European Language Resource Coordination
Connecting Europe Facility

NATIONAL INFORMATION PROCESSING INSTITUTE

European Commission

NLP MODELS, DATA SETS

NATIONAL INFORMATION PROCESSING INSTITUTE

RESOURCES AND EXPERIENCES

MAREK KOZŁOWSKI

HEAD OF THE LABORATORY OF NATURAL LANGUAGE PROCESSING

SŁAWOMIR DADAS

DEPUTY HEAD OF THE LABORATORY OF INTELLIGENT INFORMATION SYSTEMS

21 September 2022

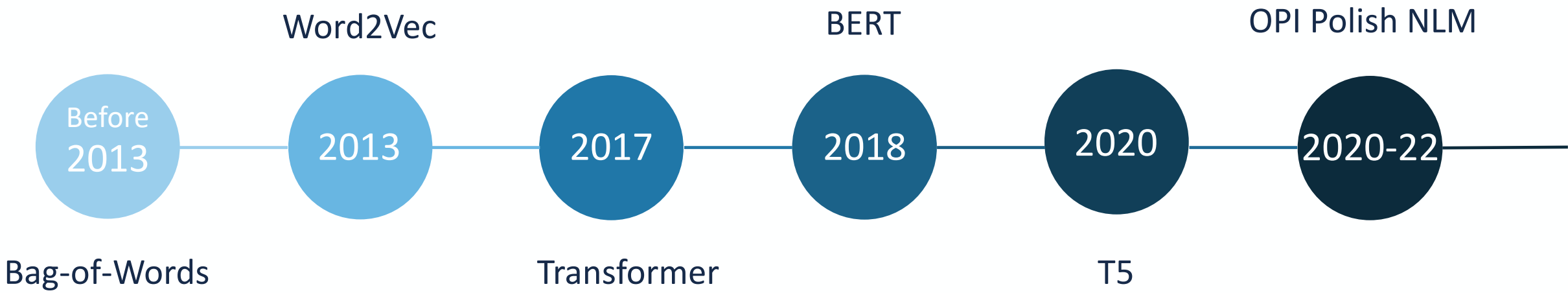
Dr. Marek Kozłowski

Marek Kozłowski is the Head of the Laboratory of Natural Language Processing at the National Information Processing Institute in Warsaw, where he leads a team of over 30 researchers and programmers who develop software that is enriched with intelligent (primarily text and image) data processing methods. He is passionate about natural language processing, data mining, and machine learning. He has written over 40 scientific publications on semantic text processing and machine learning. Marek has participated in commercial machine learning research projects for the private sector, including at Samsung, France Telecom, Orange Labs, Millward Brown, Vive Textile Recycling, and Connectis. In 2021 Marek Kozłowski and Sławomir Dadas won the nationwide AI competition organized by GovTech and the Office of Competition and Consumer Protection (UOKiK), for which the National Information Processing Institute now provides an AI-empowered system.

Sławomir Dadas

Sławomir Dadas is a lead research engineer and the deputy head of the Laboratory of Intelligent Information Systems at the National Information Processing Institute. He is a doctoral student at the Systems Research Institute of the Polish Academy of Sciences. For several years he has been professionally engaged in the design and implementation of information systems that incorporate machine learning solutions, primarily in the field of natural language processing. His research interests include natural language processing, applications of machine learning in scientometric research, distributed computing, algorithms, and data structures.

NLP EVOLUTION



DATA – THE KEY ELEMENT OF THE UNSUPERVISED AND SUPERVISED LEARNING

- **Dedicated inhouse repositories, such as:** the Polish theses database (ORPPD), Uniform Anti-Plagiarism System databases, Polish Scientific Bibliography database (PBN), POL-on databases, etc.

- **Raw text corpora used for language model pretraining:**
  - **Base corpus** (~30GB): Polish Parliamentary Corpus, Polish Wikipedia, collections of some polish books, articles
  - **Web corpus** (~300GB): CommonCrawl sampled and cleaned in order to build significant Polish textual web resources

Neural Language Models

2017

- Prior to Deep Learning, the models were rarely generic enough to be easily customized/tailored to specific tasks or domains

- Since 2017 the Transformer revolution in NLP has accelerated
- Pre-trained models are now re-usable components that can be used as it is, or customized through next training stages (fine tuning)
- The crucial issues concerning the Transformer pre-training include:
  - Unlabeled Data – the need for massive unlabeled corpora; we used the corpora of 200+ GB
  - GPUs – distributed environments; in our case large models were trained with the use of 8x Nvidia V100
  - Pre-training time – usually large model pre-training lasts 3-5 months

2020-22

- Between 2020 and 2022, we published numerous Transformer based NLMs, such as RoBERTa, BART, GPT-2. They were of various sizes - from distil, base to large ones

2022

- In 2022, we published some sentence transformer paraphrase models, or long-formers
- We also published some simpler shallow language representations (context-free ones), such as Polish Word2vec, Glove, FastText

