Unsupervised Labor Intelligence Systems: A Detection Approach and Its Evaluation

A Case Study in the Netherlands

Giuseppe Cascavilla[1,2(✉)], Gemma Catolino[1,3], Fabio Palomba[5], Andreas S. Andreou[4], Damian A. Tamburri[1,2], and Willem-Jan Van Den Heuvel[1,3]

[1] Jheronimus Academy of Data Science, Hertogenbosch, The Netherlands
{g.cascavilla,g.catolino,d.a.tamburri}@tue.nl, W.J.A.M.v.d.Heuvel@jads.nl
[2] Eindhoven University of Technology, Eindhoven, The Netherlands
[3] Tilburg University, Tilburg, The Netherlands
[4] Cyprus University of Technology, Limassol, Cyprus
andreas.andreou@cut.ac.cy
[5] University of Salerno, Fisciano, Italy
f.palomba@unisa.it

Abstract. In recent years, job advertisements through the web or social media represent an easy way to spread this information. However, social media are often a dangerous showcase of possibly labor exploitation advertisements. This paper aims to determine the potential indicators of labor exploitation for unskilled jobs offered in the Netherlands. Specifically, we exploited topic modeling to extract and handle information from textual data about job advertisements for analyzing deceptive and characterizing features. Finally, we use these features to investigate whether automated machine learning methods can predict the risk of labor exploitation by looking at salary discrepancies. The results suggest that features need to be carefully monitored, e.g., hours. Finally, our results showed encouraging results, i.e., F1-Score 61%, thus meaning that Data Science methods and Artificial Intelligence approaches can be used to detect labor exploitation—starting from job advertisements—based on the discrepancy of delta salary, possibly representing a revolutionary step.

Keywords: Case study · Data science · Artificial Intelligence

1 Introduction

In the last decade, the number of labor exploitation victims has risen in the Netherlands[1]. Globally, the report of UNODC (United Nations Office on Drugs and Crime) in 2020 showed how Internet-based trafficking had become increasingly used for illegal activities such as labor exploitation [36]. Traffickers usually target less regulated industries and those featuring seasonal demand for workers.

[1] https://tinyurl.com/we359yhe.

J. Barzen et al. (Eds.): SummerSOC 2022, CCIS 1603, pp. 79–98, 2022.
https://doi.org/10.1007/978-3-031-18304-1_5

Vulnerable sectors include agriculture, food production, cleaning, construction, manufacturing, entertainment, hospitality, retail, transportation, distribution, and consumption supply chains [13]. Recently, there exists a growing trend of recruitment through the web, social media in primis, thus increasing the number of potential victims who can be targeted by labor exploitation [9]. The absence of geographical boundaries and the spread of open-access online domains make illicit behaviors accessible to a broad range of individuals that may facilitate crime [26]. In addition, the multijurisdictional context of the Internet is still an open challenge, thus complicating the prosecution of the perpetrators. However, technology improved getting short-term job arrangements, resulting in a growth in self-employed individuals. So, more effort is needed to prevent and address issues like labor exploitation, legal worker classification, wage, benefits required, and educate workers about their rights. United Nations did the first step to recognize common labor exploitation practices. Indeed, several institutions tried to define a list of forced labor and human trafficking indicators. They converged in a list of 67 indicators for human trafficking[2]. They are divided into six categories based on the type of recruitment, i.e., deceptive, coercive, abuse of vulnerability, exploitative conditions of work, forms of coercion, and abuse of vulnerability at the destination. These indicators can provide a complete understanding of the commonly utilized practice for forced labor.

The primary purpose of this paper is to shed light on coercive labor practices in the dutch economic sectors of unskilled labor. The goal of this work is double. On the one hand, we want to frame deceptive behaviors in a social media context and what are the risk factors involved in looking at online job advertisements on Dutch market labor. On the other hand, we want to define an approach that automatically detects deceptive practices based on the delta salary. In particular, we extracted information about job advertisements from Facebook groups and pages, focusing our attention on Dutch market labor. Then, we preprocess these data and apply topic modeling—applying Latent Dirichlet Allocation (LDA) and Latent Semantic Indexing (LSI)—to extract the most relevant feature that characterizes job advertisement, possibly compared to the risky indicators cited above. Finally, based on the features extracted above, we constructed a logistic regression model for assessing whether those can predict potential discrepancy in terms of salary—according to the Ducth National Salary Tables[3]—thus possibly indicating the risk of labor exploitation. Results showed how features like job class, external link represents info that characterizes job advertisements and should be carefully monitored. Finally, our logistic regression model showed encouraging results, i.e., 61% F-Measure, thus meaning that AI approaches can be used to detect labor exploitation announcements based on the discrepancy of delta salary. Indeed, detecting and preventing labor exploitation starting from job advertisements, thus represents a revolutionary step that can help decrease this issue. We are already collaborating with Dutch Police

[2] https://www.ilo.org/wcmsp5/groups/public/---ed_norm/---declaration/documents/publication/wcms_105035.pdf.

[3] https://www.cbs.nl/en-gb/labour-and-income.

to provide an intelligence dashboard that consists of an AI model for real-time crawling data from social media and highlights possible labor exploitation in the dutch labor market. i.e., SENTINEL Project.

Structure of the Paper. In Sect. 2, we briefly introduce the existing literature in the context of our study. Section 3 concerns the methodology of our research and Sect. 4 shows the results our experimentation. We discuss and conclude our paper in Sect. 5.

2 Related Work

This section provides a brief grounding on what has been done so far by the existing literature in labor exploitation identification.

