

**Nowy raport**

# Przegląd polskich zasobów językowych w otwartym dostępie o znaczącym potencjale dla AI w biznesie

Warszawa, grudzień 2024



Ministerstwo  
Cyfryzacji

Tu tworzymy przyszłość

**GRAi**  
GRUPA ROBOCZA  
DS. SZTUCZNEJ INTELIGENCJI

# Spis treści

<b>Autorzy</b>	<b>5</b>
<b>Wstęp</b>	<b>13</b>
Dane i ich znaczenie	14
Modele bazowe dlaczego to za mało?	16
<b>Zapotrzebowanie polskiego biznesu na językowe zasoby i zbiory danych</b>	<b>20</b>
<b>Definicje i zakres: Polskie zbiory danych i otwarte dane tekstowe</b>	<b>27</b>
Definicje	27
Zakres	29
Klasyfikacja polskich zasobów i zestawów danych	30
Klasyfikacja wg funkcji	30
Klasyfikacja wg źródeł	31
Klasyfikacja według wspólnych europejskich przestrzeni danych	32
Model dojrzałości otwartych danych	34
<b>Metodyka opisu i klasyfikacji polskich tekstowych zasobów i zestawów danych</b>	<b>35</b>
<b>Kompendium polskich zbiorów danych językowych</b>	<b>37</b>
<b>Szczegółowa analiza wybranych polskich zasobów i zbiorów danych</b>	<b>71</b>
Spichlerz	71
BAN-PL	81
PolQA	84
<b>Podsumowanie</b>	<b>90</b>
Ocena zaspokojenia potrzeb biznesowych	90
Dane językowe do specjalizowanych modeli uczenia maszynowego trenowanych pod kątem zadań NLP	91
Zestawy danych językowych wymagane do budowania i dostrajania modeli językowych	91
Proces walidacji zestawów danych	92
Rekomendacje dotyczące danych językowych	93
Kierunki rozwoju i kontynuacja prac	94
Analiza pogłębiona obszaru medycznego i farmaceutycznego (healthcare/pharma)	94
Eksploracja nowych dziedzin	96

Raport przygotowali eksperci działający na zasadach *pro publico bono* w Grupie Roboczej ds. Sztucznej Inteligencji (GRAI) przy Ministerstwie Cyfryzacji oraz zaproszeni przez nich do współpracy goście w ramach prac nad projektem pn. „Polskie zbiory danych typu open data: lista dostępnych zbiorów danych z polskiego rynku do budowania modeli”.

Ani Rada Ministrów, ani żadna osoba działająca w imieniu Rady Ministrów nie ponosi odpowiedzialności za sposób wykorzystania zamieszczonych w niniejszym materiale informacji. Wyłącznie odpowiedzialność za treści zawarte w niniejszym dokumencie ponoszą jego autorzy i zaproszeni przez nich do współpracy eksperci. Poglądy wyrażone w tym dokumencie odzwierciedlają opinię grupy ekspertów Grupy Roboczej ds. Sztucznej Inteligencji i w żadnym wypadku nie mogą być postrzegane jako oficjalne stanowisko Rady Ministrów ani jej poszczególnych członków.

Dokument może być kopiowany i wykorzystywany publicznie jedynie bez naruszania jego spójności. Prawa autorskie i majątkowe do materiałów wykorzystanych w raporcie, które pochodzą z obcych źródeł, należą o ich właścicieli.

## Szanowni Państwo!

Z ogromną satysfakcją przedstawiamy nasz najnowszy raport, który jest częścią szerszego projektu pod nazwą „Polskie zbiory danych typu open data: Lista dostępnych zbiorów danych z polskiego rynku do budowania modeli”. Publikacja ta skupia się na przeglądzie polskich zasobów tekstowych w otwartym dostępie, które posiadają istotny potencjał dla zastosowań sztucznej inteligencji (AI) w obszarze biznesu.

W dobie dynamicznego rozwoju generatywnej sztucznej inteligencji, zwłaszcza w obszarze dużych modeli językowych, biznes coraz bardziej docenia możliwości, jakie niesie wykorzystanie tych technologii. Ulepszenie procesów biznesowych, udoskonalenie i automatyzacja komunikacji z klientami oraz pracownikami – to tylko niektóre z obszarów, w których GenAI może przynieść znaczące korzyści. Jednak, aby w pełni wykorzystać te możliwości, niezbędne są odpowiednie dane – zarówno te szkoleniowe, oparte na historycznych informacjach, jak i produkcyjne, np. aktualne akty prawne w otwartym dostępie.

Nasza publikacja jest wynikiem prac Grupy Roboczej ds. Sztucznej Inteligencji. Ma na celu nie tylko stworzenie rejestru polskich zasobów tekstowych, ale również analizę istniejących luk i potrzeb w tym obszarze. Raport ten jest częścią większego projektu, który planujemy rozwijać. Włączymy do niego także inne rodzaje danych – zdjęcia, filmy, nagrania dźwiękowe – oraz rozszerzymy analizę w zależności od potrzeb szerszego grona odbiorców, wśród których znajdują się między innymi ośrodki naukowe czy strategiczne branże, np. z zakresu ochrony zdrowia czy środowiska.

W opracowaniu staraliśmy się uchwycić aktualne tendencje i wyzwania związane z otwartymi danymi oraz ich rolę w rozwoju AI. Dzięki wkładowi ekspertów z różnych dziedzin nasze badania zyskały interdyscyplinarny charakter, co pozwoliło na całościowe podejście do omawianych zagadnień. Raport rzuca światło na szeroki wachlarz możliwości dalszych badań i wypracowanych na ich podstawie rekomendacji.

Mamy nadzieję, że przedstawiony materiał stanie się cennym źródłem inspiracji i przyczyni się do wzrostu ilości badań, syntez oraz działań w obszarze sztucznej inteligencji i otwartych danych.

Z wyrazami szacunku

Autorzy

## Autorzy



### Sebastian Kondracki

Chief Innovation Officer, Deviniti

Lider SoDA AI Research Group

Twórca projektu SpeakLeash

Szef innowacji w Deviniti. To jedna z najbardziej dynamicznie rozwijających się firm, które tworzą oprogramowanie w Polsce. Lider zespołów „Polskie modele i narzędzia AI” oraz „Polskie zbiory danych typu open-data” w Grupie Roboczej ds. Sztucznej Inteligencji przy Ministerstwie Cyfryzacji. Ekspert ds. transformacji cyfrowej, zwłaszcza w obszarze jej głównych sił napędowych takich, jak: procesy innowacyjne i szeroko rozumiana sztuczna inteligencja. Autor książki „Python i AI dla e-commerce”, a także wykładowca na studiach podyplomowych w Akademii Leona Koźmińskiego i Wyższej Szkole Bankowej we Wrocławiu. Aktywny badacz wpływu „otwartości” (open data, open science, open source) na rozwój innowacyjności i kreatywności w społeczeństwie i biznesie. Współtwórca projektu SpeakLeash /'spix.lęs/ Spichlerz – zestawu narzędzi do budowy pierwszego w Polsce tak obszernego (>1 TB) zróżnicowanego zbioru danych wysokiej jakości wraz z bibliotekami umożliwiającymi łatwy dostęp i pełną otwartą replikację danych lub ich analizę. Specjalizuje się w AI: NLP (Natural Language Processing), NLG (Natural Language Generation), Generative AI, Explainable Artificial Intelligence (XAI) w biznesie, a w szczególności w e-commerce i w sektorze bankowym.



## Dr inż. Jerzy Biernacki

Chief AI Officer, Miquido

Organizator AI Waves

Związany z branżą IT od 15 lat, a od 10 z Miquido, czyli software housem, który wspiera przedsiębiorstwa w dostarczaniu innowacyjnych produktów cyfrowych. Krakowska firma nie tylko pomaga swoim partnerom w przyspieszeniu cyfrowego rozwoju, dzięki no/low-code, technologii cross-platform oraz sztucznej inteligencji, której ambasadorem jest Jerzy. Jego fascynacja AI oraz zapał do wdrażania najnowszych rozwiązań doprowadziła go do założenia i rozwoju działów R&D (badań i rozwoju) oraz AI w Miquido. Jest również organizatorem webinarów z serii AI Waves, gdzie rozmawia z ekspertami z całego świata o wykorzystywaniu potencjału sztucznej inteligencji w biznesie. Od kilku lat sprawuje kierownicze stanowisko, a przy tym aktywnie wspiera klientów w budowaniu innowacyjnych rozwiązań, w szczególności przy wykorzystaniu sztucznej inteligencji w aplikacjach. Brał udział w kilkudziesięciu projektach dla znanych firm w Polsce i zagranicą, takich jak: PZU, Play, Aviva, Onkyo czy HelloFresh. Występuje w roli eksperta na wielu znanych konferencjach branżowych, podcastach, czy webinarach. Chętnie dzieli się swoją wiedzą i spostrzeżeniami na profilu na LinkedIn, który przyciąga liczne grono odbiorców, czy w magazynach takich jak Forbes. Przez wiele lat związany był z Akademią Górniczo-Hutniczą w Krakowie, gdzie miał okazję rozwijać swoje zainteresowania badawcze, a następnie zdobył tytuł doktora informatyki.



## **Maria Filipkowska**

Data Scientist, Deviniti

Co-founder of Speakleash.org

Polonistka, doktor nauk humanistycznych w zakresie językoznawstwa, specjalistka Data Science zajmująca się przetwarzaniem języka naturalnego (NLP) z dużym doświadczeniem w branży tłumaczeniowej, obecnie pracuje jako Data Scientist w zespole R&D w Deviniti. W Spichlerzu analizuje dane językowe i bierze udział w tworzeniu zestawów instrukcji służących do ewaluacji modeli (tzw. benchmarków).



## **Krystian Bień**

Ekspert AI, IoMT, XR dla sektor ochrony zdrowia

Fundacja Platforma Przemysłu Przyszłości

Obecnie pełni funkcję Stream Lead Data Management, Analytics and Platform Design w globalnej firmie, gdzie codziennie wspiera ponad 400,000 pracowników. Skupia się na digitalizacji organizacji i wprowadzaniu innowacji. Specjalizuje się w przystosowaniu cyfrowych innowacji, takich jak Sztuczna Inteligencja (AI), Automatykacja Procesów Robotycznych (RPA), Internet Rzeczy (IoT) i Internet Rzeczy Medycznych (IoMT) oraz Rozszerzona Rzeczywistość (XR). Działa głównie w branży farmaceutycznej i medycznej, ale nie tylko. Doświadczenie to pozwoliło zdobyć wszechstronną wiedzę zarówno praktyczną, jak i teoretyczną. Wykształcenie łączące elementy techniczne i nietechniczne, umożliwia mu efektywne łączenie potrzeb tych dwóch obszarów. To przekłada się między innymi na wprowadzania innowacji w adekwatny sposób w zależności od poziomu organizacji, do którego jest ona zaadresowana (poziom: strategiczny, taktyczny, operacyjny). Posiada również szeroką wiedzę z zakresu zarządzania, projektowania systemów informatycznych i Data Science, zdobytą na różnych kursach i studiach podyplomowych.

Jako wykładowca akademicki dzieli się swoim doświadczeniem i ekspertyzą z przyszłymi specjalistami z różnych dziedzin, nie tylko technicznych. Jest aktywnym uczestnikiem krajowych i międzynarodowych konferencji ICT w roli prelegenta i autora publikacji naukowych. Bierze udział w projektach dla znanych firm i instytucji publicznych, zarówno w Polsce, jak i za granicą, a także jako ekspert w różnych fundacjach i grupach roboczych. Regularnie dzieli się swoimi spostrzeżeniami i doświadczeniami z szerszą społecznością poprzez LinkedIn oraz inne platformy branżowe.



## **Michał Kurasiński**

Szef Działu Transformacji AI, Inteligencji Biznesowej i Analityki Danych

Polpharma

Ukończył MBA dla kadry IT na Akademii Leona Koźmińskiego i Informatykę na Politechnice Częstochowskiej. Członek Grupy Roboczej ds. Sztucznej Inteligencji (GRAI) przy Kancelarii Prezesa Rady Ministrów. Współautor publikacji na temat „IoT w polskiej gospodarce”.

Posiada ponad dwadzieścia lat doświadczenia branży IT. Specjalizuje się w digitalizacji organizacji i wprowadzaniu innowacji. Jego doświadczenie umożliwia głębokie zrozumienie procesów działających w różnych organizacjach, co pozwala na efektywne wykorzystanie nowoczesnych technologii mających na celu poprawę działalności firmy. Jego doświadczenie obejmuje adaptację cyfrowych innowacji, takich jak Sztuczna Inteligencja (AI), Analityka Danych, Automatyzacja Procesów Robotycznych (RPA), Internet Rzeczy (IoT), platformy Low Code - No Code, oraz Rozszerzona Rzeczywistość (XR). Jego główne obszary działania to branża farmaceutyczna.





## Dr inż. Paweł Tadejko

Ekspert AI, Data Scientist, wykładowca akademicki

Wydział Informatyki Politechniki Białostockiej,

Polski Klaster Budowlany

Pracownik naukowo-dydaktyczny na Wydziale Informatyki Politechniki Białostockiej. Ekspert w zespole GRAI (Ministerstwo Cyfryzacji) w obszarach kompetencji przyszłości, cyfrowych i AI. Jako Chief Digital Officer w Polskim Klastrze Budowlanym odpowiada za doradztwo w tematach AI, transformacji cyfrowej i digital twin. Ekspert ds. transformacji cyfrowej przedsiębiorstw. Brał udział m.in. w projektach Szkoła Lidera 4.0 (w zakresie przygotowania liderów transformacji w firmach, transformacja cyfrowa, sztuczna inteligencja) i konkursach Fabryka Przyszłości (audyt dojrzałości cyfrowej przedsiębiorstw). Obszary specjalizacji to transformacja cyfrowa (w szczególności w obszarze „dane i analityka”, sztuczna inteligencja, big data (cloud i edge/fog computing) i cyfrowy bliźniak (digital twin), ze szczególnym uwzględnieniem branży budowlanej.

Head of Research and Development startupu tworzącego jedną z najbardziej znanych na świecie platform z obszaru komunikacji wewnętrznej [workai.com](https://workai.com). Trzykrotnie wyróżniony tytułem "Lista 100", za działania na rzecz edukacji cyfrowej, przygotowanej przez Porozumienie na Rzecz Umiejętności Cyfrowych w Polsce (PTI, Miasta w Internecie). Ze sztuczną Inteligencją związany od czasu pracy nad doktoratem i obrony w 2010. Doktorat z zastosowań AI w analizie EKG i klasyfikacji arytmii. Organizuje i kieruje studiami podyplomowymi "Data Science", „Marketing Internetowy”, czy "Projektowanie UX" na Politechnice Białostockiej.

Od ponad 20 lat pomaga korporacjom, firmom i startupom oswajać nowe technologie, jako analityk biznesowy, badacz UX, naukowiec i software/solution architect – głównie w projektach transformacji cyfrowej, Smart City i e-Administracji. Aktywny działacz podlaskiego Klastra Technologicznego InfoTECH i współzałożyciel Technikum Programistycznego InfoTECH w Białymstoku.



## Aleksander Obuchowski

Co-founder, CTO

TheLion.AI

Współzałożyciel grupy badawczej TheLion.ai tworzącej projekty open-source w obszarze sztucznej inteligencji w medycynie, gdzie pracuje nad projektami takimi jak Eskulap – polski medyczny model języka – i UMIE – uniwersalny enkoder obrazów medycznych. Wykładowca Akademicki, opiekun specjalizacji Sztuczna Inteligencja i doktorant na Polsko-Japońskiej Akademii Technik Komputerowych. Naukowo zajmuje się przetwarzaniem języka naturalnego (w szczególności modelami języka) i sztuczną inteligencją w medycynie. Finalista Forbes 25 under 25 w kategorii Nowe technologie.



