Teaching Machines to Read

Computational Linguistics is the intersection of Computer Science and Human Language. Our goal: build systems to process, analyse, and generate language.

Text Categorisation (TC) — A foundational CL task: automatically assigning text a label or category.

Real-World Applications: Spam filtering, sentiment analysis, news tagging.

> "Teaching machines to read, label, and understand language."

Why Categorise Text? The Data Explosion

The Challenge

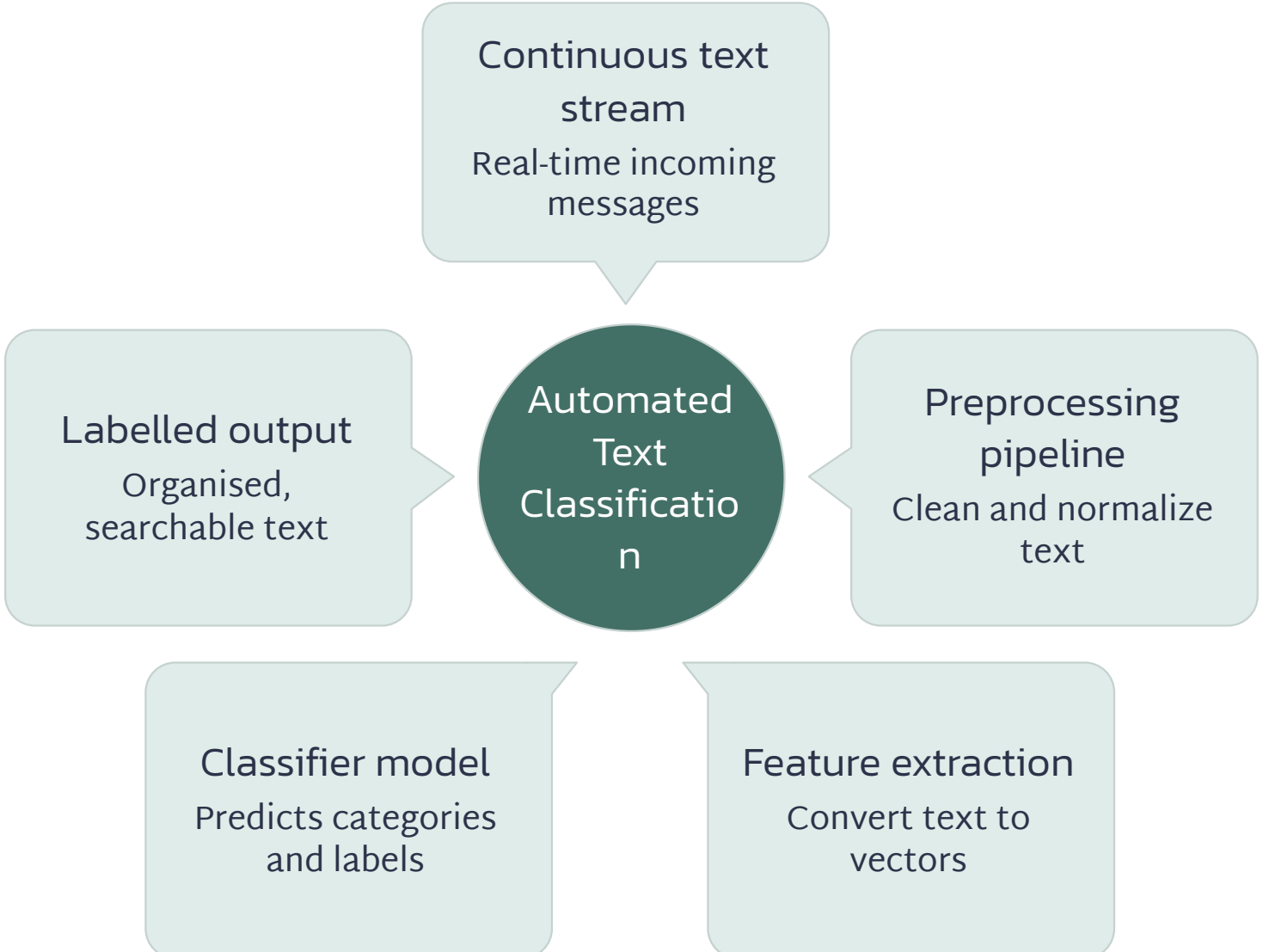
Millions of messages, reviews, and articles generated every second. Impossible for humans to process or organise at scale.

