Prepared by group 1

NLP-Based Insights Discovery for Industrial and Service Improvement

10 March, 2025

Introduction

Industries generate vast amounts of unstructured maintenance data, making it difficult to extract useful insights. This study explores how Natural Language Processing (NLP) can analyze maintenance reports to improve asset performance and service efficiency.

Research Question:

How can NLP be used to analyze maintenance reports for industrial asset and service improvement?

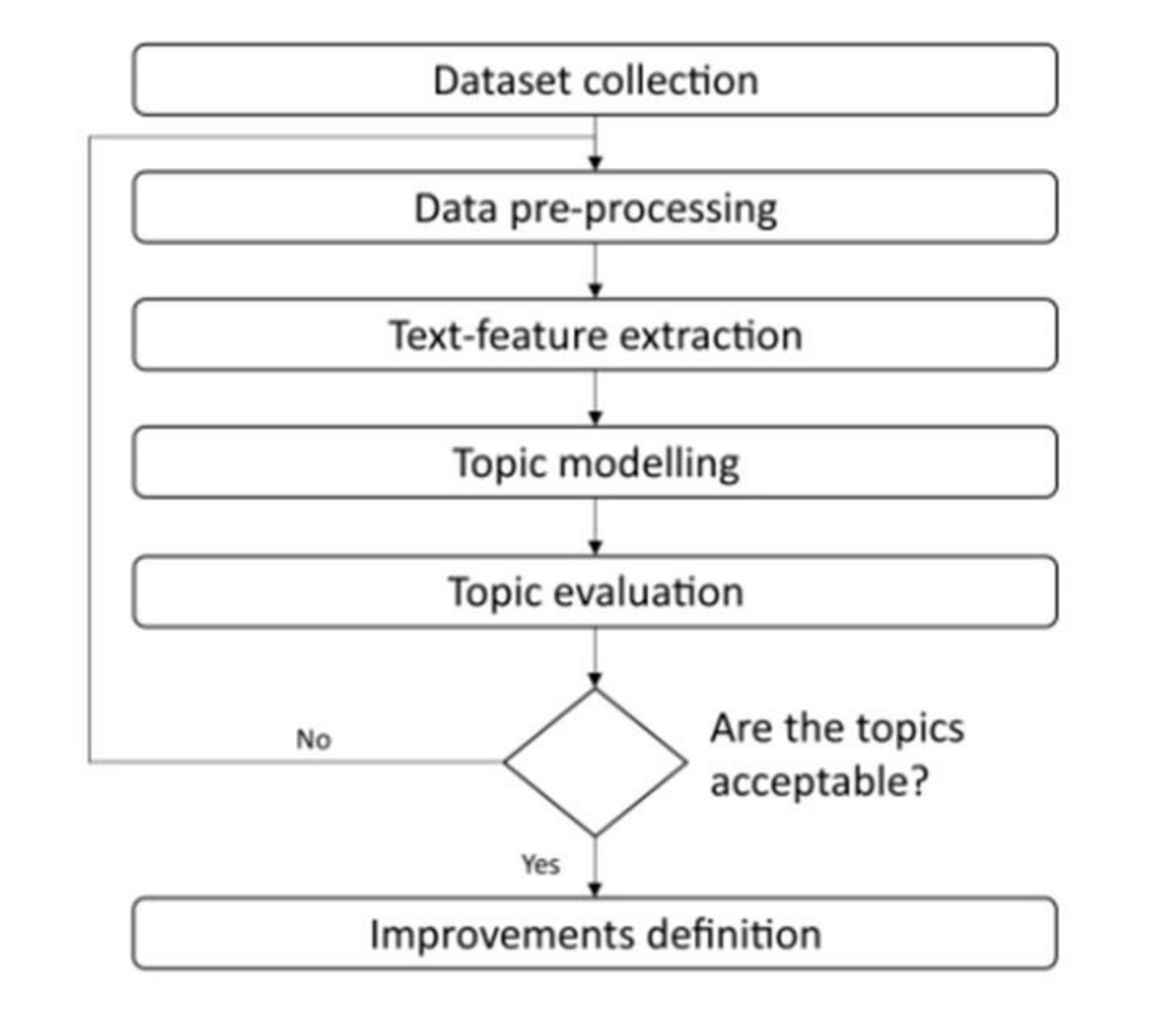
Objectives:

