Data Extraction for Sustainability Reporting

ENFUSE

Inspiring & Fusing - Enterprises & Students

Understanding the Challenge

SMEs often lack the resources to comply with new sustainability reporting regulations. 4See, a company focused on helping SMEs, identified the need for an automated, accurate, and scalable method to extract data from employee contracts. This research aims to develop such a solution using machine learning and data mining techniques.

This research focuses on the use of BERT to extract valuable data from unstructured documents like employee contracts to help SMEs comply to the sustainability reporting standards like the ESRS. This is done in collaboration with 4see and the ENFUSE competition. CRISP-DM methodology and BERT were used for extracting data from employee contracts. Two techniques were tested namely entity extraction and question answering available on the Hugging Face Platform.

1

Findings

Question-answering more effective than entity extraction

2

Performance

Performance was poor with CUAD but good with SQuAD2.0

3

Impact

Benefit SMEs by helping them meet sustainability standards efficiently and cost effectively

4

Future

Improve speed and accuracy of extraction.

1

Business Understanding

Clearly define the problem, requirements, and sustainability goals to guide the development of the solution.

2

Data Preparation

Convert the unstructured contract documents into a format suitable for natural language processing.

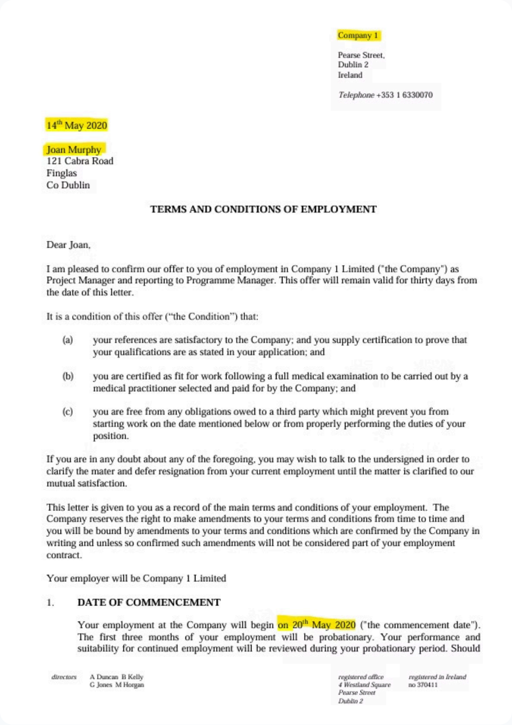
3

Modelling

Explore and compare different BERT-based approaches, including entity extraction and question-answering, to determine the most effective method.

Business Understanding

PDF or Word Documents



Company 1

Pearse Street,
Dublin 2
Ireland

Telephone +353 1 6330070

14th May 2020

Joan Murphy
121 Cabra Road
Finglas
Co Dublin

TERMS AND CONDITIONS OF EMPLOYMENT

Dear Joan,

I am pleased to confirm our offer to you of employment in Company 1 Limited ("the Company") as Project Manager and reporting to Programme Manager. This offer will remain valid for thirty days from the date of this letter.

It is a condition of this offer ("the Condition") that:

(a) your references are satisfactory to the Company; and you supply certification to prove that your qualifications are as stated in your application; and

(b) you are certified as fit for work following a full medical examination to be carried out by a medical practitioner selected and paid for by the Company; and

(c) you are free from any obligations owed to a third party which might prevent you from starting work on the date mentioned below or from properly performing the duties of your position.

If you are in any doubt about any of the foregoing, you may wish to talk to the undersigned in order to clarify the mater and defer resignation from your current employment until the matter is clarified to our mutual satisfaction.

This letter is given to you as a record of the main terms and conditions of your employment. The Company reserves the right to make amendments to your terms and conditions from time to time and you will be bound by amendments to your terms and conditions which are confirmed by the Company in writing and unless so confirmed such amendments will not be considered part of your employment contract.

