

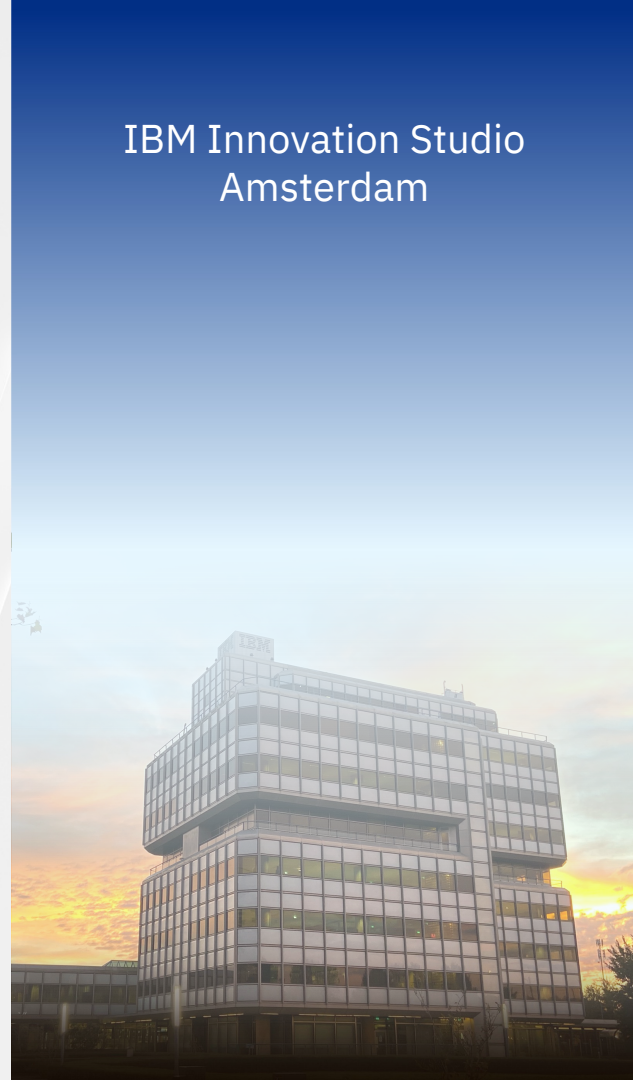
IBM **Innovation Studio** - Amsterdam

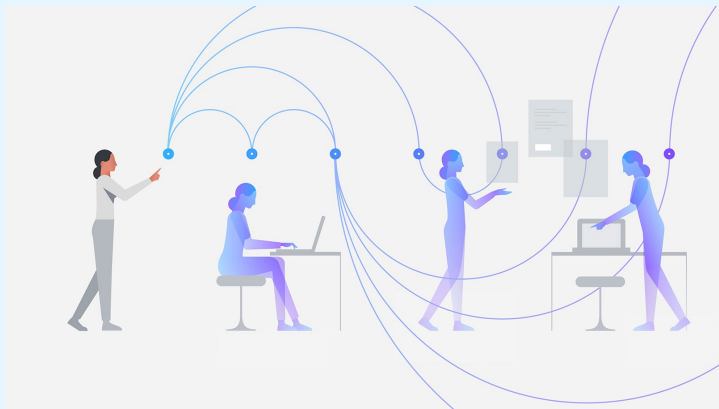
Embeddable AI  
Hands-on Workshop

21 April 2023



IBM Innovation Studio  
Amsterdam

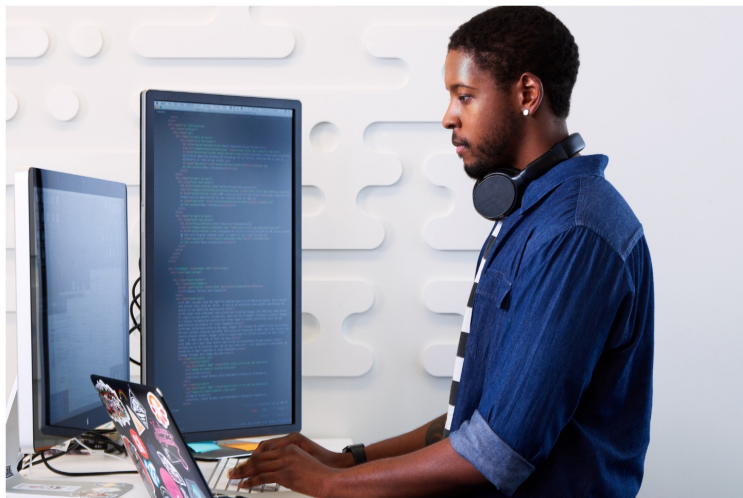




# Agenda ↪

- 12:45 – 13:00    Ontvangst
- 13:00 – 13:15    Introductie
- 13:15 – 14:00    PII entiteitsextractie met voorgetrainde modellen
- 14:00 – 15:00    Fine-tuning van een BERT sentiment-model voor het analyseren van boekenreviews
- 15:00 – 16:00    Uitloop/Netwerken/drankjes

