Mental health and precarious work in the Netherlands

Moin Niamat[2601479]

Vrije Universiteit, De Boelelaan 1105, 1081HV Amsterdam, the Netherlands

Abstract. Working is important to be able to pay rent, buy food and make life easier, but also to have a purpose in life. Even though work has an important positive influence on our lives, it can also have negative side effects, namely job insecurity which is linked to precarious work. The more precarious a job is, the more (mental) health issues employees experience. Furthermore, precarious work has become more prevalent in the economy. Due to this, we tried to automatically analyse the sentiment towards precarious work conditions in the Netherlands from statements made on social media, using modern Natural Language Processing (NLP) techniques. The rise of NLP techniques and the accuracy of these techniques such as BERT and RoBERTa, are used to predict the sentiment of labelled Tweets. Furthermore, we categorized the Tweets based on the Employment Precariousness Scale (EPRES). Lastly, we analysed if it is possible to determine the sentiment per dimension of precarious work and the magnitude of the sentiment. We obtained that the results are limited for predicting the sentiment towards precarious work conditions.

Keywords: NLP · Sentiment · Twitter · Precarious work.

1 Introduction

In the current society it is necessary to work, because we live in a capitalistic society, in which we need money to survive. Given the importance of work, insecurity about your job can create serious worries. What if your employer suddenly decides to fire you and all of a sudden you can not pay your rent. This kind of job insecurity causes health disparities (Landsbergis, Grzywacz, and LaMontagne 2014). Research has shown that the more precarious a job the more health issues employees have [1]. Employees with a permanent contract can not get fired on the spot without a good reason so for these people, their health will be less influenced since they will experience less job insecurity. However, for people without a permanent contract, this might be a lurking threat, which can cause mental and physical health problems (Kivimäki et al. 2003). Working circumstances like these can be classified as precarious work. Precarious work is a broad term used for refers to non-standard or temporary employment which may be not well paid, insecure or unprotected, or more general precarious work is every type of work with which a household cannot be supported (Fudge and Owens 2006). The term precarious work is used to describe all insecure and unstable conditions related to work. Not having a permanent contract is just one of the many aspects of precarious work. Some other examples of precarious work are: irregular working hours, low income and no or little respect from your boss. Diener and Biswas-Diener 2002 showed that money has a large impact on happiness. For example, if people have (more) money to spend on buying something nice and have enough clothes and food, people are happier. An other important thing you need money for to survive, is to pay rent. These are necessities of life that can be paid with your salary. Sorauren 2000 explained that money has influence on the motivation to work, however there are also other non-monetary related motivators. Furthermore, work also helps to give life a purpose but it also gives someone an identity and social life (Ward and King 2017). But just like everything, also work comes with insecurities. What if your employer suddenly decides to fire you and all of a sudden you can not pay your rent. This kind of job insecurity causes health disparities (Landsbergis, Grzywacz, and LaMontagne 2014). Research has shown that the more precarious a job the more health issues employees have [2]. Employees with a permanent contract can not get fired on the spot without a good reason so for these people, their health will be less influenced since they will experience less job insecurity. However, for people without a permanent contract, this might be a lurking threat, which can cause mental and physical health problems (Kivimäki et al. 2003). Working circumstances like these can be classified as precarious work. Precarious work is a broad term used for refers to non-standard or temporary employment which may be not well paid, insecure or unprotected work, or more general precarious

[1] http://www.voorverbeteringvatbaar.be/interview-slechte-job-slecht-voor-je-gezondheid/

[2] http://www.voorverbeteringvatbaar.be/interview-slechte-job-slecht-voor-je-gezondheid/

work is every type of work for which you are not able to support an household (Fudge and Owens 2006). Not having a permanent contract is just one of the many aspect of precarious work. The term precarious work is used to describe all insecure and unstable conditions related to work. Some other examples of precarious work are: irregular working hours, low income and no or little respect from your boss.
Over the years, precarious work has become more prevalent in the economy [3]. It is estimated that 60% of the contracts that are initiated between 2007 and 2011 did not follow the standard of "fixed job, fixed working hours" (OECD 2015). This implies that more then half the employees that got a contract during 2007 and 2011, lives with the idea they experience job insecurity. This makes them work as hard as they can and maybe even harder then they would work, if they had a permanent contract, even though this was not asked for by the employee (Green 2008). The Dutch Central Bureau for Statistics (CBS), showed that from 2013 till 2021, the number of non-permanent employees increased over time (CBS 2022).
Several studies have shown that insecure working conditions can have an impact on mental health, causing a depression, burn-out or feelings of unhappiness, compared to people with a permanent contract. One of these studies was conducted by Virtanen et al. (2007) which concluded that people with non-permanent contracts generally have more mental health issues, compared to people with a permanent contract.

Social media use has been increasing for the past decades, people are increasingly using it to share achievements but also to express their feelings and displeasure about subjects (Auxier and Anderson 2021). In 2008, 100 million people used Facebook, whereas in 2018, 2.26 billion people were on Facebook [4]. We assume that people affected by precarious work might use social media to express their thoughts and feelings about about their working conditions. Nowadays people tend to (over)share everything what is on their mind on social media (Pennington 2020). One social media platform often used is Twitter. On Twitter, users can express themselves within a limit of 280 characters. Twitter also lets users use so called hashtags. These hashtags allows users to categorize their tweets making it easier classify tweets. This makes it possible to find the subcategories of tweets you are looking for. This research will utilise the hashtag feature of Twitter to scrape for tweets that are related to precarious work in the Netherlands.

