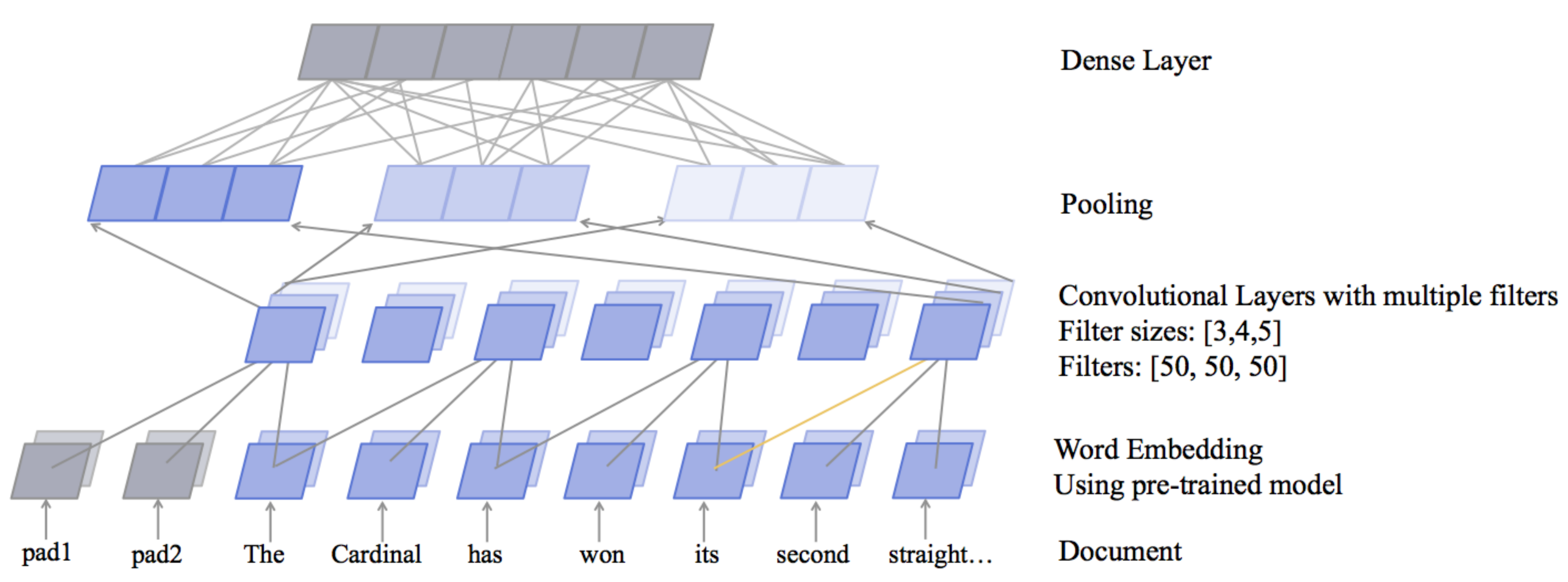


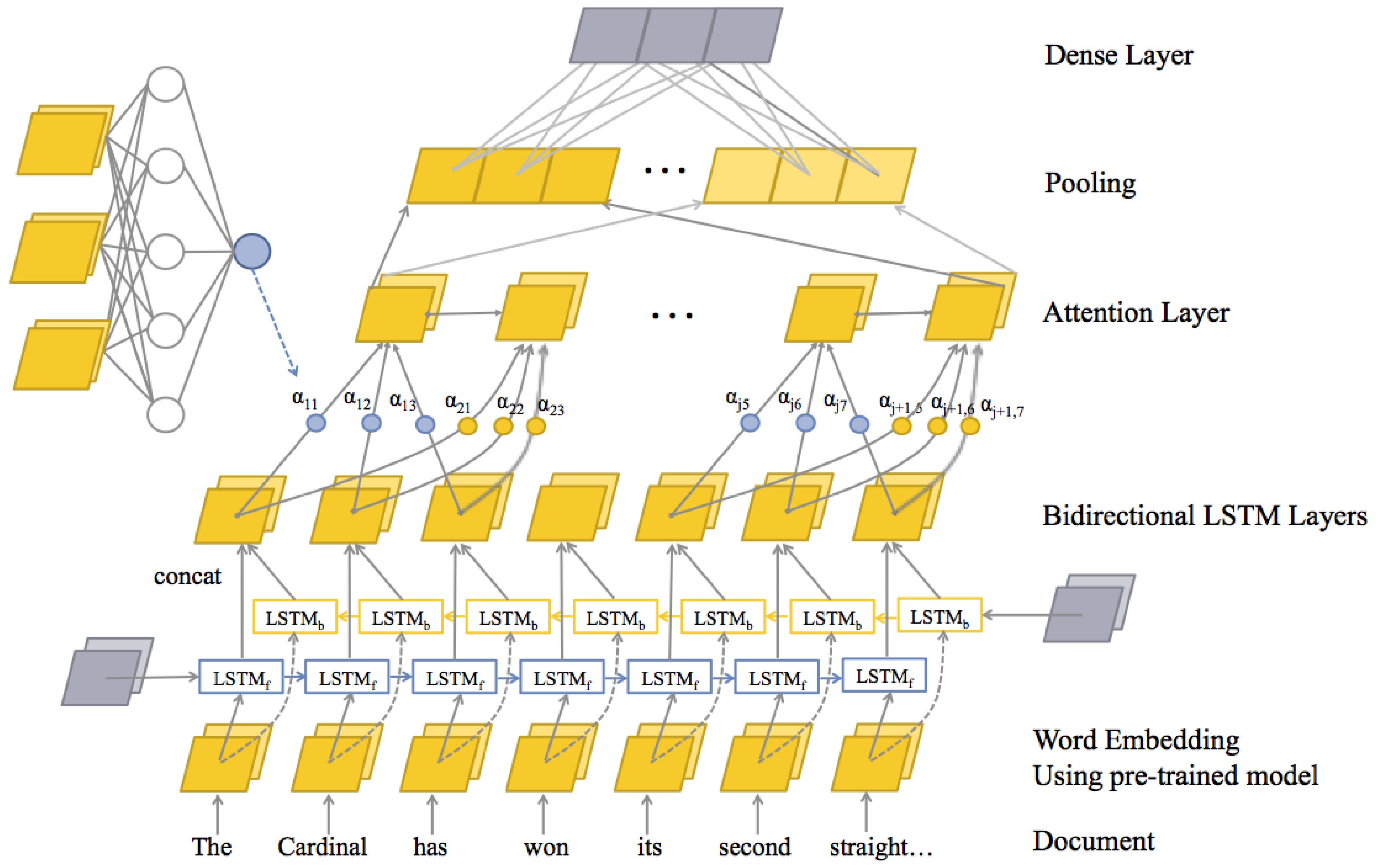
Fig 1. t-SNE visualization for different classification tasks