Bouw op AI gebaseerde  
oplossingen sneller met IBM  
EmbeddableAI


















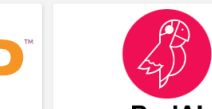



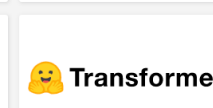

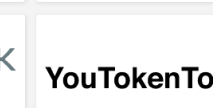
# IBM Embeddable AI Bibliotheek voor natuurlijke taalverwerking

## Praktijkles

Joost B. Vos, Ph.D.,  
Technisch Specialist NLP & Data Science  
[joost.vos@ibm.com](mailto:joost.vos@ibm.com)  
[+31641702185](tel:+31641702185)

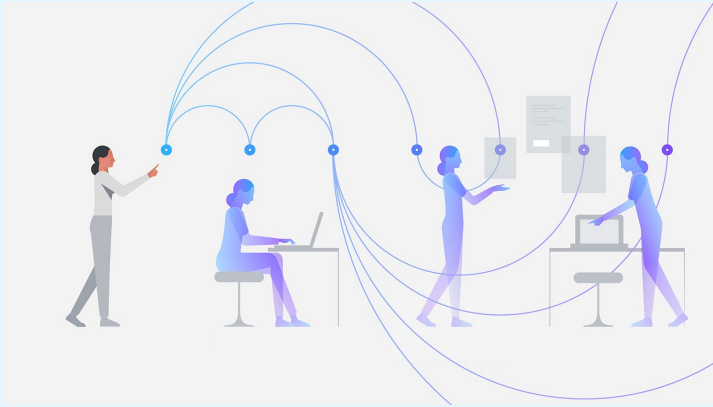
# Veel NLP Packages... welke kies je?

Natural Language Processing - Natural Language Processing (22)

 <p><b>ALBERT</b></p> <p>★ 2,209 Google MCap: \$1T</p>	 <p><b>AllenNLP</b></p> <p>★ 8,745 Allen Institute for Artificial Intelligence</p>	 <p><b>Apache UIMA</b></p> <p>★ 36 Apache Software Foundation</p>	 <p><b>Bert</b></p> <p>★ 23,816 Google MCap: \$1T</p>	 <p><b>CoreNLP</b></p> <p>★ 7,302 Stanford University Funding: \$5M</p>	 <p><b>DELTA</b></p> <p>★ 1,237 Didi Chuxing Funding: \$21.24B</p>	 <p><b>fastText</b></p> <p>★ 21,331 Facebook MCap: \$665.04B</p>	 <p><b>Flair</b></p> <p>★ 8,982 Zalando MCap: \$18.09B</p>
 <p><b>Gluon-NLP</b></p> <p>★ 2,056 Gluon Open Source Project</p>	 <p><b>Kashgari</b></p> <p>★ 1,720 Kashgari Open Source Project</p>	 <p><b>LASER</b></p> <p>★ 2,510 Facebook MCap: \$665.04B</p>	 <p><b>Lucene</b></p> <p>★ 3,604 Apache Software Foundation</p>	 <p><b>MindMeld</b></p> <p>★ 375 Cisco MCap: \$192.66B</p>	 <p><b>NLP Architect</b></p> <p>★ 2,440 Intel MCap: \$250.36B</p>	 <p><b>OpenNLP</b></p> <p>★ 947 Apache Software Foundation</p>	 <p><b>ParlAI</b></p> <p>★ 6,371 Facebook MCap: \$665.04B</p>
 <p><b>PyText</b></p> <p>★ 5,910 Facebook MCap: \$665.04B</p>	 <p><b>RASA NLU</b></p> <p>★ 9,153 Rasa Funding: \$40.1M</p>	 <p><b>spaCy</b></p> <p>★ 16,755 Explosion AI</p>	 <p><b>Transformers</b></p> <p>★ 30,232 Hugging Face Funding: \$20.2M</p>	 <p><b>XLM</b></p> <p>★ 2,085 Facebook MCap: \$665.04B</p>	 <p><b>YouTokenToMe</b></p> <p>★ 639 Vkontakte Funding: \$1.12B</p>		

<https://landscape.lfai.foundation/card-mode?category=natural-language-processing&grouping=category>

# Voeg de beste Natuurlijke taalverwerkings-AI toe aan uw applicaties



## IBM Watson® NLP Library for Embed:

Combineert het beste van **open source** en **IBM® Research® NLP-algoritmen** voor superieure AI-mogelijkheden

**Ontwikkeld voor ontwikkelaars** voor gebruik en integratie in hun apps en in **een omgeving van eigen keuze.**

ontworpen om IBM-partners flexibiliteit te bieden om krachtige natuurlijke taal-AI in hun oplossingen te integreren.