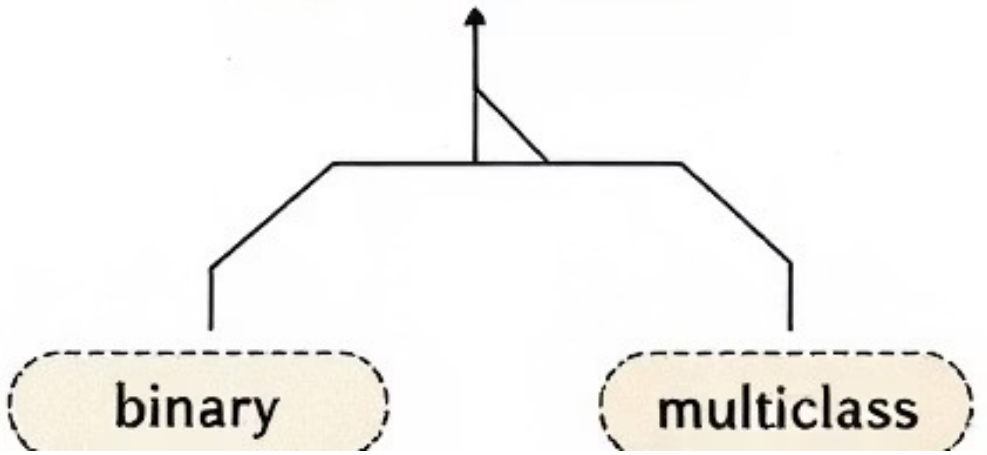
The Need

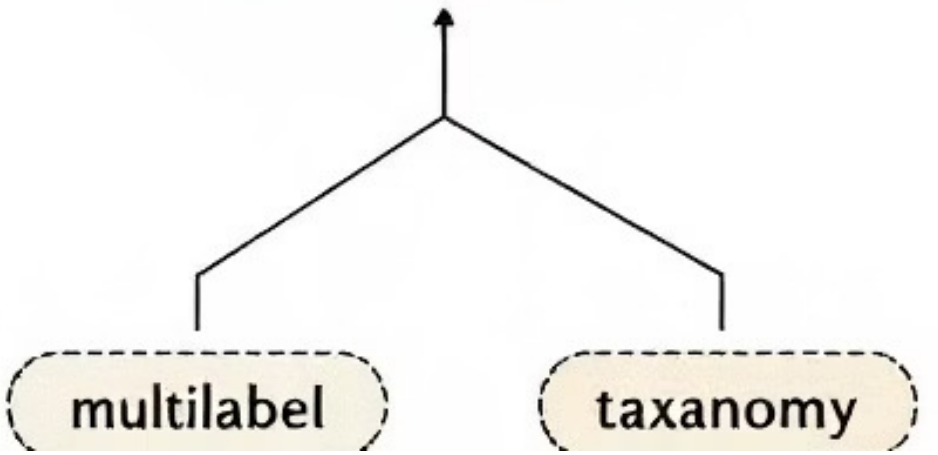
Automation is mandatory to manage, organise, and derive value from data at scale.

The Value

Business: instant customer service routing. Social media: brand perception tracking. Medicine: faster clinical note retrieval.







Types of Text Categorisation

1	2	3
Binary Classification Two categories only. Example: Spam or Not Spam. The simplest form of TC.	Multi-Class Classification Many categories, text belongs to exactly one. Example: Sports, Politics, or Technology.	Multi-Label Classification Text can belong to zero, one, or many categories simultaneously. Example: "AI in Healthcare" → {Tech, Health}.

Step 1: Text Preprocessing

Goal: Clean, normalise, and tokenise text for the machine to understand.

1 **Lowercasing & Punctuation Stripping**

Ensures "Hello!" equals "hello" for consistency.

2 **Stopword Removal**

Eliminate low-value words (the, a, is) that add noise without meaning.