Models

Multi-pooled CNN



LSTM with Attention



- For many, sequence modeling is synonymous with recurrent networks.
- Yet convolutional architectures may outperform recurrent networks on topic classification, because convolutional networks are good at finding local features, i.e. keywords.
- For comparison, we employ both CNN and RNN.

Naïve Bayes Benchmark

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)} \qquad p(y|x) = p(x_1|y)p(x_2|y) \ldots p(y)$$

Results

Table 2. Test Accuracies

Model	Guardian	Economist
Naïve Bayes	68%	55%
LSTM (w/o attention)	70%	81%
LSTM (w/ attention)	82%	72%*
CNN	**87%**	**85%**

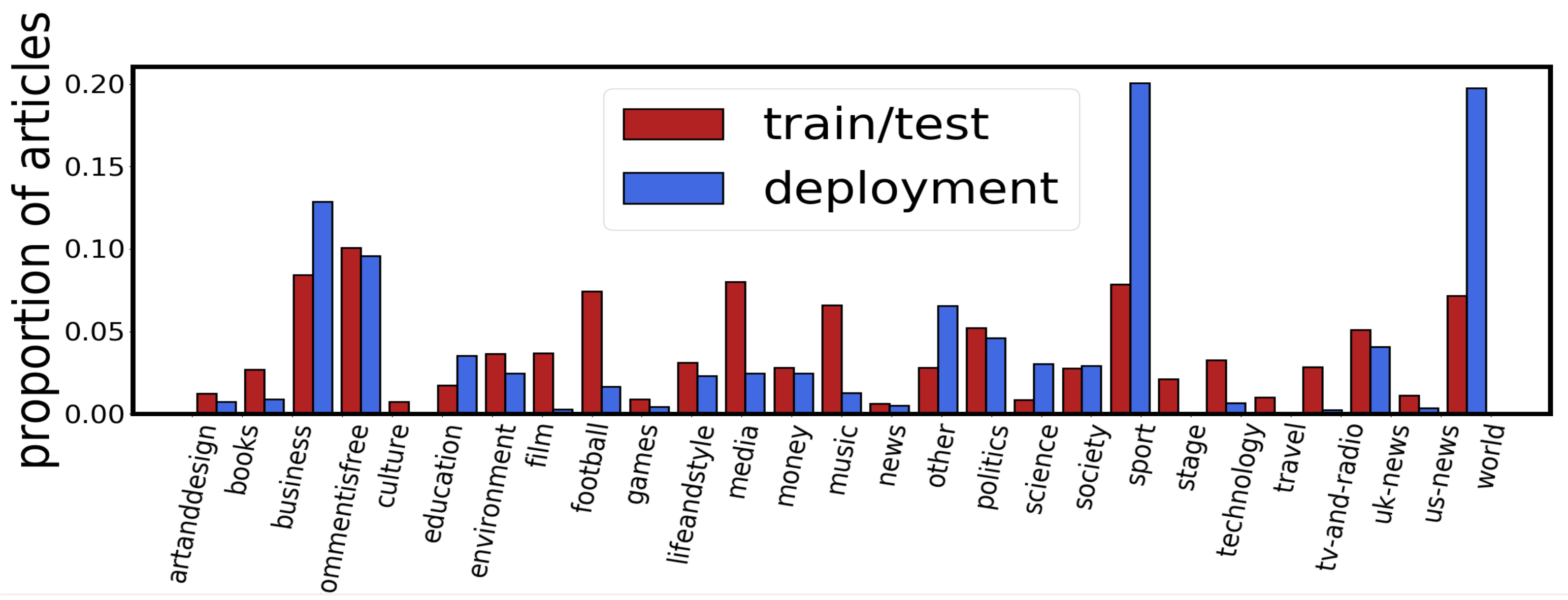


Fig 2. Topics distribution in *the Guardian*

Table 3. State of Art Document Classification Accuracies

	# Label	Accuracy
Yoon Kim (2014)	2 - 5	75%~95%
Stanford 20Newsgroups	20	88%
Zichao Yang (2016)	5	70%
Duyu Tang (2015)	5	66%

Improvements

- Character level CNNs
- More rigorous implementation of attention in CNNs
- Semi-supervised learning

References

- Yoon Kim. Convolutional neural networks for sentence classification. *EMNLP* 2014.
- Yang, Zichao, et al. Hierarchical attention networks for document classification. *NAACL HLT 2016*.
- Tang, Duyu, Bing Qin, and Ting Liu. Document modeling with gated recurrent neural network for sentiment classification. *EMNLP 2015*