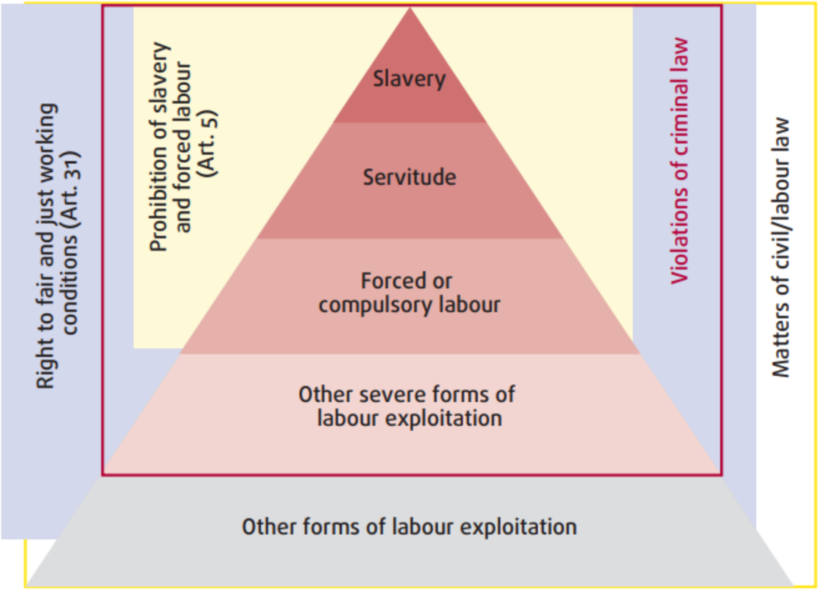
UNSUPERVISED LABOR INTELLIGENCE SYSTEMS: A DETECTION APPROACH AND ITS EVALUATION

A.S. Andreou, G. Cascavilla, G. Catolino, F. Palomba, D.A. Tamburri, W.J. Van Den Heuvel

JADS
Jheronimus
Academy
of Data Science

Introduction to the problem

- Labour exploitation is one of the main problem in the EU listed by the Council of Europe
- It consists in several levels of criminal practices (slavery, servitude, forced labor etc.) and forms (working conditions, low salary, human rights violations, type of abuse)
- Internet can play an important role because of the absence of geographical boundaries, lower recruitment cost and the facility of reaching larger pool of potential candidates


Right to fair and just working conditions (Art. 31)
Prohibition of slavery and forced labour (Art. 5)
Slavery
Servitude
Forced or compulsory labour
Other severe forms of labour exploitation
Violations of criminal law
Matters of civil/labour law
Other forms of labour exploitation

Note: Victims of all forms of exploitation set out in Figure 3 may also be victims of trafficking whenever the elements of the trafficking definition in Article 2 of the Anti-Trafficking Directive, as covered by Member State law, are met.

Source: FRA, 2015

Scope of the research

Definition

Labour exploitation is an illegal offense that involves human abuse and causes unfair market competition.

Focus of interest

Social media job announcement for unskilled job in the Netherlands in vulnerable sectors indicated by the existing literature

Contribution

Tackling the challenge to automatically spot potential labour exploitation by using indicators stated by the existing literature

