Model Input	Training Words			Training Vocab (Unique Words)		
	Count	% Full Text	Avg. per Article	Count	% Full Text	Avg. per Article
Full Text	49.7M	100	672	87,541	100	260
Lead Paragraph	7.5M	15	104	40,309	48	61
Headlines	0.6M	1	8	11,150	15	6
Nouns	13.3M	27	180	64,998	80	110
Lemmas	49.7M	100	672	75,648	94	241

Table 1: Model Inputs

Desk	General Descriptor	Online Sections	Taxonomic Classifier
Foreign Desk	• Immigration and Refugees • Jews • Music • Religion and Churches	• World	• Top/News/World/Europe • Top/News/World/Countries and Territories/Austria • Top/Features/Arts/Music • Etc. (8 others)
Book Review Desk	• Books and Literature	• Arts • Books	• Top/Feature/Arts • Top/Features/Books • Top/Features/Books/Book Reviews
Classified		• Paid Death Notices	• Top/Classifieds/Paid Death Notices

Table 2: Possible Model Outputs

Cleaned Label Name	Original Label Name	Articles
book review	Book Review Desk	1,750
	Book Review Dest	1
business & financial	Business Desk	4
	Business World Magazine	8
	Business/Financial Desk	6,078
	Business/Financial Desk;	1
	Business/Financial desk	1
	Business/FinancialDesk	5
	Business\Financial Desk	2
	E-Business	6
	E-Commerce	17
	Financial Desk	11,276
	Financial Desk;	16
	Money & Business/Financial Desk	3
	Money and Busines/Financial Desk	1
	Money and Business/Financial Desk	939
	Moneyand Business/Financial Desk	1
	Small Business	12
	SundayBusiness	82
	The Business of Green	3
cars	Automobiles	122
	Cars	28

Table 3: Label Cleaning Examples

Model Input	**Model Type**			**Best Model**
	MNB	**LR**	**CNN**	
Full Text	**0.729**	**0.815**	0.190	LR
Lead Paragraph	0.682	0.731	**0.248**	LR
Headlines	0.622	0.647	0.173	LR
Nouns	0.713	0.786	0.185	LR
Lemmas	0.727	0.811	0.189	LR
Best Input	Full Text	Full Text	Lead Para	**Full Text in LR**

Table 4: Model Accuracies on Test Data

Model Input	**Vocab Size**	**Padding Size**
Full Text	207,070	500
Lead Paragraph	99,210	139
Headlines	30,658	11
Nouns	165,531	378
Lemmas	193,694	500

Table 5: CNN Vocabulary and Padding Sizes