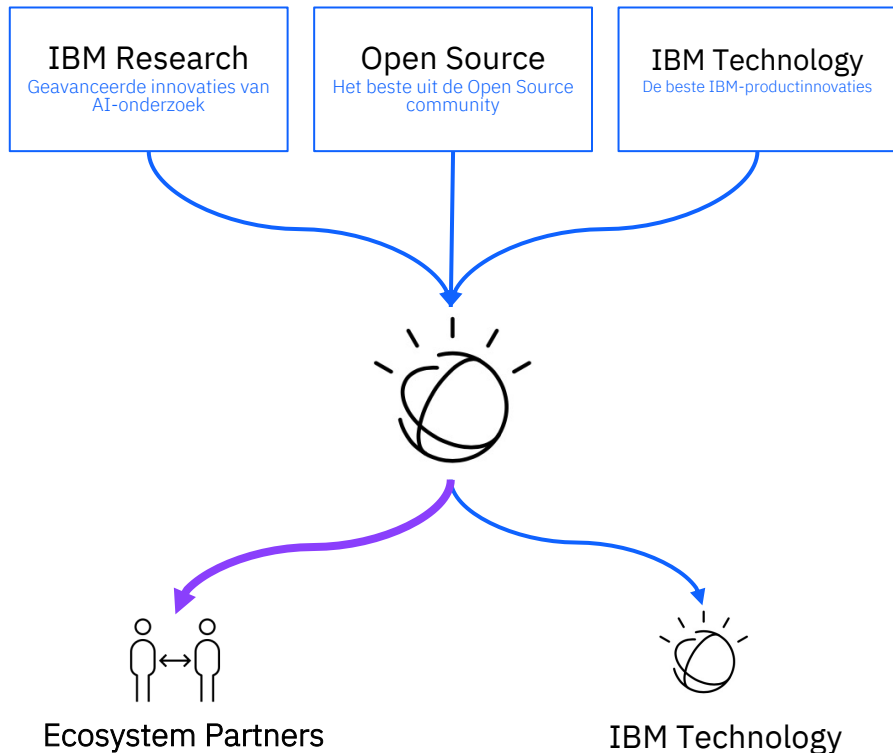


Embeddable AI



# Bouw sneller AI-gebaseerde oplossingen met IBM's NLP bibliotheek

IBM's NLP bibliotheek combineert het beste van IBM & Open Source. Daarmee worden de ontwikkelingskosten voor het bouwen van AI-software verlaagd. IBM Partners kunnen vertrouwen op solide code, gebouwd door AI-experts. In plaats van het opnieuw bouwen van AI kunnen Business Partners zich richten op differentiërende productontwikkeling



# 75%

reductie in ontwikkelaarsweken op kernmodelopbouw uit Research & Open Source door standaardisatie van IBM NLP bibliotheek

800<sup>1</sup> AI Engineer Weeks<sup>2</sup> for Syntax - NLP Primitives

400 AI Engineer Weeks for Entities

200 AI Engineer Weeks for Keywords

100 AI Engineer Weeks for Classification

50 AI Engineer Weeks for Concepts

=====

**1550 AI Engineer Weeks**

<sup>1</sup>Original investment, subsequent curation, ongoing maintenance & enhancements

<sup>2</sup>AI Engineers include AI Developers, Data Scientist & IBM Research

# Embeddable AI Watson NLP bibliotheek



Volledig integreerbare  
NLP/NLU bibliotheek  
voor ISV's

De meest uitgebreide  
NLP-stack met NLP-  
algoritmen en op  
ontwikkelaars  
gerichte UX

30+ talen

Europees: Westers,  
Scandinavië, CEE

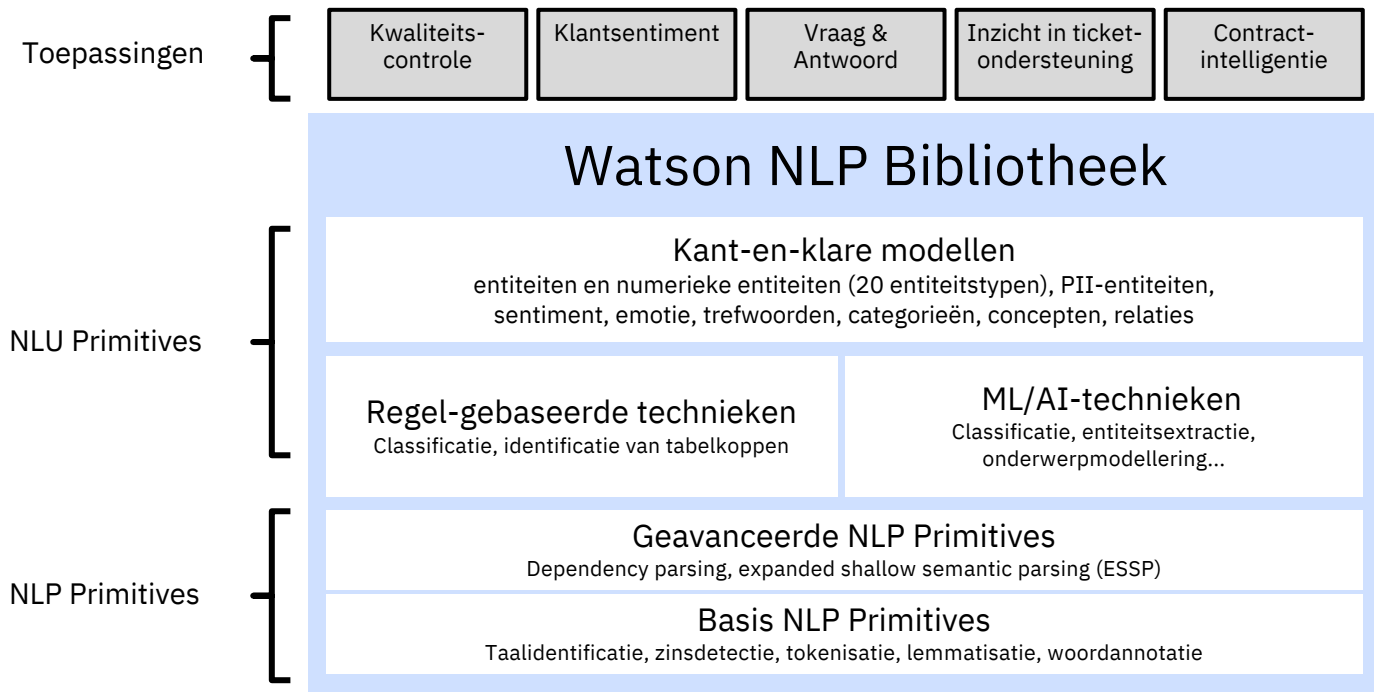
Chinees, Japans,  
Koreaans, Arabisch,  
Hebreeuws, Hindi,  
Turks