MODELS – BERT-BASED

- How we trained and tuned the Polish RoBERTa models:
  - Collect and pre-process a large corpus of Polish documents (200+ GB)
  - Train different transformer-based language models from scratch (BERT-based ones)
  - Fine-tune the models on a dozen of Polish linguistic tasks; for example, the KLEJ is a set of nine evaluation tasks for the Polish language understanding created by Allegro
  - Release the pre-trained models to the public

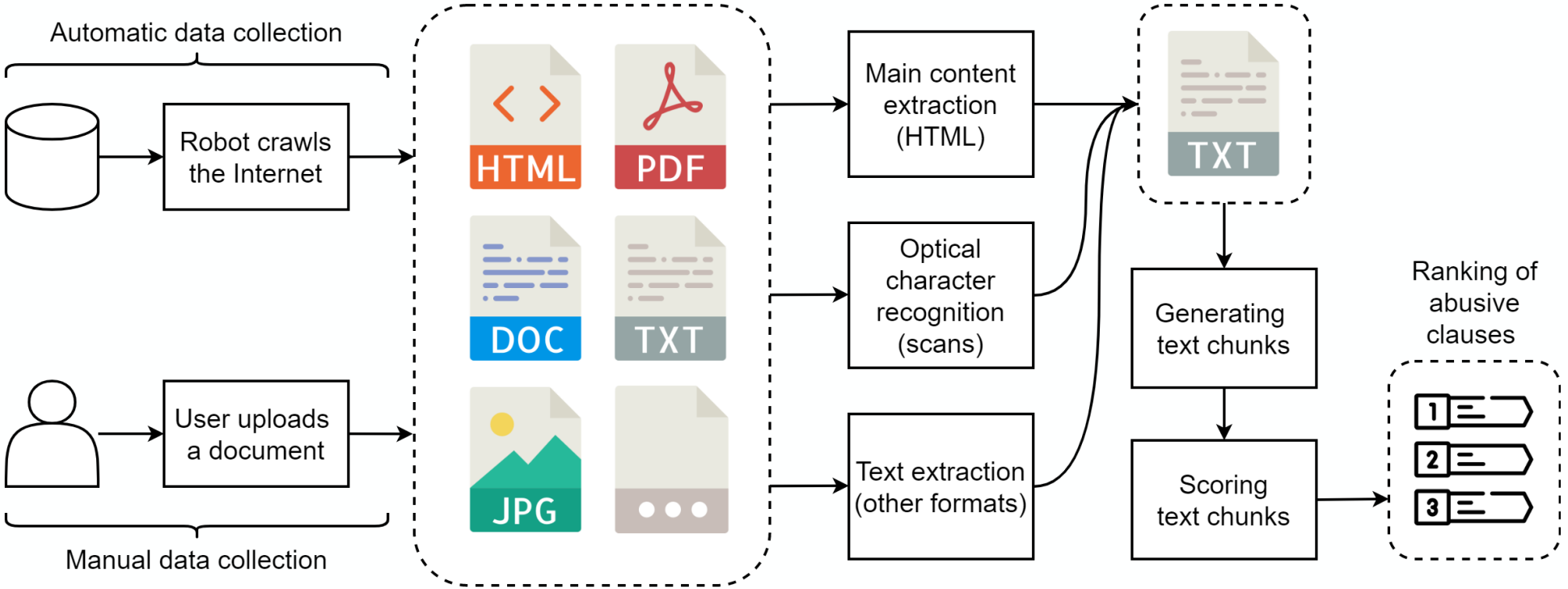
Model	L / H / A*	Batch size	Update steps	Corpus size	KLEJ Score**	Fairseq	Transformers
RoBERTa (base)	12 / 768 / 12	8k	125k	~20GB	85.39	v0.9.0	v3.4
RoBERTa-v2 (base)	12 / 768 / 12	8k	400k	~20GB	86.72	v0.10.1	v4.4
RoBERTa (large)	24 / 1024 / 16	30k	50k	~135GB	87.69	v0.9.0	v3.4
RoBERTa-v2 (large)	24 / 1024 / 16	2k	400k	~200GB	88.87	v0.10.2	v4.14
DistilRoBERTa	6 / 768 / 12	1k	10ep.	~20GB	84.55	n/a	v4.13

SYSTEMS WITH NLP COMPONENTS - JSA

https://www.gov.pl/web/ncbr-en/uniform-anti-plagiarism-system

SYSTEMS WITH NLP COMPONENTS - ARBUZ

In 2021, UOKiK organized an artificial intelligence and natural language processing competition for the development of an intelligent tool capable of scanning consumer contracts for provisions that violate consumer rights. Our instuitute won the competition and is now responsible for the development of a system that analyses contracts/regulations and detects potential violations.