## Paweł Kiszczak

Data Scientist, Deviniti

Data Scientist / ML Engineer, Speakleash.org

Data Scientist w dziale R&D w firmie Deviniti oraz członek zespołu Fundacji Spichlerz, odpowiedzialnej za stworzenie modelu językowego Bielik. na co dzień pracuje nad najnowszymi rozwiązaniami z zakresu sztucznej inteligencji i szuka sposobu wykorzystania jej w biznesie. Jako członek fundacji z kolei odpowiada za zbieranie i obróbkę danych, ich analizę jak i za rozwój aplikacji webowych. Autor wielu narzędzi do wewnętrznego użytku z zakresu crawlingu, przetwarzania danych oraz organizacji pracy. na swoim koncie posiada kilkadziesiąt gigabajtów danych, które wzbogaciły zbiory dostępne w projekcie Spichlerz.



## Adrian Gwoździej

Senior Data Scientist, Bank Pekao S.A., [Speakleash.org](https://speakleash.org)

Inżynier uczenia maszynowego z 5-letnim doświadczeniem w obszarze bankowości i cyberbezpieczeństwa. Pasjonat sztucznej inteligencji zaangażowany w projekt Spichlerz, gdzie odpowiada za budowanie modeli AI, które klasyfikują treści oraz określają ich jakość. Zajmuje się także tworzeniem zaawansowanych narzędzi do pozyskiwania danych z Internetu i ich przetwarzania. Na co dzień pracuje w domenie NLP.



## Igor Ciuciura

Software Developer, SII

Data Scientist / ML Engineer, [Speakleash.org](https://speakleash.org)

Programista, który zawodowo specjalizuje się w tworzeniu rozwiązań z obszaru testowania i optymalizacji działania urządzeń sieciowych w SII. w Spichlerzu zajmuje się wszelkimi procesami, w które zaangażowane są dane pozyskiwane w ramach projektu. Współtwórca oprogramowania oraz licznych narzędzi z dziedziny sztucznej inteligencji, które stanowią o technicznym rozwoju Spichlerza.



## Michał Chodorek

Adwokat, Partner, kancelaria KRK  
Kieszewska Rutkowska Kolasiński

Specjalista w zakresie prawa Life Sciences i prawa  
nowych technologii w branżach

Healthcare i Life Sciences. Doradza klientom w projektach e-Health i m-Health, badaniach klinicznych, w projektach dotyczących rozwoju wyrobów medycznych (w tym AI / *software as a medical device*), pozyskiwania i wykorzystywania danych medycznych – w tym dla potrzeb trenowania algorytmów AI / ML, wtórnego wykorzystywania danych pochodzących od uczestników badań klinicznych, rozwoju i wdrażania narzędzi AI/ML wspierających decyzje diagnostyczne i terapeutyczne. Pracuje dla wiodących międzynarodowych firm farmaceutycznych i medycznych działających w Polsce i wielu organizacji branżowych. Autor publikacji dotyczących klasyfikacji oprogramowania jako wyrobu medycznego, wykorzystywania AI w badaniach klinicznych i procesie badania i rozwoju leków i przetwarzania danych osobowych w ramach badań klinicznych. Współautor międzynarodowych publikacji dotyczących Digital Healthcare i ochrony tajemnicy przedsiębiorstwa. Członek Stowarzyszenia na rzecz Dobrej Praktyki Badań Klinicznych i Grupy Roboczej ds. Sztucznej Inteligencji (GRAI).

## Wstęp

Zaawansowane technologie cyfrowe, w tym zwłaszcza sztuczna inteligencja (z ang. *Artificial Intelligence* – AI<sup>1</sup>), stają się częścią naszego życia. Przyciągają uwagę i rozbudzają wyobraźnię, a przez to stają się nośnikiem ważnych zmian społecznych, politycznych i gospodarczych. Zmieniają sposób, w jaki podmioty rynkowe podejmują decyzje i wchodzi w interakcję ze swoimi zewnętrznymi interesariuszami (np. pracownikami, klientami, społeczeństwem)<sup>2</sup>. Pozwalają na kreację nowych modeli biznesowych, wdrażanie nowych narzędzi i strategii konkurencyjnych, ulepszenie dotychczasowych procesów. Wpływają na nasze potrzeby, oczekiwania, zachowania oraz postawy i w ten sposób łączą nas w sieć cyfrowych powiązań i interakcji. Wszystko to dzieje

się wokół nas, a zarazem poza nami, i tworzy nasze nowe cyfrowe otoczenie<sup>3</sup>.

W ciągu ostatniego roku duże firmy technologiczne z USA udostępniły narzędzia generatywnej sztucznej inteligencji (GenAI), takie jak ChatGPT, Gemini, GitHub Copilot, Stable Diffusion, co zapoczątkowało dynamiczne zmiany w gospodarce – ich zakres i skutki poznamy dopiero w przyszłości<sup>4</sup>. Poruszenie na rynku na niespotykaną wcześniej skalę zostało wywołane przez dwa czynniki: powszechny dostęp do tych narzędzi oraz możliwość ich zastosowania do wielu codziennych zadań. Ten przewrót technologiczny zwiększył zapotrzebowanie na zasoby, w których istotne miejsce zajmują dane.

<sup>1</sup> W języku polskim używa się też niekiedy skrótu SI (od tłumaczenia tego terminu: sztuczna inteligencja).

<sup>2</sup> A.Kaplan, M.Haenlein. (2019), *Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations,*

*illustrations, and implications of artificial intelligence*, „Business Horizons”, vol. 62, no. 1, s. 15-25.

<sup>3</sup> D. Kaczorowska-Spychalska, Ł.Sułkowski (2021), *Determinants of the adoption of AI wearables – practical*

*implications for marketing*, „Human Technology”, vol. 17, no 3, s. 294-320.

<sup>4</sup> M. Chui et.al (2023), *The economic potential of generative AI: The next productivity frontier*, McKinsey & Company [dostęp: 03.12.2023]

# Dane i ich znaczenie

---

Ludzkość nigdy wcześniej nie zbierała ani nie przetwarzała tak dużych ilości danych. Przeszliśmy od rewolucji informacyjnej do rewolucji danych pochodzących z wielu źródeł, obiektywnych i dostępnych w czasie rzeczywistym<sup>5</sup>. Skutkiem tego zjawiska jest nieustanny przepływ informacji pomiędzy niemal nieograniczoną liczbą nadawców i odbiorców. Dążymy do przekładania każdego elementu rzeczywistości na dane. Nie tylko na potrzeby konkretnego, zaplanowanego procesu czy projektu, ale z myślą o wykorzystaniu ich w przyszłości w sposób, o którym, na etapie ich gromadzenia, nawet nie jesteśmy w stanie pomyśleć<sup>6</sup>. Analiza danych pozwala na odkrycie szeregu wzorców i zależności. Dzięki temu poszczególne gospodarki, przedsiębiorstwa i społeczeństwo mogą ewoluować od analogowej przeszłości, poprzez cyfrową teraźniejszość, ku wyzwaniom autonomicznej przyszłości<sup>7</sup>.

Zarządzanie danymi jako wspólnym zasobem to najbardziej perspektywiczna strategia dla nowoczesnych gospodarek (*środek dla wielu aktorów i do wielu celów*)<sup>8</sup>. Zgodnie z raportem *Open Data Maturity 2021*, Polska znalazła się na czwartym miejscu wśród państw zaliczanych do kategorii *Trendsetters*<sup>9</sup>. To grupa krajów, które realizują obecnie bardzo zaawansowaną politykę otwartych danych na wszystkich szczeblach swojej administracji. Udostępniają one wysokiej jakości informacje dostosowane do potrzeb użytkowników, wychodzą z inicjatywą i wspierają powstawanie ekosystemów otwartych danych, charakteryzujących się wysokim poziomem interakcji i ponownego ich

---

<sup>5</sup> A. Poniewierski (2020), *SPEED. Bez granic w cyfrowym świecie*, [www.speednolimits.com](http://www.speednolimits.com), Warszawa s. 54.

<sup>6</sup> Ł. Iwasiński (2016), *Spoleczne zagrożenia danetyzacji rzeczywistości*, [w:] M. Przystek-Samokowa, B. Sosińska- Kalata, Z. Wiorogórska (red.), *Nauka o informacji w okresie zmian. Informatologia i humanistyka cyfrowa*, Wydawnictwo SBP, Warszawa, s. 135-146.

<sup>7</sup> B. Gregor, D. Kaczorowska-Spychalska (2020), *Technologie cyfrowe w biznesie. Przedsiębiorstwa 4.0 a sztuczna inteligencja*, PWN, Warszawa, s. 9.

<sup>8</sup> B. Wawrzyniak, M. Musidłowska, J. J. Zygmuntowski (2022), *Uwolnić potencjał danych. Zarządzanie danymi jako zasobem współdzielonym*, KPRM, Warszawa, s. 11.

<sup>9</sup> Raport dokonuje oceny dojrzałości cyfrowej poszczególnych krajów, biorąc pod uwagę cztery kryteria: realizowaną politykę w zakresie otwartych danych, ocenę funkcji jakie posiadają krajowe portale otwartych danych, w tym tych umożliwiających użytkownikom dostęp do danych i wsparcie zachodzących interakcji, wpływ w oparciu o działania mające na celu monitorowanie i pomiar ponownego wykorzystania otwartych danych, a także jakość (meta-)danych. D. van Hesteren, L. van Knippenberg (2021), *Open Data Maturity Report*, Luxembourg: Publications Office of the European Union, s. 3.

wykorzystania. Proces ten podlega pomiarowi z zastosowaniem metodologii oceny wpływu w różnych dziedzinach<sup>10</sup>.

Wśród kluczowych zaleceń dla tej grupy państw, znalazły się między innymi kwestie związane z podnoszeniem ilości i jakości udostępnianych danych i ich prezentacji, takie jak:

1. ich zatwierdzenie,
2. przekształcenie na różne alternatywne formaty,
3. lepsze zrozumienie potrzeb odbiorców,
4. umieszczanie linków do źródeł danych udostępnianych w czasie rzeczywistym,
5. umożliwienie użytkownikom komentowania i oceny zestawów danych z możliwością ich wprowadzenia w algorytmy wyszukiwania
6. pogłębianie współpracy z różnymi grupami interesariuszy (np. zespoły otwartych danych, uniwersytety, instytucje badawcze) w celu opracowania ram i wskaźników pomiaru i połączenie i wsparcie ekosystemów otwartych danych,
7. udoskonalenie strategii w zakresie otwartych danych.

Wydaje się, że współcześnie konieczne jest głębokie osadzenie działalności w perspektywie danych i zmiana dotychczasowego sposobu myślenia w tym zakresie, aby ludzie i firmy<sup>11</sup>, branże, sektory, grupy społeczne, poszczególne gospodarki i społeczeństwa czerpały korzyści. Całkowita wolność dostępu do danych powinna jednak dotyczyć danych publicznych czy danych pochodzących ze środowiska naturalnego człowieka. Udostępnianie danych firm, szczególnie sektora MŚP (małych i średnich przedsiębiorstw), czy też tych dotyczących zdrowia obywateli – ze względu na swoją specyfikę – powinno być ograniczone, zgodnie z podejściem wypracowanym dla danego sektora<sup>12</sup>. Tworząc bowiem technologie i je wykorzystując, możemy w istotnym stopniu zmieniać świat, aktywnie się do niego adaptować lub nawet go kreować<sup>13</sup>.

---

<sup>10</sup> D. van Hesteren, L. van Knippenberg (2021), Open Data Maturity Report, Luxembourg: Publications Office of the European Union, s. 94.

<sup>11</sup> A. Castrounis, (2019), AI for people and business, O'Reilly Media, Inc., USA, s. 11.

<sup>12</sup> M. Borowik, L. Maśniak, R. Kroplewski, H. Romaniec (2017), [Przemysł + Gospodarka oparta o dane, Ministerstwo Cyfryzacji](#) [dostęp: 28.10.2022].

<sup>13</sup> E. Brynjolfsson, A. McAfee (2016), *The second machine age*, John Wiley & Sons, New York-London, s. 362.

# Modele bazowe dlaczego to za mało?

---

Modele leżące u podstaw innowacji z zakresu GenAI to tzw. *modele bazowe* (z ang. *foundation models*). Są to modele ogólnego zastosowania, które następnie można dostroić (z ang. *fine-tuning*) do wymagań lokalnych (np. do danego języka – przeważająca większość tych modeli powstaje dla języka angielskiego) lub do konkretnych zadań (np. konwersacja, analiza wydźwięku, rozpoznawanie nazw własnych itp.). Dostrajanie również generuje koszty, wymaga kolejnych zestawów danych, odpowiednio oznaczonych i przygotowanych pod kątem wybranych zastosowań, a za dostęp do modeli bazowych najczęściej trzeba płacić.

To jednak nie wszystko. Pamiętajmy, że każda społeczność, która posługuje się jednym językiem i zamieszkuje określony teren, ma wspólną historię, ma też jedyne w swoim rodzaju i sobie właściwe dziedzictwo kulturowe. Znajduje ono odzwierciedlenie w języku (danych tekstowych), obrazach (danych graficznych), dźwiękach (danych audio i audiowizualnych). Jeśli zatem użyjemy modelu utworzonego przez specjalistów z obcego kręgu kulturowego i wytrenowanego na danych pochodzących głównie z tego kręgu, wraz z nim (nieświadomie) możemy przejąć obcy nam dotąd sposób myślenia o świecie<sup>14</sup>.

Na koniec kilka przykładów. Jeśli poprosimy Stable Diffusion<sup>15</sup> o wygenerowanie zdjęcia curry – potrawy indyjskiej (znanej na całym świecie, popularnej też w USA, a zatem przypuszczalnie występującej w wielu zbiorach danych wizualnych) – możemy otrzymać obraz, w którym wielu z nas rozpozna coś bliskiego swojemu doświadczeniu.

---

<sup>14</sup> [O pożyczaniu innych światów czyli po co nam polski LLM](#), Ministerstwo Cyfryzacji [dostęp: 03.12.2023].

<sup>15</sup> To otwarty model do generowania obrazów, którego kod jest publicznie dostępny, można go znaleźć na stronie [github Stability-AI/Generative models](#)





**Ilustracja 1.** Wygenerowany obraz przedstawiający potrawę indyjską.

Źródło: obraz wygenerowany przez model Stable Diffusion

Jeśli jednak to samo narzędzie poprosimy o wygenerowanie przykładowego żurku, rezultat może nas zaskoczyć. W tym przypadku zapytanie składało się z jednego słowa: „żurek”. Próba sprecyzowania polecenia, np. „tradycyjny polski żurek” oraz „tradycyjny polski żurek w glinianej misce, podany na rustykalnym drewnianym stole” również nie przyniosła zadowalających rezultatów (wygenerowane obrazy przedstawiały odpowiednio zestaw owoców i pieczony drób w brytfannie).



**Ilustracja 2.** Wygenerowany obraz przedstawiający polski żurek

Źródło: obraz wygenerowany przez model Stable Diffusion

Komercyjny model DALL-E 3 (firmy OpenAI) z podobnym zadaniem poradził już sobie trochę lepiej, ale nadal otrzymane zdjęcie jest niejednoznaczne. Oto, co można uzyskać, wpisując „żurek” w polu zapytań.



