Small Data and Data Centric AI: Case Study from the Master's Program in Artificial Intelligence at Sofia University

Maria Nisheva-Pavlova [1, 2] and Bilyana Dobreva [1]

[1] *Faculty of Mathematics and Informatics – Sofia University St. Kliment Ohridski, 5 James Bourchier Blvd., Sofia, 1164, Bulgaria*
[2] *Institute of Mathematics and Informatics – Bulgarian Academy of Sciences, 8 Acad. Georgi Bonchev Str., Sofia, 1113, Bulgaria*

Abstract

Recently, the term "small data" has become essential in the field called "data centric AI". While big data is used for different types of correlation analysis, small data is the real source for finding causal relationships between the objects studied. The paper discusses the experience in creating small datasets and transfer learning, gained in the Master's program in Artificial Intelligence at the Faculty of Mathematics and Informatics at Sofia University, focusing on some good examples of student projects.

Keywords

Big data, small data, data centric AI, transfer learning, question answering system

1. Introduction

After the initial wave of research and technological developments related to *big data*, the interest in the so-called *small data* and especially in the methodologies for creating appropriate small datasets and their use in the field of *data centric artificial intelligence* is constantly growing. Correctly constructed small data are commonly used by people in decision-making in various areas of particular public importance. The creation and use of suitable small datasets, along with the application of proper kinds of transfer learning, is the basis of data centric artificial intelligence. In recent years, a number of successful projects (mostly pre-diploma and diploma projects) of students from the Master's program in artificial intelligence at the Faculty of Mathematics and Informatics at Sofia University are addressing this issue.

Information Systems & Grid Technologies: Fifteenth International Conference ISGT'2022, May 27–28, 2022, Sofia, Bulgaria
EMAIL: marian@fmi.uni-sofia.bg (M. Nisheva-Pavlova); bddobreva@uni-sofia.bg (B. Dobreva)
ORCID: 0000-0002-9917-9535 (M. Nisheva-Pavlova)

2. Small data and its importance for the creation of decision support systems

According to the most popular informal definition, "small data is data that is 'small' enough for human comprehension. It is data in a volume and format that makes it accessible, informative and actionable" [1]. A more formal definition of small data has been given by Allen Bonde: "Small data connects people with timely, meaningful insights (derived from big data and/or "local" sources), organized and packaged – often visually – to be accessible, understandable, and actionable for everyday tasks" [2].

As many authors note, small data is what people usually think of as data. While big data can be understood as high-volume raw data coming from heterogeneous sources (e.g. social media publications, customer transactions, etc.) which is difficult to comprehend and manage, small data is produced from raw data by cleaning and reducing it into small, visually-appealing objects representing particular aspects of large datasets.

From the point of view of their applicability for the creation of different types of data analytics systems, the clearest dividing line between 'big' and 'small' data can be formulated as follows [3]: "Big Data is all about *finding correlations*, but Small Data is all about *finding the causation*, the reason why". More precisely, big data is important is all cases of building medium-term and long-term policies and strategic decisions.

From the other side, small data refers to definite and specific attributes of datasets, which can be used to analyze the current situation in depth and to make adequate personalized decisions. Therefore, small data is best placed to support decision-making at the current time.

For example, clinicians favor small data over big data for healthcare assessments as well as for building personalized prediction and decision-making models (see Figure 1).

Big Data Model	Small Data Model
What can be the effect of immunization programs?	Is my child's immunity to diseases taken care of?
Where do some of the healthiest people in the world live	Is my diabetes medication working as expected
Are there any generic factors to identify a disease	Am I susceptible to X disease?

Figure 1: Comparison of the applicability of Big Data and Small Data models in healthcare [4]

Also, there are various types of cases in which a particular person or organization needs quick and instant analysis of the available data and there is no need to use big data analytical tools for the purpose.

3. Data centric AI

The concept of data centric artificial intelligence, which has recently been actively involved in research and applied development, refers to *building AI systems with quality data*. The data centric AI approach is based on the idea to focus on ensuring that the data used clearly show what the developed AI system needs to learn.

As Andrew Ng notes in his popular interview for IEEE Spectrum [5], "data centric AI is *the discipline of systematically engineering the data needed to successfully build an AI system*". So, if until recently the dominant idea was to focus on improving the code, nowadays it is more effective for a lot of applications to consider that the quality of code is generally a solved problem and the focus should be moved to finding approaches to improve the data [5].

In particular, instead of working directly with a large amount of raw and noisy data, it is better to make at the beginning appropriate efforts to improve the consistency of the data and in this way to achieve a significant improvement in productivity. Especially for big data applications, the common approach has been: "If the data is noisy, let's just get a lot of data and the algorithm will average over it" [5]. But the data centric approach assumes to try to develop tools that point on data inconsistencies and give an effective way to overcome most of them in order to get a truly high performing system.

