FORTRA

WHITE PAPER (Titus)

Smarter Data Protection with Machine Learning

Minimize Risk and Increase Confidence with More Accurate and Efficient Data Identification, Classification, and Policy Enforcement.

With the exponential growth of unstructured data, organizations are evaluating how to deliver better privacy protection and stronger defenses against cyberthreats. Concerns around highly visible data breaches around the world have fueled a strong demand for higher standards of reliability and transparency around the protection and handling of personal data. Governmental agencies and other organizations have also started to draft and enforce tighter data security regulations across industries.

Protecting sensitive personal and business data is a mission-critical task. Successful data protection solutions require a defined set of rules around how data is kept and handled, but many organizations struggle to accurately identify the vast amounts of data that moves through day-to-day workflows. It's difficult to protect something if you don't know what or where it is.

Manual processes for identifying, classifying, and protecting data are cumbersome when dealing with large volumes of data. In addition, many employees are unsure how to treat the various types of data they encounter, others are simply unfamiliar with the content. When processes around data handling and classification become time-consuming and unwieldy, most people find workarounds that can lead to data leakage and other security risks.

Not only is accurate handling important for the security of critical business and personal information, but it's also important for meeting regulatory requirements such as the Global Data Protection Regulation (GDPR). Noncompliance penalties can be up to €20 million, or 4 percent annual global turnover — whichever is higher — of an organization's worldwide annual revenue.

Digital tools that include machine learning capabilities, however, can help organizations meet all of these challenges efficiently and accurately. This paper explores how machine learning offers a way for organizations to improve data categorization and classification to better protect their sensitive data and comply with information rights regulations.

How Machine Learning Can Help

Machine learning is a type of artificial intelligence that uses analytical models to give computer systems the ability to learn from data, without being explicitly programmed or fed that data. Perhaps, as you are reading this, alarm bells are going off because machine learning has a reputation for requiring huge computing farms and millions of training samples. No need to worry — all machine learning is not created equal.

There are different types of machine learning, including supervised, unsupervised and deep learning, and each can be used in a number of different ways. This paper deals specifically with using supervised machine learning for data categorization.

Supervised machine learning is ideal for data categorization as it helps avoid the potential risk of bad actors misinforming the model. Data classification can be done by an analytical model rather than by individual user input, which has the potential for greater error.

In addition, supervised machine learning requires only a couple hundred data samples to create a model instead of the millions of samples required with other types of machine learning. Supervised machine learning also puts organizations in charge of building their own training model to fit their unique business and protect their proprietary and sensitive information.

Machine Learning Categorization

Machine learning categorization can help users make better decisions about how information is handled with automated or recommended data classifications based on organization-specific categories and policies. This technology leverages proven algorithms to build a company-centric model and predict different categories of email and documents. Based on the model, the program either suggests or automatically applies classifications on unknown documents, enabling greater security and staff productivity.

Some types of data — social security numbers, banking information, personal health information — are obviously highly sensitive. Other types are not as clear cut. What about business plans, product documentation and other types of intellectual property? The lack of consistent identification, classification and protection for similar types of data poses a real security risk. That's why context is so important and why the combination of people and technology is ideal for keeping information protected. Machine learning categorization can identify different types of data, such as email or documents, as well as words and unique phrases within each data type that help provide the context. For example, the wording in a press release would be quite different from that found in product documentation.

Using supervised machine learning for categorization allows organizations to efficiently train a model, or corpus, for data identification, without requiring massive resources or data. Basically, you feed your software categorized examples and it learns to recognize those similarities in documents to inform classification suggestions. This methodology reduces risk by improving accuracy and efficiency of data identification, classification and protection. Eventually, when accuracy confidence levels increase, machine learning can provide automatic data identification and classification for an additional layer of security.

For example, let's say someone forwards an email to an external stakeholder, not realizing a draft financial statement is attached. Based on a predefined analytical model, data classification technologies embedded in the user's email workflow would alert the sender that they are about to send out sensitive and not-yet-classified information. The program would also suggest a potential classification, along with an estimated accuracy level. The user can either select "Send anyway" or check to be sure that the classification is correctly set on the asset attached to the email.

As organizations deploy their machine learning models and gather data on when users override or trust the algorithms, administrators can further refine their models and increase confidence. As confidence levels in predictions rise, organizations typically opt to automate more classifications, improving the overall efficiency of data protection.