**Ilustracja 3.** Wygenerowany obraz przedstawiający polski żurek

Źródło: obraz wygenerowany przez model DALL-E 3

Podobnie rzecz ma się z literaturą: oto, co przeczytamy, kiedy poprosimy ChatGPT o opowiadanie legendy o Złotej Kaczce:

**You**  
opowiedz legendę o złotej kaczce

**ChatGPT**  
Istnieje kilka różnych wersji legendy o Złotej Kaczce, ale jedna z najbardziej znanych pochodzi z baśni Hansa Christiana Andersena i nosi tytuł "Złota Kaczka". Oto krótkie streszczenie tej baśni:

W pewnym królestwie żył biedny chłopiec imieniem Hans, który pracował u zamożnego rolnika. Jednego dnia rolnik powierzył mu zadanie strzeżenia stada kaczek, a wśród nich była jedna, zwyczajna kaczka, która wydawała się inna niż reszta. Gdy chłopiec zbliżył się do kaczki, ta nagle zamieniła się w piękną Złotą Kaczkę.

Złota Kaczka była zaklęta, a jedynym sposobem na jej uwolnienie było znalezienie księżniczki, która była gotowa poślubić Hansa. Chłopiec wyruszył w podróż, a na swojej drodze spotkał różne postacie, w tym starą kobietę, która udzieliła mu pomocy. Gdy Hans dotarł do zamku, księżniczka zgodziła się poślubić biednego chłopca.

W efekcie tego aktu miłości i dobra, Złota Kaczka straciła swoje złote pióra, a zamiast nich pojawiło się piękne pióropusze. Księżniczka i Hans żyli długo i szczęśliwie, a Złota Kaczka była dowodem na to, że dobro, pokora i miłość mogą przełamać klątwy i przynieść szczęście.

**Ilustracja 4.** Prośba wygenerowania legendy o złotej kaczce w ChatGPT

Źródło: chat.openai.com (GPT-3.5 Turbo)

Warszawska legenda o Złotej Kaczce jest opisana np. na stronie [Go to Warsaw](#) [dostęp 19.11.2024]. Być może bajka Brzydkie kaczątko Hansa Christiana Andersena jest lepiej reprezentowana w danych treningowych, stąd to nazwisko się pojawiło w odpowiedzi.

Nawet reguły ortograficzne mogą w GenAI się zmienić. Pamiętajmy, że w języku polskim, w tytule książki, zapisujemy wielką literą tylko pierwsze słowo (o ile pozostałe nie są nazwami własnymi), natomiast w angielskim w każdym słowie tytułu stosuje się wielkie litery:

**You**  
przetłumacz tytuł książki: So Long, and Thanks for All the Fish

**ChatGPT**  
"Tak Długo i Dzięki za Wszystkie Ryby"



**Ilustracja 5.** Prośba o przetłumaczenie tytułu książki w ChatGPT

Źródło: chat.openai.com (GPT-3.5 Turbo)

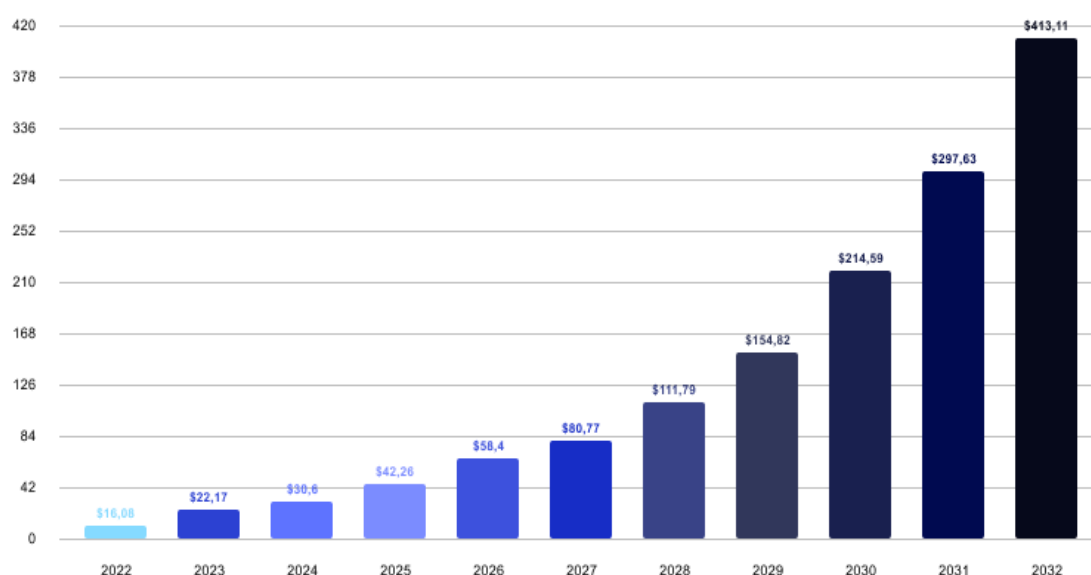
Poprawne tłumaczenie (i poprawny zapis) tytułu to: *Cześć, i dzięki za ryby* (autor: Douglas Adams).



## Zapotrzebowanie polskiego biznesu na językowe zasoby i zbiory danych

Wybór tego tematu jako pierwszego do raportu cząstkowego (tj. przeglądu polskich zasobów tekstowych, w ramach prac nad projektem „Polskie zbiory danych typu open data: Lista dostępnych zbiorów danych z polskiego rynku do budowania modeli”) nie jest przypadkowy. Dane te odgrywają kluczową rolę w budowaniu i dostrajaniu modeli wykorzystywanych w dziedzinie przetwarzania języka naturalnego (NLP). w kontekście biznesowym efektywność, dostępność oraz koszty związane z przetwarzaniem języka naturalnego mają istotne znaczenie, bo większość komunikacji biznesowej – zarówno z klientami, pracownikami, jak i partnerami – opiera się na tekście.

Wartość globalnego rynku przetwarzania języka naturalnego



**Wykres 1.** Wartość globalnego rynku przetwarzania języka naturalnego

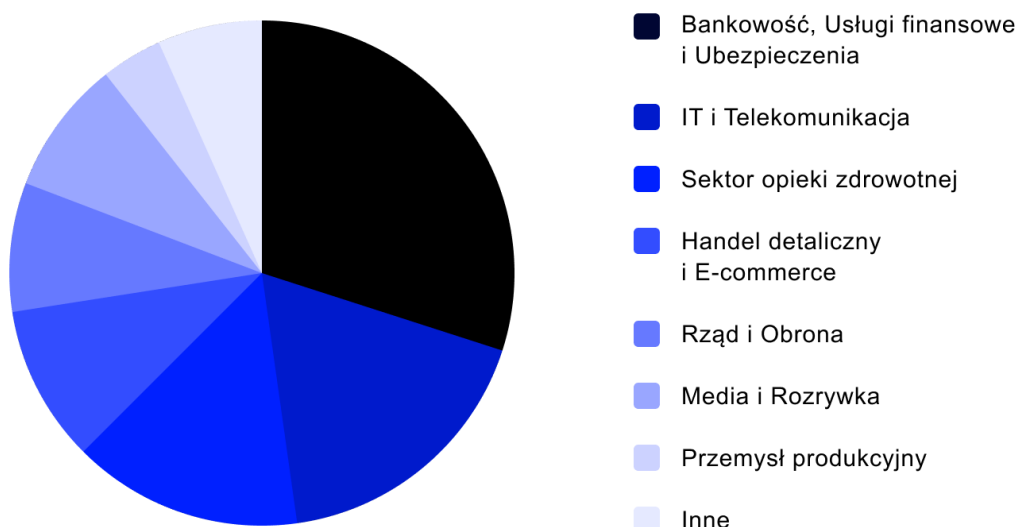
Źródło: [Natural Language Processing Market Size, Share, and Trends 2024 to 2034](#), Precedence Research [dostęp: 03.12.2023]

Znaczenie technologii przetwarzania języka naturalnego (NLP) w perspektywie rynkowej zostało wyraźnie podkreślone [w badaniach przeprowadzonych przez Precedence Research](#).

To firma wyspecjalizowana w analizach rynkowych i usługach doradczych. Według ich ostatniej analizy globalny rynek NLP osiągnął wartość 16,08 miliarda USD w 2022 roku

i, według prognoz, ma wzrosnąć do około 413,11 miliarda USD do 2032 roku. Oznacza to złożoną roczną stopę wzrostu (CAGR) na poziomie 38,4% w okresie od 2023 do 2032 roku. Te dane obrazują dynamicznie rosnące zapotrzebowanie na rozwiązania i usługi NLP w sektorze biznesowym. Poniższy wykres „Udział branż w globalnym rynku NLP” przedstawia proporcje, w jakich poszczególne sektory gospodarcze przyczyniają się do ogólnej wartości rynku NLP.

### Udział branż w globalnym rynku NLP



**Wykres 2.** Udział branż w globalnym rynku NLP

Źródło: [Global Natural Language Processing Market Size](#), KBV Research [dostęp: 03.12.2023]

Wyraźnie widać, że największy udział mają branże, w których kluczowe jest zarządzanie komunikacją z szeroko rozproszonymi klientami i w których usługi często świadczone są cyfrowo. Są to sektor BFSI oraz IT i telekomunikacja. Sektor opieki zdrowotnej również wykazuje znaczący udział, co odzwierciedla zwiększone zastosowanie NLP w obszarach zarządzania informacjami o pacjentach oraz w diagnostyce. Mniejsze, lecz nadal ważne segmenty rynku, takie jak handel detaliczny i e-commerce, rząd i obrona, media i rozrywka oraz przemysł produkcyjny, potwierdzają wszechstronność NLP. Pozostałe sektory łącznie stanowią istotny segment, który odzwierciedla zastosowanie NLP w szerokiej gamie dziedzin o mniejszym udziale rynkowym.

## Komentarz eksperta



### Dr inż. Jerzy Biernacki

Chief AI Officer, Miquido  
Organizator AI Waves

## Przetwarzanie języka naturalnego w sektorze BFSI

Specyfika sektora BFSI wymaga przetwarzania dużej ilości dokumentów tekstowych. Wiele zadań, zarówno w ramach wewnętrznych procesów, jak i w interakcjach z klientami, wymaga efektywnego przetwarzania języka naturalnego. w związku z tym, aby zautomatyzować te procesy lub tworzyć rozwiązania skierowane do obsługi klienta, niezbędne są modele językowe, które nie tylko bardzo dobrze radzą sobie z językiem polskim, ale także często ze specyfiką branży. Tekstowe zbiory danych w języku polskim są w tym kontekście kluczowe dla tworzenia solidnych rozwiązań opartych o modele językowe. bez odpowiedniej ilości i jakości danych tekstowych nie jest możliwy efektywny fine-tuning modeli językowych (istotna część cyklu rozwoju tych modeli) oraz budowanie rozwiązań opartych o architekturę RAG (z ang. Retrieval Augmented Generation).

W obszarze wsparcia i obsługi klienta, wykorzystanie modeli językowych pozwala na tworzenie zaawansowanych chatbotów, które udzielają informacji o produktach i usługach w naturalny sposób. Mogą one przykładowo zostać wykorzystane do streszczania lub tłumaczenie klientom wniosków, regulaminów czy umów. Takie wsparcie jest nieocenione dla klientów, którzy nie są obeznani z branżową terminologią lub chcą szybko uzyskać konkretną informację, bez konieczności przeglądania kilkudziesięcio- stronicowego dokumentu. Ponadto tego typu rozwiązania umożliwiają wsparcie dla obcokrajowców, w tym imigrantów, którzy chcą korzystać z usług finansowych i ubezpieczeniowych dostępnych w Polsce.

W zakresie automatyzacji procesów wewnętrznych, modele językowe przyczyniają się do efektywniejszego przetwarzania wniosków oraz dokumentów, co przekłada się na szybsze i dokładniejsze podejmowanie decyzji oraz lepszą organizację pracy. Skrócenie czasu procesowania dokumentów przekłada się również realnie na satysfakcję klientów.

Modele językowe mogą zostać wykorzystane również do budowania rozwiązań, które zwiększają bezpieczeństwo. Przykładami takich rozwiązań są systemy wykrywające ataki socjotechniczne, oszustwa czy próby wyłudzeń w Internecie. Przeciwdziałanie tego typu zagrożeniom zwiększa także zaufanie klientów do instytucji finansowych.

## Komentarz eksperta



### Krystian Bień

Ekspert AI, IoMT, XR dla sektor ochrony zdrowia

Fundacja Platforma Przemysłu Przyszłości

## Przetwarzanie języka naturalnego w sektorze Healthcare/MedTech

Przetwarzanie języka naturalnego (NLP) odgrywa kluczową rolę w sektorze opieki zdrowotnej i MedTech i pomaga w zarządzaniu i interpretacji ogromnych ilości danych medycznych. Dzięki NLP możliwe jest efektywne analizowanie notatek lekarskich, wyników badań i historii chorób, co przekłada się na lepsze zrozumienie potrzeb pacjentów i szybsze podejmowanie decyzji klinicznych. NLP znacząco przyspiesza proces tworzenia i zarządzania dokumentacją medyczną, co jest kluczowe dla efektywności pracy personelu medycznego. Dzięki temu technologia ta nie tylko usprawnia pracę placówek medycznych, ale również przyczynia się do lepszego zarządzania czasem i zasobami. Ponadto wykorzystanie NLP w tworzeniu interaktywnych chatbotów medycznych ułatwia pacjentom dostęp do informacji o objawach, leczeniu, czy poradach zdrowotnych. Jest to szczególnie ważne w zarządzaniu chorobami przewlekłymi oraz w sytuacjach, gdy pacjent potrzebuje szybkiej i rzetelnej porady. NLP pomaga również w analizie i monitorowaniu opinii pacjentów, co jest nieocenione dla poprawy jakości usług i zrozumienia ich potrzeb. Poprzez analizę informacji zwrotnej, placówki medyczne mogą dostosowywać swoje usługi do oczekiwań i wymagań pacjentów. Dodatkowo modele językowe mogą być wykorzystane do wykrywania potencjalnych oszustw i nadużyć w dokumentacji medycznej, co ma istotne znaczenie dla ochrony zasobów systemu opieki zdrowotnej. W ten sposób NLP przyczynia się do zwiększenia bezpieczeństwa i efektywności w sektorze zdrowia. Kolejnym ważnym aspektem jest tłumaczenie i dostępność językowa. NLP umożliwia tłumaczenie informacji medycznych na różne języki, co jest niezbędne w międzynarodowym środowisku medycznym oraz w obszarach o wielokulturowej populacji.

Podsumowując, przetwarzanie języka naturalnego w sektorze opieki zdrowotnej i MedTech przyczynia się nie tylko do usprawnienia procesów i obniżenia kosztów, ale również do znaczącej poprawy jakości opieki nad pacjentami.

Określenie najczęstszych zadań NLP w biznesie zależy od specyfiki branży. Na przykład, w mediach dominują generatory tekstu, automatyczna korekta i stylizacja tekstu, podczas gdy sektory zorientowane na klienta takie jak bankowość, często stosują narzędzia do analizy komunikacji i sentymentu. Oto zarys najbardziej powszechnych zastosowań NLP w biznesie:

1. Analiza sentymentu – szczególnie przydatna do oceny informacji zwrotnej od klientów.
2. Klasyfikacja tekstu – wykorzystywana do filtrowania e-maili i innych form komunikacji.
3. Rozpoznawanie mowy – podstawa działania asystentów głosowych.
4. Tłumaczenie maszynowe – narzędzia tłumaczące tekst dla łamania barier językowych.
5. Rozpoznawanie i ekstrakcja informacji – kluczowe w rekrutacji i analizach rynkowych.
6. Generowanie tekstu – automatyzacja tworzenia raportów i content marketingu.
7. Autokorekta i autouzupełnianie – ułatwienie w wyszukiwaniu informacji.
8. Chatboty – usprawnienie obsługi klienta.
9. Knowledgeboty – zarządzanie wiedzą firmową.
10. Monitorowanie mediów społecznościowych – analiza trendów i reakcji rynkowych.
11. Streszczenia językowe – efektywne podsumowywanie spotkań i dokumentów.
12. Wsparcie rozwoju produktów i usług np. identyfikacja luk funkcjonalnych w produktach na podstawie analizy opinii klientów, generowanie pomysłów na nowe funkcjonalności oparte na rzeczywistych potrzebach i preferencjach klientów poprzez analizę komunikacji z klientem.