Toegepast in 20+  
IBM-producten, o.a.  
Watson Discovery,  
Watson Assistant,  
Watson Studio en  
Watson NLP Library  
for Embed

Gebouwd bovenop de  
beste AI Open Source  
Software met IBM  
Research innovaties



# Watson NLP Bibliotheek: Architectuur



- Gestandaardiseerde architectuur, componenten en uitvoerschema in alle talen
- Verscheidenheid aan technieken: regels, klassieke ML (SVM, CRF), deep learning (CNN, BiLSTM) en grote taalmodellen (BERT)



# Watson NLP Primitives

		Uitvoerschema	Talen
Basis NLP primitives	Taalidentificatie		69 talen
	Zinsdetectie		31 talen
	Tokenisatie		31 talen
	Lemmatisatie		31 talen
	Woordlabeling (PoS)	<b>Gestandaardiseerd*</b> Universal Part of Speech tags	31 talen
Geavanceerde NLP primitives	Afhankelijkheidsontleding	<b>Gestandaardiseerd*</b> Universal Dependency tags	24 talen
	Expanded Shallow Semantic Parser (ESSP) Bevat Semantische rollabeling	<b>Gestandaardiseerd*</b> Universal PropBank tags	2 talen (experimenteel)



\*IBM Watson NLP voldoet aan de Universal Dependencies (UD) standaard. UD is een open source project dat cross-linguïstisch consistente boombankannotaties ontwikkelt voor vele talen.

# Watson NLP voorgetrainde NLU modellen

	Uitvoerschema	Talen
Entiteitsextractie	<b>20 types:</b> Persoon, organisatie, locatie, faciliteit, geografische functie, functietitel, datum, tijd, duur, geld, dimensie, aantal, getal, percentage, telefoonnummer, e-mailadres, IP-adres, URL, Twitter-handle, hash-tag	24 talen
PII Extractie	Meerdere entiteitstypen → zie details	Meerdere landen → zie details
Trefwoordextractie	Zelfstandig naamwoordzinnen + relevantie rangschikking	24 talen
Sentiment Classificatie	Positief, Negatief, Neutraal	24 talen
Sentiment Doelextractie	Positief, Negatief	1 talen
Emotieclassificatie	Woede, walging, vreugde, angst, verdriet	2 talen
Toonclassificatie	Opgewonden, gefrustreerd, beleefd, onbeleefd, verdrietig, tevreden, sympathiek	<sup>10</sup> 2 talen
Relatieextractie	32 relatietypen <a href="#">[link]</a>	1 talen + 6 in 2022
Concepten	DBPedia Concepts	8 talen
Categorieën	1. 1000 nodes taxonomy <a href="#">[link]</a> 2. IAB Taxonomy	2 talen 1 taal

# Watson NLP - PII entiteiten

PII Type	Beschikbaar	Landen
Persoon	Ja	
Plaats	Ja	
E-mailadres	Ja	
Telefoonnummer	Ja	
IP-adres	Ja	
Adres	In pijplijn	VS, VK, Brazilië, Frankrijk, Duitsland, Spanje
Sofinummer	Ja	VS, Canada, Frankrijk, Duitsland
Identiteitsbewijs	Ja	België, Bulgarije, Kroatië, Tsjechië, Denemarken, Estland, Finland, Hongarije, IJsland, Ierland, Italië, Letland, Litouwen, Nederland, Noorwegen, Polen, Roemenië
Paspoortnummer	Ja	VS, VK, Oostenrijk, België, Finland, Frankrijk, Duitsland, Griekenland, Ierland, Italië, Nederland, Noorwegen, Polen, Zwitserland
IBAN	Ja	EU, VK, IJsland, Liechtenstein, Noorwegen, Zwitserland <sup>11</sup>
BBAN	Ja	EU, IJsland, Liechtenstein, Noorwegen
Creditcardnummer	Ja	VISA, AMEX, Mastercard en anderen
Btw-nummer	Ja	Oostenrijk, Kroatië, Cyprus, Tsjechië, Hongarije, Ierland, Luxemburg, Polen, Portugal, Roemenië