2.1 Labour Exploitation Identification

The online recruitment of an exploitable workforce takes part on employment websites, online agencies, and social networks [19]. The existing academic literature experimented with strategies to infer deceptive recruitment for labor exploitation [38]. The state of the art experimented with strategies to infer deceptive recruitment for labor exploitation. Volodko et al. [38] addressed the problem showing the indicators for labor exploitation by the existing literature may be commonplace characteristics of online job advertisements for people looking for jobs abroad. They manually labeled all the job advertisements from the most famous Lithuanian website. They experimented with Poisson regression to test if the characteristics of one advertisement can give enough information to predict the number of labor trafficking indicators present.

Kejriwal et al. [23] developed a search engine to address the problem of collecting evidence about labor exploitation but, at the same time, minimizing investigative effort. The system exploits ontologies, knowledge graphs, and word embedding to extract information from Open and Dark Web for human trafficking identification. In addition, they used several strategies such as keyword strategy to extract information to create an investigation schema that helps the graph algorithm analyze the crawled web corpus.

Tong et al. [35] introduced a multi-modal deep learning algorithm to detect suspected human trafficking advertisements automatically. The approach uses both text and images and shows a high accuracy compared to models that use one of the sources. Nevertheless, the approach is hardly interpretable, especially when evaluating the impact of the features in a different context. Zhu et al. [41] proposed a language model-based approach for creating a phrase dictionary for identifying human trafficking indicators in adult service ads. The model showed a good performance and a reasonable interpretation of the keywords retrieved as potential trafficking indicators, thanks to the pipeline developed for automatically detecting and extracting data from potential fraudulent websites. This pipeline also detects and clusters human trafficking activities into unknown criminal organizations.

Siddiqui et al. [33] highlighted the importance of pre-processing tasks when dealing with unstructured or semi-structured text in order to separate relevant snippets of information from the unorganized text and find a way to improve the decision-making process in regard criminal fight.

The state of art on labor exploitation identification mainly concerns sexual trafficking [7]. Sweileh et al. [34] showed how labor trafficking is underrepresented compared to sex trafficking. One reason is that indicators of sexual exploitation are more discernible and less ambiguous in the online textual context of working offers Di Nicola et al. [12]. Burbano et al. [4] make a corpus in Spanish language from social media text and build a predictive model in order to identify automatically.

2.2 Social Media Topic Detection

A recent challenge in research is to detect topics from online social networks. These topics are mainly connected to disaster events, urban planning, public health, political or marketing studies [24]. The open challenge is to interpret a massive volume of unstructured data [20], but without knowing what should be the final pattern as in the information retrieval method [22]. Since the quantity of data available in social media is exponentially growing [6] there is a need to recognize the necessity to employ tools for automatic topic discovery. Thus, the goal is to detect topics that are high-level patterns of textual data. For this reason, topic models represented a powerful techniques for discovering hidden text patterns [18]. The idea behind topic modeling is to create a thematic structure that defines a determined amount of underlying concepts through an efficient process that takes less representation space and noise and, consequently, can manipulate large amounts of data without human supervision.

Latent Dirichlet Allocation (LDA) is the dominant topic modeling technique in this particular field of research [37]. Shahbazi et al. [32] collected contents from different social media to conduct a semi-automatic process. Rohani et al. [30] addressed the problem to detect topics from a large variety of semantic text by proposing a topic modeling technique based on LDA. Statistical topic modeling based on LDA is also effective in crime prediction. Gerber et al. [14] showed that the combination of the standard approach—based on kernel density estimation—with additional Twitter features improved spatial-temporal crime prediction in one city in the United States. Once assessed the probability of each word belonging to a certain topic is, the topic modeling process evaluates each topic's cohesion in each neighborhood.

Social network such as Twitter has been used for extracting any information: Wang et al. [39] showed the possibility to predict hit-and-run crime incidents; Godin et al. [15] provided a method for recommending hashtags for tweets in a fully unsupervised approach based on LDA; Cordeiro et al. [8] improved tweet event description by extracting latent topics using LDA from the tweets text for each hashtag signal obtained after wavelet analysis; Prier [28] detected tobacco-related topics in order to provide a better understanding about public health problems in the United States. Cvijikj et al. [10] proposed a trend detection

algorithm that can collect data from Facebook and detect disruptive events and popular topics in a near-real-time interval since Facebook does not provide real-time streaming access as the other social media. One problem with discovering topics from social media is the granularity that every topic can have once determined the number of topics in a corpus. Deng et al. [11] proposed a three-level LDA topic model combined with keyword matching and coherence analysis to identify topics and sub-topics and provide a good level of interpretability and a better understanding of the evolution that any topic can have over time. Keyword matching can also be done through the use of co-occurrences between pairs of the discussion topics in a key graph-based model approach [25], or through the use of algorithms such as Rapid automatic extraction algorithm (RAKE) [21].

3 Methodology

In this section, we present the methodology of our research.

3.1 Research Questions

The aim of this work is *to understand what indicators can be employed to detect labor exploitation in online job offers and define an approach to detect possibly labor exploitation alerts through salary discrepancy.* To this end, we defined the following research questions:

RQ1 - *Which are the most common features that characterize deceptive online job advertisements?*
RQ2 - *Can we use a logistic regression analysis to detect deceptive online job post practise?*

To answer **RQ1**, we collected data from Facebook public groups and pages on Dutch market labor. We filtered out all the posts that did not match the online job posting that had not been recalled by the existing literature demonstrated in Sect. 2. After scraping and gathering the data, we probed and extracted meaningful features for our further analysis. Indeed, we used NLP techniques to extrapolate meaningful information from the unstructured text. Next, we performed topic modeling analysis using LDS and LSI to explore the most common and insightful features that characterize job advertisements for spotting potential labor exploitation from social media job postings.

Based on the feature extracted above, we answer **R2** constructing a logistic regression model for identifying potential discrepancies between the salary proposed by the online announcements and the national Dutch working wage calculated by the Dutch"Centraal Bureau voor de Statistiek" (CBS)[4].

[4] https://www.cbs.nl/en-gb/labour-and-income.