Wszystkie wspomniane techniki NLP wymagają zestawów danych tekstowych, często oznaczonych ręcznie przez człowieka. Przykładowo, treści e-maili są klasyfikowane do odpowiednich kategorii, a opinie klientów o usługach i produktach są etykietowane jako pozytywne, negatywne lub neutralne. Istotna jest domena danego sektora: np. w kategorii wiertarek ręcznych stwierdzenie „produkt jest bardzo ciężki” może być negatywne, podczas gdy w kategorii młotów kowalskich może być pozytywne. Kluczowe jest też dopasowanie językowe danych: dla analizy opinii w języku polskim potrzebny jest zestaw danych właśnie w tym języku.



W 2017 roku opublikowany został pionierski artykuł „*Attention Is All You Need*”<sup>16</sup>, który pokazał światu architekturę uczenia głębokiego zwaną *transformer*, przeznaczoną dla modeli przetwarzających dane sekwencyjne, takie jak zdania czy akapity. Ta innowacja zrewolucjonizowała branżę NLP, nadając narzędziom zdolności do zaawansowanej interpretacji tekstu, np. rozpoznawania ironii i sarkazmu oraz wydobywanie informacji z materiałów o niskiej jakości, takich jak skany dokumentów czy transkrypcje audio. w efekcie wzrosło zapotrzebowanie na nowe rodzaje danych – duże, tematycznie zróżnicowane zbiory tekstowe wysokiej jakości (do trenowania między innymi dużych modeli językowych) oraz zestawy instrukcji (do dostrajania modeli bazowych).

Duże modele językowe są zwykle trenowane na rozbudowanych, wielojęzycznych zbiorach danych, co sprawia, że większość z nich jest *language-agnostic*. Oznacza to, że mogą one przetwarzać teksty w różnych językach, w tym angielskim, niemieckim, włoskim i polskim. Jednocześnie nie są specjalnie dostosowane do konkretnego z nich. Taka uniwersalność pozwala na szerokie zastosowanie technologii w różnorodnych kontekstach językowych. Jednak w specyficznych dziedzinach taka generalizacja może prowadzić do nieścisłości czy zapożyczeń językowych. na przykład, model korygujący polski tekst branżowy może nieodpowiednio używać angielskich terminów lub błędnie formatować śródtytuły. w związku z tym, przy narzędziach typu AI writing assistant, istotne jest dostosowanie modelu do konkretnych potrzeb językowych danej branży, co podkreśla rynkową potrzebę posiadania specjalistycznych zestawów danych dla każdej branży, najczęściej w formie instrukcji (na przykład tekst przed korektą i po korekcie i stylizacji domenowej).

Rozwój dużych modeli językowych (LLM) i generatywnej sztucznej inteligencji (GenAI) sprawił, że biznes z entuzjazmem przyjął systemy typu *question answering*. Są one idealne do tworzenia chatbotów i knowledgebotów, które mogą odpowiadać na pytania klientów dotyczące produktów i wspierać ich w procesie wyboru, a także pomagać pracownikom w zrozumieniu wewnętrznych procedur czy regulacji. Najczęściej wykorzystywaną architekturą w takich zastosowaniach jest model RAG (z ang. *Retrieval Augmented Generation*), opracowany przez laboratorium Meta AI<sup>17</sup>. RAG łączy model językowy, używany do generowania odpowiedzi, z dostępem do pamięci zawierającej bazy danych lub

---

<sup>16</sup> A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin (2017), [Attention Is All You Need](#) [dostęp: 08-01-2024]

<sup>17</sup> P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. Yih, T. Rocktäschel, S. Riedel, D. Kiela (2021), Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks

indeksy dokumentów, aby uzyskać dodatkowe informacje. System, po otrzymaniu zapytania, najpierw używa komponentu pobierania danych do znalezienia najważniejszych treści, a następnie wykorzystuje je do stworzenia dokładniejszej odpowiedzi. Kluczową cechą RAG jest pamięć, która może zawierać zarówno wewnętrzne dokumenty firmy, jak i zewnętrzne źródła, np. aktualne akty prawne czy istotne publikacje medyczne. To otwiera nowe możliwości biznesowe, zwłaszcza w kontekście wykorzystania otwartych zasobów tekstowych. Pozwala na tworzenie odpowiedzi z pewnych źródeł, dzięki czemu wyniki są bardziej precyzyjne, kontekstualnie zgodne oraz przede wszystkim obarczone znacznie niższym ryzykiem podania nieprawdziwych informacji (halucynacji).



## Definicje i zakres: Polskie zbiory danych i otwarte dane tekstowe

Przed rozpoczęciem analizy polskich zasobów tekstowych w otwartym dostępie warto zdefiniować kluczowe terminy związane z tym tematem. Zrozumienie pojęć takich jak otwarty dostęp, otwarte dane, zestawy danych, oraz kryteriów, które decydują o tym, czy zestaw danych jest uznawany za *polski*, jest niezbędne dla pełnego zrozumienia kontekstu i zakresu naszej publikacji.

### Definicje

---

**Otwarte dane** zgodnie z definicją zawartą na stronie WWW „[Open Definition](#)”, administrowaną przez Open Knowledge Foundation (OKF), to: „Dane i treści, których można swobodnie używać, modyfikować i udostępniać każdemu w dowolnym celu”<sup>18</sup>. [Rozwijając tę definicję, wg OKF](#) kluczowymi aspektami otwartości są:

1. **Dostępność i dostęp** – dane muszą być dostępne w całości. Koszt dostępu do nich nie może być wyższy niż rozsądne koszty ich reprodukcji, najlepiej jeśli możliwe jest ich pobranie przez Internet. Dane muszą być również dostępne w wygodnej i łatwej do modyfikacji formie.
2. **Ponowne użycie i redystrybucja** – dane muszą być dostarczane na warunkach, które umożliwiają ich ponowne użycie i redystrybucję, w tym mieszanie z innymi zestawami danych. Dane muszą być czytelne dla maszyn.
3. **Powszechny udział** – każdy musi mieć możliwość użycia, ponownego wykorzystania i redystrybucji danych – nie powinno być dyskryminacji w zakresie dziedziny działalności lub w zakresie osób czy grup. Przykładowo ograniczenia typu „non-commercial”, które uniemożliwiałyby komercyjne wykorzystanie, lub ograniczenia używania danych tylko w określonych celach (np. wyłącznie w edukacji), nie są dozwolone<sup>19</sup>.

---

<sup>18</sup> [The Open Definition \(portal Open knowledge\)](#) [dostęp: 03.12.2023]

<sup>19</sup> [Open Knowledge Library](#) [dostęp: 03.12.2023]

Definicja otwartych danych stworzona przez OKF nie skupia się na źródle pochodzenia danych, w szczególności nie dzieli ich na dane pochodzące z sektora publicznego i sektora prywatnego (biznesu).

W kontekście sektora publicznego definicję otwartych danych znajdziemy [w ustawie z 11 sierpnia 2021 r. o otwartych danych i ponownym wykorzystywaniu informacji sektora publicznego \(Dz. U. 2021 poz. 1641\)](#)<sup>20</sup>, gdzie dane otwarte są zdefiniowane jako „informacje sektora publicznego udostępniane lub przekazywane w postaci elektronicznej, bezwarunkowo lub z uwzględnieniem warunków, o których mowa w rozdziale 3, kompletne, aktualne, w wersji źródłowej, w otwartym i niezastrzeżonym formacie przeznaczonym do odczytu maszynowego, które są przeznaczone do bezpłatnego ponownego wykorzystywania na tych samych zasadach dla każdego użytkownika, bez konieczności potwierdzania tożsamości przez użytkownika”.

Informacje sektora publicznego są zdefiniowane w ustawie (Dz. U. poz. 1641) jako „każda treść lub jej część, niezależnie od sposobu utrwalenia, w szczególności w postaci papierowej, elektronicznej, dźwiękowej, wizualnej lub audiowizualnej, będąca w posiadaniu podmiotu zobowiązanego”. Podmioty zobowiązane według ustawy (Dz. U. poz. 1641) obejmują szeroki zakres instytucji i organizacji. Należą do nich nie tylko jednostki sektora finansów publicznych (m.in organy administracji rządowej i samorządowej, jednostki budżetowe, samorządowe zakłady budżetowe, oraz inne jednostki, które są finansowane z budżetu państwa lub samorządu terytorialnego), ale również między innymi podmioty, które świadczą określone usługi publiczne, w tym w sektorze transportu lotniczego, drogowego czy morskiego. Ustawa (Dz. U. poz. 1641) wyłącza jednak niektóre instytucje, takie jak jednostki publicznej radiofonii i telewizji czy Polską Agencję Prasową.

**Polskie zestawy danych w otwartym dostępie** to zbiory danych charakteryzujące się polskim kontekstem, czyli zawierające informacje związane z językiem polskim, polskim społeczeństwem, kulturą lub innymi aspektami życia w Polsce. do przykładów należą dane tekstowe w języku polskim, badania dotyczące polskiej populacji, obrazy związane z polskim dziedzictwem kulturowym, czy inne materiały powiązane z Polską. Te zbiory mogą również zawierać dane w różnych językach, ale muszą oferować możliwość łatwego filtrowania lub wyodrębnienia treści specyficznie o kontekście polskim, zdefiniowanym powyżej.

---

<sup>20</sup> [dostęp 03.12.2023]

Zbiory te są dostępne w ramach otwartego dostępu, zgodnego z definicją otwartych danych. Zestawy danych mogą pochodzić zarówno z sektora publicznego, jak i prywatnego.

**Polskie otwarte dane tekstowe** stanowią specyficzną podkategorię w ramach polskich zestawów danych w otwartym dostępie, zawierającą dane w formie tekstowej. Podkategoria ta dotyczy szerokiego zakresu materiałów – od literatury po transkrypcje rozmów, artykuły prasowe, dokumenty historyczne i więcej. Kluczową cechą tych zbiorów jest ich otwartość, co oznacza, że mogą być swobodnie używane w dowolnym celu, w tym do rozwoju i treningu modeli językowych (również dużych).

## Zakres

---

Zakres tego raportu został zawężony specyficznie do polskich otwartych danych językowych tekstowych.



# Klasyfikacja polskich zasobów i zestawów danych

---

## Klasyfikacja wg funkcji

- 1. Zbiór czystych tekstów do uczenia nienadzorowanego i dużych modeli językowych (LLM)** – zbiory te są kluczowe w procesie tworzenia dużych modeli językowych, takich jak GPT czy BERT. Składają się z obszernych ilości nieprzetworzonych tekstów i umożliwiają modelom uczenie się rozumienia języka, jego struktur i kontekstów bez potrzeby stosowania etykiet. Jako przykład takiego zestawu danych można wymienić [„The Pile”](#)<sup>21</sup>.
- 2. Dane tekstowe z adnotacjami** – zbiory te składają się z tekstów, które zostały uzupełnione o specyficzne informacje, takie jak etykiety kategorii, identyfikacja nazwanych encji (NER) czy sentymenty. Są one podstawą do uczenia nadzorowanego w aplikacjach, takich jak rozpoznawanie nazwanych encji (NER), klasyfikacja tekstu (na przykład określanie sentymentu jako pozytywny lub negatywny) oraz rozpoznawanie intencji użytkownika. Przykładem takiego zbioru danych może być zestaw opinii z IMDB zawierający oceny sentymentu, dostępny na stronie [Large Movie Review Dataset](#).
- 3. Zestawy instrukcji** – zbiory te zawierają instrukcje lub pytania wraz z odpowiedziami lub przewidzianymi działaniami. Wykorzystuje się je do trenowania modeli pod kątem zadań polegających na rozumieniu instrukcji i realizowaniu działań w oparciu o treść tekstową. Stosuje się je w różnych kontekstach, takich jak dostrajanie dużych modeli językowych (LLM), tworzenie chatbotów, rozwijanie asystentów głosowych czy systemów odpowiadających na pytania. Przykładem takiego zbioru danych jest [„Alpaca Data” dostępny na platformie Hugging Face](#).
- 4. Dane dialogowe i konwersacyjne** – zbiory te obejmują transkrypcje rozmów i dialogów. Są wykorzystywane do trenowania modeli w celu zrozumienia kontekstu rozmowy, generowania odpowiedzi i prowadzenia naturalnych konwersacji. Stosowane są przede wszystkim w systemach dialogowych, chatbotach i asystentach głosowych, gdzie umożliwiają płynne i intuicyjne interakcje z użytkownikami.
- 5. Dane multimodalne** – zbiory te łączą tekst z innymi rodzajami danych, jak obrazy, dźwięk czy wideo, i są wykorzystywane do trenowania modeli, które rozumieją zależności między różnymi typami informacji. Takie modele znajdują zastosowanie

---

<sup>21</sup> [dostęp 19.11.2024]

w rozpoznawaniu obrazów opisanych słowami, generowaniu tekstowych opisów do obrazów czy automatycznym tworzeniu napisów do filmów. Dane multimodalne pozwalają na bardziej złożone i zaawansowane zadania przetwarzania informacji oraz integrują różne źródła danych w jednym modelu.

## Klasyfikacja wg źródeł

Dostępne dane można również klasyfikować według źródeł. Ten podział wydaje się wartościowy, ponieważ informacja o tym, skąd tekst pochodzi, daje wstępne wyobrażenie o typie i jakości tych danych. Posługując się kategoriami [z projektu CLARIN](#)<sup>22</sup>, możemy wyróżnić następujące typy źródeł:

1. **Prasa (i książki)** – te zbiory zawierają zazwyczaj długie teksty o dobrej jakości językowej, są też źródłem wiedzy o zmianach w systemie językowym, ponieważ zawierają dane z różnych epok. Przykładem takiego zbioru jest [projekt Gutenberg](#).
2. **Korpusy dyskursu parlamentarnego** – to zbiory języka mówionego: zapis wypowiedzi w sytuacji oficjalnej, w konwersacji ograniczonej pewnymi regułami. Mogą one być podstawą do analizy dyskursu, źródłem wiedzy historycznej, a także socjolingwistycznej.
3. **Korpusy komunikacji internetowej** – w porównaniu z dwoma poprzednimi zbiorami ten zawiera dane stosunkowo nowe, tj. głównie z ostatnich 20 lat. Należą do niego fora internetowe, blogi, czaty, ale także *Wikipedia*. To zbiór zróżnicowany pod względem jakości oraz formy i stylu komunikacji. Te dane są źródłem wiedzy o współczesnym języku i jego różnych odmianach, zawierają jednak dużo szumu.
4. **Korpusy równoległe** – gromadzą dane wielojęzyczne (np. Eur-LEX), znajdują zastosowanie w modelach tłumaczeniowych.
5. **Mieszane** – ta kategoria nie była uwzględniona w cytowanym artykule, ale widzimy potrzebę jej dodania, ponieważ w naszym katalogu pojawiają się zbiory, które łączą w sobie dane z różnych wymienionych wyżej źródeł (multizbiory, np. SpeakLeash).