Your employer will be Company 1 Limited

1. **DATE OF COMMENCEMENT**

Your employment at the Company will begin on 20th May 2020 ("the commencement date"). The first three months of your employment will be probationary. Your performance and suitability for continued employment will be reviewed during your probationary period. Should

directors A Duncan B Kelly
G Jones M Horgan

registered office
4 Westland Square
Pearse Street
Dublin 2

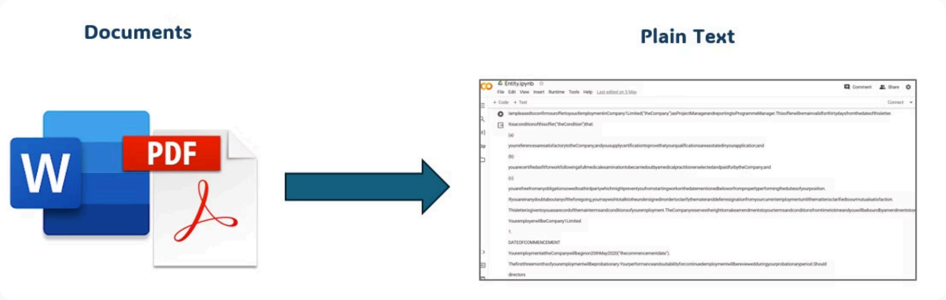
registered in Ireland
no 370411

Extracted data required:

- First Name
- Surname
- Employee's Address
- Company Address
- Position of job
- Rate of Pay
- Hours of work
- Break time
- Manager/ Supervisor name
- Date of contract signed
- Terms and Conditions of employment

Goals

- Extraction of the information to excel documents
- Sustainable
- Cost-effective
- Apply a user-friendly approach



```
IampleasedtoconfirmourofferloyouofemploymentinCompany1Limited("theCompany")asProjectManagerandreportingtoProgrammeManager.Thisofferwillremainvalidforthirtydaysfromthedateofthisletter.
Itisaconditionofthisoffer("theCondition")that:
(a)
yourreferencesaresatisfactorytotheCompany,andyousupplycertificationtoprovethatyourqualificationsareasstatedinyourapplication;and
(b)
youarecertifiedasfitforworkfollowingafullmedicalexaminationtobecarriedoutbyamedicalpractitionerselectedandpaidforbytheCompany;and
(c)
youarefreefromanyobligationsowedtoathirdpartywhichmightpreventyoufromstartingworkonthedatementionedbeloworfromproperlyperformingthedutiesofyourposition.
Ifyouareinanydoubtaboutanyoftheforegoing,youmaywishtotalktotheundersignedinordertoclarifythematteranddeferresignationfromyourcurrentemploymentuntilthematterisclarifiedtoourmutualsatisfaction.
Thisletterisgiventoyouasarecordofthemaintermsandconditionsofyouremployment.TheCompanyreservestherighttomakeamendmentstoyourtermsandconditionsfromtimetotimeandyouwillbeboundbyamendmentstoy
YouremployerwillbeCompany1Limited
1.
DATEOFCOMMENCEMENT
YouremploymentattheCompanywillbeginon20thMay2020("thecommencementdate").
Thefirstthreemonthsofyouremploymentwillbeprobationary.Yourperformanceandsuitabilityforcontinuedemploymentwillbereviewedduringyourprobationaryperiod.Should
directors
```

Data preparation

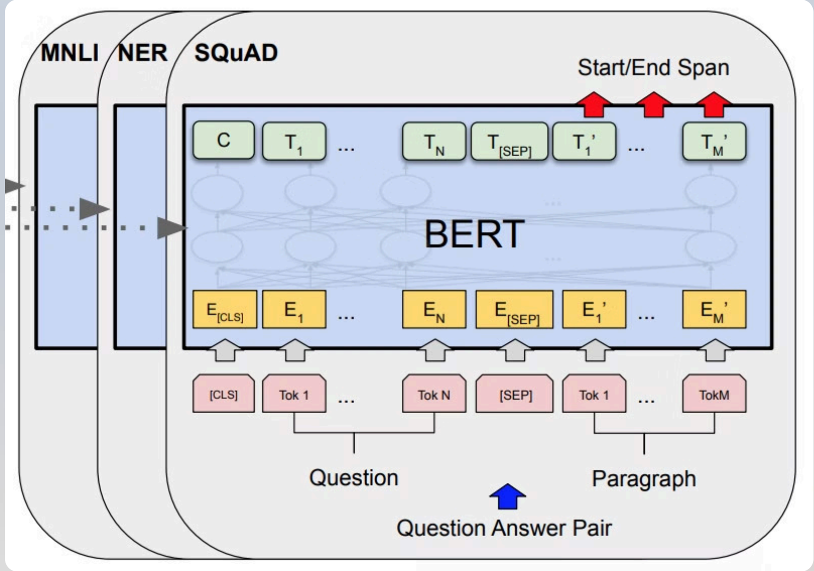
- **PyMuPDF** → text_from_pdf function
- **python-docx** → text_from_dpcx function