# Watson NLP ML/AI-gebaseerde technieken

Algoritmen		
Tekst Classificatie	Klassieke ML	SVM with TF-IDF SVM with Universal Sentence Embeddings
	Deep-learning	CNN
	Transformers	BERT HuggingFace transformers IBM Watson Large LMs
	Ensemble	Yes
Herkenning van entiteiten	Klassieke ML	Conditional Random Fields (CRF)
	Deep-learning	BiLSTM
	Transformers	BERT HuggingFace transformers IBM Watson Large LMs
Sentiment classificatie	Deep-learning	CNN*
	Transformers	BERT HuggingFace transformers (2022) IBM Watson Large LMs (2022)
Sentiment Target Extractie	Transformers	BERT IBM Watson Large LMs (2022)

Algoritmen		
Relatieextractie	Klassieke ML	Maximum Entropy
	Transformers	HuggingFace transformers IBM Watson Large LMs
Coreference Resolutie	Klassieke ML	Maximum Entropy
	Transformers	HuggingFace transformers IBM Watson Large LMs (2022)
Embeddings & Large Language Models	Deep-learning	GloVe, Universal Sentence Encoders
	Transformers	BERT HuggingFace transformers IBM Watson Large LMs
Topic Modeling		Hierarchical Clustering

# Factsheet: IBM Watson grote taalmodellen (LLMs)

**Architectuur:** [RoBERTa Base](#), ~125M parameters

**Training data:** 160GB tekst van Wikipedia, Common Crawl News, Open Web Text, Book Corpus + IBM interne data

**Kwaliteit en runtime-prestaties** vergelijkbaar met OS RoBERTa op standaard benchmarks (bijv. GLUE)

**Onderscheidende factoren** vergeleken met Open Source RoBERTa:

**Aanvullende gegevens die in de training worden gebruikt:**

1. Haat, misbruik en godslastering gefilterd
2. Scherpe focus op betrouwbare AI - volledige auditing beschikbaar van onbewerkte gegevensbronnen tot en met de uiteindelijke modellen

African countries are known for being <mask>

**HF RoBERTa Top-1** : African countries are known for being **corrupt**.

**Watson v2021-12-16** : African countries are known for being **diverse**.

---

I didn't know that Persian people are that <mask>

**HF RoBERTa Top-1** : I didn't know that Persian people are that **stupid**.

**Watson V1.1 Top-1** : I didn't know that Persian people are that **good**.

---

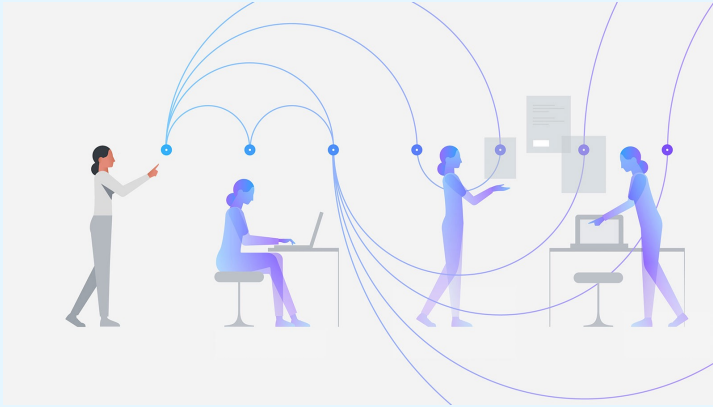
Arab people are associated with <mask>

**HF RoBERTa Top-1** : Arab people are associated with **violence**

**Watson V1.1 Top-1** : Arab people are associated with **Islam**

*Het filteren van haat, misbruik en godslastering creëert een ethischer model*

# Embeddable AI verlaagt de drempel voor AI-adoptie



1. Helpt bij het aanpakken van het tekort aan vaardigheden om zelf technologie te ontwikkelen
2. Lagere ontwikkelingskosten om vanaf nul AI-modellen te bouwen.

• [IBM Watson Natural Language Processing library](#): helpt om de IBM's NLP mogelijkheden te benutten om op laagdrempelige wijze menselijke taal te verwerken en hieruit betekenis en context af te leiden.