The goal of this research is to analyse whether it is possible to use modern Natural Language Processing (NLP) techniques to determine the sentiment of people toward precarious work on social media. We aim to better understand how employees experience precarious work and the influence it has on them and their health, by analysing Tweets related to this. We focused on Tweets, because there are free available and easy applicable libraries to scrape tweets. By doing

[3] https://www.liverpool.ac.uk/heseltine-institute/blog/precarious-work-is-on-the-rise/
[4] https://ourworldindata.org/rise-of-social-media

so this research hopes to aid in the improvement of working conditions and mental health in the Netherlands. The main research question is: *Can the sentiment towards precarious work conditions in the Netherlands be determined using NLP techniques?* To help answer the main research question two sub questions have been formulated. The first sub questions is: *Can we correctly select and categorize the tweets based on the Employment Precariousness Scale (EPRES) formulated by Vives et al. 2010 using NLP techniques ?* The Employment Precariousness Scale defines six categories of precarious work. By grouping the tweets into these clusters a better understanding of the different aspects of precarious work can be obtained. The second sub question is: *Is it possible to determine the sentiment per dimension of precarious work and the magnitude of the sentiment?* To obtain the tweets, we will scrape the Dutch tweets. Then we will label a subset of these tweets to train or model on. After which NLP techniques will be applied to process the tweets. To the best of our knowledge, this type of analysis has not been done.

We will first review previously done research. Then we will present our experimental design. This is followed by the analysis of the results in section four. To conclude this research we will then explain the conclusions and discuss these. In this last section we will also explain the limitations of this research.

2 Related work

For this research, it is important to understand precarious work and what type of research has been done. Furthermore, it is important to understand Natural Language Processing since this will be the approach used to answer the research question.

2.1 Background on Natural Language Processing (NLP)

Natural Language Processing is the foundation of our research, hence we first conducted research on NLP. NLP is defined as the sub field where Artificial Intelligence and Computer Science and Linguistic, are combined and making computers understand the words (spoken or written) by humans (Chopra, Prashar, and Sain 2013). The development of NLP can be divided into three phases. The earliest phase is called symbolic NLP. Natural Language Processing (NLP), was introduced by Turing and Haugeland 1950. In this article, he recommended to use the Turning test as criterion of intelligence. This test consists of a task which contains automated interpretation and generation of natural language.
After Alan Turing's research, Searle 1980 introduced the second phase: symbolic NLP. He did this by conducting the Chinese room experiment. This experiment went as follows: given a collection of rules, in this case a Chinese phrasebook, with questions and matching answers, the computer is instructed to pursue NLP tasks by implementing the rules to the confronted data. During the 60's, many natural language processing systems have been developed. One of the most famous ones is ELIZA, written by Weizenbaum 1966, which is a simulation of a Rogerian psychotherapist. ELIZA uses almost no information on the human thoughts or emotions, however, they found that ELIZA was surprisingly able to sometimes provide a human-like interaction.
Until the 80's, most of the NLP systems where constructed based on a complex set of rules. From the late 80's however, a revolution in NLP took place due to the introduction of machine learning algorithms for NLP. This second phase of NLP is also referred to as Statistical NLP. This revolution was caused by both the continuous increase in computational power (Moore 1965) and the decreasing dominance of the theories of linguistics (Ten Hacken 2007). Chomskyan's theories discouraged the use of corpus linguistics which control the machine-learning approach to language processing (ibid.). During the mid 90's, together with the growth of the world wide web, an increasing amount of unannotated language data became available. This led to more and more focus on semi and un-supervised learning algorithms (Hinton and Sejnowski 1999). These type of algorithms use data which has not been annotated. Hence, leading in a more difficult learning algorithms compared to supervised learning.
In the last decade, machine learning methods, such as, representation learning and deep neural network have broadly used in NLP. This is the current state of the field, also known as Neural NLP. The rise of this was partly caused by the many results that showed that these techniques (Goldberg 2016, Goodfellow, Bengio, and Courville 2016) can provided state-of-the-art results for many

natural language tasks, such as language modeling (Jozefowicz et al. 2016) and parsing (Charniak et al. 2016, Vinyals et al. 2015). Language modelling is can be looked at as a probability distribution over multiple sequences of words (Jozefowicz et al. 2016). Given a sequence of words of length n, a probability is assigned by the language model, to the complete sequence. These results were especially important in healthcare, since NLP can be used to analyse text in digital medical records.

2.2 Current NLP algorithms

One approach often used for sentiment analysis is TextBlob, which is an simple Python library supporting complex analysis and operations on textual data (Gujjar and HR 2021). For lexicon-based approaches, it is able to define sentiment by using the intensity of each word within a sentence and its semantic orientation and intensity (Bonta and Janardhan 2019). To do this, TextBlob needs a a predefinded dictionary that classifies words as positive or negative. It then assigns scores to all words individually, and then calculates the final sentiment.