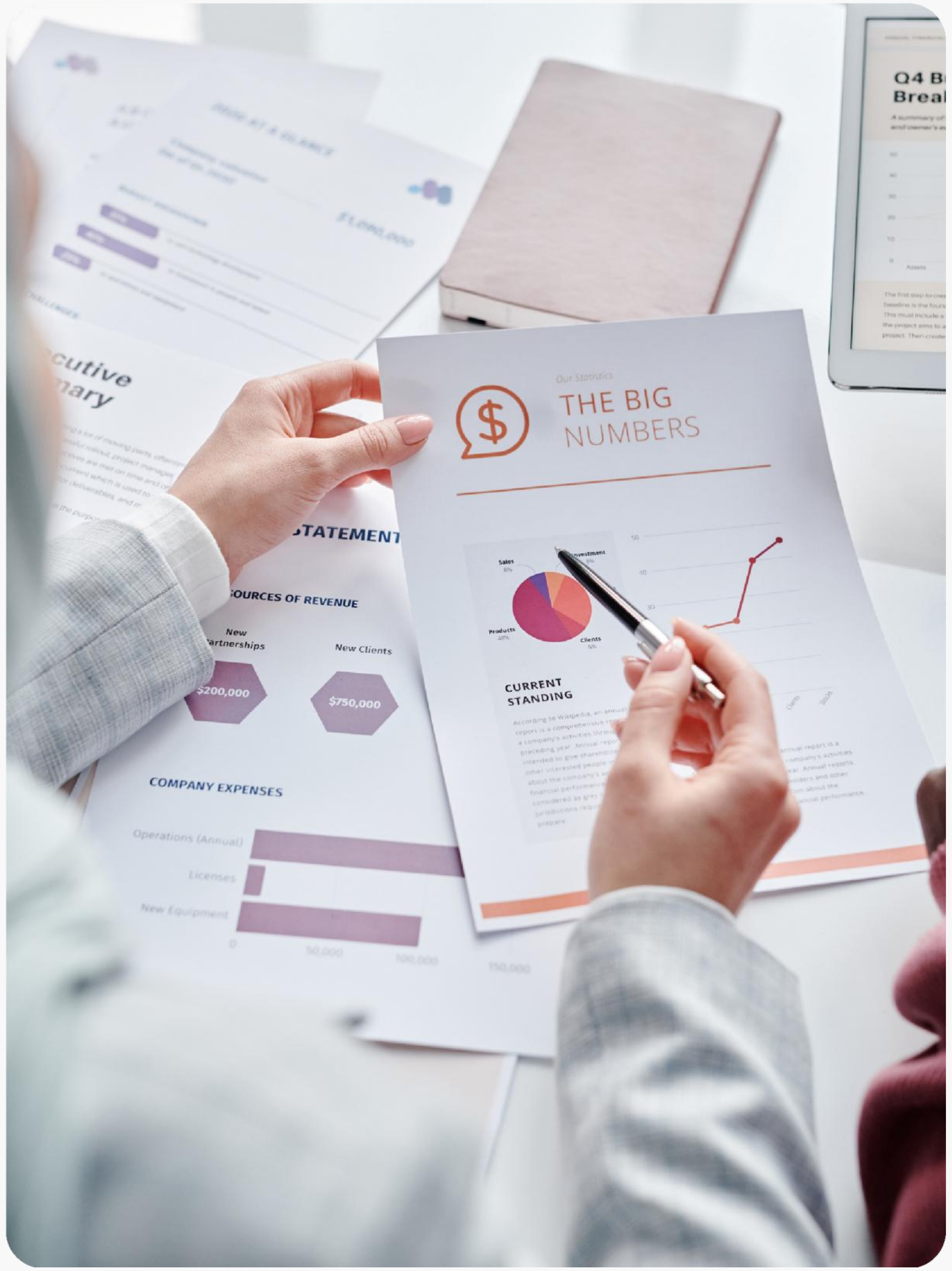
- Extract insights from unstructured maintenance data.
- Identify common failures and inefficiencies.
- Enhance decision-making for maintenance strategies.

Relevance & Significance:

Traditional manual analysis of reports is inefficient. NLP bridges this gap by unlocking hidden knowledge, improving efficiency, reducing downtime, and optimizing service quality.