Following the data centric AI paradigm, a significant number of pre-diploma and diploma projects in various application areas are being developed in the Master's program in Artificial Intelligence at the Faculty of Mathematics and Informatics at Sofia University St. Kliment Ohridski. Among the most significant of them is the project for a virtual health assistant called Medico-Help [6], developed in 2021.

Medico-Help is a web-based expert system that functions as an intelligent chatbot, capable

- to automatically collect data from trusted websites,
- to build and extend automatically a medical knowledge base and to search in it,
- to generate hypotheses for medical diagnoses based on symptoms.

As an initial version of the knowledge base of Medico-Help, a small standardized ontology for human diseases[2], developed at the School of Medicine at the University of Maryland has been used. The system has a module for auto-

[2] https://disease-ontology.org

mated collection of specialized data from trusted sources on the Internet. The role of such a source in the pilot version of Medico-Help is played by MedIndia[3]. The new data retrieved from the documents provided by MedIndia are analyzed and used to gradually enrich the domain knowledge base of Medico-Help. Information about new drugs and additional symptoms is also periodically added for this purpose. The available version of the knowledge base is used to generate answers to the user questions, most often in the form of assumptions about diagnoses corresponding to the indicated symptoms as well as suggestions about possible treatment regimens. Each diagnosis assumption includes information about the disease such as description, related symptoms, synonyms and drugs. The virtual assistant can also draw the user's attention to possible other related symptoms that might be missed.

4. Transfer learning

A popular approach in deep learning that supports the implementation of the principles of data centric AI is *transfer learning* where pre-trained models are used as a first approximation of the solution of primarily computer vision and natural language processing (NLP) tasks.

Jason Brownlee characterizes transfer learning is a "machine learning method where a model developed for a task is reused as the starting point for a model on a second task" [7].

There are many advantages of using transfer learning instead of a machine learning model built from scratch. The most significant of them are [8]:

- a transfer learning model needs less data as compared to a model build from scratch,
- a transfer learning model needs less computation power,
- a transfer learning model requires less time because most of the heavy work is already done on the pre-trained model and only a relatively small part is done by the new model.

The common approach to transfer learning in the field of deep learning is the Pre-trained Model Approach [9]. Its implementation consists of three main stages:

- *Select Source Model.* A proper pre-trained model is chosen from the set of available models. Lately many research institutions release freely available models on challenging datasets that can be included in the pool of candidate models from which to make choice.
- *Reuse Model.* The chosen pre-trained model is then used as the starting point for a model on the current task of interest.

[3] https://www.medindia.net

- *Tune Model.* The new model may need to be refined on the input-output data pairs available for the task of interest.

Nowadays it is popular to perform transfer learning on natural language processing problems in which text is used as input or output.

For these types of problems, an appropriate *word embedding* – a mapping of words to a high-dimensional real-valued vector space where different words with a similar meaning have a similar vector representation – is usually constructed and used [10].

There are many efficient techniques for learning this kind of word representations, e.g. Embedding Layer, Word2Vec, GloVe [7]. It is a common practice for research and development organizations to release models, pre-trained on large corpora of text documents under a permissive license.

A good illustration of the principles of transfer learning is the natural language processing technique supported by the recently popular Bidirectional Encoder Representations from Transformers (BERT) [10]. BERT is a method for generating a common language model that can understand natural language. The generated language model can then be used even without additional training. BERT has achieved some of the best results in many NLP tasks.

BERT is pre-trained on a very large corpus of non-annotated texts on the task of language modeling (15% of words are masked and BERT is trained to predict them from the context). The other task on which the model is pre-trained, is the task of predicting the next sentence. As a result of the training process, BERT learns appropriate contextual embeddings of words. After the preliminary training with non-annotated data on various tasks, BERT can be fine-tuned with fewer resources and smaller datasets to optimize its work on specific tasks. For fine tuning the model is first initialized with the parameters for pre-training and then all parameters are fine-tuned using annotated data from the further tasks. There are particular fine-tuned models for each of these tasks, although they are initialized with the same pre-training parameters[4], e.g. BERT-Large, Uncased (Whole Word Masking); BERT-Large, Cased (Whole Word Masking); BERT-Base, Multilingual Cased (New); BERT-Base, Chinese, etc.

5. An example: Intelligent system for answering specialized questions about COVID-19

The intelligent system for answering specialized questions about COVID-19 was designed and implemented in 2021–2022 as a diploma project for the completion of the Master's program in artificial intelligence at Sofia University [11]. It may be considered as a good example of application of the principles of data

[4] https://github.com/google-research/bert

centric AI, particularly of the transfer learning methodology in solving problems in information retrieval, natural language processing and knowledge discovery in text.