3 **Normalisation: Stemming & Lemmatisation**

Stemming: crude chopping (running → run).
Lemmatisation: linguistic root word (ran → run). More accurate.

4 **Tokenisation**

Breaking text into discrete units (words, subwords, or characters).

Step 2: Feature Extraction—Words to Numbers

The Challenge: Machine learning models only understand numbers (vectors).

Bag-of-Words (BoW)

Document is a 'bag' of words; word order is ignored. Vector = [Count of word 1, Count of word 2, ...]

TF–IDF

Weights words by importance. Formula: tfidf(t,d) = tf(t,d) × log(N/df(t))

Word Embeddings (Word2Vec, GloVe)

Dense, continuous vectors capturing semantic meaning. Similar words map to nearby vectors. Example: King – Man + Woman ≈ Queen.

The Classic ML Pipeline & Baseline Models

Classic Models (Baselines):

- Naïve Bayes

  Fast, probabilistic, strong baseline for text classification.

- Logistic Regression

  Simple linear model for binary and multi-class problems.

- Support Vector Machine (SVM)

  Finds the maximal separation boundary between classes.

Transition: We've prepared our text and features. Now, let's see how these machines actually learn, starting with the mathematics of Naïve Bayes.

Probabilistic Model: Naïve Bayes

Core Idea: Classification based on Bayes' Theorem. Assumption: features (words) are independent given the class (the "Naïve" part).

Formula Core:

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)}$$

Where c is class, x is features. For text: $P(c|words) \propto P(words|c)P(c)$

- Strengths

  Interpretable, computationally efficient, works well with limited data.

- Limitations

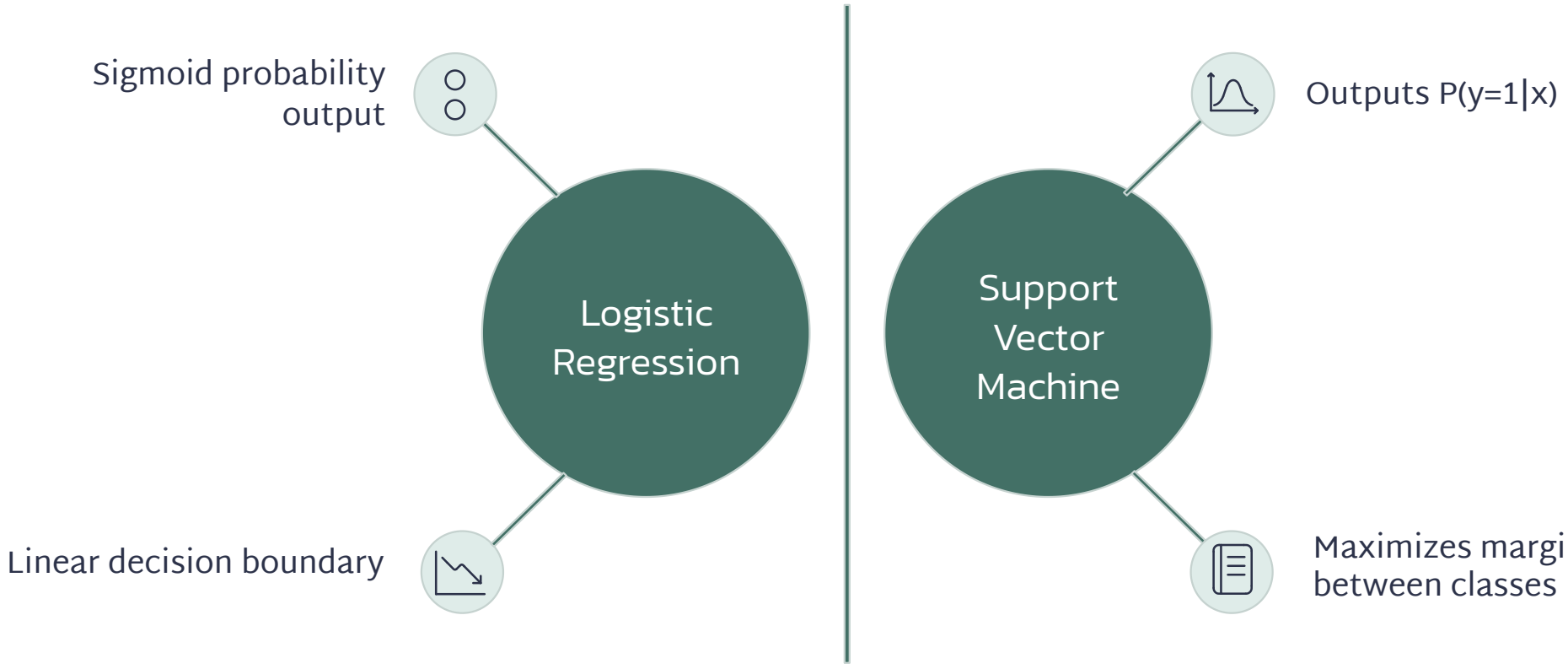
  Independence assumption rarely holds; struggles with feature interactions.