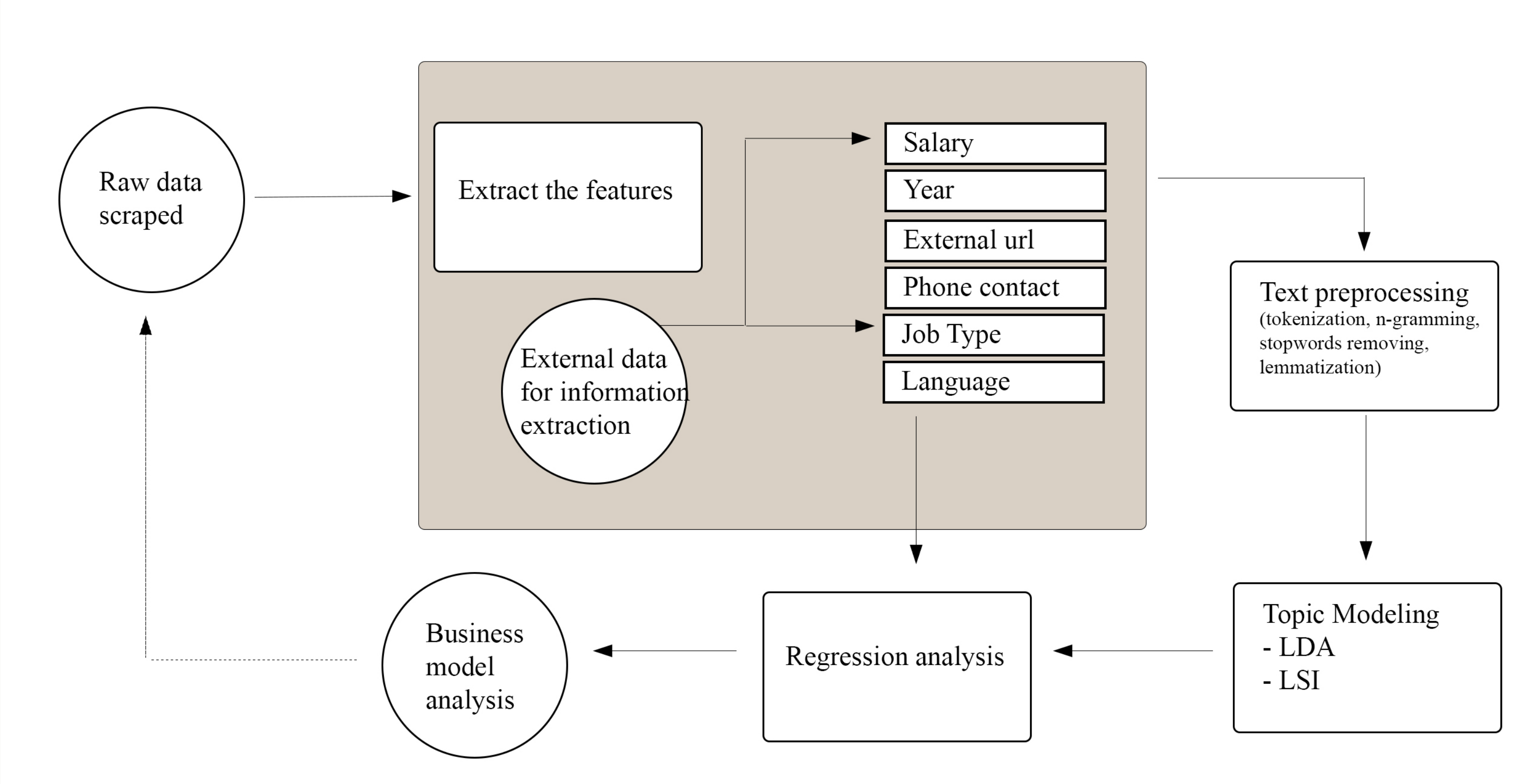
Literature Review

- **Sexual trafficking is the most dominant focus of research, while labour trafficking is under-researched**
- **Limited availability of data and few source of them used**
- **Dealing with unstructured data in different way in order to identify labour exploitation** (Kejriwal and Szekely (2017); Tong et al (2017); Volodko et al (2020))
- **Rich literature regarding social media topic detection: topic modeling (LDA) is the most common method** (Rohani (2017); Godin et al.(2013)) **and often in combination with other methods like hree-level LDA + keywords or hashtags + unsupervised LDA** (Shahbazi and Byun (2020); Wang and Brown (2012); Deng et al. (2020))

Problem Formulation

1. Which are the most common features that characterize deceptive online job advertisements?
2. Can we use a logistic regression analysis to detect deceptive online job post practice?