Valence Aware Dictionary and sEntiment Reasoner (VADER) is an other approach often used for sentiment analysis. It is an open source, rule lexicon based analyzing pre-built library that can be found within NLTK (library of NLP) (Maitra and Sandhya 2022). VADER is designed specifically to analyse sentiments in social media, and uses the combination of both the sentiment lexicon and a list of lexical features which are labeled by their semantic orientation (either positive or negative) (Hutto and Gilbert 2014). It then calculates the sentiment of the text and returns the probability on the given input sentence if it is positive, negative or neutral. VADER is able to analyse data from most social media platforms such as Facebook and Twitter.

Flair is an other tool which can also be found as a Python library just like VADER and TextBlob and it is also an open source sentiment analysis tool. Flair is directly build on PyTorch and the developers of Flair have released different pre-trained models that can be used for different tasks (Wolf et al. 2020; Kula et al. 2020). One of Flair‘s pre-trained models is a model designed for sentiment analysis which is trained on a IMDB dataset (Shaikh et al. 2022). This model is simple to use for making predictions. In addition to this, it is also possible to use Flair to train a classifier using your own dataset (Akbik et al. 2019).

One of the most recent NLP tools is Bidirectional Encoder Representations from Transformers (BERT), (Devlin, Chang, et al. 2018). BERT, made by Google, is machine learning technique, based on transformers, is specially designed for pre-training NLPs (Devlin and Chang 2018). BERT continuously learns unsupervised, from unlabeled text and keeps improving itself when its used. Furthermore, it is able to understand the human language in its natural spoken manner. Google has been using BERT for its search engine. Furthermore, since the late 2020s BERT has been implemented in almost every

English-language query. In addition to this, BERT has become universal NLP experiments (Rogers, Kovaleva, and Rumshisky 2020). This is why we will dive deeper into BERT to understand how we can implement it in our research.

2.3 Bidirectional Encoder Representations from Transformers (BERT)

Since BERT has state-of-the-art performance on natural language, and out performed any other known NLP on eleven NLP tasks, it is interesting to implement it into the research (Devlin, Chang, et al. 2018). Furthermore, it is simple and empirically powerful language representation model. ibid., made this model to improve the fine-tuning based approaches. Furthermore, they explained why it is important to perform bidirectional pre-training for language representations. The reasons why it has state-of-the-art performance, however, is not well understood (Kovaleva et al. 2019, Clark et al. 2019).
BERT is assembled in a two step framework, namely pre-training and fine-tuning. The pre-training phase is used to train the model on an unlabeled dataset and over different pre-training tasks. During the second phase the model is fine-tuned and initiated with the pre-trained parameters, after which all these parameters are fine-tuned by implementing the labeled dataset from the downstream tasks (Devlin, Chang, et al. 2018). These downstream tasks each have separate fine-tuned models, despite being initiated with the same pre-trained parameters.

2.4 Robustly optimized BERT pre-training Approach (RoBERTa)

In the paper of Adoma, Henry, and W. Chen 2020, they showed that the RoBERTa model of Liu et al. 2019, has higher accuracy scores, compared to BERT (Devlin, Chang, et al. 2018), XLNet (Yang et al. 2019), and DistilBERT (Sanh et al. 2019). RoBERTa is, just like BERT, a self-training method. It is based on the same framework as BERT, however some modifications are applied to the pre-training phase which improved end-task performance. For example for BERT, the masking is done only one time at the data preparation step. BERT takes each sentence and masks it in ten different manners, hence at training only those ten variations, of each sentence will be observed. RoBERTa on the other hand performs masking at training which leads to each sentence being incorporated into a minibatch, get mask and hence the number of variations for each sentence is unbounded (Liu et al. 2019). Furthermore, this new model is trained with dynamic masking, full-sentences without NSP loss, larger byte-level BPE and large mini-batches. Since RoBERTa is an improvement on BERT, we are interested in both models and want to compare the results.

2.5 Dutch versions of RoBERTa

Since we are interested in analysing Dutch Tweets and precarious work in the Netherlands, the Dutch version of RoBERTa will be used, which is called RobBERT. RobBERT is an improvement on the Dutch version of BERT, which is

BERTje, developed by Vries et al. 2019. This model works the same as BERT with the only difference being that BERT is trained on Dutch tokens only on Wikipedia text, whereas BERTje is constructed based on a large and diverse dataset of 2.4 billion tokens.
The Dutch version of RoBERTa is RobBERT, developed by Delobelle, Winters, and Berendt 2020. The main difference between RoBERTa and RobBERT is that they used BERTje instead of BERT but they applied the same improvements as Liu et al. 2019 did on BERT. Since both BERTje and RobBERT gave state-of-the-art results and because they are open source and are available in Dutch, we will use both models to train our data on and compare the results.

2.6 Precarious Employment

Besides the NLP methods mentioned above, we also needed to obtain some insight in precarious work, especially how to analyse it. Padrosa et al. 2021 used Employment Precariousness Scale for Europe, also known as EPRES-E as main variable and operationalized this using 13 proxy indicators. These proxy indicators where sorted into six dimensions. They followed this approach since it is not possible to compare the precarious employment (PE) between countries due to the lack of internationally meaningful measures for the PE. It is however important to be able to understand this phenomenon and to learn more using country-specific information. This is why they tried to assess the measurement invariance (MI) for the Employment Precariousness Scale for Europe (EPRES-E). This is an adjustment EPRES constructed in the European Working Conditions Survey (EWCS) (Puig-Barrachina et al. 2014). In their results they showed that different scoring patterns in each dimension leads to similar EPRES-E scores. This implies that the PE is constructed using different sources inside the six dimensions for each country based on its broader sociopolitical trajectories. Due to this, they concluded that EPRES-E can be used to compare countries however, the score of each dimension should be expressed together with the overall EPRES-E score. This gave us the possibility to translate the operationalisation used by EPRES-E, to Dutch and then use these translations in combination with the dimensions mentioned earlier.