3.2 Data Collection

In order to extract the information about job advertisements, we chose Facebook. The reason behind our choice is related to its popularity. Indeed, it is the world's most widely used social media platform, especially in the context of non-sexual labor work. For the data collection, we used a scraper written in *Python* as programming language[5]. For identifying Facebook groups and pages that post unskilled job offers in the Netherlands, we defined a query with simple keywords that would capture the context of our research e.g., job, offer. Then we double-checked the results to check whether a group or a page contained some job offers, e.g., post every week. Out of more than 200 groups, we kept 20 of them. To scrape the job offer posts of every group, we increased the number of `pages` in a range from 200 to 500 depending on the limit of the posts that one group had, and set the `posts_per_page` as 500 in order to avoid to lose data from groups that mainly contain short text. The number of entries scraped was initially 10301. To decrease the number of entries, we applied the criteria listed below. For ensuring 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 or 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

To guarantee the last criterion, we removed duplicates once we merged all the posts from different sources of groups and pages. We used the *cosine similarity* to filter out further posts that show strong similarities with each other. Nonetheless, we cannot exclude missing relevant posts. The scraper provided a JSON file as output in which the text, the post id, and the timestamp are stored. We then proceed to pre-process the qualified posts. The dataset is available online using the online appendix [1].

3.3 Data Preparation

Data preprocessing is essential since we deal with unstructured data, i.e., posts, that need to be remodeled to be input for the topic modeling analysis. In addition, raw data extracted from the collection phase need to be transformed into an understandable format. Therefore, we deployed Regular Expression (RegEx) to extract relevant features:

- 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;

[5] https://github.com/kevinzg/facebook-scraper.

– For each contact information, we yielded a positive value if the expression found the pattern in the string and a negative instead;

Afterward, we employed a language detector to recognize which language was used in the job announcements, i.e., *Java* library ported from Google's language detection and recreated to *Python*[6]. Since this language detection algorithm is non-deterministic, it is not always reliable. Therefore, the first and the second authors of this paper manually double-checked half of the posts to check whether the language detected was the correct one. Moreover, the detection can result in ambiguous and not comprehensive as some posts can contain not only one language. Therefore, we gave priority to labels with a different language other than English and/or Dutch. Finally, we translated data into English to use them as input for the topic modeling task. We employed Google translate API for doing this task[7]. Posts written on social media might contain typos/errors. Hence, we extended the translation with a spelling corrector[8]. Furthermore, we identified any possible duplicates, filtered out them. **2873** represent the final number of posts we got.

Subsequently, we extrapolated the salary offered in each post. Once again, regular expressions came in handy when dealing with numeric characters with a specific meaning according to their position in the text. We developed a heuristic approach to get only the digits that represent the salary and nothing else, such as the phone number or the date:

– We first retrieved all the words in a consecutive or closed position to every number for every post;
– We selected only the words that related to a money offer such as *euro*, *gross*, *hourly*;
– We retrieved the digit in a new feature, and we kept all the keywords necessary to calculate the salary and convert it hourly and into the gross form

Since we want to have an accurate conversion from a net wage to a gross one, we collected the data from a gross/net converter website for each amount of hourly wage and each year[9]. Then, we matched the net wage with the salary obtained from the preprocessing and the year with the one from the timestamp retrieved with the scraper; finally, we obtained the gross salary. As for the posts that did not explicitly specify whether the salary was grossly or net, we reasonably assume it as gross wage. We kept the timestamp as the next feature. In particular, we consider the year as the most informative part of the timestamp. Moreover, we found five main languages, with varying degrees, consistently present in the dataset, five different type of job post and the amount of post on each year. These features are not related to the salary, hence we defined them as "other" as shown in Table 1.

[6] https://github.com/shuyo/language-detection.
[7] https://py-googletrans.readthedocs.io/en/latest/.
[8] https://textblob.readthedocs.io/en/dev/.
[9] https://thetax.nl.

Table 1. Top 5 frequency of the other features

Language	Count	Job type	Count	Year	Count
Polish	1459	Manufacturing	976	2020	803
English	815	Transportation and storage	603	2021	795
Dutch	447	Wholesale and retail trade	448	2019	575
Romanian	80	Construction	354	2018	259
Lithuanian	61	Agriculture, forestry and fishing	205	2016	198

The type of job represents the last feature. We performed a heuristic approach to extract this information together with manual labeling:

- We took into consideration the type of job that, according to the existing literature, is most likely to be at risk for labor exploitation;
- We considered the United Nations' classification of the job sector[10], and we extracted all the keywords related to the type of job previously considered;
- We then matched these keywords with each post in our dataset, filtering out all the rest of the words.

Once we had only these relevant keywords, we could manually label each post within the appropriate job sector.

3.4 [RQ1]. Topic Modeling for Deceptive Online Job Advertisements

After extracting the data, we tried to find the essential information that characterizes online job advertisements using topic modeling, possibly spotting potential signals of labor exploitation. The usage of topic modeling showed promising results in uncovering hidden communities of tweets in social media [30]. We exploited two topic modeling techniques, Latent Dirichlet Allocation (LDA) and Latent Semantic Indexing (LSI).

LDA is an unsupervised learning that views documents as bags of words. It is a generative probabilistic model of a corpus based on a three-level hierarchical Bayesian model. The probabilistic topic model estimated by LDA consists of two tables (matrices). The first table describes the probability or chance of selecting a particular part when sampling a particular topic (category). The second table describes the chance of selecting a particular topic when sampling a particular document or composite. Indeed, it determines the proportion of a collection of topics for each document of corpus-based on the distribution of the keywords [2]. Once the number of topics is given, the document's topic distribution is reorganized. Finally, the keywords are distributed inside the topics to have the ideal output of topic-keywords structure.