Discriminative Models: Regression & Margins

Logistic Regression: A linear model using the Sigmoid function to output a probability between 0 and 1.

$$P(y=1|x) = \sigma(w^T x + b) = \frac{1}{1 + e^{-(w^T x + b)}}$$

Support Vector Machine (SVM): Finds the optimal hyperplane maximising margin between classes.



Discriminative Advantage

Directly models decision boundaries rather than class distributions.

Practical Impact

Often outperforms generative models on well-separated datasets.

The Deep Learning Revolution

CNNs for Text (Local Patterns)

Filters capture n-grams and key phrases. Effective for short-range dependencies.

RNN / LSTM (Sequential Context)

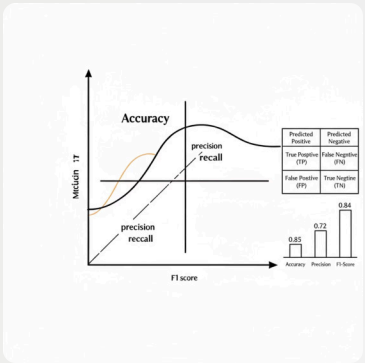
Process text sequentially; hidden state h_t captures context:

$$h_t = \sigma(W x_t + U h_{t-1} + b)$$

Transformers (BERT) – The Game Changer

Self-Attention mechanism: each word weights importance of all others. Formula: $\text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$

Transformers revolutionised NLP by capturing long-range dependencies and enabling massive parallelisation.



Evaluation Metrics & The Path Forward

Confusion Matrix: Tabulates actual vs. predicted labels (TP, TN, FP, FN).

Accuracy

Overall correctness across all classes.

$$\frac{TP + TN}{TP + TN + FP + FN}$$

Precision

Trustworthiness of positive predictions.

$$\frac{TP}{TP + FP}$$

Recall

Model's ability to find all actual positives.

$$\frac{TP}{TP + FN}$$

F1 Score

Harmonic mean; standard single metric balancing precision and recall.

$$2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

The Future: From ambiguity and sarcasm to data imbalance, domain adaptation, and ethical bias mitigation. Transfer learning, zero-shot classification, explainable AI, and multi-modal models represent the evolution of text categorisation toward human-level understanding.

Challenges & Limitations

Ambiguity & Sarcasm

Text understanding is complex; models struggle with nuanced language, sarcasm, and figurative speech.

Data Imbalance

Skewed datasets lead to models biased towards majority classes, underperforming on rare categories.

Domain Adaptation

Models trained on one domain often perform poorly when applied to different textual contexts.

Ethical Issues

Bias in training data can lead to unfair or discriminatory predictions, requiring careful mitigation strategies.

Modern Trends: The Future is Transferable

01

Transfer Learning

Leveraging pre-trained models on large datasets (like BERT) and fine-tuning them for specific tasks, dramatically reducing training time and data requirements.

02

Zero-Shot Classification

Enabling models to classify text into categories they haven't seen during training, based on semantic understanding.

03

Explainable AI (XAI)

Developing methods to understand why a model makes a particular classification, increasing transparency and trust.

04

Multi-Modal Models

Integrating text with other data types (e.g., images, audio) to provide richer context and more comprehensive understanding.

> "From simple word counts to billion-parameter transformers—text categorisation continues to be at the heart of Computational Linguistics."