2.7 Sentiment analysis and text mining

As mentioned before, Twitter has been increasingly used by people to share their thoughts and feelings. Due to this many studies have been conducted on these Tweets, using text mining. Text mining is a process that is driven by information obtained from the written resources, such as articles, books, emails and websites (Allahyari et al. 2017). In text mining there are three different perspectives: data mining, information extraction and Knowledge Discovery in Databases (ibid.). Furthermore, text mining has three main processes: 1) structuring the input text, 2) pattern deriving inside the structured data, 3) evaluating and interpretation of the output.
Text mining and sentiment analysis can handle unstructured data, unlike other

"classical" data mining methods (Oza and Naik 2016). Lexicon is the primary sentiment analysis technique that classifies text data into predefined sentiment class sets. To calculate the scores, a sentiment lexicon (dictionary containing words and there corresponding sentiment scores) is used (Sun, Luo, and J. Chen 2017). Wang et al. 2012, stated a real-time sentiment analysis system to classify political tweets for the 2012 presidential elections in the USA. For this they used 36 million tweets. The model has an 59% accuracy for predicting the sentiment of the political tweet.

3 Experimental Design

Using the results obtained from the literature research we will set up the design. First we construct a dataset of Tweets by scraping 9073 Tweets Dutch Tweets based based on combinations of the following dutch words:

- Precair + Werk
- Precair + Job
- Precair + Baan
- Precair + Contract
- Onzeker + Werk
- Onzeker + Job
- Onzeker + Baan
- Onzeker + Contract
- Kwetsbaar + Werk
- Kwetsbaar + Job
- Kwetsbaar + Baan
- Kwetsbaar + Contract

To scrape Twitter, we used the programming language `python` and the scraping package `snscrape`. The Dutch Tweets are then checked by the two other researchers to make sure they are applicable for this research. Next, we will scrape the Tweets to clean them. Furthermore, all hyperlinks, emoticons, punctuation need to be removed and all the words will be changed into lowercase. In addition to this, the 1000 Tweets will be labelled based on dimensions of precarious work as done by Padrosa et al. 2021. These 1000 labelled Tweets are then used to train our model, such that our model is able to classify the other Tweets based on the dimensions. In the best case, the model will be able to correctly divide the dimension of precarious work as stated by ibid. To do this labeling the following labels are used:

- Not relevant
- Temporariness
- Disempowerment
- Vulnerability
- Exercise of rights
- Unpredictability of working times
- Wages
- Precarious Tweet but no clear dimension
- Consequence of precarious work

3.1 Classification and sentiment analysis model

After this labelling, the models are trained. The first model that is trained, is the one for making predictions on the dimensions. After this the second model, for making predictions on sentiment analysis, will be trained. Both models are trained based on the Dutch Social Dataset. This model is able to determine the

sentiment of the Tweets which is also an important aspect for our research. After the data curation part, the NLP models were constructed. For this research two separate models are needed to make predictions, one to make predictions on the dimension and one for sentiment analysis. To construct both models, we will use RobBERT as base.

3.2 Methods

We will train two RoBERT models to make the predictions we are interested in and to perform sentiment analysis. For sentiment analysis we need to clean the Tweets for which we used regular expressions also known as rational expression. Regular expressions is a technique that searches for patterns of a string of characters and filters this out of the text (Mitkov et al. 2003). These patterns are often used to find a string or to find and replace a string. The package we used for this is the `RegEx` (Van Rossum 2020)
The design of BERT allows for fine tuning the pre-trained BERT model using only an output layer which provides state-of-the-art models Devlin, Chang, et al. 2018. Furthermore, BERT was constructed using two simultaneously pre-training tasks, namely Masked Language Modeling and Next Sentence Prediction. When we finetune the model, the output layers of these pre-training tasks are replaced by a new target task classifier.
There are two versions of BERT: a "base" version and a "large" version. The difference between these two versions is that the base version contains 12 encoder layers each having 12 heads and 768 dimensions, hence there are in total 110 million parameters. The large versions is extended version of the base version, containing 24 encoders instead of 12. In addition to this, it contains 16 heads and 1024 dimensions, so there are 240 million parameters in total.
BERT has to first create the embedding representation for a single sentence or pair of sentences, depending on the task we are interested in, to be able to process the input.
It starts by encoding each sentence, also referred to as "sequence", as a series of tokens by applying the WordPiece algorithm (Wu et al. 2016). This algorithm is designed for subword tokenization. It breaks down words into smaller parts if no correspondence in its vocabulary is found. It is possible that parts are broken down up to the character level which implies that an embedding representation can be created, for any present word in the input.
Next, the most often occurring n-grams are merged together until we reached a pre-set vocabulary size. Besides, WordPiece also takes the probability the probability into account when merging characters, instead of the frequency. BERT has a vocabulary of 30 thousand tokens. In addition to the basic tokens, BERT is also able to assign some special tokens, CLS, SEP, PAD and MASK. Furthermore, it contructs two extra embeddings during the pre-processing step. The first one is Segment Embedding, that states if a token is related to a sentence or not. The second one is Positional Embedding which is assigned to every token to retain the structural information of a sequence. This is important since elements

of the sequence are fed to the model simultaneously.

