# ONLINE JOB ADVERTISEMENTS DEDUPLICATION USING LARGE LANGUAGE MODEL

JAKUB ŻEREBECKI, MIKOŁAJ TYM

# Web Intelligence Deduplication Challenge

- Challenge was announced by European Statistics Awards
- The Deduplication Challenge was focused on identifying potential duplicates of job postings published on the web
- Companies often publish job advertisements on different web portals
- Posting advertising the same jobs must be identified and removed using automatic and robust solutions to avoid double counting

# Dataset

- The competition dataset contain 112,000 online job advertisements, retrieved from around 400 websites active in the European Union
- The competition organizers have taken authentic job advertisements and created full, semantic, temporal, partial duplicates across different languages
  - Thus, organizers created a **synthetic** dataset for the competition
- 12.5B possible combinations

# Considered duplicates

- Full
- Semantic
- Temporal
- Partial
- Non-duplicate

# Full duplicates

- Two job advertisements are both exactly the same, i.e. they have the same job title and job description
- They may have differing sources and retrieval dates

# Semantic duplicates

- Two job advertisements advertise the same job position and include the same content in terms of the job characteristics
  - The same occupation, education or qualification requirements
- They may be expressed differently in natural language or in different languages

# Temporal duplicates

- Temporal duplicates are semantic duplicates with varying advertisement retrieval dates

# Partial duplicates

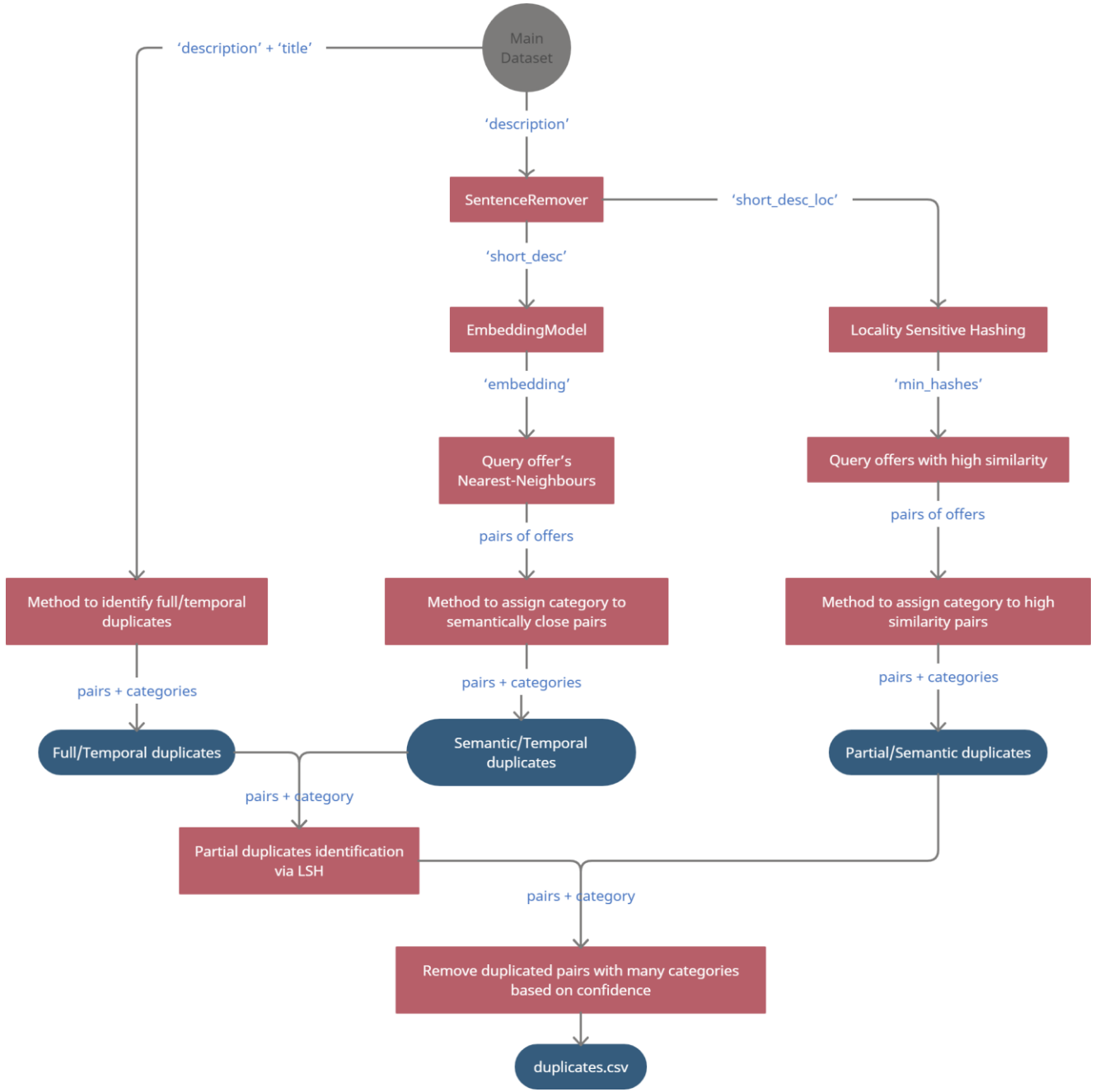
- Two job advertisements describe the same job position but do not necessarily contain the same characteristics
  - One job advertisement contains characteristics that the other does not
- Partial duplicates can be identified by searching the parent offer
  - It is common that one job advertisement (parent) contains all the information, while another advertisement (child) with missing words from the parent offer’s text is placed on another website

