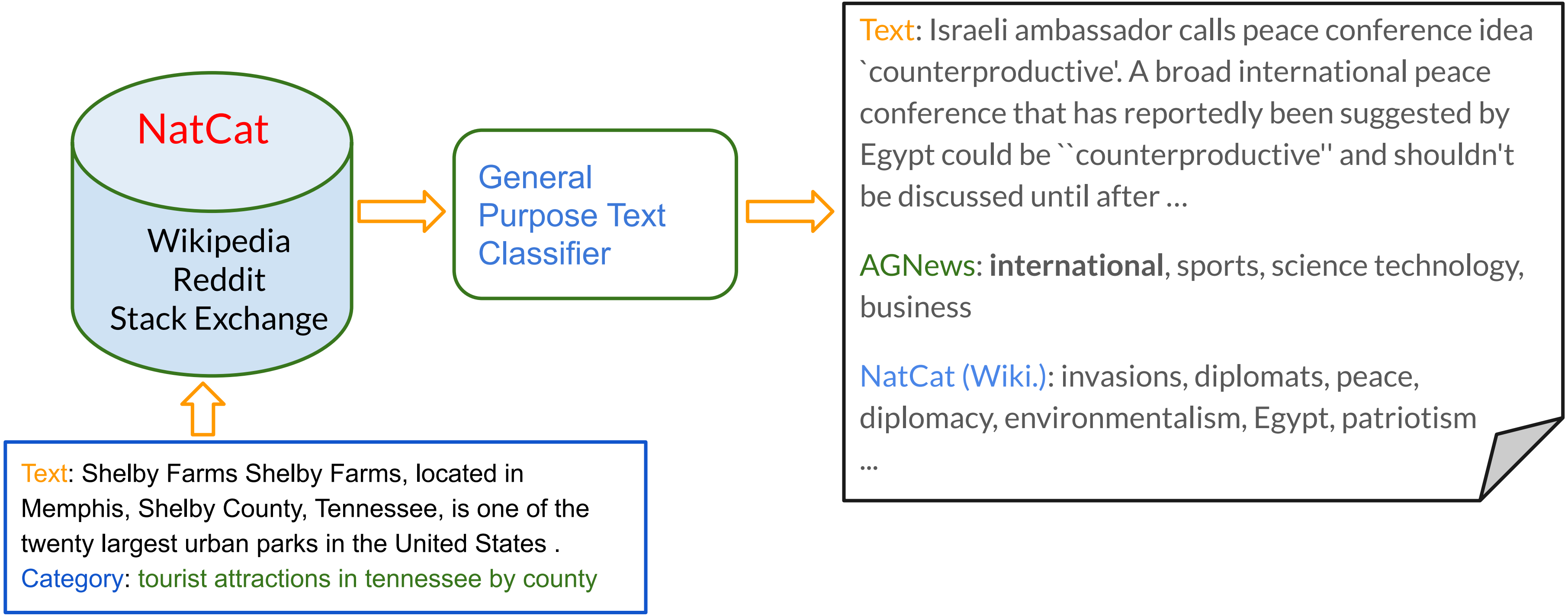
NatCat: Weakly Supervised Text Classification with Naturally Annotated Resources

Zewei Chu[1], Karl Stratos[2], Kevin Gimpel[3]

[1]University of Chicago [2]Rutgers University [3]Toyota Technological Institute at Chicago



Text: Shelby Farms Shelby Farms, located in Memphis, Shelby County, Tennessee, is one of the twenty largest urban parks in the United States .
Category: tourist attractions in tennessee by county

Text: Israeli ambassador calls peace conference idea `counterproductive'. A broad international peace conference that has reportedly been suggested by Egypt could be ``counterproductive'' and shouldn't be discussed until after ...

AGNews: **international**, sports, science technology, business

NatCat (Wiki.): invasions, diplomats, peace, diplomacy, environmentalism, Egypt, patriotism

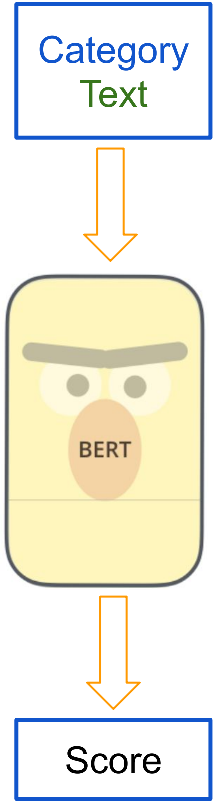
...

NatCat Resource Statistics

	Wikipedia	Stack Exchange	Reddit
# categories	1,730,447	156	3,000
# documents	2,800,000	2,138,022	7,393,847
avg. # cats. per doc.	86.9	1	1
mode # cats. per doc.	46	1	1
avg. # words per doc.	117.9	58.6	11.4

Strong performance of NatCat trained general purpose classifiers on CatEval tasks

	Topical (Acc)	Sentiment (Acc)	Multi Label (LRAP)	all
BERT	63.3	53.8	41.0	53.3
+ensemble	**63.8**	54.6	**41.4**	54.3
RoBERTa	61.3	55.8	40.5	53.4
+ensemble	62.4	**59.7**	41.2	**55.6**
ESA	47.1	41.2	29.7	40.2



Category
Text

Score