They did this through playing around with the pre-training setup of BERT. Their main adjustment was made on the NSP task. They trained the model on a bigger batch size and longer sequences. This was done by using Dynamic Masking and a larger dataset. Furthermore, Liu et al. 2019 explained that the NSP was not able to understand the meaning of long-distance dependencies. In contrast to what Devlin, Chang, et al. 2018 stated, experiments show that removing the NSP in the pre-training part, gives the same or even better results on down stream tasks. In addition, increasing the bigger batch size improves the capability of learning of the model. To be more specific, they change the one million steps with size 256 sequences to 125 thousand steps of 2 thousand sequences. This increased the perplexity score of the MLM task. Additionally, the Dynamic Masking method, gave better results than the Static Masking used in BERT. The last adjustment they made was using a better dataset to pre-train RoBERTa.

4 Results

Next, the results, obtained from training the models are stated. First the results from labelling the data will be explained. Then, the results from models for the dimensions will be shown. Lastly, the results obtained from the model for sentiment analysis are shown.

4.1 Data analysis

As mentioned in the previous section, the data worked with contained 1000 labeled Tweets. The distribution of the labels is shown in Figure 1. The labeling is done by the other researchers, based on their expertise they have in this field. From this we obtained that most of the Tweets are labeled as: "Not relevant", "Temporariness" and "Precarious tweet but no clear dimension". The unlabelled dataset contained 8073 tweets.

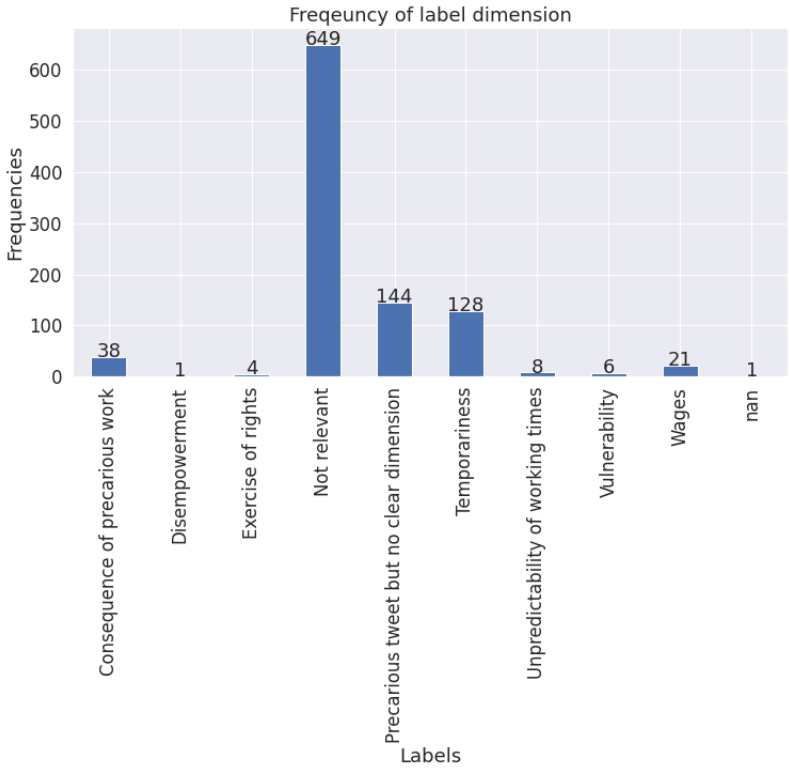

Fig. 1: Barplot of frequency after labeling dimensions

4.2 Hyperparamets

The parameters used for RobBERT model for classification are:

- Number of Epochs: 10
- Batch size: 32
- Padding: 128
- Learning rate: $2 \times e^{-}5$
- ϵ $1 \times e^{-}8$
- Optimizer: AdamW
- Output layer: 8

The parameters used for RobBERT model for sentiment analysis are:

- Number of Epochs: 4
- Batch size: 32
- Padding: 335
- Learning rate: $2 \times e^{-}5$
- ϵ $1 \times e^{-}8$
- Optimizer: AdamW
- Output layer: 3

The padding is decided based on the length of the longest tweet. The batch size of 32 was recommended by Devlin, Chang, et al. 2018 for efficient GPU memory usage. For the learning rate inspiration was taken form ibid. They fine tuned the model using the following learning rates: 5e-5, 4e-5, 3e-5, and 2e-5. After testing all these options 2e-5 was picked because it performed the best for our case. The ϵ was kept as the base value which is 1e-8. Changing this parameter did not change the results in any significant way. For optimizer AdamW was used just like in ibid. The output layers were decided based on the number of possible labels we had, for classifications of the dimensions this was 8 and for the sentiment this was 3.

4.3 Dimensions

The labels given to the Tweets are in words, but to be able to make predictions, this needed to be changed into numbers. This is how our model contains 8 output nodes.

- Consequence of precarious work = 0
- Disempowerment = 1
- Exercise of rights = 2
- Not relevant = 3
- Precarious tweet but no clear dimension = 4
- Temporariness = 5
- Unpredictability of working times = 6
- Vulnerability = 7
- Wages = 8

From Table 1, we obtain that the model is only able to predict labels 3,4 and 5. These are *Not relevant*, *Precarious tweet but no clear dimension* and *Temporariness*, respectively. This means that the model never predicts any of the other dimensions. Some examples of tweets and their assigned dimensions by the model can be seen in Table 5.