Made with Gamma

Modelling

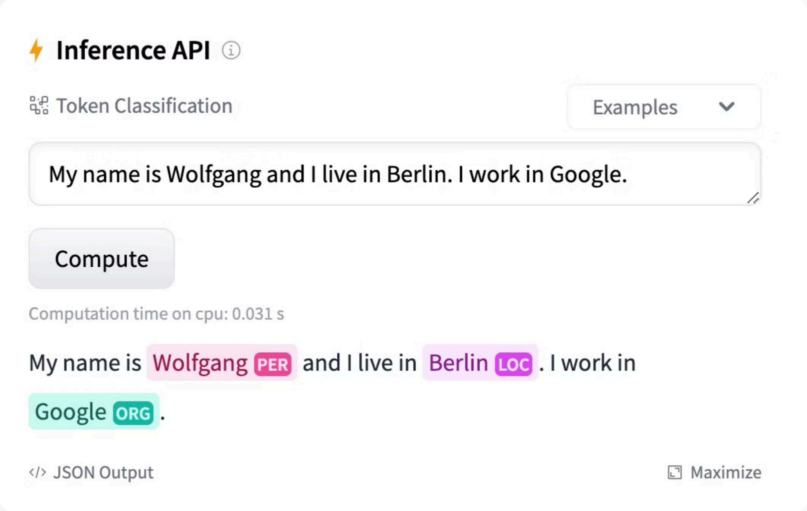
Bidirectional Encoder Representations from Transformers (BERT)

- □ Named Entity Recognition (NER)
- □ Question-Answering using Stanford Question Answering Dataset (SQuAD)



Exploring Entity Extraction

Initial experiments used a BERT-based Named Entity Recognition (NER) model to extract data points like names and addresses from the contract texts. However, the results were unsatisfactory, with the extracted entities being fragmented and incomplete.



My name is Wolfgang and I live in Berlin. I work in Google.

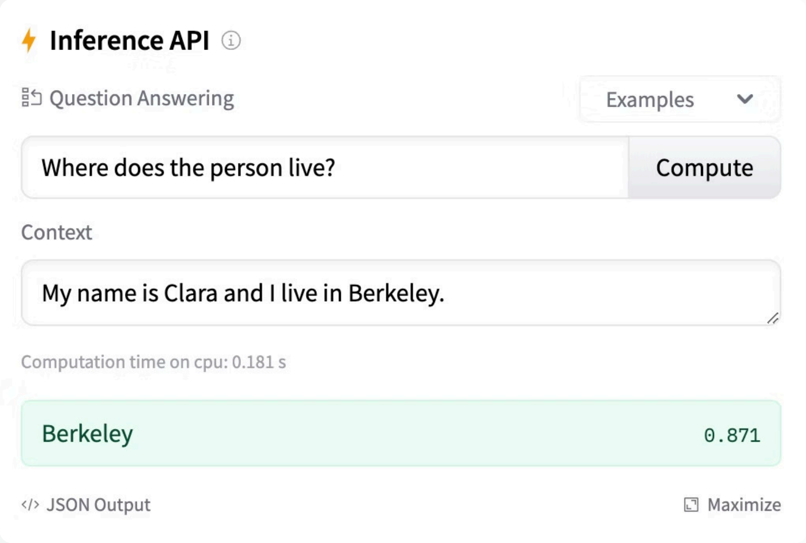
Computation time on cpu: 0.031 s

My name is Wolfgang PER and I live in Berlin LOC . I work in Google ORG .

bert-base-NER

bert-base-NER is a fine-tuned BERT model that is ready to use for **Named Entity Recognition** and achieves **state-of-the-art performance** for the NER task. It has been trained to recognize four types of entities: location (LOC), organizations (ORG), person (PER) and Miscellaneous (MISC) [1].