The development of the system was motivated by the popular COVID-19 Open Research Dataset Challenge of Kaggle[5]. It uses the Covid-19 Open Research Dataset (CORD-19), released in 2020 by the Allen Institute for AI (AI2) in cooperation with other leading institutions [12]. CORD-19 is a large and growing collection of more than 1,000,000 publications and preprints on Covid-19 and previous coronaviruses such as SARS and MERS. It integrates papers and preprints from several sources, collected by Semantic Scholar (see Figure 2). Paper documents are processed to extract full text. Metadata are harmonized by the Semantic Scholar team at AI2.

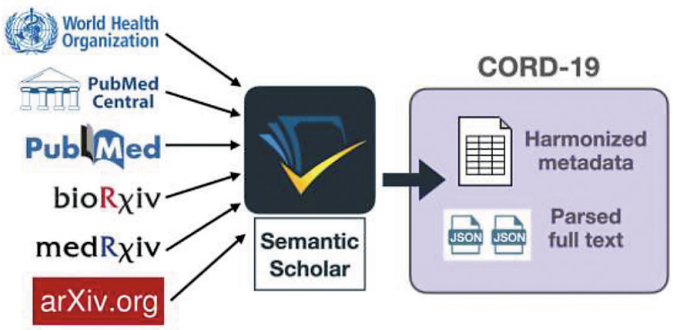

Figure 2: Data sources and structure of CORD-19 [12]

In the process of developing the system, preliminary preparation of the data was performed, which includes recognition of the language of each of the available papers and selection of those in English, followed by tokenization of the abstracts and texts of the selected papers. This results in the actual working version of the dataset, the content of which is used to generate the answers to the user questions.

When it receives a question from the user, the system first determines the rank of each paper in the dataset relative to the user question. The Okapi BM25 best matching ranking algorithm[6] is used for this purpose and the first five papers with the highest ranks are selected.

The next step is to use the BERT Large Uncased Whole World Masking model, pre-trained and fine-tuned on the Stanford Question Answering Dataset[7]. The texts of the five selected papers and the user question are submitted to it. As

[5] https://www.kaggle.com/datasets/allen-institute-for-ai/CORD-19-research-challenge

[6] https://nlp.stanford.edu/IR-book/html/htmledition/okapi-bm25-a-non-binary-model-1.html

[7] https://rajpurkar.github.io/SQuAD-explorer

a result of the execution of BERT, the generated answers to the user question are returned. Each answer contains data about the author(s) and the title of the respective paper, its estimated BERT score and BM25 score and a brief description of the essence of the results, presented in it, in their most general and abstract formulation.

The system was successfully tested on the questions from Round #1 of the cited competition of Kaggle. Figure 3 shows the results of the search for answers to the question "What do we know about vaccines and therapeutics?" (Task 3 from Round #1 of CORD-19 Challenge) and Figure 4 shows the results for the question "What has been published about medical care?" (Task 5 from Round #1 of CORD-19 Challenge).

Task: What do we know about vaccines and therapeutics?

	Title	Authors	Answer	BERT Score	BM25 Score
0	Value of Immunizations during the COVID-19 Eme...	Stefanati, Armando; d'Anchera, Erica; De Motol...	definition therapeutic protocols	0.012529	8.782067
1	In vitro testing of combined hydroxychloroquin...	Andreani, Julien; Le Bideau, Marion; Duflot, I...	36 analog quinine known inhibit acidification ...	0.000814	8.183944
2	Cell and animal models of SARS-CoV-2 pathogene...	Leist, Sarah R.; Schäfer, Alexandra; Martinez,...	cell animal models	0.112626	7.528309
3	A Targeted Vaccine against COVID-19: S1-Fc Vac...	Herrmann, Andreas; Maruyama, Junki; Yue, Chany...	sarscov2 available critically needed	0.023573	7.429836
4	Network graph representation of COVID-19 scien...	Cernile, George; Heritage, Trevor; Sebire, Nei...	timely access view urgency outbreak	0.078380	7.132842

Figure 3: Results for Task 3 from Round #1 of CORD-19 Challenge of Kaggle

Task: What has been published about medical care?