Label	0	1	2	3	4	5	6	7	8
Amount	0	0	0	7284	717	72	0	0	0

Table 1: Predictions of the dimensions

4.4 Sentiment Analysis

Next, the 54268 Tweets were used as test set to validate our model on the remaining 54269 samples. After this the Precision, Recall, F1-score and Support were calculated, to give us an insight on the performance of the model. These scores are shown in Table 2, and show that neutral and positive have, an overall higher score compared to negative, which implies that the model is not as good as in predicting negative as well as it is in predicting neutral and positive.

	Precision	Recall	F1-score	Support
Negative	0.86	0.69	0.77	8424
Neutral	0.82	0.95	0.88	31657
Positive	0.91	0.70	0.79	41188
Accuracy			0.84	54269
Macro average	0.86	0.78	0.81	54269
Weighted average	0.85	0.84	0.84	54269

Table 2: Classification metrics of sentiment analysis based on Dutch_Social

Next, we focused on the scraped Tweets and made predictions using the trained model. The following predictions were obtained from training this model:

Label	Negative	Neutral	Positive
Amount	4666	3063	1344

Table 3: The predicted sentiment of scraped Tweets

	Precision	Recall	F1-score	Support
Negative	0.61	0.65	0.63	152
Neutral	0.47	0.41	0.44	111
Positive	0.26	0.27	0.26	37
Accuracy			0.52	300
Macro average	0.44	0.45	0.44	300
Weighted average	0.51	0.52	0.51	300

Table 4: Classification metrics of sentiment analysis based on labeled sentiment

The determine the performance of the model on the scraped Tweets, 300 Tweets were labeled. This was again done by the supervisors. These Tweets are then used to compare it with the results from Table 3 to determine the performance of the trained model. The Accuracy score for the labelled dataset is equal to 0.52. This is shown in Table 4. Some examples of tweets and their assigned sentiment by the model can be seen in Table 6.

4.5 Example output form the models

Tweet	Dimension
Hoe kwetsbaar is een wielrenner op de baan? Paar keer geluk gehad op onze training vandaag @tomboonen1. Toch nog een mooie training gehad	Not relevant
Er hangt een vervelende sfeer onder het personeel. Mensen zijn onzeker over de toekomst en horen links en rechts dat mensen ontslagen worden. Ook mogelijke inhouding van 20% op het salaris is onderwerp van gesprek. Weinig gezelligheid op het werk #COVID19NL	Precarious tweet but no clear dimension
ik wil niet opzoek naar een nieuwe baan en solliciteren en geen loon krijgen en onzeker zijn over alles aaaa ik haat dit	Temporariness

Table 5: Example of tweets and their assigned dimensions

Tweet	Sentiment
tot verbijstering van het Wegenerpersoneel is hun baan onzeker nu dagblad de Pers failliet blijkt.Steeds meer werkelozen gaan tenonder!	Negative
Veel vragen over de Wet Werk en Zekerheid Een ding is wel zeker Met deze Wet is veel onzeker geworden #socakkoord #BTcontracten @OR_POSTNL	Neutral
For the record, ik houd gelukkig van mijn baan als postbezorger, want ik geniet van het buiten zijn, bomen en planten bekijken ondertussen. Toch zit je wel in een structuur vast, stel dat je ander werk wil gaan doen, dan is dat knap moeilijk als Wajonger en autist...#onzeker	Positive

Table 6: Example of tweets and their assigned sentiment

5 Conclusion

5.1 Discussion and Conclusion

To answer the research question: *Can the sentiment towards precarious work conditions in the Netherlands be determined using modern NLP techniques?* and sub questions: *Can we categorize the Tweets based on the Employment Precariousness Scale (EPRES) mentioned in Vives et al. 2010 using modern NLP techniques ?* and *Is it possible to determine the sentiment per dimension and the magnitude of the sentiment?* we will discuss the results. These results, shown in the previous section, gave us the following conclusions. First, we look at the model that predicts the dimensions. We obtain that only three out of nine dimensions is predicted. The model is only able to predict *Not relevant*, *Precarious tweet but no clear dimension* and *Temporariness*. As shown in Figure 1, the most probable reason why the model cannot predict the other labels is due to the other labels occurring less.
For the model constructed for the sentiment analysis we obtained the following results. In Table 2, we find that the the model is best able to predict "Neutral". This is obtained by the highest scores for the F1-score and the Recall. We need to keep in mind that the F1-score has as downside that it does not take into account the true negatives. If we look at the predictions made on the scraped Tweets dataset we obtain that mostly "Negative" is predicted. This is shown in Table 3. If we look at the accuracy score of the predictions of the scraped Tweets, we obtain a lower accuracy score compared to the accuracy score given in Table 1. This lower accuracy score could be due to the difficulty of labeling the Tweets. The supervisors told us that some Tweets were ambiguous or only contained a hyperlink or were just retweets. However, the accuracy score of 0.52 implies that we have a model that performs better than randomly assigning one of the three labels (which would give us an accuracy score of ≈ 0.33)
We obtain that not all dimensions can be predicted and we do not have a very high accuracy score.