LSI is the second method that we implemented. It attempts to solve the issues of lexical matching by retrieving information using statistically determined conceptual indexes rather than individual words. It represents a method that maps

[10] https://unstats.un.org/unsd/publication/seriesm/seriesm_4rev4e.pdf.

documents into a latent semantic space [31]. Since this new space has fewer semantic dimensions than the original one, this technique works as a dimensionality reduction. A truncated singular value decomposition (SVD) evaluates the structure of the words in each document given. The vectors created from the truncated SVD are then used for retrieval. The result is that these vectors produce a more reliable performance in understanding the meaning compared to the individual phrases since they can handle the synonymy problems.

Data Preprocessing. Before deploying the topic modeling, we preprocessed our data. First, we considered the unstructured text filtered by the keywords from the previously extrapolated features. Then, we prepared the text data for the preprocessing tasks.

The first operation is to correct wrongly translated words: they are translated without a proper or reasonable meaning concerning the context in the text. For this reason, we created an initial list of the unique words from the raw corpus of text, and then we detected the language of each word. The words that were not detected in English were discarded or corrected based on a manual check. Then, we ordered the words by frequency, checking if some uncommon words appeared in an unexpected frequency, and replaced them with the correct word.

Another step performed consisted of removing the stopwords, i.e., the most common words such as articles, pronouns, and prepositions. We included as stopwords words such as *'work'*, *'job'* and *'Netherlands'* since they often appeared, not adding any additional information to the text. Removing this low-level information from the text also helped reduce the number of tokens used and streamline the following steps. We removed punctuations, and we tokenized each post in a list of words with the use of the Gensim library[11]. We computed the bigrams and trigrams, i.e., the sets of two and three adjacent words. We tried different value of the n-gram parameters' function `min_counts` and `threshold` to achieve an optimal combination of n-grams words. We performed a lemmatization of the word since we wanted to produce words that could be easily readable and recognizable. Finally, we created the dictionary and the corpus using `id2word`, which maps the unique id of each word to a token.

Applying Topic Modeling. Once we completed the preprocessing steps, we trained the LDA model. The initial task is to set the number of topics. In order to define the optimal number of the topic, we ran a grid search setting the minimum and the maximum number of topics and the `step` as 1. We gradually reduced the number of topics to 15 once we measured the coherence score, i.e., the degree of semantic similarity between high scoring words in the topic, for each amount of topics that the LDA model was generating.

As for LSI, We replicated the same preprocessing implementation for LDA. We used Gensim library to implement the model[12]. In this case, we explored different values for the hyperparameters step, i.e., `chunksize` and `decay`.

[11] https://pypi.org/project/gensim/.
[12] https://radimrehurek.com/gensim/models/lsimodel.html.

`Chunksize` indicates the number of posts used in each training chunk, and it can affect the speed of the training, while `decay` value gives a weight of existing observations relatively to new ones.

Once we had the topics, we evaluated the goodness of the model. We used two evaluation metrics: the perplexity score, which captures the level of generalization of the model, and the coherence score.

The perplexity score is a statistical measure that estimates the distribution of words in the documents and tells how the model can represent the statistics of the held-out data [2]. Since it has been proved that this metric may not yield human interpretable topics and it can be not positively correlated with the human judgments [5], we include other metrics along with the perplexity score. The metric that evaluates the coherence score is *'c_v'*. It measures the score using a normalized pointwise mutual information (NPMI), and the cosine similarity once obtained the co-occurrence between words [29]. Due to the space limitation, we do not report the formula.

3.5 [RQ2]. Building a Logistic Regression Model

This RQ aims to predict deceptive behavior, considering as dependent variables the most tangible value that we can get from an online job announcement: the salary. To obtain our dependent binary variable, we relied on data regarding the employment, working hours, and wages in the Netherlands[13]. In particular we procured the hourly wage per class of job and per year. We calculated the difference between these values and the salary extracted from our dataset.

The predictive variable was named *delta salary*: when the salary offered in a job post is larger than the one displayed by the national statistic agency, we attributed a positive value, we gave a negative value otherwise.

In order to classify this variable, we deployed a logistic regression model. This technique can estimate the probability of occurrence, including making a connection between features and the likelihood of specific outcomes. The features that we considered are the ones obtained from the data preparation phase and the topics obtained from the LDA topic model, and they are:

- Topic
- Year
- Language
- Job Class
- Presence of phone contact
- Presence of external url

In this way, we wanted to discover whether it is possible to find a hidden pattern between potential labor exploitation indicators and if there is a chance to improve the accuracy of the investigation in this area of research. At the same time, having a tangible asset as the salary identification can help to assess the economic impact better and better understand the business model employed.

[13] https://opendata.cbs.nl/statline/#/CBS/en/dataset/81431ENG/table?ts=1637795200937.

Before running the model, we initially explored the data and checked possible class unbalancing. Unfortunately, the negative value overcame the positive ones. However, the ratio was not excessively unbalanced (70%–30%). Moreover, oversampling with artificial data could deteriorate the quality of the dataset. On the other side, undersampling could discard potentially meaningful data and undermine the model's accuracy. We then proceeded to encode the variables. We first converted the categorical variable such as the *job class*, the *language* and the *topic* into one hot encoded variables, *time* into ordinal encoded variable and *phone contact* and *external url* into simple dummy variables. As a consequence, the number of variables exponentially increased. Thus, we measured a possible correlation among the variables and performed a recursion feature elimination to select the meaningful variables and avoid high dimensionality problems. We used the Sklearn library to implement it[14].

The recursion feature elimination selects features by recursively evaluating smaller groups of features. First, the estimator is trained on the original set of features to determine the importance of each feature. Then, the least significant feature is removed from the current group of features. The process is repeated until the given set of features is attained. We decided to keep half of the features out of those previously made. We divided the dataset into training and testing data, finally implementing the model. We removed the features that exhibit p-values higher than 0.05 and re-run the model. Then we removed topic features to compare with the previous model and evaluated whether the model showed any difference in terms of explainability. We then evaluated our final model to predict the accuracy. We also wanted to see what is the accuracy of both of the binary predictive classes. Thus we calculated the precision, recall, and F1-score. We finally evaluated the sensitivity/precision trade-off using the ROC curve.