Implementation



Raw data scraped

Extract the features

External data for information extraction

Salary

Year

External url

Phone contact

Job Type

Language

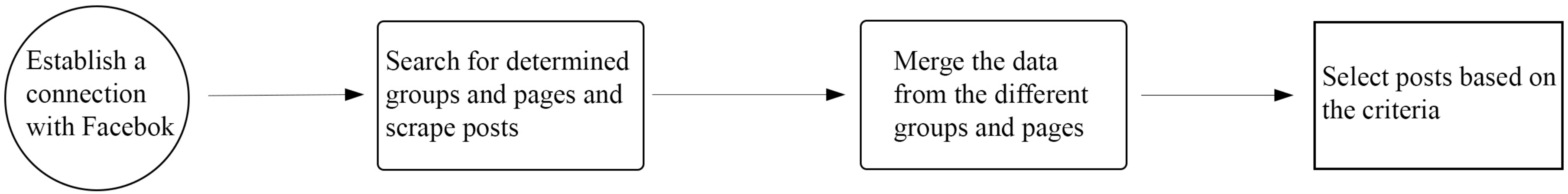
Text preprocessing
(tokenization, n-gramming, stopwords removing, lemmatization)

Topic Modeling
- LDA
- LSI

Regression analysis

Business model analysis

- Data Collection
- Data Preprocess
- Topic modeling
- Regression analysis



Establish a connection with Facebok

Search for determined groups and pages and scrape posts

Merge the data from the different groups and pages

Select posts based on the criteria

Data Collection and Preprocessing

Criteria for the collection of text

Removing duplicates

Ad-hoc RegEx functions to extract features from the unstructured text

Language detection and translation

Hourly gross wage calculation with the help of external website

Data Collection

- **Facebook groups and pages that post unskilled job offers in the Netherlands**
- **20 Facebook groups involved**
- **10301 entries scraped. For data quality extraction we defined the following criteria:**
    - The post is from groups and/or pages that have a clear mention of job announcements for the Netherlands or the Benelux region
    - The post contains at least 100 characters
    - The post is from a group/page that shows some activity in 2021 and has two posts per month or an average of one post per week
    - The post is unique and not a duplicate

Data Preparation

- We assigned a label as a new key for the contact information for each job offered by matching ad hoc regular expressions in strings of text;
- We considered as a piece of contact information two details: external website and phone contact;
- For each contact information, we yielded a positive value if the expression found the pattern in the string and a negative instead;
- language detector to recognize which language was used in the job announcements
- 2873 the final number of post

Topic Modeling

- Preprocessing (text translation, stopwords, tokenization, n-grams)
- Latent Dirichlet Allocation (LDA) and Latent Semantic Indexing (LSI) experimentation
- LDA is a generative probabilistic model of a corpus based on a three-level hierarchical Bayesian model. It determines the proportion of a collection of topics for each document of corpus-based on the distribution of the keywords
- LSI Topic modeling is a text mining technique which provides methods for identifying co-occurring keywords to summarize large collections of textual information. It helps in discovering hidden topics in the document, annotate the documents with these topics, and organize a large amount of unstructured data
- n-gram threshold to predict the next word in the sentence, alpha and eta parameters for LDA
- Chunksize for the N. of documents to be used in each training chunk and decay for LSI
- Coherence score to have a measure of the N. of topics and Perplexity as evaluation metrics