---

<sup>22</sup> D. Fišer, J. Lenardič, T. Erjavec (2018), [CLARIN's Key Resource Families](#) [dostęp: 03.12.2023]

## Klasyfikacja według wspólnych europejskich przestrzeni danych

Wspólne europejskie przestrzenie danych (z ang. [Common European Data spaces](#))<sup>23</sup> to kluczowa inicjatywa w ramach polityki Komisji Europejskiej na lata 2019-2024, priorytetowo traktującej „Europę dostosowaną do ery cyfrowej”. Strategia europejska dotycząca danych, opublikowana w 2020 roku, ma na celu stworzenie wspólnych europejskich przestrzeni danych w strategicznych sektorach społeczeństwa i dziedzinach o interesie publicznym.

Przestrzenie te mają być suwerennymi, godnymi zaufania i interoperacyjnymi środowiskami, w których dane mogą swobodnie przepływać wewnątrz sektorów i między nimi, z zachowaniem zasad i wartości europejskich. UE finansuje kilka inicjatyw związanych z przestrzeniami danych, takich jak Data Spaces Support Centre i Smart Open-source Middleware (SIMPL) w ramach programu Digital Europe i Horizon Europe 1.

Czternaście początkowych wspólnych europejskich przestrzeni danych to:

1. **Produkcja.** Przemysłowa przestrzeń danych, która ma wspierać konkurencyjność i wydajność przemysłu UE.
2. **Zielony Ład.** Przestrzeń danych Zielonego Ładu, która umożliwi wykorzystanie głównego potencjału danych do wsparcia działań priorytetowych Zielonego Ładu w takich kwestiach, jak zmiana klimatu, gospodarka o obiegu zamkniętym, zanieczyszczenie, różnorodność biologiczna i wylesianie.
3. **Mobilność.** Przestrzeń danych w zakresie mobilności, która ma zapewnić Europie czołową pozycję w rozwoju inteligentnego systemu transportowego.
4. **Zdrowie.** Przestrzeń danych w zakresie zdrowia, niezbędna dla postępu w zapobieganiu chorobom, ich wykrywaniu i leczeniu, a także dla świadomych i opartych na dowodach decyzji, które mają na celu ulepszenie systemów opieki zdrowotnej.
5. **Finanse.** Przestrzeń danych finansowych w celu stymulowania innowacji, przejrzystości rynku, zrównoważonych finansów, a także dostępu europejskich przedsiębiorstw do finansowania i bardziej zintegrowanego rynku.

---

<sup>23</sup> Komisja Europejska, Shaping Europe's Digital Future, [Common European Data spaces](#) [dostęp: 08.01.2024]



6. **Energia.** Przestrzeń danych dotyczących energii, zbieranych w celu promowania większej dostępności i międzysektorowego udostępniania danych w sposób zorientowany na klienta, bezpieczny i godny zaufania.
7. **Rolnictwo.** Przestrzeń danych dotyczących rolnictwa, która ma na celu poprawę wyników w zakresie zrównoważonego rozwoju i konkurencyjności sektora rolnego poprzez przetwarzanie i analizę danych.
8. **Administracje publiczne.** Przestrzeń danych dla administracji publicznych, których celem jest poprawa przejrzystości i rozliczania wydatków publicznych oraz jakości wydatków, zwalczania korupcji, zarówno na poziomie unijnym, jak i krajowym.
9. **Umiejętności.** Przestrzeń danych dotyczących umiejętności, aby lepiej dopasować umiejętności między systemami kształcenia i szkolenia a potrzebami rynku pracy.
10. **Dziedzictwo kulturowe.** Przestrzeń danych dla dziedzictwa kulturowego wspiera cyfrową transformację sektora kulturowego w Europie. Pozwoli instytucjom dziedzictwa kulturowego w całej Europie na dzielenie się zdigitalizowanymi treściami dziedzictwa kulturowego, z wysokiej jakości metadanymi.
11. **Język.** Przestrzeń danych językowych ma na celu zbudowanie niezawodnego repozytorium zasobów językowych społeczeństw europejskich.
12. **Media.** Przestrzeń danych medialnych ma umożliwić organizacjom medialnym aktywnym we wszystkich podsektorach mediów współpracę przez udostępnianie i uzyskiwanie dostępu do danych w sposób oparty na wzajemnych korzyściach.
13. **Badania naukowe i innowacje.** Przestrzeń danych z obszaru badań naukowych i innowacji, która połączy razem dane powstałe na potrzeby programów badawczo-wdrożeniowych w państwach UE.
14. **Turystyka.** Przestrzeń danych turystycznych zapewni firmom turystycznym dostęp do informacji z całej UE. Da możliwość dostosowania oferty do oczekiwań turystów, propozycji usług do nowych grup turystycznych, przewidywania dużego napływu turystów. Tym samym umożliwi bardziej efektywne planowanie zasobów i tworzenie nowych możliwości biznesowych.

# Model dojrzałości otwartych danych

---

Ocena Dojrzałości Otwartych Danych (ODM) jest przeprowadzana co roku w celu zmierzenia postępów krajów europejskich w promowaniu i ułatwianiu dostępności oraz ponownego wykorzystywania informacji sektora publicznego. Metodyka opiera się na czterech wymiarach dojrzałości otwartych danych i obejmuje:

1. poziom rozwoju krajowych polityk, które promują otwarte dane,
2. funkcje i dane udostępniane na krajowych portalach danych,
3. jakość metadanych na krajowych portalach danych oraz,
4. inicjatywy, które mają na celu monitorowanie ponownego wykorzystania i wpływu otwartych danych.

Przegląd wyników państw członkowskich UE, wyniki dla wszystkich uczestniczących krajów oraz linki do pobrania raportu, dokumentu metodologicznego i innych danych można pobrać ze specjalnej podstrony<sup>24</sup> [data.europa.eu](http://data.europa.eu). Istotne jest, że metodyka jest dobrze udokumentowana i zawiera wymiar dotyczący jakości danych.

Z punktu widzenia przeglądu polskich zasobów w otwartym dostępie, kluczowym elementem tej metodyki wydaje się ocena dojrzałości jakości danych, opisana w niej jako „Dimension 4: Open data quality”. Miara jakości otwartych danych ocenia środki przyjęte przez administratorów danych – w celu zapewnienia systematycznego i terminowego gromadzenia metadanych oraz mechanizmy monitorowania – w celu zapewnienia publikacji zgodnych ze standardem metadanych DCAT-AP (DCAT Application Profile<sup>25</sup>) i kilkoma wymogami dotyczącymi jakości wdrażania. Aby ocenić te elementy, wymiar składa się z czterech wskaźników: (1) aktualność i kompletność metadanych, (2) monitorowanie i mierzenie, (3) zgodność z DCAT-AP oraz (4) wdrażanie i dane powiązane.

Ten raport zawiera ocenę jakości i przydatności wybranych zbiorów danych w formie komentarza do każdego zbioru, która spełnia po części kryterium (1) aktualność i kompletność metadanych. w przyszłości, przy kolejnych aktualizacjach raportu, autorzy będą starali się zawrzeć szerszą ocenę jakości wyspecyfikowanych zbiorów.

---

<sup>24</sup> [Open Data Maturity in Europe](https://data.europa.eu) | EU Data Portal | [data.europa.eu](http://data.europa.eu) [dostęp: 19.12.2023]

<sup>25</sup> [DCAT-AP for data portals in Europe](https://publications.ec.europa.eu/) - EU Vocabularies - Publications Office of the EU, [dostęp: 19.12.2023]

## Metodyka opisu i klasyfikacji polskich tekstowych zasobów i zestawów danych

Autorzy raportu skupili się na stworzeniu ewidencji polskich zasobów tekstowych i ich wstępnej klasyfikacji. Kluczowym aspektem zestawienia jest uwzględnienie licencji i autorstwa danych, co ma zasadnicze znaczenie w kontekście nadchodzącego AI ACT. Ten akt prawny wymaga od twórców modeli AI udokumentowania źródeł danych oraz zapewnienia, że nie naruszają one praw autorskich. Ewidencja zawiera kilka kluczowych parametrów, takich jak:

1. **Nazwa:** identyfikator zestawu danych.
2. **URL:** link do źródła danych lub strona WWW zasobu.
3. **Autor/Opiekun:** osoba lub organizacja odpowiedzialna za zestaw danych.
4. **Licencja:** typ licencji, pod którym dane zostały udostępnione.
5. **Zastosowanie:** potencjalne użycie zestawu danych w aplikacjach czy badaniach.
6. **Rozmiar:** objętość zbioru danych w języku polskim.
7. **Format/Dostęp:** format danych (API, pliki, PDF, JSON) oraz sposób dostępu do nich.
8. **Mapowania European Data Space<sup>26</sup>:** *DataSets (Health, Industrial & Manufacturing, Agriculture, Culture, Mobility, Green Deal, Security, Public Administration, Media)*

Celem autorów raportu jest to, aby dotarł on do szerokiej gamy odbiorców – od przedsiębiorców i studentów po instytucje i ośrodki naukowe, które tworzą i udostępniają zasoby tekstowe. Promocja raportu ma na celu nie tylko edukację, ale również zachęcenie do dalszego poszerzania rejestru o nowe, wartościowe zestawy danych.

---

<sup>26</sup> [Common European Data Spaces: Challenges and Opportunities](#), Springer Nature Link [dostęp 19.11.2024]

## Komentarz eksperta



### Michał Chodorek

Adwokat, Partner, kancelaria KRK Kieszkowska  
Rutkowska Kolasiński

Specjalista w zakresie prawa Life Sciences i prawa  
nowych technologii w branżach

### Zestawy danych a AI ACT / Data ACT

AI Act nakłada na producentów systemów AI istotne wymogi dotyczące danych treningowych, testowych i walidacyjnych, w szczególności w przypadku systemów AI wysokiego ryzyka (high-risk AI). Wymogi te dotyczą również modeli wykorzystujących przetwarzanie języka naturalnego (NLP).

Zarządzanie zbiorami danych treningowych, testowych i walidacyjnych w trakcie całego cyklu życia systemu AI musi być prowadzone zgodnie z odpowiednimi procedurami, które gwarantują nie tylko odpowiednią jakość i reprezentatywność tych danych, ale również rozliczalność i dokumentowanie podjęcia poszczególnych kroków, decyzji i zmian w odniesieniu do tych zbiorów danych. Odpowiednie zarządzanie zbiorami danych musi obejmować w szczególności:

1. podejmowanie decyzji projektowych dot. tych zbiorów danych,
2. procesy gromadzenia danych,
3. przygotowywanie danych (pre-processing) – adnotowanie, etykietowanie, czyszczenie, aktualizację, czy wzbogacanie,
4. formułowanie założeń, szczególnie w odniesieniu do informacji, które dane mają mierzyć i reprezentować,
5. ocenę dostępności, ilości i przydatności potrzebnych zbiorów danych,
6. ocenę możliwych stronniczości (bias) występujących w danych oraz odpowiednie mechanizmy wykrywania, zapobiegania i ich zabezpieczenie przed atakami hakerskimi.

Wszystkie powyższe procesy uwzględniane są w systemie zarządzania jakością (QMS) i systemie zarządzania ryzykiem (RMS) systemu AI, a wyniki ich realizacji, w tym szczegółowe informacje dotyczące właściwości zbiorów danych treningowych, testowych i walidacyjnych, sposobu ich pozyskania i pre-processingu, są uwzględniane w jego dokumentacji technicznej.

Przygotowane Kompendium polskich zbiorów danych językowych powinno w istotnym stopniu ułatwić producentom systemów AI realizację powyższych obowiązków.

## Kompendium polskich zbiorów danych językowych


Nazwa	SpeakLeash
Logo	
URL	<a href="https://speakleash.streamlit.app/">https://speakleash.streamlit.app/</a>
Autor/Opiekun	Fundacja SpeakLeash
Licencja	Licencja zależna od źródła danych
Zastosowanie	<ol style="list-style-type: none"> <li>1. Trenowanie dużych modeli językowych</li> <li>2. Dostrajanie dużych modeli językowych</li> <li>3. Tworzenie dedykowanych zestawów danych (np. pytanie-odpowieź, instrukcje itp.)</li> </ol>
Rozmiar	<b>2,5 TB</b> (stan na 04.11.2024)
Format/Dostęp	<a href="#">Pakiet Python</a> , pliki w formacie ZST/JSONL/TXT
Mapowanie na wspólne europejskie przestrzenie danych	Język, Dziedzictwo kulturowe
Komentarz	Największy zbiór polskich danych tekstowych, wzorowany na The Pile. Oprócz danych tekstowych zawiera także metadane administracyjne (licencja, data stworzenia zbioru, źródło) oraz metryki jakościowe i ilościowe opisujące każdy z dokumentów w zbiorze. Całość dopełnia klasyfikacja do jednej z ponad 110 kategorii.

Nazwa	Projekt OSCAR
Logo	
URL	<a href="https://oscar-project.org">https://oscar-project.org</a>
Autor/Opiekun	Inria, ALMAAnaCH, Praire Institute, DFKI, OpenGPT-X
Licencja	Creative Commons CC0
Zastosowanie	<ol style="list-style-type: none"> <li>1. Trenowanie dużych modeli językowych</li> <li>2. Tworzenie dedykowanych zestawów danych</li> </ol>
Rozmiar	<b>147,1 GB dla j. polskiego, cały zbiór &gt; 5 TB</b> (stan na 05.11.2024)
Format/Dostęp	Za pośrednictwem portalu <a href="#">Hugging Face</a>
Mapowanie na wspólne europejskie przestrzenie danych	Język, Dziedzictwo kulturowe
Komentarz	<p>OSCAR jest korpusem wielojęzycznym, zawierającym w sobie dane tekstowe w ponad 150 językach. Zdecydowanym liderem jest język angielski (3,4 TB), zaraz za nim plasuje się język chiński. Sporo większy udział od języka polskiego mają na przykład włoski (259,4 GB), niemiecki (594,7 GB) i – co ciekawe – także węgierski (150,1 GB).</p>

Nazwa <b>CulturaX</b>	
<b>URL</b>	<a href="https://huggingface.co/datasets/uonlp/CulturaX">https://huggingface.co/datasets/uonlp/CulturaX</a>
<b>Autor/Opiekun</b>	The University of Oregon NLP Group
<b>Licencja</b>	m4C, OSCAR
<b>Zastosowanie</b>	<ol style="list-style-type: none"> <li>1. Trenowanie dużych modeli językowych (LLM)</li> <li>2. Tworzenie dedykowanych zestawów danych</li> </ol>
<b>Rozmiar</b>	<b>około 419 GB</b> (stan na 05.12.2023)
<b>Format/Dostęp</b>	Za pośrednictwem portalu Hugging Face.
<b>Komentarz</b>	To pokaźny, wielojęzyczny zbiór danych, który zawiera 6,3 biliona tokenów w 167 językach, dostosowany do rozwoju modeli LLM. Wysoką jakość danych osiągnięto przy pomocy technik i narzędzi, takich jak MinHash, deduplikacja, oczyszczanie danych czy filtrowanie szumów.

