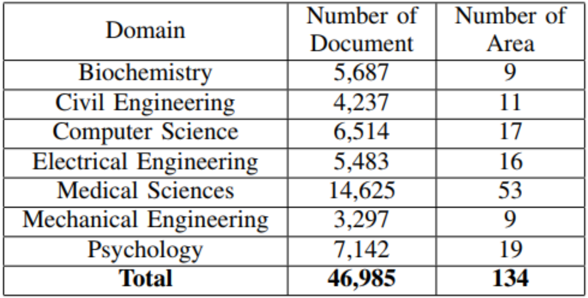
HDLTex: Hierarchical Deep Learning for Text Classification

2017

Data

- Data pairs (d_i, Y_i)
  - document $d_i = (w_{i,1}, w_{i,2}, \dots, w_{i,J_i})$ where J_i is the length of the document i, $w_{i,j}$ is the word embedding vectorization of word j in document i.
  - label $Y_i = (k, l)$ where $k \in \{1, \dots, K\}$ and $l \in \{1, \dots, L_k\}$

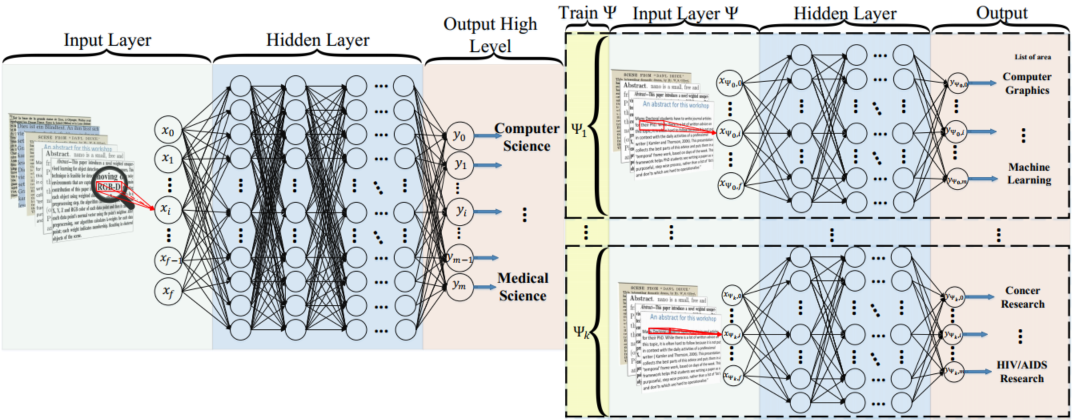
TABLE I: Details of the document set used in this paper.



Domain	Number of Document	Number of Area
Biochemistry	5,687	9
Civil Engineering	4,237	11
Computer Science	6,514	17
Electrical Engineering	5,483	16
Medical Sciences	14,625	53
Mechanical Engineering	3,297	9
Psychology	7,142	19
Total	**46,985**	**134**

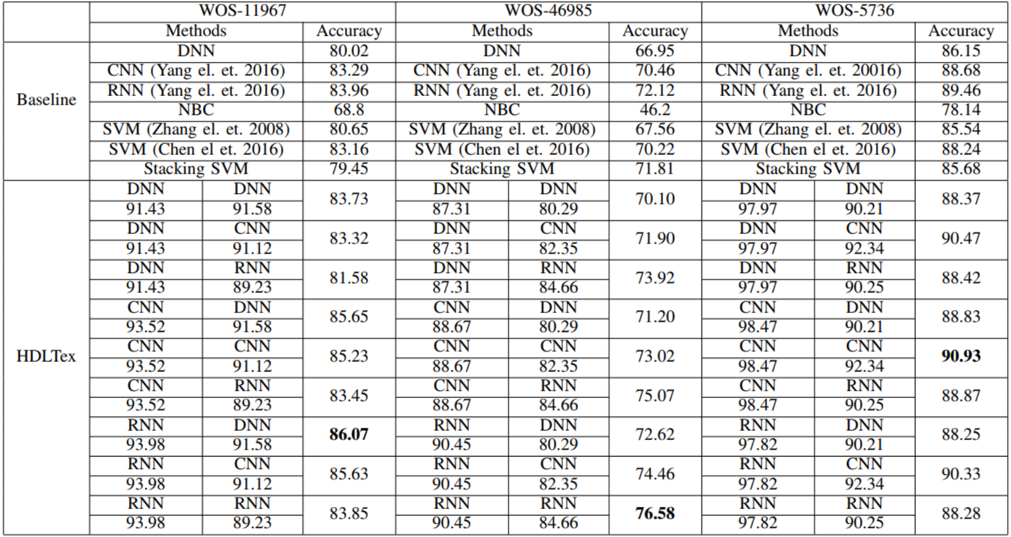
  - $K = 7$, $L_1 = 9, L_2 = 11, \dots, \sum_{k=1}^{K} L_k = 134$.

HDLTex



- Hierarchical deep networks:
  - For a document d_i, the parent-level network predicts the parent-level label $k \in \{1, \ldots, K\}$
  - Then the m-th child level network predicts the child-level label $l \in \{1, \ldots, L_k\}$

HDLTex



	WOS-11967			WOS-46985			WOS-5736		
	Methods		Accuracy	Methods		Accuracy	Methods		Accuracy
Baseline	DNN		80.02	DNN		66.95	DNN		86.15
	CNN (Yang el. et. 2016)		83.29	CNN (Yang el. et. 2016)		70.46	CNN (Yang el. et. 20016)		88.68
	RNN (Yang el. et. 2016)		83.96	RNN (Yang el. et. 2016)		72.12	RNN (Yang el. et. 2016)		89.46
	NBC		68.8	NBC		46.2	NBC		78.14
	SVM (Zhang el. et. 2008)		80.65	SVM (Zhang el. et. 2008)		67.56	SVM (Zhang el. et. 2008)		85.54
	SVM (Chen el et. 2016)		83.16	SVM (Chen el et. 2016)		70.22	SVM (Chen el et. 2016)		88.24
	Stacking SVM		79.45	Stacking SVM		71.81	Stacking SVM		85.68
HDLTex	DNN	DNN	83.73	DNN	DNN	70.10	DNN	DNN	88.37
	91.43	91.58		87.31	80.29		97.97	90.21	
	DNN	CNN	83.32	DNN	CNN	71.90	DNN	CNN	90.47
	91.43	91.12		87.31	82.35		97.97	92.34	
	DNN	RNN	81.58	DNN	RNN	73.92	DNN	RNN	88.42
	91.43	89.23		87.31	84.66		97.97	90.25	
	CNN	DNN	85.65	CNN	DNN	71.20	CNN	DNN	88.83
	93.52	91.58		88.67	80.29		98.47	90.21	
	CNN	CNN	85.23	CNN	CNN	73.02	CNN	CNN	**90.93**
	93.52	91.12		88.67	82.35		98.47	92.34	
	CNN	RNN	83.45	CNN	RNN	75.07	CNN	RNN	88.87
	93.52	89.23		88.67	84.66		98.47	90.25	
	RNN	DNN	**86.07**	RNN	DNN	72.62	RNN	DNN	88.25
	93.98	91.58		90.45	80.29		97.82	90.21	
	RNN	CNN	85.63	RNN	CNN	74.46	RNN	CNN	90.33
	93.98	91.12		90.45	82.35		97.82	92.34	
	RNN	RNN	83.85	RNN	RNN	**76.58**	RNN	RNN	88.28
	93.98	89.23		90.45	84.66		97.82	90.25	