5.2 Limitations and further research

Before we discuss the results, we want to explain that finding a design to construct the dataset. The first approach was to scrape Dutch Tweets based on hashtags that purely focus on precarious work at universities. For scraping the following hashtags were used: `#WOinActie`,
`#NormaalAcademischPeil` and `#alarmdag`. Since we are interested in determining the sentiment per dimension and the magnitude of the sentiment toward precarious work circumstances in the dutch academic field, we used the hashtags in the scraped Tweets to obtain additional Tweets that could also be relevant. From this could then construct a dataset containing Dutch Tweets about precarious work at universities. We then analysed the Tweets and obtained how often specific words occur in the Tweets which we can then set as keywords.

In addition to this we also analysed the Dutch language to see how often these words occur in it and then compare it with how often this word occurred in the scraped Tweets dataset. The keywords are then be used to further enhance the Tweets dataset. However, we were not able to answer our main research question with this approach nor apply NLP on it, and thus it could not be used for our research. As a result, it was not possible to make predictions and perform sentiment analysis. Since this approach, was focusing more on the causes of precarious work, i.e. globalization and rapid technological changes [5], rather than precarious work itself. The obtained Tweets did not relate to precarious work rather than it was focusing on the causes of precarious work, which made it not possible to answer our research question using this dataset. To be able to analyse on precarious work itself, we used the operationalisation. This was first translated to Dutch. We used these questions to determine the most important keywords and after which we scraped the Tweets based on these keywords. This approach however also did not gave us the information needed for our research. This was one of the most time consuming and most problematic part of this research and needs further improvement. Furthermore, improvements on the dataset should be done, such as labeling more data or scraping a wider variety of Tweets, since more data means more observations the model can learn from and hence improve the predictability of the model. An other improvement could be oversampling the less occurring dimensions by doing this the class imbalance will decrease. Due to the time constraint we did not succeed in applying this. In addition to this, since not all dimensions can be predicted and we do not have a very high accuracy score, further research should be done and improvement should be made to be able to use NLP techniques to determine the sentiment towards precarious work conditions. For example, improving the dataset by using a different labelling approach for the Tweets.

[5] https://archive.discoversociety.org/2018/04/03/precarious-work-understanding-the-new-employment-relations/#: :text=The%20recent%20rise%20of%20precarious,all%20collective%20arrangements%20that%20might

References

Adoma, Acheampong Francisca, Nunoo-Mensah Henry, and Wenyu Chen (2020). "Comparative analyses of bert, roberta, distilbert, and xlnet for text-based emotion recognition". In: *2020 17th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)*. IEEE, pp. 117–121.

Akbik, Alan et al. (2019). "FLAIR: An easy-to-use framework for state-of-the-art NLP". In: *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics (demonstrations)*, pp. 54–59.

Allahyari, Mehdi et al. (2017). "A brief survey of text mining: Classification, clustering and extraction techniques". In: *arXiv preprint arXiv:1707.02919*.

Auxier, Brooke and Monica Anderson (2021). "Social media use in 2021". In: *Pew Research Center* 1, pp. 1–4.

Bonta, Venkateswarlu and Nandhini Kumaresh2and N Janardhan (2019). "A comprehensive study on lexicon based approaches for sentiment analysis". In: *Asian Journal of Computer Science and Technology* 8.S2, pp. 1–6.

CBS (2022). *Werkzame beroepsbevolking; positie in De Werkkring*. URL: `https://www.cbs.nl/nl-nl/cijfers/detail/82646NED?dl=7165`.

Charniak, Eugene et al. (2016). "Parsing as language modeling". In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2331–2336.

Chopra, Abhimanyu, Abhinav Prashar, and Chandresh Sain (2013). "Natural language processing". In: *International journal of technology enhancements and emerging engineering research* 1.4, pp. 131–134.

Clark, Kevin et al. (2019). "What does bert look at? an analysis of bert's attention". In: *arXiv preprint arXiv:1906.04341*.

Delobelle, Pieter, Thomas Winters, and Bettina Berendt (2020). "Robbert: a dutch roberta-based language model". In: *arXiv preprint arXiv:2001.06286*.

Devlin, Jacob and Ming-Wei Chang (2018). "Open sourcing BERT: State-of-the-art pre-training for natural language processing". In: *Google AI Blog* 2.

Devlin, Jacob, Ming-Wei Chang, et al. (2018). "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *arXiv preprint arXiv:1810.04805*.

Diener, Ed and Robert Biswas-Diener (2002). "Will money increase subjective well-being?" In: *Social indicators research* 57.2, pp. 119–169.

Fudge, Judy and Rosemary Owens (2006). *Precarious work, women, and the new economy: The challenge to legal norms*. Bloomsbury Publishing.

Goldberg, Yoav (2016). "A primer on neural network models for natural language processing". In: *Journal of Artificial Intelligence Research* 57, pp. 345–420.

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville (2016). *Deep learning*. MIT press.

Green, Francis (2008). "Temporary work and insecurity in Britain: a problem solved?" In: *Social Indicators Research* 88.1, pp. 147–160.

Gujjar, J Praveen and Prasanna Kumar HR (2021). "Sentiment analysis: Textblob for decision making". In: *Int. J. Sci. Res. Eng. Trends* 7.2, pp. 1097–1099.