Nazwa PL Wiki	
Logo	 <b>WIKIPEDIA</b> Wolna encyklopedia
URL	<a href="https://dumps.wikimedia.org/plwiki/">https://dumps.wikimedia.org/plwiki/</a>
Autor/Opiekun	Wikimedia
Licencja	Creative Commons CC 4.0
Zastosowanie	<ol style="list-style-type: none"> <li>1. Trenowanie dużych modeli językowych</li> <li>2. Tworzenie dedykowanych zestawów danych (np. QA)</li> </ol>
Rozmiar	<b>około 2 GB</b> (stan na 05.12.2023)
Format/Dostęp	Za pośrednictwem domeny <a href="https://dumps.wikimedia.org/">https://dumps.wikimedia.org/</a>
Mapowanie na wspólne europejskie przestrzenie danych	Język
Komentarz	Polska Wikipedia to prawie 1,6 miliona unikalnych haseł, co stanowi świetną bazę wiedzy oraz doskonale źródło materiału do tworzenia zestawów danych zawierających pary pytań i odpowiedzi (QA).




Nazwa	PoEmo2.0
Logo	
URL	<a href="https://huggingface.co/datasets/clarin-pl/polemo2-official">https://huggingface.co/datasets/clarin-pl/polemo2-official</a>
Autor/Opiekun	CLARIN-PL
Licencja	CC BY-NC-SA 4.0
Zastosowanie	<ol style="list-style-type: none"> <li>1. Budowa modeli do analizy sentymentu (wydźwiku, emocji)</li> <li>2. Ewaluacja modeli i ocena jakości ich działania</li> </ol>
Rozmiar	<b>27,7 MB</b> (stan na 05.12.2023)
Format/Dostęp	Za pośrednictwem portalu Hugging Face
Mapowanie na wspólne europejskie przestrzenia danych	Język
Komentarz	<p>PoEmo 2.0 to zbiór danych internetowych recenzji konsumenckich z czterech zakresów tematycznych: medycyny, hotelarstwa, produktów i szkolnictwa. Jest on adnotowany ręcznie na poziomie pełnych recenzji i pojedynczych zdań. Aktualna wersja, czyli 2.0, obejmuje 8 216 recenzji zawierających 57 466 zdań. Każdy tekst i zdanie zostały ręcznie opatrzone adnotacjami dotyczącymi sentymentu w schemacie 2+1, co daje łącznie 197 046 adnotacji. Około 85% recenzji pochodzi z dziedziny medycyny i hotelarstwa. Każda recenzja jest opatrzona jedną z czterech etykiet: pozytywną, negatywną, neutralną lub niejednoznaczną.</p>

Nazwa	Wroclaw Corpus of Consumer Reviews Sentiment (WCCRS)
Logo	
URL	<a href="https://clarin-pl.eu/dspace/handle/11321/700">https://clarin-pl.eu/dspace/handle/11321/700</a>
Autor/Opiekun	CLARIN-PL
Licencja	CC BY-NC-SA 4.0
Zastosowanie	Budowa modeli do analizy sentymentu, emocji
Rozmiar	<b>15,2 MB</b> (stan na 05.12.2023)
Format/Dostęp	Za pośrednictwem <a href="#">domeny CLARIN-PL</a> , pliki w formacie TXT
Mapowanie na wspólne europejskie przestrzenie danych	Język
Komentarz	Wrocławski Korpus Recenzji Konsumentckich to korpus polskich recenzji z adnotacją sentymentu na poziomie całego tekstu oraz na poziomie zdań dla następujących dziedzin: hotelarstwo, medycyna, produkty i szkolnictwo. Adnotacja sentymentu dotyczy tylko dla hoteli i medycyny. Zbiór zbliżony swoją zawartością do PolEmo 2.0.


Nazwa	Narodowy Korpus Języka Polskiego (NKJP)
Logo	
URL	<a href="http://nkjp.pl/index.php?page=9&amp;lang=0">http://nkjp.pl/index.php?page=9&amp;lang=0</a>
Autor/Opiekun	IPI PAN, PWN, IJP PAN, UŁ
Licencja	Brak danych
Zastosowanie	Trenowanie dużych modeli językowych
Rozmiar	<b>2,4 GB</b> (stan na 05.12.2023)
Format/Dostęp	Za pośrednictwem <a href="#">domeny NKJP</a> bądź przez Hugging Face (opublikowane są tam zbiory danych bazujące na NKJP)
Mapowanie na wspólne europejskie przestrzenie danych	Język, Dziedzictwo kulturowe, Media,
Komentarz	<p>Lista źródeł korpusu zawiera nie tylko klasykę literatury polskiej, ale też prasę codzienną i specjalistyczną, nagrania rozmów, teksty ulotne i internetowe. Zróżnicowanie tematyczne i gatunkowe tekstów, dbałość o reprezentację rozmówców obu płci, w różnym wieku i z różnych regionów są dla wiarygodności korpusu równie ważne jak jego wielkość. Jest często punktem wyjścia dla subkorpusów specjalizowanych, takich jak: NKJP-POS, NKJP-NER czy NKJP1M dostępnych na Hugging Face.</p>

Nazwa <b>Kompleksowa Lista Ewaluacji Językowych (KLEJ)</b>	
Logo	
URL	<a href="https://klejbenchmark.com/tasks/">https://klejbenchmark.com/tasks/</a>
Autor/Opiekun	Allegro Machine Learning Research Team, AGH
Licencja	Zależna od zbioru (GNU GPL v.3, CC BY-NC-SA 4.0, CC BY-SA 3.0)
Zastosowanie	Ewaluacja modeli i ocena jakości ich działania
Rozmiar	<b>około 50 MB</b> (stan na 05.12.2023)
Format/Dostęp	Za pośrednictwem <a href="#">domeny KLEJ Benchmark</a> bądź przez Hugging Face
Mapowanie na wspólne europejskie przestrzenie danych	Język
Komentarz	KLEJ, a właściwie KLEJ Benchmark, to zestaw zadań ewaluacyjnych dla modeli języka polskiego. Jego zadaniem jest rozwój NLU, stąd też wymaga przemyślanego doboru zadań. Pozwala to na skupienie się na różnych cechach, długościach i gatunkach tekstu, co w połączeniu ze zróżnicowaną trudnością zadań pozwala na wiarygodną ocenę. Cały zestaw składa się z 9 zadań polegających między innymi na analizie sentymentu, klasyfikacji NER, wykrywaniu hejtu w social mediach, odpowiadaniu na pytania czy parafrazowaniu.

Nazwa	MAUPQA
Logo	
URL	<a href="https://huggingface.co/datasets/ipipan/maupqa">https://huggingface.co/datasets/ipipan/maupqa</a>
Autor/Opiekun	IPI PAN
Licencja	CC BY-SA 4.0
Zastosowanie	<ol style="list-style-type: none"> <li>1. Trenowanie dużych modeli językowych</li> <li>2. Dostrajanie dużych modeli językowych</li> </ol>
Rozmiar	7,21 GB (stan na 05.12.2023)
Format/Dostęp	Za pośrednictwem portalu Hugging Face
Mapowanie na wspólne europejskie przestrzenie danych	Język
Komentarz	<p>MAUPQA to kolekcja 14 zbiorów danych do wyszukiwania polskich dokumentów. Większość z nich jest generowana maszynowo lub tłumaczona maszynowo z języka angielskiego. w sumie jest to ponad 1 milion pytań, 1 milion pozytywnych i 7 milionów negatywnych (<i>hard-negative</i>) par pytanie-odpowieź. Jest to świetne źródło danych zarówno do trenowania LLM, jak i ich dostrajania.</p>

Nazwa	PoIQa
Logo	
URL	<a href="https://huggingface.co/datasets/ipipan/polqa">https://huggingface.co/datasets/ipipan/polqa</a>
Autor/Opiekun	IPI PAN
Licencja	CC BY-SA 4.0
Zastosowanie	<ol style="list-style-type: none"> <li>1. Trenowanie dużych modeli językowych</li> <li>2. Dostrajanie dużych modeli językowych</li> </ol>
Rozmiar	3,38 GB (stan na 05.12.2023)
Format/Dostęp	Za pośrednictwem portalu Hugging Face
Mapowanie na wspólne europejskie przestrzenie danych	Język
Komentarz	<p>PoIQa to pierwszy polski zbiór danych zawierających pytania otwarte i odpowiedzi na nie. Składa się on z 7 tysięcy pytań, 87 525 ręcznie oznaczonych fragmentów z odpowiedzią oraz ponad 7 milionów fragmentów, w których potencjalnie znajduje się odpowiedź. Zbiór danych może być wykorzystywany do trenowania zarówno wyodrębnienia zdań z dłuższych fragmentów tekstu, jak i tworzenia definicji pojęć abstrakcyjnych. Tekst w języku polskim pochodzi z programu telewizyjnego Jeden z Dziesięciu (pytania) i polskiej wersji Wikipedii (fragmenty).</p>

Nazwa	PoIEval
<b>Logo</b>	
<b>URL</b>	<a href="https://github.com/poleval">https://github.com/poleval</a>
<b>Autor/Opiekun</b>	IPI PAN
<b>Licencja</b>	Brak danych
<b>Zastosowanie</b>	Ewaluacja modeli i ocena jakości ich działania
<b>Rozmiar</b>	Brak danych
<b>Format/Dostęp</b>	Za pośrednictwem repozytoriów na GitHub / na stronie PoIEval
<b>Mapowanie na wspólne europejskie przestrzenie danych</b>	Język
<b>Komentarz</b>	PoIEval to konkurs wzorowany na popularnej formule Kaggle, który stawia przed uczestnikami zadania związane z przetwarzaniem języka polskiego. do tej pory odbyło się 6 edycji, począwszy od 2017 roku.

Nazwa	BAN-PL
Logo	
URL	<a href="https://github.com/ZILiAT-NASK/BAN-PL/">https://github.com/ZILiAT-NASK/BAN-PL/</a>
Autor/Opiekun	ZILiAT-NASK
Licencja	CC BY 4.0
Zastosowanie	<ol style="list-style-type: none"> <li>1. Dostrajanie dużych modeli językowych</li> <li>2. Ewaluacja modeli i ocena jakości ich działania</li> </ol>
Rozmiar	4,9 MB (stan na 05.12.2023)
Format/Dostęp	Za pośrednictwem repozytorium na GitHub
Mapowanie na wspólne europejskie przestrzenie danych	Język
Komentarz	<p>Zbiór (repozytorium) zawiera pierwszy publicznie dostępny. Zestaw treści uznanych za obraźliwe oraz szkodliwe z serwisu Wykop.pl (nazywanego również polskim Redditem). Całość składa się z 24 tysięcy zanonimizowanych komentarzy przypisanych do jednej z kategorii: „szkodliwe” bądź „neutralne”, po 12 tysięcy w każdej z kategorii.</p>




Nazwa	SCWAD-CDSCorpus
Logo	
URL	<a href="http://git.nlp.ipipan.waw.pl/Scwad/SCWAD-CDSCorpus">http://git.nlp.ipipan.waw.pl/Scwad/SCWAD-CDSCorpus</a>
Autor/Opiekun	IPI PAN
Licencja	CC BY-NC-SA 4.0
Zastosowanie	<ol style="list-style-type: none"> <li>1. Dostrajanie dużych modeli językowych</li> <li>2. Ewaluacja modeli i ocena jakości ich działania</li> </ol>
Rozmiar	1,8 MB (stan na 05.12.2023)
Format/Dostęp	Za pośrednictwem repozytorium na <a href="#">GitHubie</a>
Mapowanie na wspólne europejskie przestrzenie danych	Język
Komentarz	<p>Korpus składa się z 10 tysięcy par zdań w języku polskim, które zostały ręcznie oznaczone pod kątem powiązań semantycznych i zależności. Zbiór danych może być wykorzystany do oceny modeli kompozycyjnej semantyki dystrybucyjnej języka polskiego. Zbiór danych został zaprezentowany na konferencji ACL 2017. Szczegółowy opis zasobu znajduje się w publikacji Wróblewskiej i Krasnowskiej-Kieraś (2017).</p>

Nazwa	Wolne Lektury
Logo	
URL	<a href="https://wolnelektury.pl/">https://wolnelektury.pl/</a>
Autor/Opiekun	Fundacja Wolne Lektury
Licencja	Brak danych
Zastosowanie	<ol style="list-style-type: none"> <li>1. Trenowanie dużych modeli językowych</li> <li>2. Tworzenie zestawów danych</li> </ol>
Rozmiar	Brak danych
Format/Dostęp	<a href="#">Za pośrednictwem API</a> , (różnorodność formatów (TXT, PDF, EPUB, MOBI))
Mapowanie na wspólne europejskie przestrzenie danych	Język, Dziedzictwo kulturowe, Media
Komentarz	<p>W zbiorach Wolnych Lektur znajduje się ponad 6,5 tysięcy utworów, w tym wiele lektur szkolnych zalecanych do użytku przez Ministerstwo Edukacji i Nauki, które trafiły już do domeny publicznej. Są tam też tłumaczenia klasyki na wolnej licencji. Wszystkie utwory zamieszczone w bibliotece Wolne Lektury są udostępniane na wolnych licencjach, dzięki czemu każdy może je zgodnie z prawem bezpłatnie je przeglądać, odsłuchiwać, pobierać na swój komputer, a także udostępniać innym i cytować.</p>

Nazwa	
<b>System Analizy Orzeczeń Sądowych (SAOS)</b>	
Logo	
URL	<a href="https://www.saos.org.pl/">https://www.saos.org.pl/</a>
Autor/Opiekun	ICMMiK UW
Licencja	Brak danych
Zastosowanie	<ol style="list-style-type: none"> <li>1. Trenowanie dużych modeli językowych</li> <li>2. Tworzenie zestawów danych</li> </ol>
Rozmiar	około <b>6.5GB</b> (stan na 05.12.2023)
Format/Dostęp	<a href="#">Za pośrednictwem API</a>
Mapowanie na wspólne europejskie przestrzenie danych	Język, Administracje publiczne
Komentarz	Serwis SAOS prezentuje dane dostępne w Portalu Orzeczeń Sądów Powszechnych, serwisu orzeczeń Sądu Najwyższego, serwisu Trybunału Konstytucyjnego oraz serwisu Krajowej Izby Odwoławczej. na dzień 03.12.2023 w serwisie było obecnych ponad 480 tysięcy orzeczeń.

Nazwa	1000 Nowel
Logo	
URL	<a href="https://clarin-pl.eu/dspace/handle/11321/312">https://clarin-pl.eu/dspace/handle/11321/312</a>
Autor/Opiekun	CLARIN-PL
Licencja	CC BY 4.0
Zastosowanie	Trenowanie dużych modeli językowych
Rozmiar	<b>156,9 MB</b> (stan na 05.12.2023)
Format/Dostęp	Za pośrednictwem <a href="#">domeny CLARIN-PL</a>
Mapowanie na wspólne europejskie przestrzenie danych	Język, Dziedzictwo kulturowe
Komentarz	Korpus tekstów literackich, który może służyć jako zbiór wzorcowy do kategoryzacji tekstów. Zawiera 1000 powieści napisanych w języku polskim lub przetłumaczonych na język polski przez różnych autorów. Każdy tekst jest przechowywany jako osobny plik .txt.