As another example, consider a business acquisition. Information related to this acquisition could include stock symbols, purchase price details and the public announcement date. The entire acquisition project might be code named "Project Chicago." All of these details are often called "material nonpublic information" (MNPI). It would be difficult for anything but a human or machine learning technologies to distinguish the MNPI associated with this acquisition from details related to a flight to Chicago, since they both would include the word "Chicago," a dollar amount, dates and a symbol that could be difficult to understand as a stock symbol versus an airline confirmation code.

Machine learning categorization can make such distinctions and then automatically classify data based on rules or policies.

How It Works

Machine learning categorization integrates with tools users are already familiar with and builds on existing data policies, which makes the learning curve very short.

First a designated data steward seeds a training model, or corpus, of unstructured data with accurate data categories. The starting point is with clearly defined historical data for each category type, including email organized in different categories of mailbox folders and documents organized in file folders in a file directory. Data for the training model can also be "crowdsourced" from many employees using emails collected and tagged with categorization metadata in a single mailbox folder or using documents collected and tagged with categorization metadata in a single file folder.

Typically, the data steward creates and edits a config file to help with model training and evaluation. The training model can be created with anywhere from 50 to 250 data examples in each category to train and evaluate the corpus.

The data steward uses machine learning categorization tools to process the sample set — evaluating, interpreting results and refining the corpus where necessary. The refined model is then integrated into user workflows.

Running seamlessly behind the scenes, machine learning categorization predicts the category of unknown data and helps users make better decisions when handling information in their day-to-day workflow. This technology gives users additional data protection and support without additional effort or training.

Machine learning categorization enables smarter data identification by providing a more consistent, accurate and efficient way to identify, classify and secure data within the flow of work. Intelligent content analysis is trained to recognize company-specific categories and apply the appropriate policies.

Alerts allow users to respond to suggested categories, either accepting them or designating another. These responses trigger a feedback and reporting mechanism within the machine learning system, which enables organizations to refine and enhance their data models to continuously improve accuracy and adapt to changing needs.

An organization's policy engine uses the machine learning categorization runtime with model files to predict an email or document's categorization and provide a confidence level. Users evaluate the text of an email or document against the model's suggestions to ultimately define a categorization and confidence level. The system responds by mapping it to the suggested classification based on the policy defined.

As confidence and acceptance of machine learning categorization grows, some organizations may choose to automate aspects of the process, eliminating manual responses to suggestions.

The Nitty Gritty Matters

Content identification is key to handling information correctly. But false positive identification can detract from the effectiveness of the solution. Some technologies out there can help check content, but they often work from a predefined library or regular expression builder that identifies a combination of mathematical or logical operators, constants, functions, table fields, controls and other properties. These methods can work well for finding specific patterns of characters; however, they are not comprehensive. Other technologies run content checking for just a few seconds, and anything not checked within that time frame is skipped.

Machine learning categorization technologies that work from user-defined models are more comprehensive and more sensitive to the unique context of a particular organization.

Conclusion

Machine learning categorization gives organizations greater confidence in the accuracy and consistency of their data classification systems — and ultimately greater peace of mind that their sensitive data is protected and that the company is in compliance with security regulations.

Establishing a policy-driven foundation to help facilitate the identification and classification of sensitive data at creation, in motion or at rest allows organizations to apply the right level of protection. With a configurable policy management platform, organizations can automatically apply policies based on classification and categorization. Titus Intelligent Protection, powered by machine learning, enhances the effectiveness of data security programs by improving the accuracy and efficiency of data identification, minimizing risk and increasing confidence.

Titus solutions integrate easily with an organization's existing software systems, giving them the freedom to deploy the technologies that best fit their business requirements without disrupting user workflows. Security solutions work better together, enabling consistent policy enforcement and unlocking the value of an organization's technology investments.

FORTRA

Fortra.com

About Fortra

Fortra is a cybersecurity company like no other. We're creating a simpler, stronger future for our customers. Our trusted experts and portfolio of integrated, scalable solutions bring balance and control to organizations around the world. We're the positive changemakers and your relentless ally to provide peace of mind through every step of your cybersecurity journey. Learn more at fortra.com.