Methodology





1.Dataset collection

2. Data pre-processing

- Dataset translation.
- Tokenization
- Stop-Words Removal
- Part-of-Speech (POS) tagging and Lemmatization
- N-grams

Note:

Corpus = All maintenance reports

Document = One maintenance report

Token = Words in the report (e.g., "engine", "failure", "oil")

3. Text-feature extraction

- Term Frequency-Inverse Document Frequency (TF-IDF)

$$\boldsymbol{TF}(\boldsymbol{t},\boldsymbol{d}) = \frac{(Number\ of\ occurrences\ of\ term\ t\ in\ document\ d)}{(Total\ number\ of\ terms\ in\ the\ document\ d)}$$

$$\boldsymbol{IDF}(\boldsymbol{t},\boldsymbol{D}) = \log_e \frac{(Total\ number\ of\ documents\ in\ the\ corpus)}{(Number\ of\ documents\ with\ term\ t\ in\ them)}$$

$$TF\text{-}IDF(t,d,D) = \boxed{TF(t,d)} \times \boxed{IDF(t,D)}$$

4.Topic modelling

- LDA (Latent Dirichlet Allocation) Algorithm

  LDA is a topic modeling algorithm that identifies hidden topics in a collection of documents by analyzing word patterns and their co-occurrences, helping to categorize and summarize large text datasets automatically.

5.Topic evaluation

- Coherence

$$C(T) = \sum_{i=1}^{N} \sum_{j=i+1}^{N} \log \frac{P(w_i, w_j)}{P(w_i)P(w_j)}$$

Where:

- $C(T)$ = Coherence of a topic T
- N = Number of top words in the topic
- w_i, w_j = Words in the topic
- $P(w_i, w_j)$ = Probability of both words appearing together in a reference corpus
- $P(w_i), P(w_j)$ = Individual probabilities of words

- Perplexity

$$PP(D) = \exp\left(-\frac{1}{N} \sum_{d=1}^{D} \sum_{w \in d} P(w|d) \log P(w|d)\right)$$

Where:

- N = Total words in the dataset
- $P(w_i|w_1, ..., w_{i-1})$ = Probability of a word given previous words (for language models)
- $P(w|d)$ = Probability of a word in a document (for topic models)

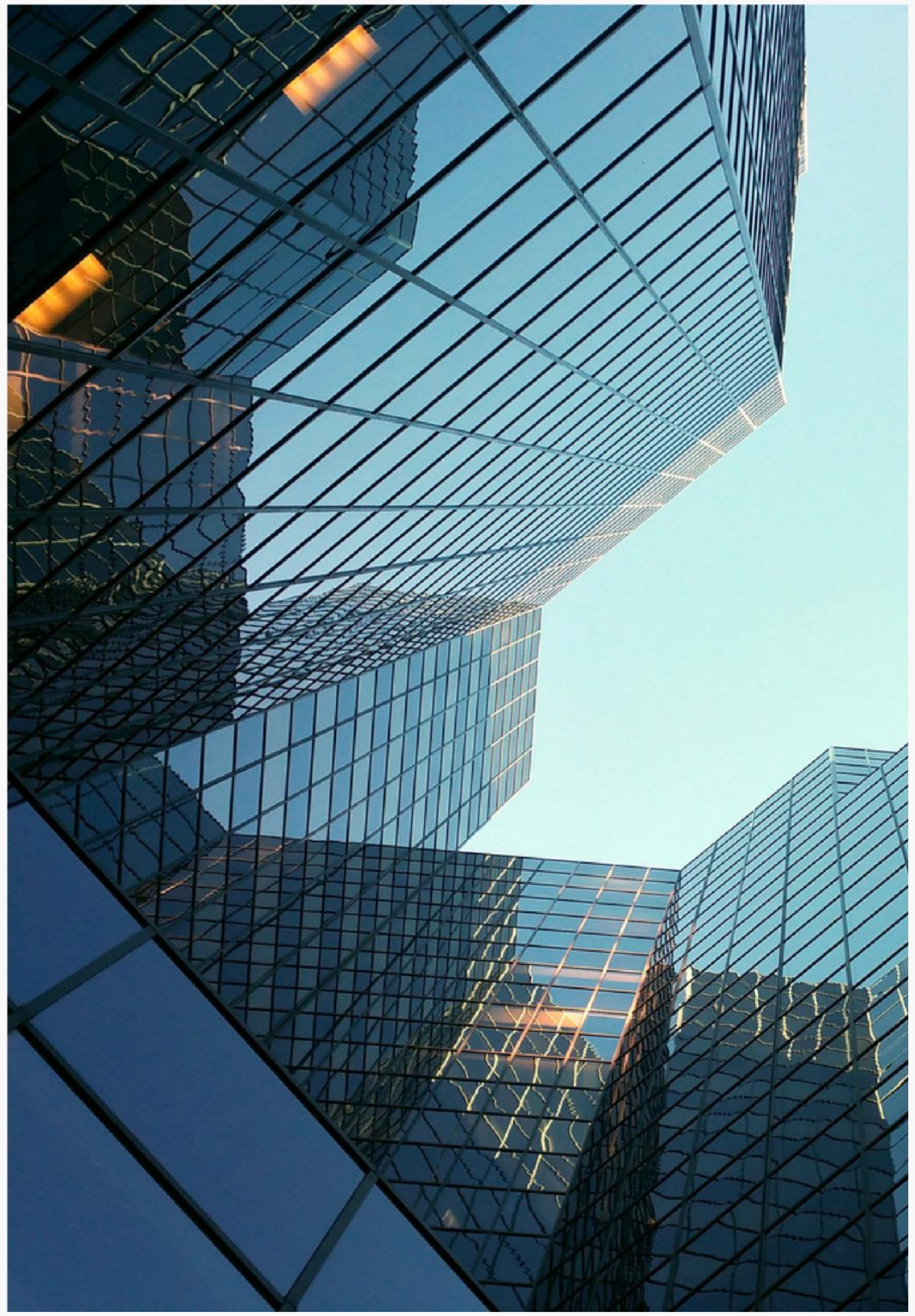
Higher Coherence Score, Lower Perplexity Score= Good model