<b>Nazwa</b>	<b>Repozytorium Centrum Otwartej Nauki (CEON)</b>
<b>Logo</b>	
<b>URL</b>	<a href="https://depot.ceon.pl/">https://depot.ceon.pl/</a>
<b>Autor/Opiekun</b>	ICM UW
<b>Licencja</b>	Udzielana przez ICM UW
<b>Zastosowanie</b>	<ol style="list-style-type: none"> <li>1. Trenowanie dużych modeli językowych</li> <li>2. Tworzenie zestawów danych</li> </ol>
<b>Rozmiar</b>	<b>1,15 GB</b> (stan na 05.12.2023)
<b>Format/Dostęp</b>	Za pośrednictwem <a href="#">domeny CEON</a> w formacie PDF
<b>Mapowanie na wspólne europejskie przestrzenie danych</b>	Język, Dziedzictwo kulturowe, Badania naukowe i innowacje
<b>Komentarz</b>	<p>Repozytorium Centrum Otwartej Nauki jest adresowane do całego środowiska naukowego w Polsce. w repozytorium udostępnione są materiały naukowe, takie jak: artykuły, książki, materiały konferencyjne, raporty czy rozprawy doktorskie. Dostęp do wszystkich materiałów zgromadzonych w repozytorium jest otwarty. Repozytorium jest zgodne z opracowanym przez Open Archive Initiative (OAI) protokołem pobierania metadanych, dzięki czemu dokumenty w nim udostępnione mogą być łatwo znalezione przez serwisy gromadzące informacje o zasobach naukowych w formie cyfrowej. w swoich zbiorach CEON oferuje ponad 21 tysięcy publikacji.</p>


Nazwa	BEIR-PL/msmacro-pl
Logo	 <span data-bbox="726 456 1129 504">Politechnika Wroclawska</span>
URL	<a data-bbox="564 584 1294 613" href="https://huggingface.co/datasets/clarin-knext/msmarco-pl">https://huggingface.co/datasets/clarin-knext/msmarco-pl</a>
Autor/Opiekun	CLARIN-PL / PWr
Licencja	Brak danych
Zastosowanie	<ol style="list-style-type: none"> <li data-bbox="564 824 1121 853">1. Trenowanie dużych modeli językowych</li> <li data-bbox="564 875 1118 904">2. Dostrajanie dużych modeli językowych</li> <li data-bbox="564 927 991 956">3. Tworzenie zestawów danych</li> </ol>
Rozmiar	<b>1,26 GB</b> (stan na 05.12.2023)
Format/Dostęp	Za pośrednictwem portalu <a data-bbox="911 1088 1094 1117" href="https://huggingface.co">Hugging Face</a>
Mapowanie na wspólne europejskie przestrzenie danych	Język
Komentarz	<p>Zbiór msmarco-pl jest częścią większego zbioru o nazwie BEIR-PL, służącego do porównywania modeli LLM. Ten konkretny zestaw zawiera ponad 180 tysięcy pytań i odpowiedzi wygenerowanych ręcznie w serwisie Bing, w sumie około 8,8 miliona wierszy. Oryginalnie stworzony w języku angielskim, został przetłumaczony na język polski przy pomocy Tłumacza Google.</p>

Nazwa	Polona
Logo	
URL	<a href="https://polona.pl/">https://polona.pl/</a>
Autor/Opiekun	Biblioteka Narodowa
Licencja	Brak danych
Zastosowanie	<ol style="list-style-type: none"> <li>1. Trenowanie dużych modeli językowych</li> <li>2. Tworzenie zestawów danych</li> </ol>
Rozmiar	Brak danych
Format/Dostęp	Za pośrednictwem strony, pliki w formacie PDF (zeskanowane dokumenty)
Mapowanie na wspólne europejskie przestrzenie danych	Język, Dziedzictwo kulturowe, Media, Badania naukowe i innowacje
Komentarz	Polona.pl to prawie 4 miliony publikacji dostępnych online oraz przez cyfrową wypożyczalnię Academica w 4 tysiącach bibliotek w całej Polsce. Większość materiałów przygotowana jest w formie skanów.


Nazwa		WiktorS/Polish-News
URL	<a href="https://huggingface.co/datasets/WiktorS/polish-news">https://huggingface.co/datasets/WiktorS/polish-news</a>	
Autor/Opiekun	WiktorS @ Hugging Face	
Licencja	Apache-2.0	
Zastosowanie	<ol style="list-style-type: none"> <li>1. Trenowanie dużych modeli językowych</li> <li>2. Tworzenie zestawów danych</li> <li>3. Ewaluacja modeli i ocena jakości ich działania</li> </ol>	
Rozmiar	<b>624 MB</b> (stan na 05.12.2023)	
Format/Dostęp	Za pośrednictwem portalu Hugging Face	
Mapowanie na wspólne europejskie przestrzenie danych	Język, Media	
Komentarz	<p>Zbiór danych zawierający ponad 250 tysięcy artykułów pochodzących z <a href="#">polskiego serwisu informacyjnego</a>. Jak podał autor, głównym celem zebrania tych danych było wykorzystanie ich do stworzenia modelu opartego na architekturze typu transformer, którego zadaniem było podsumowanie przetwarzanego tekstu.</p>	



Nazwa	Projekt Gutenberg
<b>Logo</b>	
<b>URL</b>	<a href="https://www.gutenberg.org/">https://www.gutenberg.org/</a>
<b>Autor/Opiekun</b>	Project Gutenberg Literary Archive Foundation (PGLAF)
<b>Licencja</b>	Small Print License (Project Gutenberg-tm License)
<b>Zastosowanie</b>	Trenowanie dużych modeli językowych
<b>Rozmiar</b>	Brak danych
<b>Format/Dostęp</b>	Za pośrednictwem <a href="#">strony projektu Gutenberg</a>
<b>Mapowanie na wspólne europejskie przestrzenie danych</b>	Język, Dziedzictwo kulturowe, Media
<b>Komentarz</b>	Projekt Gutenberg udostępnia swoje zbiory danych w postaci elektronicznych wersji książek, których prawa autorskie już wygasły. Swoje zbiory szacują na ponad 70 tysięcy pozycji, z czego mniej niż 50 to pozycje polskojęzyczne.


Nazwa		Open Subtitles-PL
Logo		
URL	<a href="https://www.opensubtitles.org/pl">https://www.opensubtitles.org/pl</a>	
Autor/Opiekun	Brano (założyciel Open Subtitles Organisation) oraz społeczność portalu	
Licencja	Brak treści objętych prawem autorskim	
Zastosowanie	<ol style="list-style-type: none"> <li>1. Trenowanie dużych modeli językowych</li> <li>2. Tworzenie zestawów danych</li> </ol>	
Rozmiar	<b>1,39 GB</b> (stan na 05.12.2023)	
Format/Dostęp	Za pośrednictwem REST API bądź bezpośrednio z <a href="#">portalu Opensubtitles</a>	
Mapowanie na wspólne europejskie przestrzenie danych	Język, Dziedzictwo kulturowe	
Komentarz	Ogólnodostępna baza napisów do filmów i seriali telewizyjnych w wielu językach, rozwijana jest przez międzynarodową społeczność OpenSubtitles. Zbiór zawiera niemal 7 milionów tekstów, z czego 332 280 to napisy polskojęzyczne.	


Nazwa	Biblioteka Nauki
Logo	 <b>Biblioteka Nauki</b>
URL	<a href="https://bibliotekanauki.pl/">https://bibliotekanauki.pl/</a>
Autor/Opiekun	Platforma Otwartej Nauki
Licencja	Zależne od zbioru, zwykle rozwinięcie CC BY
Zastosowanie	<ol style="list-style-type: none"> <li>1. Trenowanie dużych modeli językowych</li> <li>2. Dostrajanie dużych modeli językowych</li> <li>3. Tworzenie zbiorów danych</li> </ol>
Rozmiar	Brak danych
Format/Dostęp	Bezpośrednio z portalu <a href="#">Biblioteka Nauki</a> , pliki w formacie PDF
Mapowanie na wspólne europejskie przestrzenie danych	Język, Dziedzictwo kulturowe, Badania naukowe i innowacje
Komentarz	<p>Biblioteka Nauki prowadzona jest w Interdyscyplinarnym Centrum Modelowania Matematycznego i Komputerowego Uniwersytetu Warszawskiego w ramach Platformy Otwartej Nauki. Udostępnia w sposób otwarty pełne teksty artykułów publikowanych w polskich czasopismach naukowych i pełne teksty wybranych książek naukowych wraz z metadanymi z takich dziedzin jak nauki ścisłe, inżynierijno-techniczne, przyrodnicze, społeczne, humanistyczne, medyczne i o zdrowiu. w swoich zasobach posiada ponad 545 tysięcy artykułów, 1600 czasopism oraz blisko 1500 książek.</p>

<b>Nazwa</b> Korpus Dyskursu Parlamentarnego (z ang. Polish Parliamentary Corpus, PPC)	
<b>Logo</b>	
<b>URL</b>	<a href="http://clip.ipipan.waw.pl/PPC">http://clip.ipipan.waw.pl/PPC</a>
<b>Autor/Opiekun</b>	CLARIN-PL / IPI PAN
<b>Licencja</b>	CC BY-SA 4.0
<b>Zastosowanie</b>	<ol style="list-style-type: none"> <li>1. Trenowanie dużych modeli językowych</li> <li>2. Dostrajanie dużych modeli językowych</li> </ol>
<b>Rozmiar</b>	<b>1,7 GB</b> (wersja bez adnotacji), <b>34,4 GB</b> (wersja z komentarzami językowymi) (stan na 05.12.2023)
<b>Format/Dostęp</b>	Bezpośrednio z portalu IPI PAN
<b>Mapowanie na wspólne europejskie przestrzenie danych</b>	Język, Dziedzictwo kulturowe, Administracje publiczne
<b>Komentarz</b>	<p>Korpus Parlamentarny Rzeczypospolitej Polskiej (PPC) to duży zbiór dokumentów z prac polskiego Sejmu i Senatu poddanych analizie językowej. Opiera się na polskim korpusie sejmowym, którego zebranie było współfinansowane w ramach projektu CESAR, później rozbudowywanym przy wsparciu projektów CLARIN-PL, MARCELL oraz ParlaMint. Zawiera ponad 800 milionów segmentów.</p>

Nazwa Summarization – Polish Summaries Corpus	
Logo	
URL	<a href="https://huggingface.co/datasets/allegro/summarization-polish-summaries-corpus">https://huggingface.co/datasets/allegro/summarization-polish-summaries-corpus</a>
Autor/Opiekun	Allegro
Licencja	Brak danych
Zastosowanie	<ol style="list-style-type: none"> <li>1. Dostrajanie dużych modeli językowych</li> <li>2. Ewaluacja modeli i ocena jakości ich działania</li> </ol>
Rozmiar	<b>271 MB</b> (stan na 05.12.2023)
Format/Dostęp	Za pośrednictwem portalu Hugging Face
Mapowanie na wspólne europejskie przestrzenie danych	Język
Komentarz	Zbiór danych udostępniony przez Allegro w celu uczenia i dostrajania modeli językowych, które specjalizują się w tworzeniu podsumowań tekstu. Zawiera ponad 21 tysięcy par danych typu źródło-podsumowanie podzielonych na zestaw treningowy, testowy oraz walidacyjny.


Nazwa	Polish Question-Passage Pairs
Logo	
URL	<a href="https://huggingface.co/datasets/allegro/polish-question-passage-pairs">https://huggingface.co/datasets/allegro/polish-question-passage-pairs</a>
Autor/Opiekun	Allegro
Licencja	Brak danych
Zastosowanie	<ol style="list-style-type: none"> <li>1. Dostrajanie dużych modeli językowych</li> <li>2. Ewaluacja modeli i ocena jakości ich działania</li> </ol>
Rozmiar	<b>3,61 MB</b>
Format/Dostęp	Za pośrednictwem portalu Hugging Face
Mapowanie na wspólne europejskie przestrzenie danych	Język
Komentarz	<p>Zbiór udostępniony przez Allegro do celów uczenia i dostrajania modeli językowych specjalizujących się w odpowiadaniu na pytania użytkownika. Zawiera około 10,5 tysiąca zestawów pytanie-odpowiedź prawda/fałsz.</p>


Nazwa	EUR-Lex
Logo	
URL	<a href="https://eur-lex.europa.eu/homepage.html?locale=pl">https://eur-lex.europa.eu/homepage.html?locale=pl</a>
Autor/Opiekun	Unia Europejska, Urząd Publikacji Unii Europejskiej
Licencja	CC BY 4.0
Zastosowanie	<ol style="list-style-type: none"> <li>1. Trenowanie dużych modeli językowych</li> <li>2. Tworzenie zestawów danych</li> </ol>
Rozmiar	Brak danych
Format/Dostęp	Za pośrednictwem REST API, SPARQL, SOAP bądź bezpośrednio ze strony internetowej
Mapowanie na wspólne europejskie przestrzenie danych	Język, Administracje publiczne
Komentarz	Baza aktów prawnych Unii Europejskiej z tłumaczeniami we wszystkich językach państw członkowskich Unii. Baza zawiera ponad 8 milionów dokumentów.

Nazwa		EUREKA – System Informacji Celno-Skarbowej
Logo		
URL	<a href="https://eureka.mf.gov.pl/">https://eureka.mf.gov.pl/</a>	
Autor/Opiekun	System Informacji Celno-Skarbowej Eureka	
Licencja	Brak danych	
Zastosowanie	<ol style="list-style-type: none"> <li>1. Trenowanie dużych modeli językowych</li> <li>2. Tworzenie zestawów danych</li> </ol>	
Rozmiar	około <b>9,7 GB</b> (stan na 05.12.2023)	
Format/Dostęp	Dokumenty w formacie HTML, możliwy eksport do innych formatów	
Mapowanie na wspólne europejskie przestrzenie danych	Język, Administracje publiczne	
Komentarz	<p>System Informacji Celno-Skarbowej EUREKA to zbiór informacji podatkowych i celnych, który zawiera między innymi interpretacje indywidualne i ogólne, wiążące informacje akcyzowe na temat ich stawek, informacje o wydaniu opinii zabezpieczających lub odmowie ich wydania, odpowiedzi na interpelacje oraz wybrane orzeczenia sądów. Zbiór ten zawiera w sobie ponad 450 tysięcy dokumentów.</p>	





Nazwa	ISAP – Internetowy System Aktów Prawnych
Logo	
URL	<a href="https://isap.sejm.gov.pl/">https://isap.sejm.gov.pl/</a>
Autor/Opiekun	Internetowy System Aktów Prawnych / Ośrodek Informatyki Kancelarii Sejmu
Licencja	Brak danych
Zastosowanie	<ol style="list-style-type: none"> <li>1. Trenowanie dużych modeli językowych</li> <li>2. Dostrajanie dużych modeli językowych</li> <li>3. Tworzenie zestawów danych</li> </ol>
Rozmiar	około <b>1,7 GB</b> (stan na 05.12.2023)
Format/Dostęp	Pliki w formacie PDF
Mapowanie na wspólne europejskie przestrzenie danych	Język, Administracje publiczne
Komentarz	<p>Baza ISAP zawiera opisy bibliograficzne i teksty aktów prawnych opublikowanych na podstawie ustawy z dnia 20 lipca 2000 roku o ogłaszaniu aktów normatywnych i niektórych innych aktów prawnych (Dz.U. 2019 poz. 1461) w wydawnictwach urzędowych: Dzienniku Ustaw i Monitorze Polskim, wydawanych przez Rządowe Centrum Legislacji. Wielkość zbioru szacowana jest na ponad 83 tysiące dokumentów.</p>

Nazwa		RPL – Rejestr Produktów Leczniczych
Logo		
URL	<a href="https://rejstry.ezdrowie.gov.pl/rpl">https://rejstry.ezdrowie.gov.pl/rpl</a>	
Autor/Opiekun	Ministerstwo Zdrowia / Centrum e-Zdrowia	
Licencja	Brak danych	
Zastosowanie	<ol style="list-style-type: none"> <li>1. Trenowanie dużych modeli językowych</li> <li>2. Dostrajanie dużych modeli językowych</li> <li>3. Tworzenie zestawów danych</li> </ol>	
Rozmiar	Brak danych	
Format/Dostęp	Dostęp za pośrednictwem <a href="#">domeny Rejestr produktów leczniczych</a> , pliki w formacie PDF (warstwa tekstowa obecna, rzadziej skany)	
Mapowanie na wspólne europejskie przestrzenie danych	Język, Zdrowie	
Komentarz	<p>Rejestr produktów leczniczych zawiera informacje o wszystkich produktach leczniczych, które zostały przeznaczone dla ludzi i zwierząt oraz zostały dopuszczone do obrotu na terenie Polski. Oprócz typowych druków informacyjnych, w tej bazie znajdują się także kompletne zestawienia środków leczniczych, wzory opakowań czy charakterystyki. Jednakże najciekawszymi pod kątem pracy z dużymi modelami językowymi są ulotki, decyzje do pozwolenia o dopuszczeniu czy też streszczenia RMP (plan zarządzania ryzykiem dla produktu leczniczego). Cała baza zawiera niemalże 22 tysiące rekordów (stan na 05.12.2023).</p>	