Classification Analysis

Logistic Regression for Binary Classification

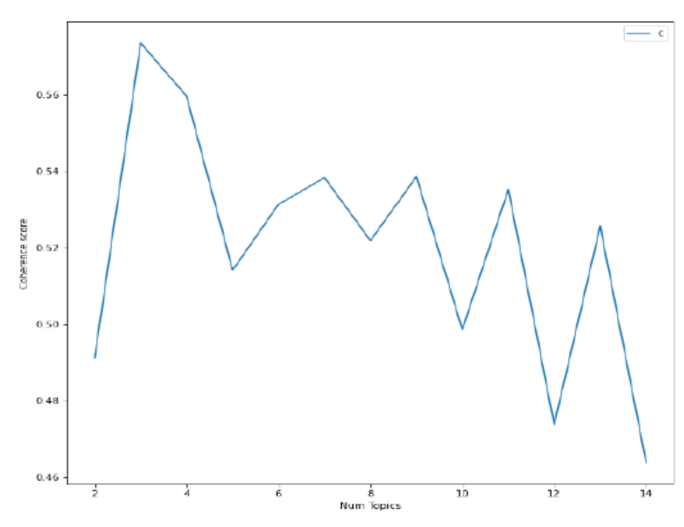
Predictive variable

- *Delta salary*

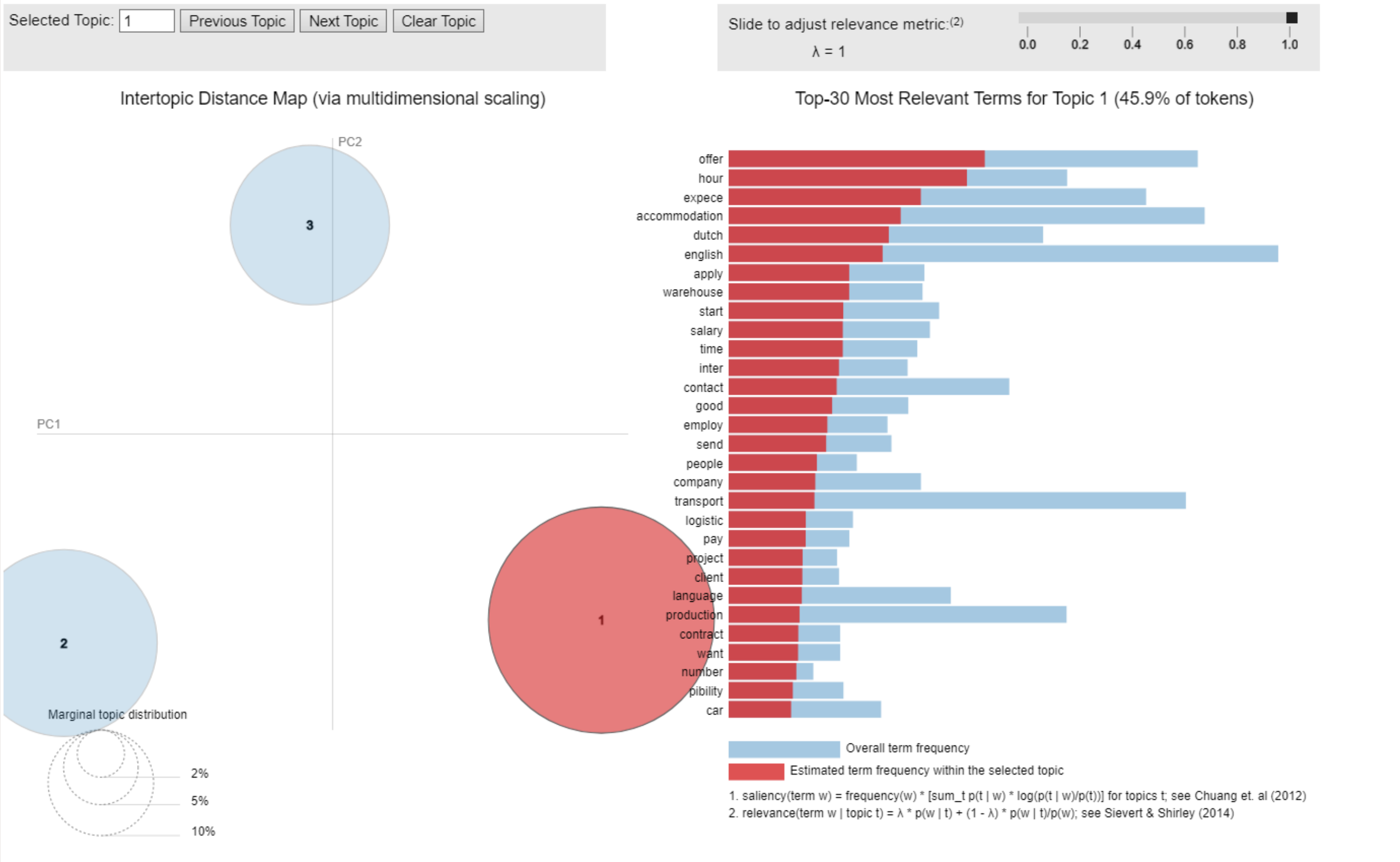
Features

- Topic
- Year
- Language
- Job class
- Presence of phone contact
- Presence of external URL