6.Definition of Improvement Actions

- Once topics are finalized, the next step is to assign topics to the documents and analyze how they are distributed.

Realtime application case and interpretation

Company A



- Italian Manufacturing Company
- have 350 dataset collected over 7 yrs in 3 different languages
- Problem :

Unstructured data and lack of resources

Realtime application case and interpretation

- After the dataset was gatered then the data was configured
- after the configuration was done lda model was used for topic evaluation
- lda is a probabilistic model that groups words into a topic based on co-occurance in document.

Realtime application case and interpretation

- after the topic was determined then the data pre-processing steps like stopword removal and pos tagging.
- coherence score observed until satisfied.
- improved score from 0.33 to 0.65 in this case
- top topics shown in the table

Data Analysis

Topic	Keywords
Communication	communication, leave, damage, perform, recommend, card, review, report, course, result
Engine	engine, request, number, movement, finish, indicate, motor, screw, reformatte, note
Format	rod, air, mold, valve, cam, unit, wheel, blow, adjust, bearing
Mart	change, mart, belt, pass, form, com, transport, chain, die, serial
Resistor	area, seem, counter, hour, help, resistor, keep, parcel, expect, conditioner
Oven	try, oven, send, fan, block, console, want, say, install, miss
Software	program, version, place, update, fault, monitor, proceed, discard, mst, hath
Terminal	device, relay, wire, error, pc, create, terminal, power, panel, lose
Ring	ring, fiber, serco, decide, fiber_optic, assembly, measure, interruption_ring, problem, image
Mold	mold, close, lock, pin, backlash, locking, min, ndolo, ntc, observation
Belt	need, run, belt, instal, temperature, day, chain, pack, part
Bottle	speed, filler, decrease, carousel, fill, stock, problem, bottle, blow, blower
Film	film, print, track, ndolo, ntc, kick, lanea, min, inrush, nom
Power	replace, power_supply, fact, message, pilz, command, lanea, min, ndolo, ntc
Cosmos	make, stop, note, start, take, climb, button, board, year, cosmos

Key Findings of the Study

"The study’s results show that NLP-based analysis can effectively extract useful insights from maintenance logs. Some key findings include:"

- Recurring Maintenance Issues:

The NLP model successfully identifies frequently reported problems across different assets.

- Failure Pattern Recognition:

It detects trends in equipment failures, helping engineers predict potential breakdowns.

- Operational Insights:

NLP provides insights into asset performance and highlights areas that need frequent maintenance.

Discussion

- Maintenance Efficiency: Identified common failures (e.g., communication issues) and led to process improvements.
- Technician Skill Gaps: Topic-based analysis revealed areas for targeted training.
- Asset Design Issues: Failure clustering helped redesign components for better reliability.
- Customer Service Enhancement: Identified customer errors, leading to training and consultancy services.

Conclusion

- Need for better text pre-processing (e.g., spell-checking, removing jargon).
- Expanding the dataset (including handwritten reports) could improve accuracy.
- Further analysis of minor topics could offer additional insights.

Thank You