Document Classification: Machine Learning Vs. Rule-Based Methods

Data is growing at a phenomenal rate, for every type of organization. With that expansion comes a renewed need to make sure that sensitive personal and business information isn't exposed or in violation of the strict new compliance laws that are being enacted around the world.

In the first part of this series, I introduced the challenge of identifying and understanding the sensitivity of information within unstructured and structured data across the enterprise storage silos. How are companies defining what sensitive data is? What kind of technologies can they rely on to identify this information to help make sure that they aren't exposed to serious privacy breaches and compliance violations?

As you can guess, some pieces of information are more difficult to detect than others. The following sections will illustrate this point more precisely by outlining some approaches to detect, map, and categorize sensitive data: rule-based methods, and machine learning methods: supervised learning and unsupervised learning (clustering).

Document Classification with Rule-Based Methods

When it comes to detecting sensitive data, the naïve and straightforward approach is based on defining exact patterns of numbers, characters, or specific terms, also known as regular expressions. Once those definitions are in place, every piece of data containing such patterns can be detected directly. This practice is quite common when it comes to matching ID numbers, credit card numbers, and email addresses, for instance.

However, this method by itself is usually insufficient, for several reasons:

- **Manual effort**—Tailoring terms and rules requires a lot of human labor.
- **False positives**—Not every 9-digit number is a Social Security Number.
- **Lack of context**—The surrounding sentences are ignored when determining the validity of any given match. For example, a rule-based method will identify a telephone number, even one that appears in an ad for office furniture and is obviously not a threat to anyone's privacy.

Some data management solutions rely solely on this type of partial detection, while claiming to use "advanced artificial intelligence." That's just not accurate.

Document Classification with Machine Learning Methods

Document Categorization with Supervised Learning

To achieve a holistic and meaningful data mapping, the ability to automatically categorize files according to their content is a huge milestone. This is made possible with a specific type of machine learning: supervised learning.

Supervised Learning is the machine learning task of inferring a mapping between data inputs and outputs based on ground-truth samples of input-output pairs.

How it works in a nutshell: Natural Language Processing techniques such as bag of words and word embedding enable the transformation of each arbitrary piece of text into a fixed-size numerical vector representation. Then, machine and deep learning algorithms are trained with the input of labeled data samples—sets of texts (documents) and their corresponding labels (categories).

In other words, a supervised machine learning pipeline uses a small set of labeled data in an attempt to generalize the essence of the categorization task in order to correctly classify even never-seen-before sets of documents.

Out-of-the-Box Categorization
Many types of documents are common in organizations: invoices, NDAs, resumes etc. For an AI system to detect and identify such documents correctly:

- The documents' content (text or image) can be scanned and interpreted.
- No custom specification is required from the organization.

For this purpose, a ready-to-use machine learning based categorization mechanism can be trained to detect dozens of predefined categories.

Custom Categorization
In addition to typical document categories, every organization is also likely to have its own unique way to divide categories of interest. There's a flexible solution to do this that only requires each organization to initially train a supervised learning model that learns to differentiate and categorize according to the unique organizational policy, using uploaded documents as data samples for the model. Later, more categories and samples can be added to update the model's predictive ability over time. Yet, such a solution may introduce some further challenges for the organization:

- How to figure out which categories exist?
- Where to obtain training samples?

Any experienced data scientist would tell you that collecting and labeling training data is one of the most crucial yet demanding steps in developing machine learning solutions. Not surprisingly, even with frameworks that only require data samples, CIOs (Chief Information Officers) in charge of managing and protecting their enterprises' data sometimes find it difficult to gather a representative set of samples of documents from different departments.

But what if there was a way to effortlessly detect and group together potential samples that exist in the actual data? Just a peek at the results enables CIOs to identify document groups that they never knew they had to protect.

Document Clustering with Unsupervised Learning

There is another unguided document grouping made possible with machine learning methods. A holistic and meaningful mapping of a vast number of instances can be achieved using clustering, a form of unsupervised learning.

While in supervised learning the training data is labeled with relevant classifications, in unsupervised learning models are to learn relationships between data points and classify the raw data without guidance or “ground truth” provided.

To illustrate how this mechanism works in our domain, think for a moment about all the millions, or perhaps billions of documents in your organization. If all documents were represented by points in a vector space, so that similar documents are within a close distance from one another, a set of statistical algorithms can be used to group together such similar documents.
This is just a simplified look at this method, and of course it takes a delicate process to represent documents’ meaning in numerical vectors, so that the proximity between vectors genuinely reflects the similarity between the documents’ meaning.

Final Thoughts

In this post we looked at several approaches that organizations can take to document classification. Machine learning offers methods that can greatly improve on rule-based methods to detect, map, and categorize sensitive documents on an organizational level. It should be noted that when mapping sensitive information at enterprise scale with any of these methods, there are many instances that require organizations to take into account the different possible contexts that can be attached to pieces of information. For instance, not all sensitive pieces of information come in the size of a full document.

There’s a lot more to dive into here. In the final part of this series on how AI is changing data management and compliance, I’ll introduce the use of machine and deep learning to utilize context towards automating sensitive data classification.

And if you missed it, go back and read Part 1, where I introduce the concepts behind the growing need for AI-based data and compliance tools.