Simplifying data classification for businesses using AI

ONE BEYOND

Contents

Introduction 3

Data classification can be daunting 5

How does AI information classification work? 7

Training AI to classify data: the process 9

Choosing the right algorithm 11

How to train the algorithm 13

Evaluating the effectiveness of your AI model 15

Data classification is the process of organising data into different categories or taxonomies based on characteristics. This includes things like data type, sensitivity level, or other metadata.

For a majority of businesses, irrespective of industry, data classification allows them to properly identify, secure and manage their data assets.

With the sheer amount of data that companies need to handle, effective data classification is more critical than ever - but more complex too.

However, without effective data handling, organisations are less likely to comply with data privacy regulations and leave themselves open to risks associated with data breaches or misuse. Poor data management also results in wasted data storage and slower processing.

Businesses that have a strong data classification strategy achieve responsible data governance and stewardship, allowing them to uncover the full value of their data while ensuring its security and proper handling.

Data classification can be daunting

So why is data handling and classification time-consuming and complicated? Well, imagine you're running a busy email server. Every day, millions of emails flood in on various topics - work documents, travel reservations, birthday greetings and more. It's a mess. But sorting through them to find what you need is like searching for a needle in a haystack.

This is where AI comes in. Powering data classification with AI is like hiring a team of super-efficient assistants. Each assistant can scan an email and instantly recognise its content - is it an invoice, a social media notification, or a confidential report? They then sort the emails into folders based on these classifications. This way, you can quickly find the information you need without sifting through tons of irrelevant messages.

How does AI information classification work?

AI data classification uses artificial intelligence to automatically organise and sort data into different categories based on predetermined rules or patterns. It transforms how data is managed by allowing huge amounts of information to be classified quickly and accurately, overcoming the challenges of doing this process manually which is extremely time-consuming and prone to human errors.

By leveraging AI to classify their data, organisations can efficiently retrieve and analyse the right information when they need it. This empowers businesses to make smarter, more informed decisions by getting them the correct data in a timely manner. AI data classification is becoming essential as companies look to get maximum value from the vast and continuously growing amounts of data they collect while ensuring proper data governance.

However, without an efficient training process, AI data classification systems won't be as efficient as businesses need them to be.

Training AI to classify data: the process

We've all heard the analogy 'garbage in, garbage out' and this applies to data. AI models operate using pattern recognition so if the data is messy, poorly labelled or inaccurate, the likelihood is that you'll find yourself with inaccurate data classifications.

So, before AI can be trained, the data needs to be prepared. Preparing quality data inputs can be approached in four stages.

Stage	Description	Benefit
Data Collection & Assessment	Identify data sources and potential issues.	■ Identify data sources (internal databases, surveys, etc.) ■ Check for inconsistencies or biases based on source. ■ Clean data: remove duplicates, fix typos, ensure consistent formats. ■ Use data visualisation tools to explore data for missing values, outliers, or unexpected patterns.
Data Labelling & Validation	Define categories and ensure accurate labelling.	■ Define clear classification categories for the AI. ■ Label data points with appropriate categories (manual by experts or crowdsourced). ■ Have a separate team validate labelled data for accuracy and consistency. ■ Resolve disagreements in labelling.
Addressing Data Quality Issues	Fix missing values, outliers, and bias.	■ Decide how to handle missing values (remove, impute, create separate categories). ■ Analyse outliers to determine if errors or genuine data. Remove or create specific categories. ■ Identify and mitigate bias in data sources or labelling demographics. ■ Balance data sets if necessary.
Data Preprocessing & Feature Engineering	Prepare data for the AI model.	■ Preprocess data: convert text to numbers, scale numerical data, normalise text formats. ■ Create new features from existing data that improve classification.

Choosing the right algorithm

The next step is deciding which algorithm you need to use to handle your data. There are a few factors you need to consider when choosing a data classification algorithm: number of categories, data imbalance, dimensionality, non-linear relationships, feature needs, data types, prediction speed, interpretability, and training data amount.

There's no single 'best' algorithm for all data classification tasks. Experimenting with different algorithms based on the factors mentioned above and evaluating their performance on your specific data is often the best way to find the optimal choice.

However, here are some examples of popular algorithms used for data classification and their strengths and weaknesses.

Algorithm	Strengths	Weaknesses
Logistic Regression	Simple, interpretable, good for binary classification	Limited to linear relationships, not ideal for complex data
Decision Trees	Easy to interpret, handles various data types	Can be prone to overfitting, accuracy might be lower than complex models
Support Vector Machines (SVM)	Powerful for high-dimensional data, clear separation between categories	Can be computationally expensive for large datasets
Naive Bayes	Efficient for large datasets with simple feature independence assumptions	Assumes features are independent, might not be suitable for highly correlated data
K-Nearest Neighbors (KNN)	Easy to implement, works well for some datasets	Performance can decline with high-dimensional data, requires choosing the right number of neighbours (K)

How to train the algorithm

Broadly speaking, there are three different learning approaches that categorise algorithms based on how they are trained and the type of data they work with:

- Supervised learning
- Unsupervised learning
- Reinforcement learning

Supervised learning is when the AI algorithm learns from labelled data to predict outputs for new data. It's trained on examples showing the right answer, so it can figure out the mapping from inputs to outputs. After training, it can predict outputs for data it hasn't seen before. Common uses are image classification, spam filtering and predictive modelling.

Unsupervised learning is when the AI algorithm finds patterns in data without being told the 'right answer.' Instead of learning from labelled examples, they identify similarities and differences in the data itself. Techniques include clustering similar data points together and simplifying data through dimensionality reduction. Uses include customer segmentation and anomaly detection.

Reinforcement learning is when the AI algorithm learns by trial and error which actions get the highest reward over many iterations. It gets feedback signals as rewards and penalties for the actions it takes. This allows it to solve complex decision-making tasks like game playing and robotics where no existing training data is labelled with correct outputs. The algorithm figures out the best sequence of decisions through repeated experience.

The more you feed training data to your data classification model, the better it gets at learning the patterns you want it to recognise and organise. Over time, you can identify any unexpected behaviours and adjust them to fit your goals. This way, you can fine-tune the program to act exactly how you want it to.

Evaluating the effectiveness of your AI model

To get a well-rounded picture of how your AI model performs, you need to assess how well the model handles unseen data using metrics like accuracy, precision, and recall. This reveals how well the model generalises to real-world scenarios.

Additionally, you need to consider factors like interpretability - can you understand how the model arrives at its classifications? This is essential for classification tasks that require trust and explanation.

Finally, analyse the model's performance across different categories, especially in imbalanced datasets, to ensure it's not biassed towards more frequent classes. By considering these factors, you can gain a comprehensive understanding of your model's strengths and weaknesses, allowing for targeted improvements and ensuring it meets your classification goals.

Building an effective AI model for data classification is a complex task with many challenges. It requires high-quality data to train the model and experience of selecting the most appropriate algorithm. Even after development, there's an ongoing need for monitoring and improvement.

To ensure internal teams can focus on core business functions, many organisations outsource this process to a team of specialists. At One Beyond, our team can handle data cleaning, labelling, and feature engineering, building a solid foundation for the model. We work with you to choose the optimal algorithm and guide the training process to avoid common pitfalls like bias and weak data security.

Get in touch with the team today and let's discuss your data management requirements.

WEBENQUIRIES@ONE-BEYOND.COM

01252 902704

ONE
BEYOND