proofpoint.

TECHNICAL BRIEF

Proofpoint AI Technologies That Protect Sensitive Data

This document describes the artificial intelligence (AI) technologies used in Proofpoint Intelligent Classification and Protection. We leverage both general and Proofpoint-proprietary AI technologies. In particular, we employ proprietary technologies to address data privacy management concerns and accelerate data privacy compliance.

With Proofpoint's AI-powered classification, your authorized users get a complete view of organizational data. They can see this data by confidentiality level, business category and other dimensions. This way your organization can better detect, manage and control unauthorized access and changes to sensitive data. This level of visibility and control lets you:

- Enhance your access control framework
- Improve the security of your data loss prevention (DLP) and information rights management (IRM) solutions
- Ensure data availability on demand to authorized users and third parties
- Reduce the risk of data breaches
- Protect your brand
- Facilitate safe cloud adoption

Our AI technologies can be grouped into four categories: Data Engineering, Natural Language Processing, Machine Learning and Deep Learning. Each category and their AI technologies are described in more detail below.

Data Engineering

Proofpoint's data engineering AI technologies make it easier to collect, store and analyze data. They analyze textual and numerical data from documents and databases automatically. They also detect language from the content, folders, file names and database tables. They clean and match the document data. And they tokenize and lemmatize text. To lemmatize means to sort words by grouping inflected or variant forms of the same word.

Data Engineering Technologies

TECHNOLOGY	GENERAL/PROPRIETARY	DESCRIPTION
Optical Character Recognition (OCR)	General	This converts a scanned image into normal raw text. The normal text is then cleaned and converted into "prepared data."
Feature Engineering	General	This extracts meaningful features from prepared contextual data and metadata. It transforms the features into a numerical vector used in downstream tasks. The system then deletes all source information. It only keeps numerical representations to assure the highest level of security.

Natural Language Processing

Proofpoint uses natural language processing (NLP) to transform, aggregate and generate textual data into meaningful information.

Natural Language Processing Technologies

TECHNOLOGY	GENERAL/PROPRIETARY	DESCRIPTION
Probabilistic Context-Based Modeling	Proprietary	This uses context to determine if data is personal as well as the probability that the data is personal. It also provides explainability to the extracted information.
Text Summarization	Proprietary	This is used for longer text documents or groups of documents. It extracts the most important sentences and phrases from a set of documents. It then creates salient features so that unsupervised machine learning (ML) algorithms can understand the meaning of the text document topic.

Machine Learning

Proofpoint developed a family of unsupervised machine learning (ML) technologies to label data with no human supervision. These include smart sampling, clustering and autolabeling. Unsupervised ML reduces costs associated with manual labeling. And they provide flexibility with regard to languages, classification types and data nature.

Machine Learning Technologies

TECHNOLOGY	GENERAL/PROPRIETARY	DESCRIPTION
Smart Sampling	Proprietary	This allows the AI model to process large amounts of data by choosing smaller sample sizes of data. These samples contain a balance of document types, topics and data.
Autolabeling	Proprietary	This predicts the business category and level of confidentiality for document clusters.
Data Loss Prevention	Proprietary	Our DLP-dictionary generation improves the protection and security of your DLP solutions. It mitigates data breaches. It drastically reduces the number of false positives. And it improves precision when detecting insensitive documents. Proofpoint Intelligent Classification and Protection generates the most effective combinations of keywords to match the documents related to any DLP rule. It does this through automatic keywords and keyphrase extraction, which can be either positive or negative. Dictionaries include thresholds, keywords weights and other parameters. The technology is based on NLP and ML algorithms. It can work without labeled datasets. PCBM and dictionaries are optional. And they are required only if generating dictionaries to enhance your risk management systems.
Supervised Machine Learning	Proprietary	This is a class of algorithm that assigns labels to new elements. New elements have not been seen before. They also have not yet been analyzed. A supervised learning algorithm learns from labeled training data to help predict outcomes for unforeseen data.
Active Learning	Proprietary	This is used in the classification review workflow. The workflow includes a stage where a key user can validate or challenge the model predictions. When the AI identifies new labels, the system applies these labels and computes a confidence level for each document. This level determines if the key user needs to review the document. If so, the key user reviews a sample of data output and correctly labels the documents. The process can be iterative. It can be done until a high confidence level is reached. Once a document is labeled correctly, the model is retrained with the new sample.
Transfer Learning	Proprietary	This is an approach where knowledge learned in source tasks, such as public knowledge bases, is transferred and used to improve the learning of a related business task. It helps fine-tune massive language models on domain-specific documents. It also is used to extract special types of personal data such as religion, nationality and passport. This kind of data is not learned on named entity recognition (NER) tasks.

Deep Learning

Proofpoint employs both supervised and unsupervised deep learning in its AI.

Deep Learning Technologies

TECHNOLOGY	GENERAL/PROPRIETARY	DESCRIPTION
Unsupervised Deep Learning	General	This is used in feature engineering, clustering and language modeling to perform personal data extraction, purpose of processing prediction and data linking.
Supervised Deep Learning	General	These algorithms use different artificial neural networks to perform the classification tasks by different dimensionalities. These include hierarchical business category prediction, confidentiality level and others.

About Proofpoint Intelligent Classification And Protection Information And Cloud Security

Proofpoint Intelligent Classification and Protection is our AI-powered data discovery and classification solution. It delivers petabyte-scale data classification and protection with accuracy, efficiency and speed. It takes inventory of all of your content wherever it may reside. Its AI engine analyzes and classifies that content. It can then recommend how best to prioritize its protection.

Proofpoint Intelligent Classification and Protection mitigates your data security risk. It gives you complete visibility into and control of your business-critical data. It augments your existing DLP programs. And together with Proofpoint Information Protection solutions, it can be a solid part of your people-centric strategy to protect against data loss caused by careless, compromised and malicious users.

Our Proofpoint Information and Cloud Security platform delivers people-centric visibility and access controls across email, cloud, web and endpoint. It offers world-class threat, content and behavior detection. These address DLP, insider-risk and secure access use cases. The platform is built on a scalable, modern cloud architecture. It features a unified administration and response console. And it offers sophisticated analytics to simplify operations and shorten response times.

LEARN MORE

For more information, visit **proofpoint.com**.

ABOUT PROOFPOINT

Proofpoint, Inc. is a leading cybersecurity and compliance company that protects organizations' greatest assets and biggest risks: their people. With an integrated suite of cloud-based solutions, Proofpoint helps companies around the world stop targeted threats, safeguard their data, and make their users more resilient against cyber attacks. Leading organizations of all sizes, including 75 percent of the Fortune 100, rely on Proofpoint for people-centric security and compliance solutions that mitigate their most critical risks across email, the cloud, social media, and the web. More information is available at www.proofpoint.com.