# Non-duplicates

- If specific job advertisements cannot be described as full duplicates, partial duplicates, semantic duplicates or temporal duplicates they are considered non-duplicates

# APPROACH

- Three different types of meth
  - Full duplicates identification
  - Comparison of encoded inform
  - Words similarity



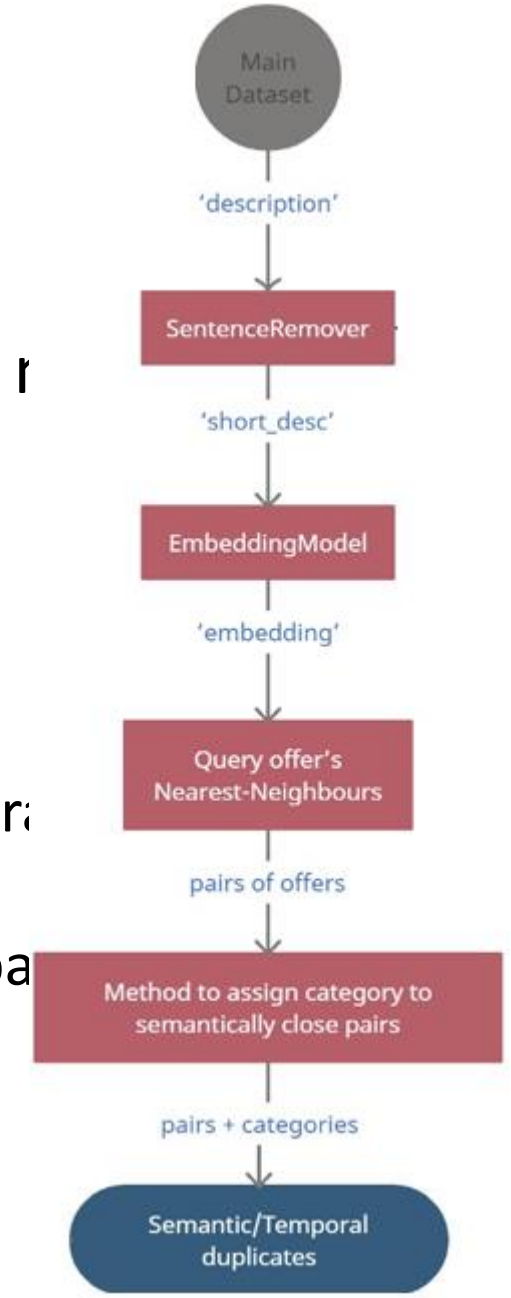
# Full duplicates identification

- The easiest to classify
- The method uses exact comparison of text
  - MD5 (maps any length job offer to fixed-size values)
  - Character-level
- Positive matches were full duplicates
- The time difference between the job posting da