Specifically, this model is a *bert-base-cased* model that was fine-tuned on the English version of the standard CoNLL-2003 Named Entity Recognition dataset. The term “CoNLL” stands for Conference on Natural Language Learning [2].



Where does the person live?

Context: My name is Clara and I live in Berkeley.

Computation time on cpu: 0.181 s

```
Berkeley    0.871
```

Leveraging Question-Answering

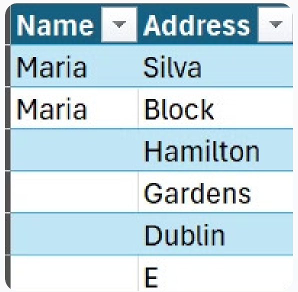
The research focuses on improvement of the results using a question-answering approach with the pre-trained "bert-large-cased-whole-word-masking-finetuned-squad" model. This model demonstrated more promising results, accurately extracting key information such as employee names, job titles, and contract details.

SQuAD 1.0 & 2.0

The current bert model used is pretrained with the wikipedia and the bookcorpus datasets on hugging face [3] [4].The above model uses the SQuAD 1.0 (Stanford Question Answering Dataset) dataset for fine tuning and that helps in the better results [5].

The SQuAD 2.0 dataset [5] was also used in an attempt to fine tune and see for any better results and it produced results similar to the squad 1.0 with no high range of improvement. The goal in the future will be to fine tune the initial model over the 2.0 dataset which will produce a model which is trained on both datasets expecting more accurate results.

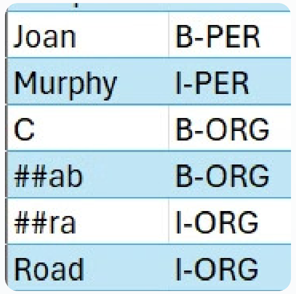
Results



Name	Address
Maria	Silva
Maria	Block
	Hamilton
	Gardens
	Dublin
	E

bert-base-NER

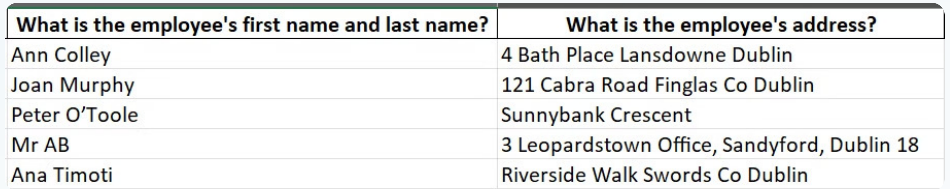
Model 1 (entity-extraction)



Joan	B-PER
Murphy	I-PER
C	B-ORG
##ab	B-ORG
##ra	I-ORG
Road	I-ORG

CNN + bert-base-NER

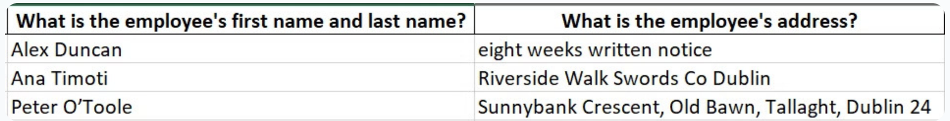
Model 2 (entity-extraction)



What is the employee's first name and last name?	What is the employee's address?
Ann Colley	4 Bath Place Lansdowne Dublin
Joan Murphy	121 Cabra Road Finglas Co Dublin
Peter O'Toole	Sunnybank Crescent
Mr AB	3 Leopardstown Office, Sandyford, Dublin 18
Ana Timoti	Riverside Walk Swords Co Dublin

bert-large-cased-whole-word-masking-finetuned-squad

Model 3 (question-answering)