4 Results

This section presents the results obtained according to the methodology described in Sect. 3.

4.1 Topic Modeling

In this section, we show the results from the experimentation presented in the methodology in the Sect. 3. We started to experiment with LDA[15]. In particular, we initially defined the optimal `threshold` of both bi-grams and tri-grams. We experimented with several sets of values. The values outside the range between 10 to 100 displayed worse scores and poorer interpretation, so we focused on those inside the range. We reported the results in Table 2. In this case, we chose 80 as the optimal choice since it showed a better coherence score, even though it was very close to the other values. Once selected the *threshold* of the n-grams,

[14] https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.RFE.html.

[15] https://github.com/giuseppecascavilla/topic_modelling.

we further kept experimenting using α and η hyperparameters. We tried several values of α and η. Since we reasonably assumed that the distribution for words in topics and documents is sparse, we expected to have a better score with a value of α and η less than 1. Table 3 shows the results of the tuning. While the optimal number of topics changes from a range of values from 2 to 6, the coherence scores are very close. The coherence score showed the same trend for the experimentation with different values of the hyperparameters. It kept increasing for the first few topics, before having a fall after the seventh topic, as we can see from the Fig. 1.

Table 2. LDA: experimentation with n-grams threshold

LDA			
Threshold	Coherence	Perplexity	N. of Topics
10	0.421	–7.924	5
50	0.552	–7.750	5
80	0.577	–7.537	4
100	0.557	–7.428	2

Once selected the *threshold* of the n-grams, we further kept experimenting using α and η hyperparameters. We tried several values of α and η. Since we reasonably assumed that the distribution for words in topics and documents is sparse, we expected to have a better score with a value of α and η less than 1. The coherence score showed the same trend for the experimentation with different values of the hyperparameters. It kept increasing for the first few topics, before having a fall after the seventh topic, as we can see from the Fig. 1.

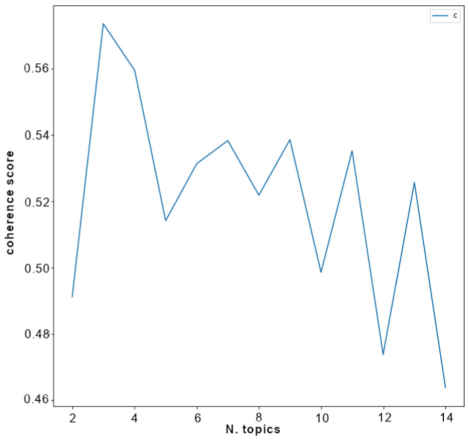

Fig. 1. Coherence scores of topics for LDA model (with $\alpha = 0.5$, $\eta = 0.01$)

The coherence scores between the optimal number of topics and the number of topics close to the optimal one do not flat out. Thus, despite the fact that the optimal number of topics often differs from one experimentation to another, the optimal range of topics remains the same. The coherence scores provides an overview of the number of topics available in the dataset.

The perplexity score changed only when the value of η was set as the lowest (0.01), and it confirms the assumption about the sparsity of the distribution of the words in each topic. Consequently, this leads to believe that the size of vocabulary for each topic is variable and topics contain uncertain word combinations. Regarding the model, the optimal number of topics have the same range of a number of the ones with the LDA experimentation, as we can see from the Table 4. The coherence scores showed a slightly better performance, especially when the *chunksize* is small and the *decay* is not more than 0.5.

Table 3. LDA: experimentation with α and η and *threshold* of bi-grams and tri-grams as 80

LDA				
α	η	Coherence	Perplexity	N. of Topics
Symmetric	*Symmetric*	0.547	–7.605	2
Auto	*Auto*	0.577	–7.536	4
0.01	*Symmetric*	0.539	–7.474	6
0.01	*Auto*	0.546	–7.498	3
0.5	*Auto*	0.546	–7.498	3
2	*Auto*	0.576	–7.905	6
0.01	0.01	0.564	–10.201	5
0.5	0.01	0.573	–10.223	3
2	0.01	0.533	–10.407	2
2	0.5	0.571	–7.554	4
2	2	0.576	–7.791	3

Table 4. LSI: experimentation with *decay* and *chunksize*

LSI			
Decay	Chunksize	Coherence	N. of Topics
0.1	10	0.627	6
0.1	50	0.591	4
0.1	100	0.559	3
0.5	10	0.624	4
0.5	50	0.605	5
0.5	100	0.622	4
1	10	0.552	3
1	50	0.530	3
1	100	0.522	10

As mentioned in Sect. 3.4, the purely quantitative metrics can limit the overall evaluation. Consequently, we looked for topics in detail through the observation-based. We considered the top 15 words of each topic, and we evaluated whether any word was shared among the topics. We made use of the Word Cloud to visualize better the top words (Fig. 2).