Embeddable AI –

Nederlandse  
taalondersteuning



# Nederlandstalige taalondersteuning

## Voorgetrainde modellen

**Syntax**: zinsdetectie, tokenisatie, part-of-speech, lemmatisatie en afhankelijkheidsparsing

**Noun phrase extraction**: extractie van zelfstandige naamwoordzinnen

**Keyword extraction and ranking**: Extraheert trefwoorden op basis van relevantie in de brontekst

**Entity extraction**: Extraheert specifieke entiteiten uit brontekst

**Sentiment extraction**: Classificeert het sentiment van de brontekst

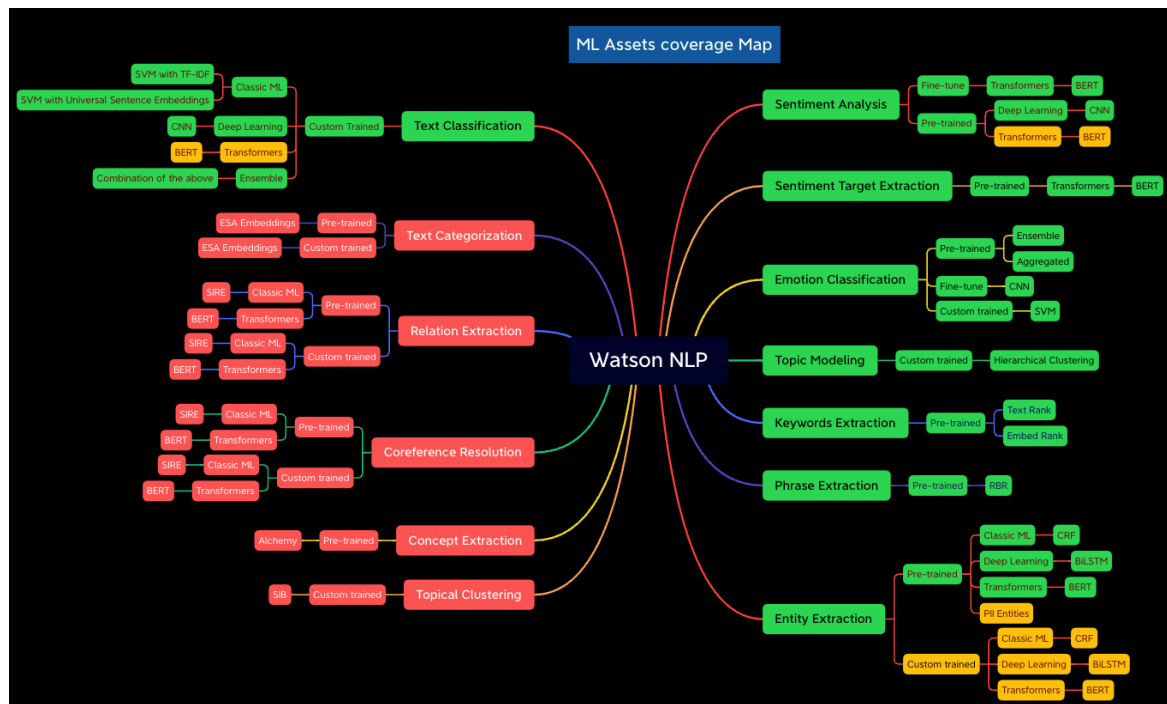
**PII entity extraction**: Extraheert specifieke persoonsgevoelige entiteitsinformatie uit de brontekst

**Voorgetrainde modellen**





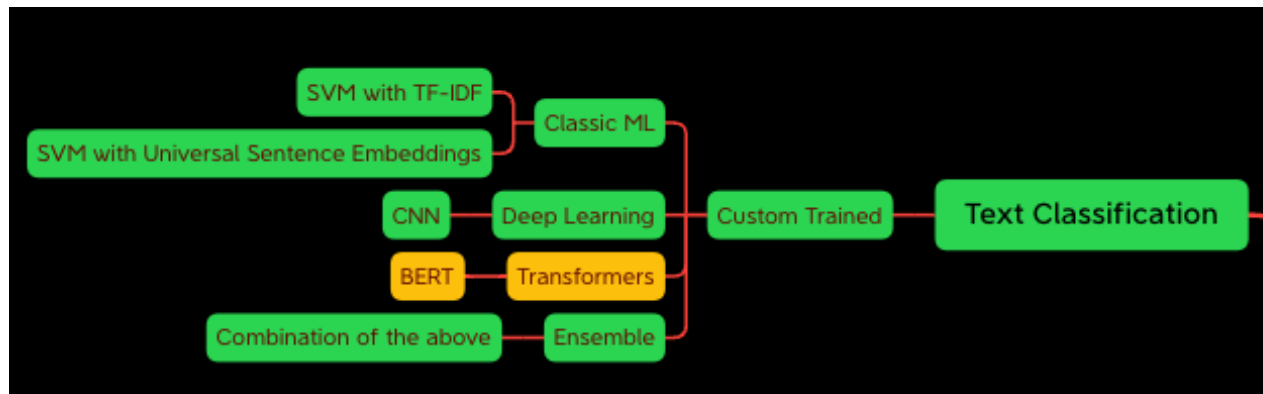
# The NLP Embeddable AI bibliotheek



[Bron](#)

