Extracting information about terrible news

A data provider extracts sensitive numerical and textual data from online news articles. This data has significant ramifications for the health and safety of much of the world’s population. About 15% of the documents they review involve circumstances where 100 people have died. In order to be able to enrich their database and act, they require data that is as “clean” and accurate as possible.

Currently, analysts manually review a large volume of articles and record relevant data. This manual method does not scale and the company needed to introduce automation to its data extraction operations.

There are several off-the-shelf Named Entity Recognition (NER) systems that can do this with varying degrees of accuracy – location is an example of an instance in which open source solutions are fairly good. However, most of the extraction tasks the company needs to perform aren’t available with off-the-shelf models.

High stakes data

The company automated a large portion of their process partnering with Idibon to create custom models for their specific data extraction needs. This custom model identified pieces of relevant data in news articles to save analysts a lot of time.

Idibon is able to deliver this state-of-the-art accuracy because it trains custom models on the same type of data that they are being used to classify. For example, in a couple of hours, analysts annotated enough documents to successfully distinguish between “people affected” and “people killed” with an accuracy above 75%. Accuracy continues to improve as more annotations come in. As Idibon prompts humans to judge, the most important documents confirm and improve the statistical models.

Going (semi-)automatic

Classifications from Idibon’s models enabled human analysts to focus their time on making these models more efficient. The information in these news articles is important enough that the data provider still wants one of their analysts to see every article; Idibon offers time savings to make this practical. In purely manual analysis, accuracy today doesn’t affect accuracy over time. With Idibon, the models improve with every document that an analyst evaluates.

Solution

Idibon automatically extracts relevant data from news articles for analysts to confirm with high accuracy and significant time savings.

Results

75%

Accuracy after two hours in distinguishing between “people affected” and “people killed”.

93%

Accuracy after two hours in identifying location of message.

60%

Reduction in analyst time after using Idibon’s technology.

A global data provider uses Idibon’s technology to extract sensitive data about health and safety from online news articles.