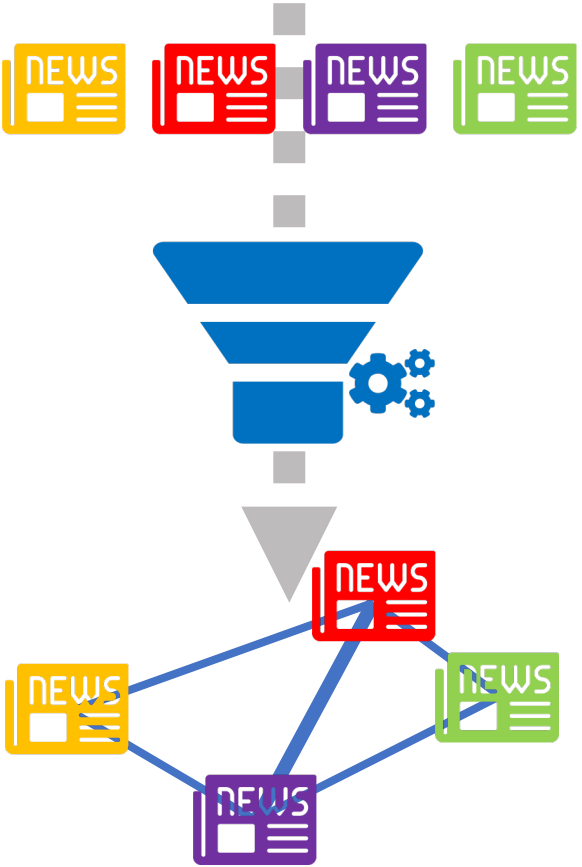

Stage	Description	Inputs	Outputs
retrieval	Articles need to be retrieved from multiple sources	Public datasets, news feeds, news aggregators	Articles (headline, body, source, URL, publication date)
processing	NLP pipeline: tokenisation, parsing, embedding	Articles	Documents with: - Parsing, segmenting - Embedded representation
linking	Similar articles are linked as they cover the same event, and similar sentences are recognised across documents	Processed articles	Similarity links between - Articles - sentences