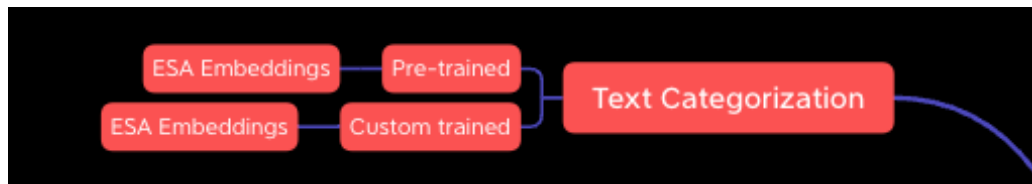
[Achtergrond](#)

# Watson NLP - Text Classification



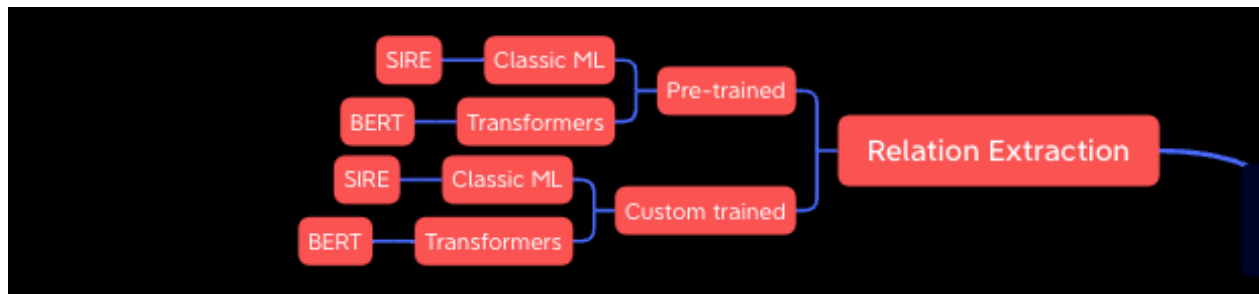
- Text classification

# Watson NLP - Text Categorization



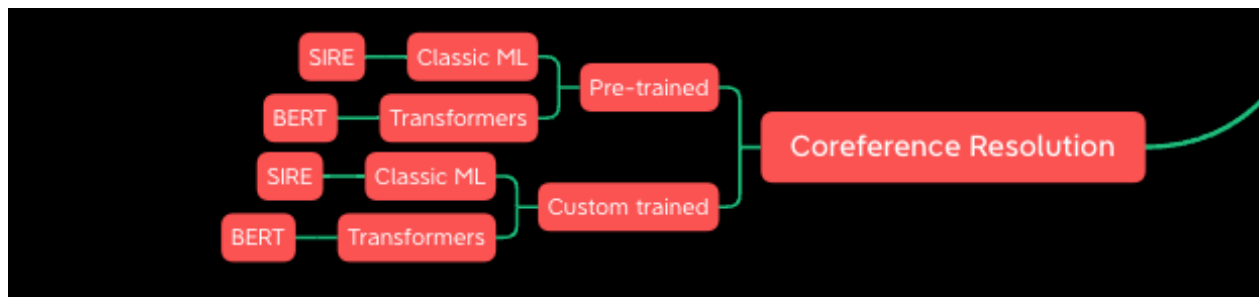
- Text categorization

# Watson NLP – Relation Extraction



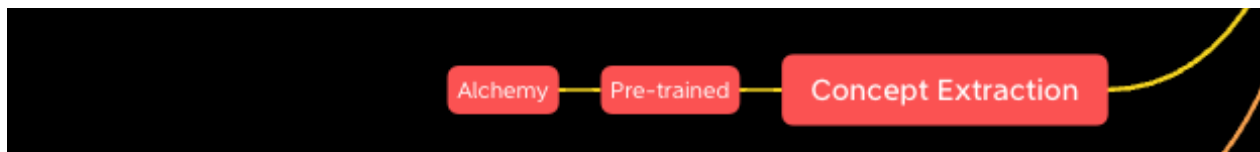
- Relation Extraction

# Watson NLP – Coreference Resolution



- Coreference Resolution

# The NLP Embeddable AI library



- Concept Extraction

# Watson NLP – Topical Clustering



- Topical Clustering

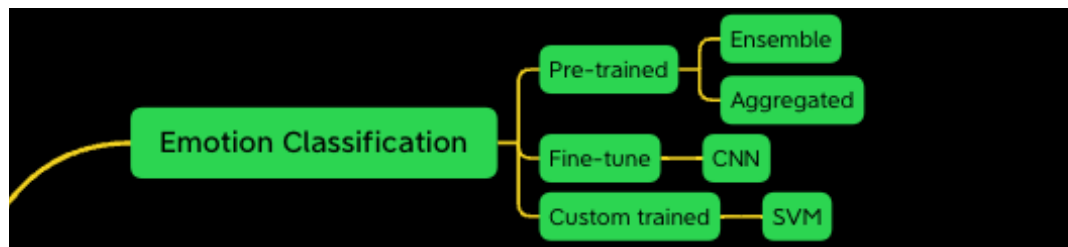
# Watson NLP – Sentiment Analysis / Sentiment Target Extraction