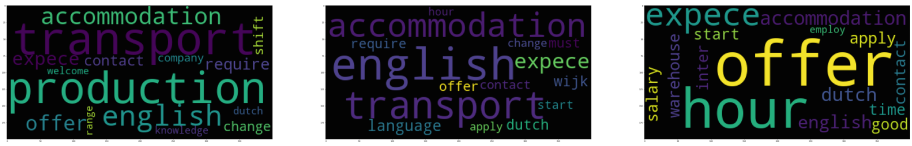

Fig. 2. Word Cloud of the 3 topics of the best LDA model

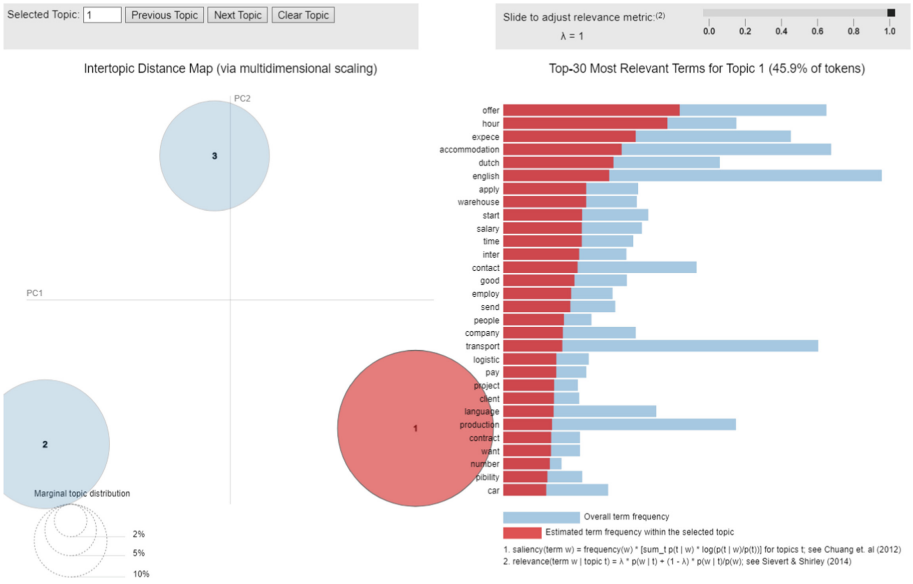

Fig. 3. Visualization of the LDA model

The majority of the topics could not identify as a determined class of job or a specific working condition. Few words were not shared among the topics, and only a few of them can be representative of a kind of job, working condition, or an evident writing pattern. The LDA best model had **3** topics, a perplexity score of **−10.223** and a coherence score of **0.573**. By observing the top 15 words, at most one word per topic, *'warehouse'* and *'shift'* for example, was relevant to differentiate the topics. We further visualized the LDA model in an intertopic distance map (Fig. 3). It provided more insight regarding the top 15 words previously analyzed. Despite most of the top words being shared in every topic,

different words had different saliency levels. For example, the word *'accommodation'* has a high value of relevance but a saliency close to 0. On the contrary, term such as *'production'* has a high relevance as well as a high saliency towards a specific topic, which makes it more informative. To access the interactive file of our LDA topic modelling please refer to [1] file name `everything.html`.

4.2 Logistic Regression Results

Regarding the regression analysis, we implemented the procedure previously described in 3.5. We used a *Python* library[16] to evaluate whether the features built can be significant to reveal if the salary offered in a job announcement is appropriate to the national values. We considered 26 features created after the econding phase. We also included topics from the LDA model.

Before the implementation, we considered reducing the dimensions to avoid using irrelevant features that would have only increased the time complexity. We used the recursive feature elimination, which is a wrapper-type algorithm that searches for a subset of features by starting with all features and subsequently removing them until a fixed number is provided [16], which in our case is 14, more than a half.

We implemented the model. Out of the 14 features that we initially had, we removed the ones with a *p-value* higher than 0.05, which indicates the statistical significance for a confidence interval of 95%. As a result, 6 variables are removed. Detailed results are available in the online appendix [1]. After this step, we wanted to check whether the model changed the significance for the remaining variables, so we took into consideration the *Pseudo-*R^2 to evaluate the change of the interpretability of the model [27]. The *Pseudo-*R^2 of the model with 14 features is 0.385, while the one of the models without significant features is 0.263. The value showed a decrease. However, it is still in a good fit range for the machine learning estimation, as it is demonstrated to be between 0.2 and 0.4 [17]. The significant variables kept by the model were 7, 3 regarding the language of the text, 3 about the type of the job and the last one is the first topic of the LDA model. We consider the Durbin Watson statistic is for the auto-correlation in the model's output. The value is –1.942, which was a sign of zero or low level of auto-correlation in the residuals.

As for 'topic' variable, its regression coefficient was negative, which indicates a negative proportional relationship between the text similar imputable to this topic and the *delta salary*. Since we wanted to evaluate the importance of the variable in the logistic model, we ran the regression without the topic. There is a slight reduction of the value of the *Pseudo-*R^2 (0.249) and a slight increase of the log-odd, from 0.440 to 0.446, which means that the topic had some importance in the prediction. Detailed results are available in the online appendix [1]

We analyzed several metrics used in a binary classification task namely accuracy, recall, precision and F1 score.

[16] https://www.statsmodels.org/dev/example_formulas.html.

As we can see from Table 5, while the negative *delta salary* had high values of precision, recall, and F1 score, the positive class had a low value of F1 score, which is caused by a low recall. Moreover, since the amount of data leans towards the negative class, the score of the macro average is lower than the one of the weighted average.

Table 5. Classification results

# 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

Detailed results are available in the online appendix [1]

5 Discussion

Our initial investigation showed a match between the labor indicators stated by the literature that we reviewed and the one from social media and how they can affect the job research in the unskilled job market. From our research study we found that human judgment still plays a big role in the evaluation and interpretation. For example, the presence of the wage can indicate that we are more likely to deal with unskilled job offers than skilled ones [3,40] By combining other indicators such as the type of job, the description, and the national salary for that job, we can assess if the offer is adequate to the national standard.

Offering a low-paid job but still above the minimum wage does not implicitly entail illegal work. We have then to discern what kind of job is more susceptible to salary discrepancy. In this regard, we need to shed light on a better evaluation of the recruitment process's weakness and what the main actors are involved for a further investigation.

Besides the salary, the richness of the information in a job description can also play an important role during the evaluation. Job announcements with short text descriptions were difficult to frame into a job category since they contained very few keywords regarding the job and, in most cases, they belonged to more than one job type. This announcement shortness problem leads the categorization ambiguous, compromising the next steps. The topic modeling phase suffered from the presence of entries with a short text. Some of them were discarded due to the lack of real information. However, others were kept since they were job offers, and ignoring them could have reduced the variety and the true representation in the social media context.