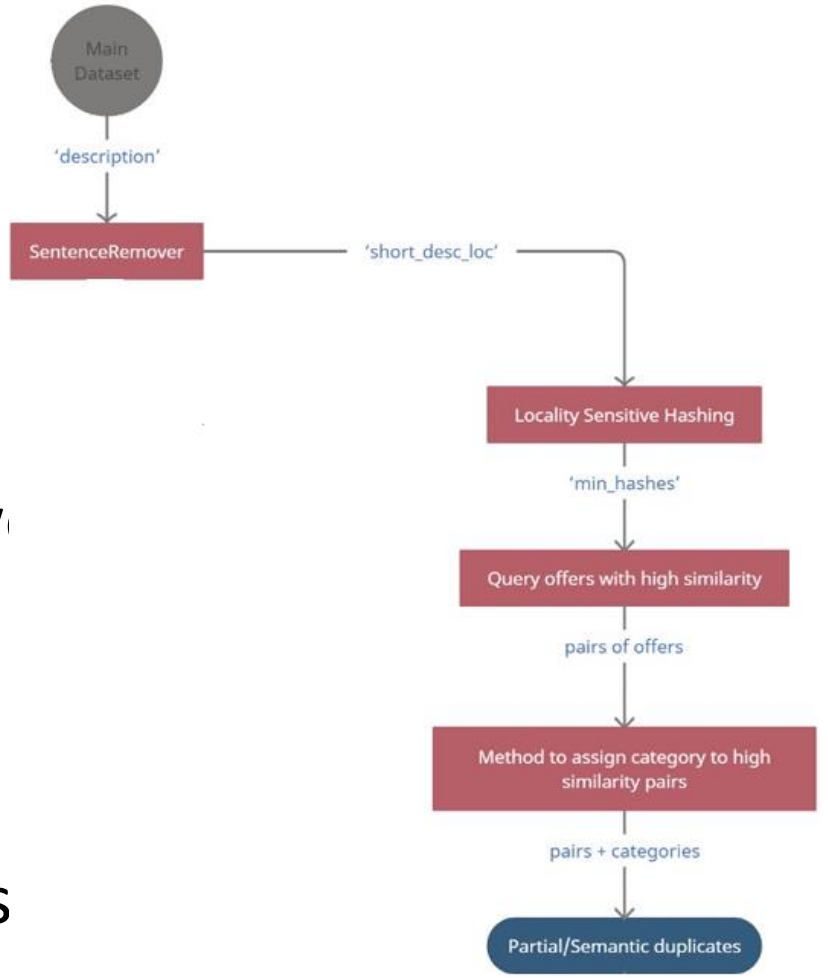
affects the classification

# Semantic duplicates identification

- Use embeddings to compare texts expressed differently in r
  language or in different languages
  - Encoder-like model to transform text
  - Similar offers have close distance
- Challenges:
  - A lot of jobs had the same informations like GDPR clauses, registra
    - Removal of common phrases at the 1st step
  - To classify 100 000 offers we need to compare above 10 billion pa
- High similarity metric means semantic duplicates



# Partial duplicates identification




- The hardest to identify
- To find partial duplicates in the same language w

on comparing text

  - Use of hash function different from MD5
  - Word level comparison
- Using this method, we could find a pairs of offers

similar then we measured if any words are missing

Parent offer

Qualifications: Proficient in at least one of programming and query languages like Python, PySpark, SQL etc Understand various Data Science, Machine Learning concepts & algorithms such as clustering, regression, classification, forecasting, neural networks, hyperparameters optimization, NLP, LLM. Very good command of English language. What we offer: Permanent employment contract Employee Assistance Program - legal, financial, and psychological consultations. Paid employee referral program.