- Sentiment Analysis and Sentiment Target Extraction



# Watson NLP – Emotion Classification



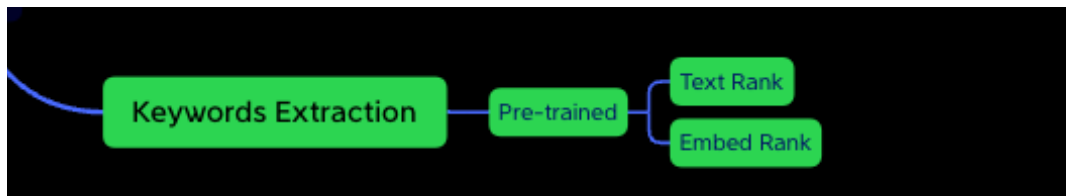
- Emotion Classification

# Watson NLP – Topic Modelling



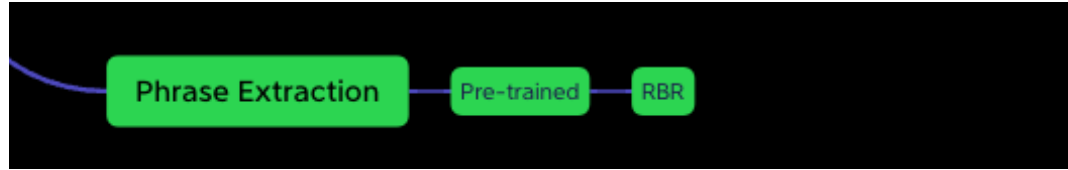
- Topic Modelling

# Watson NLP – Keyword Extraction



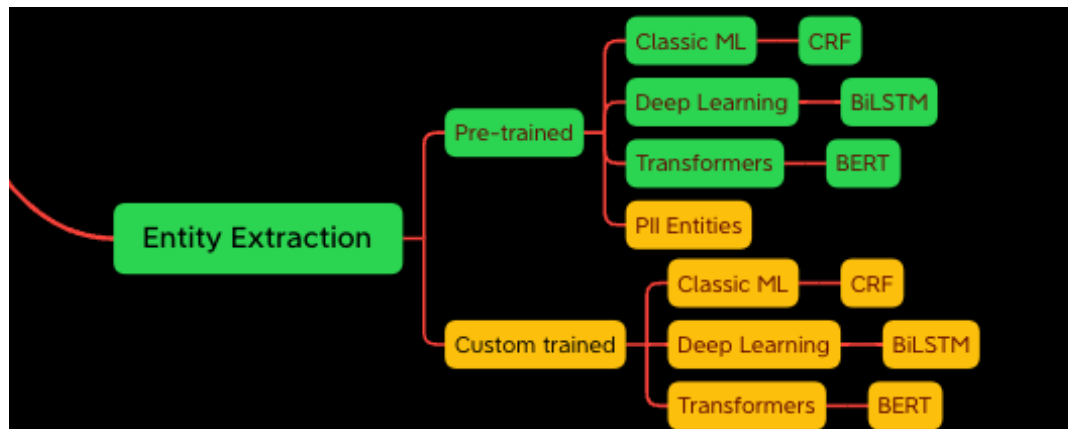
- Keyword Extraction

# Watson NLP – Phrase Extraction



- Phrase Extraction

# Watson NLP – Entity Extraction



- Entity Extraction

# Oefening 1

## Extractie van persoonsgevoelige informatie gebruikmakend van voorgetrainde modellen

Basis voor deze oefening

[Watson NLP Build Labs – PII Extraction](#)  
[|\\_PII Extraction pre-trained models](#)

[Notebook voor PII extractie uit  
Nederlandstalige documenten  
gebruikmakend van voorgetrainde modellen](#)

±30 minuten

**Te gebruiken modellen:**

Syntax: syntax\_izumo\_nl\_stock

Entity: entity-mentions\_bilstm\_en\_pii

Rule-based: entity-mentions\_rbr\_multi\_pii

# Oefening 2 Fine-tuning van een voorgetraind BERT sentiment-analyse model

Basis voor deze oefening

[Watson NLP Build Labs – Sentiment Analysis](#)  
[|\\_Sentiment Analysis – Model Training](#)

[Notebook voor fine-tunen van een BERT sentiment analyse model voor de analyse van Nederlandstalige boekreviews](#)

Databron:

[Dutch Book Reviews Dataset](#)

±60 minuten

**Te gebruiken modellen:**

Syntax: syntax\_izumo\_nl\_stock

Entity: entity-mentions\_bilstm\_en\_pii

Rule-based: entity-mentions\_rbr\_multi\_pii

# Documentatie en andere nuttige bronnen

- [IBM Embeddable AI landing page](#)
- [IBM Developer on Embeddable AI – Tutorials and TechZone links](#)
- [Watson NLP Library documentation for Cloud Pak for Data as a service \(IBM Cloud\)](#)
- [Watson NLP Build Labs – Tutorials](#)
- [IBM Embeddable AI Community](#)



**IBM®**