Hinton, Geoffrey and Terrence J Sejnowski (1999). *Unsupervised learning: foundations of neural computation*. MIT press.

Hutto, Clayton and Eric Gilbert (2014). "Vader: A parsimonious rule-based model for sentiment analysis of social media text". In: *Proceedings of the international AAAI conference on web and social media*. Vol. 8. 1, pp. 216–225.

Jozefowicz, Rafal et al. (2016). "Exploring the limits of language modeling". In: *arXiv preprint arXiv:1602.02410*.

Kivimäki, Mika et al. (2003). "Temporary employment and risk of overall and cause-specific mortality". In: *American journal of epidemiology* 158.7, pp. 663–668.

Kovaleva, Olga et al. (2019). "Revealing the dark secrets of BERT". In: *arXiv preprint arXiv:1908.08593*.

Kula, Sebastian et al. (2020). "Sentiment analysis for fake news detection by means of neural networks". In: *International conference on computational science*. Springer, pp. 653–666.

Landsbergis, Paul A, Joseph G Grzywacz, and Anthony D LaMontagne (2014). "Work organization, job insecurity, and occupational health disparities". In: *American journal of industrial medicine* 57.5, pp. 495–515.

Liu, Yinhan et al. (2019). "Roberta: A robustly optimized bert pretraining approach". In: *arXiv preprint arXiv:1907.11692*.

Maitra, Arunima and S Sandhya (2022). "Natural Language Processing in Court Order Scrutiny". In.

Mitkov, Ruslan et al. (2003). "Computer-aided generation of multiple-choice tests". In: *Proceedings of the HLT-NAACL 03 workshop on Building educational applications using natural language processing*, pp. 17–22.

Moore, Gordon (1965). "Moore's law". In: *Electronics Magazine* 38.8, p. 114.

Oza, Kavita S and Poornima G Naik (2016). "Prediction of online lectures popularity: a text mining approach". In: *Procedia Computer Science* 92, pp. 468–474.

Padrosa, Eva et al. (2021). "Comparing precarious employment across countries: measurement invariance of the employment precariousness scale for Europe (EPRES-E)". In: *Social Indicators Research* 154.3, pp. 893–915.

Pennington, Natalie (2020). "An examination of relational maintenance and dissolution through social networking sites". In: *Computers in Human Behavior* 105, p. 106196.

Puig-Barrachina, Vanessa et al. (2014). "Measuring employment precariousness in the European Working Conditions Survey: the social distribution in Europe". In: *Work* 49.1, pp. 143–161.

Rogers, Anna, Olga Kovaleva, and Anna Rumshisky (2020). "A primer in bertology: What we know about how bert works". In: *Transactions of the Association for Computational Linguistics* 8, pp. 842–866.

Sanh, Victor et al. (2019). "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter". In: *arXiv preprint arXiv:1910.01108*.

Searle, John R (1980). "Minds, brains, and programs". In: *Behavioral and brain sciences* 3.3, pp. 417–424.

Shaikh, Sarang et al. (2022). "Towards Understanding of User Perceptions for Smart Border Control Technologies using a Fine-Tuned Transformer Approach". In: *Proceedings of the Northern Lights Deep Learning Workshop*. Vol. 3.

Sorauren, Ignacio Falgueras (2000). "Non-monetary Incentives: Do people work only for money?" In: *Business Ethics Quarterly*, pp. 925–944.

Sun, Shiliang, Chen Luo, and Junyu Chen (2017). "A review of natural language processing techniques for opinion mining systems". In: *Information fusion* 36, pp. 10–25.

Ten Hacken, Pius (2007). *Chomskyan linguistics and its competitors*. Equinox London.

Turing, Alan M and J Haugeland (1950). "Computing machinery and intelligence". In: *The Turing Test: Verbal Behavior as the Hallmark of Intelligence*, pp. 29–56.

Van Rossum, Guido (2020). *The Python Library Reference, release 3.8.2*. Python Software Foundation.

Vinyals, Oriol et al. (2015). "Grammar as a foreign language". In: *Advances in neural information processing systems* 28.

Vives, Alejandra et al. (2010). "The Employment Precariousness Scale (EPRES): psychometric properties of a new tool for epidemiological studies among waged and salaried workers". In: *Occupational and environmental medicine* 67.8, pp. 548–555.

Vries, Wietse de et al. (2019). "Bertje: A dutch bert model". In: *arXiv preprint arXiv:1912.09582*.

Wang, Hao et al. (2012). "A system for real-time twitter sentiment analysis of 2012 us presidential election cycle". In: *Proceedings of the ACL 2012 system demonstrations*, pp. 115–120.

Ward, Sarah J and Laura A King (2017). "Work and the good life: How work contributes to meaning in life". In: *Research in Organizational Behavior* 37, pp. 59–82.

Weizenbaum, Joseph (1966). "ELIZA—a computer program for the study of natural language communication between man and machine". In: *Communications of the ACM* 9.1, pp. 36–45.

Wolf, Thomas et al. (2020). "Transformers: State-of-the-art natural language processing". In: *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pp. 38–45.

Wu, Yonghui et al. (2016). "Google's neural machine translation system: Bridging the gap between human and machine translation". In: *arXiv preprint arXiv:1609.08144*.

Yang, Zhilin et al. (2019). "Xlnet: Generalized autoregressive pretraining for language understanding". In: *Advances in neural information processing systems* 32.