Child offer

Qualifications: Proficient in at least one of programming and query languages like Python, PySpark, SQL etc Understand various Data Science, Machine Learning concepts & algorithms such as clustering, regression, classification, forecasting, neural networks, hyperparameters optimization, NLP, LLM. What we offer: Permanent employment contract Employee Assistance Program - legal, financial, and psychological consultations. Paid employee referral program.

# Partial duplicates cross-lingual identification

- Use embeddings to find the most similar offers to child offer

Parent offer

Qualifications: Proficient in at least one of programming and query languages like Python, PySpark, SQL etc Understand various Data Science, Machine Learning concepts & algorithms such as clustering, regression, classification, forecasting, neural networks, hyperparameters optimization, NLP, LLM. Very good command of English language. What we offer: Permanent employment contract Employee Assistance Program - legal, financial, and psychological consultations. Paid employee referral program.

Child offer

Qualifications: Proficient in at least one of programming and query languages like Python, PySpark, SQL etc Understand various Data Science, Machine Learning concepts & algorithms such as clustering, regression, classification, forecasting, neural networks, hyperparameters optimization, NLP, LLM. What we offer: Permanent employment contract Employee Assistance Program - legal, financial, and psychological consultations. Paid employee referral program.

Partial duplicate

High similarity

Sibling offer (Polish translation)

Znajomość przynajmniej jednego z języków programowania i zapytań, takich jak Python, PySpark, SQL itp.
Zrozumienie różnych koncepcji i algorytmów Data Science,
Machine Learning, takich jak klasteryzacja, regresja, klasyfikacja,
prognozowanie, sieci neuronowe, optymalizacja hiperparametrów, NLP, LLM.
Co oferujemy: Stała umowa o pracę Program pomocy pracownikom -
konsultacje prawne, finansowe i psychologiczne. Płatny program poleceń pracowników.

# COMPETITION RESULTS

# Results

- Robust methods
- We have achieved the 3rd place in Accuracy category
  - Macro F1 metric -> unweighted mean of per class F1 scores
- The 2n
  identif

Performance Rank

| # | Team | Submissions | Date of Last Submission | Full F1 ▲ | Semantic F1 ▲ | Temporal F1 ▲ | Partial F1 ▲ | Non-Duplicate F1 ▲ | Macro F1 ▲ |
|---|---|---|---|---|---|---|---|---|---|
| 1 | TwoTired | 10 | 31/03/2023 | 0.99 (1) | 0.85 (3) | 0.87 (4) | 0.77 (1) | 1.00 (1) | 0.90 (1) |
| 2 | TheClassifiers | 10 | 31/03/2023 | 0.99 (1) | 0.89 (1) | 0.92 (1) | 0.30 (3) | 1.00 (1) | 0.82 (2) |
| 3 | IDA | 10 | 31/03/2023 | 0.99 (1) | 0.84 (4) | 0.88 (3) | 0.37 (2) | 1.00 (1) | 0.82 (2) |
| 4 | Nins | 10 | 31/03/2023 | 0.99 (1) | 0.86 (2) | 0.87 (4) | 0.17 (4) | 1.00 (1) | 0.78 (3) |
| 5 | SPub.Fr | 10 | 31/03/2023 | 0.99 (1) | 0.83 (5) | 0.86 (5) | 0.17 (4) | 1.00 (1) | 0.77 (4) |
| 6 | Hyeny | 10 | 31/03/2023 | 0.99 (1) | 0.80 (7) | 0.91 (2) | 0.10 (8) | 1.00 (1) | 0.76 (5) |
| 7 | Flouss | 10 | 29/03/2023 | 0.99 (1) | 0.80 (7) | 0.82 (7) | 0.15 (5) | 1.00 (1) | 0.75 (6) |
| 8 | smrek | 10 | 31/03/2023 | 0.99 (1) | 0.72 (10) | 0.87 (4) | 0.11 (7) | 1.00 (1) | 0.74 (7) |

# Contact

- For more technical details, feel free to contact us

JAKUB

MIKOŁAJ TYM