SYSTEMS WITH NLP COMPONENTS - ANSI

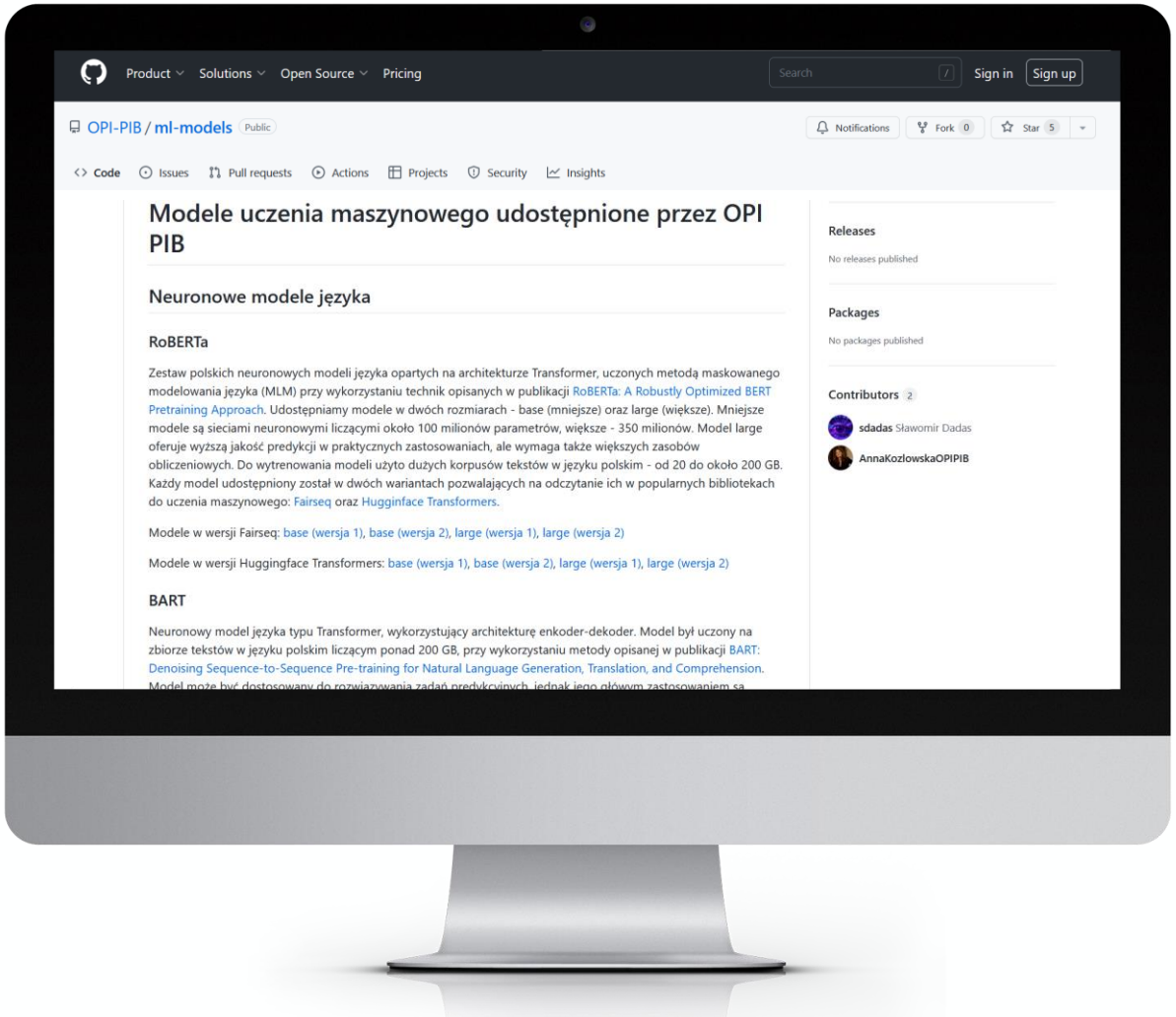
In 2022, OPI PIB and IPI PAN were granted funding as part of the Infostrateg III competition. Our responsibility is to build an information system to process data on products and users' reviews. The aim of this project is to automatically assess the quality of products based on the information available on the Internet. The project is developed for UOKiK.

- Multilingual machine learning tools supporting various languages (with the emphasis placed on Polish, English, and German).
- Multiple applications of NLP models:
  - Sentiment analysis of opinions
  - Aspect-based sentiment analysis
  - Identification of duplicated products, including cross-lingual deduplication
  - Detection of "dual quality" products

HOW TO DOWNLOAD OUR NEURAL LANGUAGE MODELS

https://github.com/OPI-PIB/ml-models:

- Pre-trained models, such as RoBERTa, BART, GPT-2, ELMo
- Models in two versions:
  - Pytorch Fairseq
  - HuggingFace Transformers
- Evaluation codes are also available
- Detailed results of the experiments can be found in the corresponding papers



European Language Resource Coordination
Connecting Europe Facility

European Commission

NATIONAL INFORMATION PROCESSING INSTITUTE

THANK YOU!

al. Niepodległości 188 B
00-608 Warsaw
Poland
www.opi.org.pl/en