Stopwords are also important factors in topic modeling. There are words such as *'english'*, *'contact'*, *'company'* or *'worker'*. These words can be considered general words in a job description and are not insightful. However, considering them

on a par with *stopwords* and removing them can lead to a complete outcome, with different topics from a less rigorous *stopwords* selection. The n-gram threshold is also a parameter that affects the output. Increasing the threshold we had a fewer number of topics and a higher coherence score, but the topics are difficult to interpret as the variety of most relevant and salient words is really low.

It was not the only time that we encountered the conflict between better performance with quantitative metrics and questionable performance with human judgment. When experimenting with LDA hyperparameters, we noticed that a better model was given with a low value of η, since it reduced the perplexity, which is connected to the model's generalization. However, the results showed topics with defined characteristics but very ambiguous from each other. Overall we can state that we Topic 1 is more related to the type of **job offer** and the amount of **hours**, hence we have also terms like **salary** offered in a job post. Topic 2 appears to be more related to the type of job, indeed in this topic we have terms like **production** and **transport** that give the idea of the main type of job offered. In the last topic, Topic 3, we find prevalence of languages like **English** and **Dutch**. Topic 3 is representative if related to the type of languages requested to work in Netherlands. The complete analysis is available in the Appendix online [1].

We aimed to capture in a so-called topic the type of job and the job description's linguistic features simultaneously. Both LDA and LSI models displayed good results with quantitative metrics, i.e., coherence and perplexity but hard interpretation and human evaluation of the data, mostly because of the lack of interpretable embedding. One problem is that we do not have clear evidence that components from one topic have a positive or negative sense. We tried to determine an explanation in this sense by including the topics as features in the regression model. In fact, we had a clear relationship between statistically significant topics and wages offered in the job announcement.

Nevertheless, metrics used along with the intertopic distance map, such as saliency and relevancy, provided more information about a single word in each topic. In this way, we discovered more insightful patterns in each topic. This last evidence showed the potential that topic modeling could achieve for this particular field of research.

As we can see from the results of our prediction model, other information can be relevant to assessing the job announcement's fairness. Different types of language might affect the salary offered. Although our dataset cannot be considered a complete representation of the job market, it is interesting that language is a significant aspect of the data. Considering that every kind of analysis needs language conformity to provide a comparison in a final evaluation, translating the text into language might risk losing essential information for the analysis.

Finally, we believe that our approach—which needs further studies—can represent the starting point for future investigation on how AI can help police to detect and prevent labor exploitation starting from job advertisements, thus representing a revolutionary step that can help decrease this issue. We are already collaborating with Dutch Police to provide an intelligence dashboard that consists of an AI model for real-time crawling data from social media and highlights

possible labor exploitation in the dutch labor market. i.e., SENTINEL Project (see the acknowledgment). These studies will also converge in providing a more general framework and tool that different countries' police can use.

6 Conclusion

In this work, we provide a promising approach for detecting and analyzing potential labor exploitation indicators in social media. First, we examined indicators of potential labor exploitation from the current literature, and we investigated their presence in real data. Then we extracted and pre-processed data from Facebook groups and pages that offer job advertisements in the dutch labor market. To extract important feature that characterizes labor, we apply topic modeling techniques, i.e., Latent Dirichlet Allocation and Latent Semantic Indexing. Then, based on the topic extracted, we constructed a logistic regression model for predicting salary discrepancy from the wage expected to the national standard. The results of our model are encouraging (F1 Score 61%), thus meaning that the Artificial Intelligence approaches should be considered for any criminal investigation. In future works, we want to consider other types of social networks and try different parameters configurations when running a machine learning model.

Acknowledgements. We thank Davide Carnevale for the work done during his master thesis. The work is supported by EU Twining DESTINI project, and, the Dutch Ministry of Justice and Safety through the Regional Table Human Trafficking Region East Brabant sponsored the project SENTINEL.

References

1. Appendix: unsupervised labor intelligence systems: a detection approach and its evaluation (2022). https://doi.org/10.6084/m9.figshare.19481339.v1
2. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)
3. Brenčič, V.: Wage posting: evidence from job ads. Can. J. Econ./Revue canadienne d'économique **45**(4), 1529–1559 (2012)
4. Burbano, D., Hernandez-Alvarez, M.: Identifying human trafficking patterns online. In: 2017 IEEE Second Ecuador Technical Chapters Meeting (ETCM), pp. 1–6. IEEE (2017)
5. Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J.L., Blei, D.M.: Reading tea leaves: how humans interpret topic models. In: Advances in Neural Information Processing Systems, pp. 288–296 (2009)
6. Chinnov, A., Kerschke, P., Meske, C., Stieglitz, S., Trautmann, H.: An overview of topic discovery in twitter communication through social media analytics. In: Americas Conference on Information System (2015)
7. Cockbain, E., Bowers, K., Dimitrova, G.: Human trafficking for labour exploitation: the results of a two-phase systematic review mapping the European evidence base and synthesising key scientific research evidence. J. Exp. Criminol. **14**(3), 319–360 (2018)