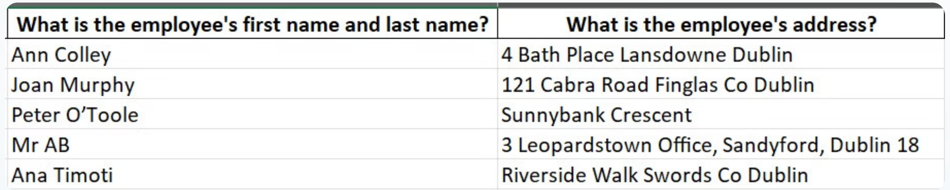


What is the employee's first name and last name?	What is the employee's address?
Alex Duncan	eight weeks written notice
Ana Timoti	Riverside Walk Swords Co Dublin
Peter O'Toole	Sunnybank Crescent, Old Bawn, Tallaght, Dublin 24

bert-base-uncased fine-tuned with SQuAD2.0

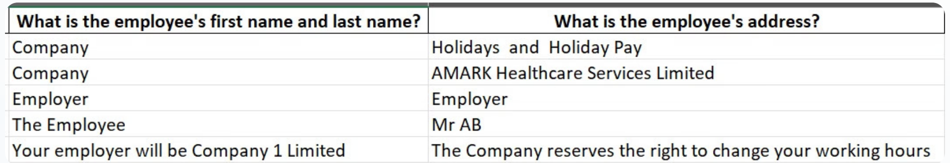
Model 4 (question-answering)

Unexpected Result



What is the employee's first name and last name?	What is the employee's address?
Ann Colley	4 Bath Place Lansdowne Dublin
Joan Murphy	121 Cabra Road Finglas Co Dublin
Peter O'Toole	Sunnybank Crescent
Mr AB	3 Leopardstown Office, Sandyford, Dublin 18
Ana Timoti	Riverside Walk Swords Co Dublin

bert-large-cased-whole-word-masking-finetuned-squad



What is the employee's first name and last name?	What is the employee's address?
Company	Holidays and Holiday Pay
Company	AMARK Healthcare Services Limited
Employer	Employer
The Employee	Mr AB
Your employer will be Company 1 Limited	The Company reserves the right to change your working hours

bert-large-cased-whole-word-masking-finetuned-squad fine-tune with CUAD

Evaluating Model Performance

bert-large-cased-whole-word-masking-finetuned-squad

Strengths

The question-answering model achieved an accuracy of 66.7% on the contract data provided by 4See.

Limitations

The model fails to understand when the text does not have the answer for a question.

Next Steps

Explore further fine-tuning and optimisation to improve accuracy of the model.

Conclusion and Future Work

1

Promising Results

The question-answering approach using BERT has demonstrated the potential to accurately extract key data from unstructured employee contracts.

2

Expansion Opportunities

Future research should explore applying the model to other types of documents, such as employee handbooks, to further expand its capabilities.

3

Continuous Improvement

Ongoing refinement of the model, including fine-tuning on relevant datasets, can enhance the accuracy and efficiency of the data extraction process.

References

[1] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. CoRR, abs/1810.04805. Available at: **http://arxiv.org/abs/1810.04805**

[2] Tjong Kim Sang, E. F., & De Meulder, F. (2003). Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 (pp. 142-147). Available at: **https://www.aclweb.org/anthology/W03-0419**

[3] Wikimedia Foundation. (n.d.). *Wikimedia Downloads*. Available online at **https://dumps.wikimedia.org**. Accessed [May 8. 2024].

[4] Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., & Fidler, S. (2015). *Aligning Books and Movies: Towards Story-Like Visual Explanations by Watching Movies and Reading Books*. Paper presented at the IEEE International Conference on Computer Vision (ICCV), December 2015.

[5] Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016). *SQuAD: 100,000+ Questions for Machine Comprehension of Text*. In J. Su, K. Duh, & X. Carreras (Eds.), Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (pp. 2383-2392). Austin, Texas: Association for Computational Linguistics. **https://doi.org/10.18653/v1/D16-1264**. Available online at **https://aclanthology.org/D16-1264**.