	Title	Authors	Answer	BERT Score	BM25 Score
0	Patient-collected tongue, nasal, and mid-turbi...	Tu, Yuan-Po; Jennings, Rachel; Hart, Brian; Ca...	health care workers described extensively	2.247430e-02	0.803355
1	D-dimer and C-reactive Protein Blood Levels Ov...	Becher, Yael; Goldman, Leonid; Schacham, Nadav...	previous reports	3.096434e-07	0.732179
2	First and second COVID-19 waves in Japan: A co...	Saito, Sho; Asai, Yusuke; Matsunaga, Nobuaki; ...	greater strain severe cases admission second w...	1.369792e-05	0.721525
3	Calling for an exponential escalation scheme i...	Wehling, Martin	rapidly vaccinating large share global population	4.819835e-07	0.711662
4	Geo temporal distribution of 1,688 Chinese hea...	Gao, Wayne; Sanna, Mattia; Wen, Chi Pang	one recent study found correlation higher cfr ...	5.233924e-05	0.705873

Figure 4: Results for Task 5 from Round #1 of CORD-19 Challenge of Kaggle

The analysis of the obtained experimental results shows that the system is relatively good at generating answers to specific questions, but it is advisable to improve the algorithm of its work by using more NLP techniques like lemmatization and dividing the texts of CORD-19 papers into separate paragraphs.

It would also be useful to enrich the dataset with which the system works with other types of documents related to COVID-19, such as technical reports, messages from governmental institutions and public organizations, etc.

Although a domain-specific corpus of data was used to create the system, the approach developed is general enough and can be applied in other areas.

6. Conclusion

Our experience in teaching AI and research and development activities in various areas of AI suggests that one of the significant challenges for data centric AI is the lack of validated methodologies – both domain-independent and domain-specific ones – for connecting small data to big data. The development of such methodologies and appropriate supporting software tools, along with the availability of a sufficient number of pre-trained machine learning models for different areas, would contribute to the rapid creation of intelligent software systems with great impact on large target groups, providing personalized services and reliable content.

6. Acknowledgements

This research is supported by Project BG05M2P001-1.001-0004 "Universities for Science, Informatics and Technologies in the e-Society (UNITe)" financed by Operational Program "Science and Education for Smart Growth", co-financed by the European Regional Development Fund.

7. References

[1] *R. Pollock, "*Forget big data, small data is the real revolution". The Guardian, 25 April 2013. URL: https://www.theguardian.com/news/datablog/2013/apr/25/forget-big-data-small-data-revolution (last visit on 31 March 2022).

[2] Small Data Group, Defining Small Data. URL: https://smalldatagroup.com/2013/10/18/defining-small-data (last visit on 31 March 2022).

[3] *C. Sarkar, "*Small Data, Big Impact!" – An Interview with Martin Lindstrom. *The Marketing Journal*, 1 May 2016. URL: https://www.marketingjournal.org/small-data-big-impact-an-interview-with-martin-lindstrom (last visit on 31 March 2022).

[4] R. Kannan, "The Importance of Small data vs Big Data for Healthcare". TRIGENT, 25 June 2019. URL: https://blog.trigent.com/the-importance-of-small-data-vs-big-data-for-healthcare (last visit on 31 March 2022).

[5] A. Ng, "Unbiggen AI". IEEE Spectrum, 9 February 2022. URL: https://spectrum.ieee.org/andrew-ng-data-centric-ai (last visit on 31 March 2022).

[6] R. Tsanova, Virtual Health Assistant. Master Thesis, Faculty of Mathematics and Informatics, Sofia University St. Kliment Ohridski, 2021 (in Bulgarian).

[7] J. Brownlee, "A Gentle Introduction to Transfer Learning for Deep Learning". Machine Learning Mastery, 16 September 2016. URL: https://machinelearningmastery.com/transfer-learning-for-deep-learning (last visit on 31 March 2022).

[8] R. Barman, S. Deshpande, S. Agarwal, U. Inamdar, "Transfer Learning for Small Dataset". Proceedings of National Conference on Machine Learning, 26th March 2019, ISBN 978-93-5351-521-8, pp. 132–137.

[9] P. Marcelino, "Transfer learning from pre-trained models". Towards Data Science, 23 October 2018. URL: https://towardsdatascience.com/transfer-learning-from-pre-trained-models-f2393f124751 (last visit on 31 March 2022).

[10] R. Horev, "BERT Explained: State of the art language model for NLP". Towards Data Science, 27 September 2021. URL: https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270 (last visit on 31 March 2022).

[11] B. Dobreva, Intelligent System for Answering Specialized Questions about COVID-19. Master Thesis, Faculty of Mathematics and Informatics, Sofia University St. Kliment Ohridski, 2022 (in Bulgarian).
[12] L. Wang et al., "CORD-19: The COVID-19 Open Research Dataset". Preprint, ArXiv, 2020;arXiv:2004.10706v2. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7251955 (last visit on 31 March 2022).