8. Cordeiro, M.: Twitter event detection: combining wavelet analysis and topic inference summarization. In: Doctoral Symposium on Informatics Engineering, vol. 1, pp. 11–16 (2012)
9. Council of Europe: Third Report on the Progress Made in the Fight Against Trafficking in Human Beings. European Commission (2020)
10. Cvijikj, I.P., Michahelles, F.: Monitoring trends on facebook. In: 2011 IEEE Ninth International Conference on Dependable, Autonomic and Secure Computing, pp. 895–902. IEEE (2011)
11. Deng, Q., Gao, Y., Wang, C., Zhang, H.: Detecting information requirements for crisis communication from social media data: an interactive topic modeling approach. Int. J. Disast. Risk Reduct. **50**, 101692 (2020)
12. Di Nicola, A., et al.: Surf and sound. The role of the internet in people smuggling and human trafficking. eCrime (2017)
13. Forte, E., Schotte, T., Strupp, S.: Serious and organised crime in the EU: The EU serious and organised crime threat assessment (SOCTA) 2017. Eur. Police Sci. Res. Bull. **16**, 13 (2017)
14. Gerber, M.S.: Predicting crime using twitter and kernel density estimation. Decis. Support Syst. **61**, 115–125 (2014)
15. Godin, F., Slavkovikj, V., De Neve, W., Schrauwen, B., Van de Walle, R.: Using topic models for twitter hashtag recommendation. In: Proceedings of the 22nd International Conference on World Wide Web, pp. 593–596 (2013)
16. Guyon, I., Weston, J., Barnhill, S., Vapnik, V.: Gene selection for cancer classification using support vector machines. Mach. Learn. **46**(1), 389–422 (2002)
17. Hensher, D.A., Stopher, P.R.: Behavioural Travel Modelling. Routledge, London (2021)
18. Hong, L., Davison, B.D.: Empirical study of topic modeling in twitter. In: Proceedings of the First Workshop on Social Media Analytics, pp. 80–88 (2010)
19. Hughes, D.M.: Trafficking in human beings in the European Union: gender, sexual exploitation, and digital communication technologies. SAGE Open **4**(4), 2158244014553585 (2014)
20. Immonen, A., Pääkkönen, P., Ovaska, E.: Evaluating the quality of social media data in big data architecture. IEEE Access **3**, 2028–2043 (2015)
21. Jeong, B., Yoon, J., Lee, J.M.: Social media mining for product planning: a product opportunity mining approach based on topic modeling and sentiment analysis. Int. J. Inf. Manage. **48**, 280–290 (2019)
22. Kasiviswanathan, S.P., Melville, P., Banerjee, A., Sindhwani, V.: Emerging topic detection using dictionary learning. In: Proceedings of the 20th ACM International Conference on Information and Knowledge Management, pp. 745–754 (2011)
23. Kejriwal, M., Szekely, P.: An investigative search engine for the human trafficking domain. In: d'Amato, C., et al. (eds.) ISWC 2017. LNCS, vol. 10588, pp. 247–262. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-68204-4_25
24. Khanjarinezhadjooneghani, Z., Tabrizi, N.: Social media analytics: an overview of applications and approaches. In: Proceedings of the 13th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2021) - Volume 1: KDIR, pp. 233–240I (2021)
25. Ko, N., Jeong, B., Choi, S., Yoon, J.: Identifying product opportunities using social media mining: application of topic modeling and chance discovery theory. IEEE Access **6**, 1680–1693 (2017)
26. Latonero, M.: Human Trafficking Online: The Role of Social Networking Sites and Online Classifieds. SSRN 2045851 (2011)

27. McFadden, D., et al.: Conditional Logit Analysis of Qualitative Choice Behavior. Academic Press, New York (1973)
28. Prier, K.W., Smith, M.S., Giraud-Carrier, C., Hanson, C.L.: Identifying health-related topics on twitter. In: Salerno, J., Yang, S.J., Nau, D., Chai, S.-K. (eds.) SBP 2011. LNCS, vol. 6589, pp. 18–25. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-19656-0_4
29. Röder, M., Both, A., Hinneburg, A.: Exploring the space of topic coherence measures. In: Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, pp. 399–408 (2015)
30. Rohani, V.A., Shayaa, S., Babanejaddehaki, G.: Topic modeling for social media content: A practical approach. In: 2016 3rd International Conference on Computer and Information Sciences (ICCOINS), pp. 397–402. IEEE (2016)
31. Rosario, B.: Latent semantic indexing: an overview. Techn. rep. INFOSYS **240**, 1–16 (2000)
32. Shahbazi, Z., Byun, Y.C.: Analysis of domain-independent unsupervised text segmentation using LDA topic modeling over social media contents. Int. J. Adv. Sci. Technol **29**(6), 5993–6014 (2020)
33. Siddiqui, T., Amer, A.Y.A., Khan, N.A.: Criminal activity detection in social network by text mining: comprehensive analysis. In: 2019 4th International Conference on Information Systems and Computer Networks (ISCON), pp. 224–229. IEEE (2019)
34. Sweileh, W.M.: Research trends on human trafficking: a bibliometric analysis using scopus database. Glob. Health **14**(1), 1–12 (2018)
35. Tong, E., Zadeh, A., Jones, C., Morency, L.P.: Combating human trafficking with deep multimodal models. arXiv preprint arXiv:1705.02735 (2017)
36. United Nations: Global Report on Trafficking in Persons 2020. UN (2021). https://books.google.nl/books?id=gGxczgEACAAJ
37. Vayansky, I., Kumar, S.A.: A review of topic modeling methods. Inf. Syst. **94**, 101582 (2020)
38. Volodko, A., Cockbain, E., Kleinberg, B.: "spotting the signs" of trafficking recruitment online: exploring the characteristics of advertisements targeted at migrant job-seekers. Trends Organ. Crime **23**(1), 7–35 (2020)
39. Wang, X., Gerber, M.S., Brown, D.E.: Automatic crime prediction using events extracted from twitter posts. In: Yang, S.J., Greenberg, A.M., Endsley, M. (eds.) SBP 2012. LNCS, vol. 7227, pp. 231–238. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-29047-3_28
40. Zhang, S.X., Cai, L.: Counting labour trafficking activities: an empirical attempt at standardized measurement. In: Forum on Crime and Society, vol. 8, pp. 37–61. United Nations (2015)
41. Zhu, J., Li, L., Jones, C.: Identification and detection of human trafficking using language models. In: 2019 European Intelligence and Security Informatics Conference (EISIC), pp. 24–31. IEEE (2019)