Model Input	Training Words			Training Vocab (Unique Words)		
	Count	% Full Text	Avg. per Article	Count	% Full Text	Avg. per Article
Full Text	5,032,755	100	684	87,541	100	264
Lead Paragraph	740,321	15	103	40,309	46	61
Headlines	57,752	1	8	11,150	13	6
Nouns	1,349,489	27	184	64,998	74	112
Lemmas	5,032,755	100	684	75,648	86	245

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	176
business & financial	Business World Magazine	1
	Business/Finance Desk	1
	Business/Financial Desk	628
	Business\Financial Desk	1
	E-Commerce	1
	Financial Desk	1,106
	Financial Desk;	2
	Money and Business/Financial Desk	79
	SundayBusiness	14
cars	Automobiles	9
	Cars	4

Table 3: Label Cleaning Examples

Model Input	Model Type			Best Model
	MNB	LR	NN	
Full Text	**0.693**	0.742	0.	LR
Lead Paragraph	0.624	0.661	0.	LR
Headlines	0.502	0.545	0.	LR
Nouns	0.664	0.702	0.	LR
Lemmas	0.685	**0.743**	0.	LR
Best Input	Full Text	Lemmas	?	**Lemmas in LR**

Table 4: Model Accuracies on Test Data