RESULTS


c
Coherence score
0.56
0.54
0.52
0.50
0.48
0.46
2
4
6
8
10
12
14
Num Topics

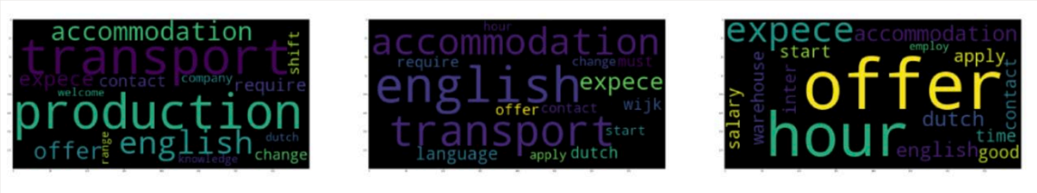
RESULTS – TOPIC MODELING



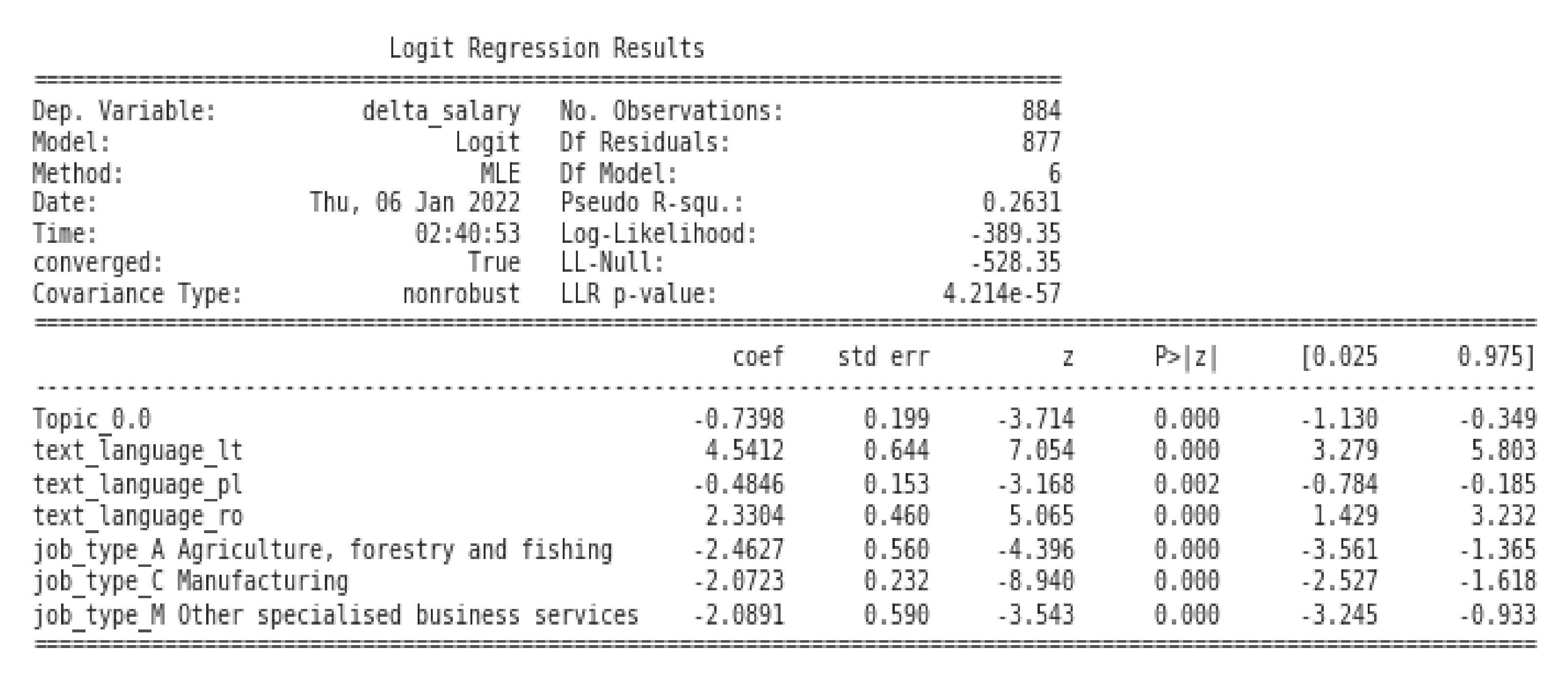
Selected Topic: 1

Slide to adjust relevance metric:(2) λ = 1

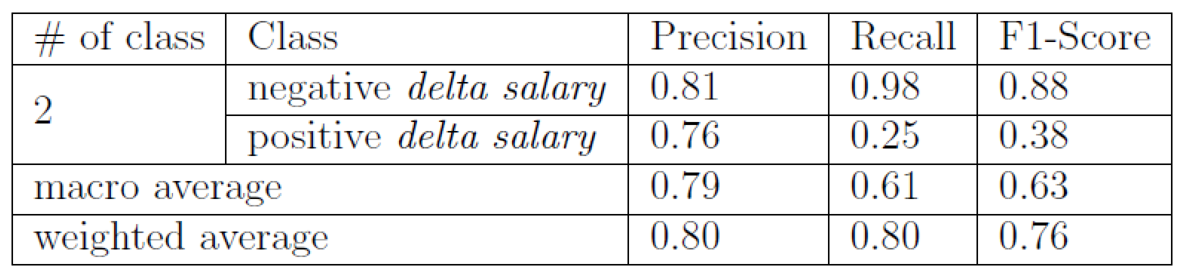
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)



RESULTS – CLASSIFICATION



```
                           Logit Regression Results
==============================================================================
Dep. Variable:           delta_salary   No. Observations:                  884
Model:                          Logit   Df Residuals:                      877
Method:                           MLE   Df Model:                            6
Date:                Thu, 06 Jan 2022   Pseudo R-squ.:                  0.2631
Time:                        02:40:53   Log-Likelihood:                -389.35
converged:                       True   LL-Null:                       -528.35
Covariance Type:            nonrobust   LLR p-value:                 4.214e-57
================================================================================================================
                                                coef    std err          z      P>|z|      [0.025      0.975]
----------------------------------------------------------------------------------------------------------------
Topic_0.0                                    -0.7398      0.199     -3.714      0.000      -1.130      -0.349
text_language_lt                              4.5412      0.644      7.054      0.000       3.279       5.803
text_language_pl                             -0.4846      0.153     -3.168      0.002      -0.784      -0.185
text_language_ro                              2.3304      0.460      5.065      0.000       1.429       3.232
job_type_A Agriculture, forestry and fishing -2.4627      0.560     -4.396      0.000      -3.561      -1.365
job_type_C Manufacturing                     -2.0723      0.232     -8.940      0.000      -2.527      -1.618
job_type_M Other specialised business services -2.0891   0.590     -3.543      0.000      -3.245      -0.933
================================================================================================================
```



# of class	Class	Precision	Recall	F1-Score
2	negative *delta salary*	0.81	0.98	0.88
	positive *delta salary*	0.76	0.25	0.38
macro average		0.79	0.61	0.63
weighted average		0.80	0.80	0.76

LIMITATIONS AND FUTURE WORK

- Necessity of ground truth data to validate the potentiality of the indicators
- Different source of data to be explored
- Social media textual data need heavy preprocess
- Images can be another important source of data for the research
- Human judgement key for the evaluation
- Scalable solution need to be proved and tested

THANK YOU