<b>Nazwa</b>	
<b>OWCA – Optimized and Well-Translated Customization of Alpaca</b>	
<b>Logo</b>	
<b>URL</b>	<a href="https://huggingface.co/datasets/emplocity/owca">https://huggingface.co/datasets/emplocity/owca</a>
<b>Autor/Opiekun</b>	Emplocity
<b>Licencja</b>	Brak danych
<b>Zastosowanie</b>	Dostrajanie dużych modeli językowych
<b>Rozmiar</b>	<b>22.2 MB</b>
<b>Format/Dostęp</b>	Za pośrednictwem portalu Hugging Face bądź repozytorium GitHub
<b>Mapowanie na wspólne europejskie przestrzenie danych</b>	Język
<b>Komentarz</b>	<p>Dataset OWCA powstał poprzez przetłumaczenie instrukcji użytych do dostrajania modelu Alpaca stworzonego przez Uniwersytet Stanforda. Tłumaczenie odbywało się maszynowo i generowane było z kilku źródeł. Efekty były następnie sprawdzane przez zawodowych tłumaczy oraz ekspertów od prac nad przetwarzaniem ludzkiego języka tak, aby zapewnić jak najwyższą jakość finalnego zestawu. Celem, który przyświecał twórcom, było umożliwienie dostrajania dużych modeli językowych rodzimym naukowcom i data scientistom, a w szczególności tym którzy planują pracować na modelu Alpaca. Nie zmienia to faktu, że dataset ten nadaje się znakomicie do dostrajania modeli bazujących na innych rozwiązaniach (np. LLaMa, GPT, Mistral). Całość składa się z ponad 51 tysięcy instrukcji.</p>

## Nazwa PoQuAD – Polish Question Answering Dataset

<b>Logo</b>	
<b>URL</b>	<a href="https://huggingface.co/datasets/clarin-pl/poquad">https://huggingface.co/datasets/clarin-pl/poquad</a>
<b>Autor/Opiekun</b>	CLARIN-PL
<b>Licencja</b>	CC BY 4.0
<b>Zastosowanie</b>	Dostrajanie dużych modeli językowych
<b>Rozmiar</b>	<b>53.5 MB</b>
<b>Format/Dostęp</b>	Za pośrednictwem portalu Hugging Face
<b>Mapowanie na wspólne europejskie przestrzenie danych</b>	Język
<b>Komentarz</b>	PoQuAD jest zbiorem danych (a konkretnie – zbiorem instrukcji) stworzonym na bazie Wikipedii (głównie mowa o sekcji „Dobry Artykuł” czy najczęściej odwiedzanych artykułach według miesięcznych rankingów, tworzonych od 2016 roku). Pozwoliło to na zebranie ponad 11 tysięcy artykułów, które zostały przekazane anotatorom. Działania te pozwoliły na stworzenie datasetu zawierającego blisko 52 tysiące instrukcji, które składają się z tytułu artykułu, jego treści oraz pytań i odpowiedzi.

Nazwa Polish GEC – Gramatical Error Correction	
Logo	
URL	<a href="https://github.com/Ermlab/polish-gec-datasets">https://github.com/Ermlab/polish-gec-datasets</a>
Autor/Opiekun	Ermlab (w ramach projektu Goodwrite)
Licencja	Apache 2.0
Zastosowanie	Dostrajanie dużych modeli językowych
Rozmiar	<b>39.8 MB</b>
Format/Dostęp	Za pośrednictwem portalu GitHub
Mapowanie na wspólne europejskie przestrzenie danych	Język
Komentarz	Jeden z najświeższych datasetów w tym zestawieniu – GEC – stworzony w ramach <a href="#">projektu Goodwrite</a> , czyli polskiej odpowiedzi na popularną w anglojęzycznym świecie aplikację Grammarly. Zestaw ten zawiera ponad 61 tysięcy przykładów par poprawnych i niepoprawnych, każda z adnotacją popełnionego błędu. w tworzenie tego datasetu zaangażowani zostali eksperci-lingwiści, odpowiedzialni za dwa zestawy. Pozostałe zostały adnotowane przez zwykłych użytkowników języka polskiego, co pozwala na zwiększenie różnorodności i wariacji całości.

## Komentarz eksperta



### Aleksander Obuchowski

Co-founder, CTO

TheLion.AI

## Na jakich danych budowane były polskie modele językowe?

W świecie polskich LLM-ów możemy zaobserwować trzy główne podejścia do pozyskiwania danych treningowych .

Najbardziej wartościowym, choć jednocześnie najtrudniejszym do pozyskania źródłem są dedykowane polskie zbiory danych. na szczególną uwagę zasługuje tu projekt Speakleash, który stał się fundamentem dla kilku znaczących modeli. Bielik-11B wykorzystał z niego ponad 20 milionów instrukcji, podczas gdy APT3-1B-Base oparł się na imponującej kolekcji 285GB danych, z czego 93% pochodziło właśnie ze Speakleash. Innym przykładem dedykowanego zbioru jest Clarin, wykorzystany w treningu modelu polish-gpt2. Te specjalistyczne zbiory danych nie tylko uczą modele języka polskiego, ale również przekazują im istotny kontekst kulturowy – element często pomijany w innych podejściach.

Druga ścieżka to wykorzystanie polskich fragmentów z dużych, wielojęzycznych zbiorów danych. Common Crawl, mimo że zawiera ogromne ilości tekstu, oferuje jedynie około 1.7% polskich treści. Co ciekawe, niektóre projekty, jak seria modeli APT3, świadomie rezygnują z tego źródła, preferując jakość nad ilość. Polska Wikipedia, choć niewielka w porównaniu z innymi źródłami, stanowi cenne uzupełnienie danych treningowych. w modelu APT3-275M-Base stanowi 4% danych, a w APT3-1B-Base - 1%. Warto też wspomnieć o zbiorze OSCAR, który posłużył do treningu modeli takich jak papuGaPT2.

Trzecie podejście polega na automatycznym tłumaczeniu anglojęzycznych zbiorów danych. Przykładowo model polka-1.1b-chat został wytrenowany na około 60,000 przetłumaczonych konwersacji. Trurl również korzysta ze zbiorów takich, jak ShareGPT, Dolly 15k czy Oasst1h.

W miarę rozwoju dziedziny obserwujemy coraz bardziej wyrafinowane podejście do doboru danych treningowych. Większe modele często łączą różne źródła, starając się wykorzystać zalety każdego z nich. Doskonałym przykładem jest APT3, który oprócz danych ze Speakleash (93%) wykorzystuje także dane z własnego crawlingu (6%) i Wikipedii (1%).

Coraz więcej modeli korzysta też z syntetycznie wygenerowanych instrukcji, np. Bielik-11B, który łączy ręcznie weryfikowane instrukcje z syntetycznymi, wygenerowanymi przez Mixtral 8x22B.

## Szczegółowa analiza wybranych polskich zasobów i zbiorów danych

Ta sekcja zawiera bardziej szczegółowe opisy trzech wybranych zbiorów danych, z których każdy jest z innego względu wartościowy dla prac nad polskim LLM-em. Pierwszy z nich – **Spichlerz** – to największy zbiór polskich tekstów, które pochodzą z rozmaitych źródeł. Ze względu na swoją różnorodność może być przydatny do tworzenia modeli bazowych. **BAN-PL** jest istotny ze względów społecznych, z uwagi na potencjalne zastosowania: może posłużyć do wykrywania mowy nienawiści i innego rodzaju niebezpiecznych treści w sieci. **PoIQa** z kolei to zbiór instrukcji (pytań i odpowiedzi), który może być dobrym źródłem danych treningowych dla modeli konwersacyjnych lub wyszukiwarek informacji.

### Spichlerz

---

Projekt Spichlerz (znany także jako SpeakLeash) powstał z inicjatywy Sebastiana Kondrackiego oraz Michała Dulemby. Wzorowany na głośnych i odnoszących sukcesy w dziedzinie NLP projektach The Pile oraz OSCAR, Spichlerz zgromadził repozytorium zawierające ponad 2,5 TB ewidencjonowanych polskich danych tekstowych. Dane te dostępne są na różnych licencjach, a do każdego zestawu dołączony jest specjalny plik (manifest), który określa jakość danych i licencję użytkowania.

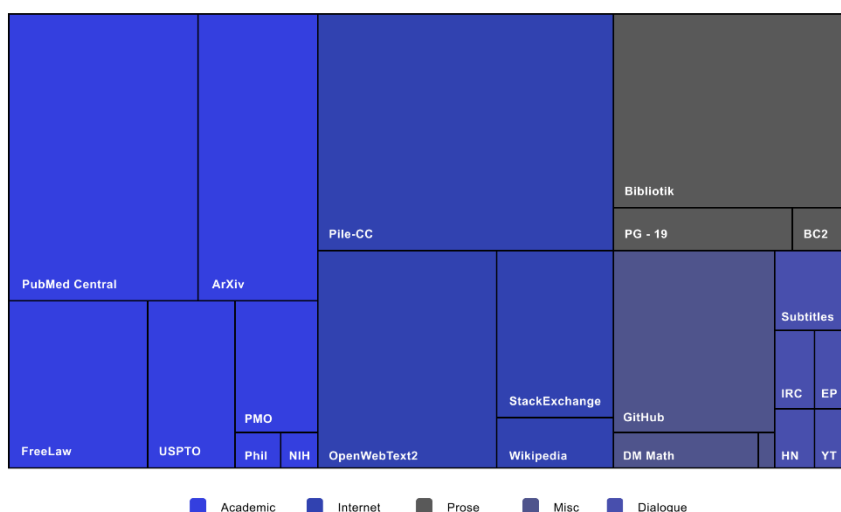
W kwietniu tego roku Spichlerz, we współpracy z Cyfronetem, wytrenował model Bielik 1.0 o 7 miliardach parametrów (niestety niedostępny do użytku komercyjnego). We wrześniu 2024 powstał model Bielik 2.0, trenowany na wybranych zbiorach tekstowych, o 11 miliardach parametrów. Bielik v2, udostępniony na licencji Apache 2.0 (czyli do użytku komercyjnego), jest uznawany za najmocniejszy model do zadań w języku polskim wśród modeli o podobnej liczbie parametrów.

Na dzień 04.11.2024 zasoby 2,5 TB Spichlerza przewyższyły pod względem objętości wspomniany projekt The Pile (825 GB).

Porównując wspomniany już dwukrotnie projekt The Pile ze Spichlerzem, możemy zaobserwować sporo ciekawych zjawisk. Wolumen obydwu zbiorów jest podobny, jednakże warto zwrócić uwagę na rozkład i proporcje danych z różnych źródeł. [W swojej pracy pt. „The](#)

[Pile: An 800GB Dataset of Diverse Text For Language Modeling](#),<sup>27</sup> twórcy z EleutherAI prezentują bardzo czytelny diagram (wykres nr 3), opisujący zebrane podzestawy danych w zależności od źródła ich pochodzenia, wielkości podzestawu czy kategorii, do której go przypisano. Dzięki temu na poniższym wykresie obserwujemy podobną objętość koloru niebieskiego (z ang. *Academic*, dane ze źródeł akademickich) oraz zielonego (ang. *Internet*, dane ze źródeł internetowych). Następne w kolejności, według ilości danych w kategorii, są podzestawy oznaczone kolorem pomarańczowym (z ang. *Prose*, dane ze źródeł literackich), szarym (z ang. *Misc*, dane z różnych źródeł, niepasujące do innych kategorii) oraz żółtym (z ang. *Dialogue*, dane bazujące na dialogach oraz czatach).

Wartość globalnego rynku przetwarzania języka naturalnego



**Wykres nr 3.** Kompozycja zbioru danych The Pile według kategorii.

Źródło: [„The Pile: An 800GB Dataset of Diverse Text For Language Modeling”](#) [dostęp 19.11.2024]

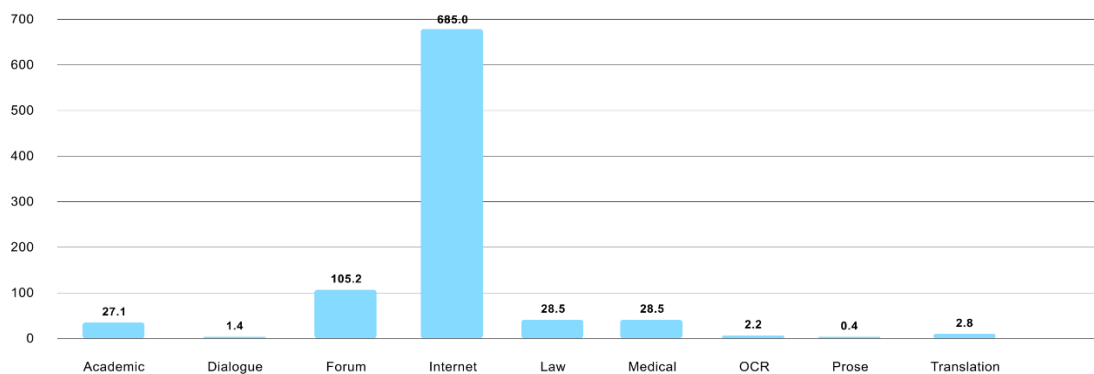
Twórcy projektu Spichlerz na swojej stronie internetowej [w sekcji „Dashboard”](#) umożliwiają podgląd postępów w pracach nad projektem w czasie rzeczywistym. Można tam znaleźć między innymi wykres obrazujący rozkład źródeł danych oraz ich objętość w podobny sposób co w przypadku EleutherAI (wykres nr 4) – różnice są widoczne już na pierwszy rzut oka. Dane z kategorii Internet (kolor zielony na wykresie nr 3) wyraźnie stanowią większość. Drugie z kolei są fora internetowe, które można sklasyfikować jako dane z kategorii Internet bądź dane dialogowe (czyli kolor żółty powyżej) według zaproponowanego przez EleutherAI klucza. Następne w kolejności są dane prawne i prace akademickie, na tym diagramie

<sup>27</sup> [dostęp 19.11.2024]



reprezentowane w polu w kolorze niebieskim. Pozostałe kategorie, mniej liczne, to dane dialogowe, medyczne, OCR (przetwarzanie optyczne zeskanowanych dokumentów w celu wydobywania tekstu), teksty literackie czy też tłumaczenia.

**Kompozycja zbioru danych Spichlerz według kategorii**  
(Stan na 12.12.2023).



**Wykres nr 4.** Kompozycja zbioru danych Spichlerz według kategorii (stan na 05.11.2024).

Źródło: [SpeakLeash / 'spix.lęs/ a.k.a. Spichlerz Dashboard](#) [dostęp 19.11.2024]

Twórcy The Pile w swojej pracy porównują również wszystkie posiadane zbiory w formie tabelarycznej, gdzie znaleźć można informacje o nazwie, wielkości czy udziale zbioru danych w tworzonym przez nich modelu. Obrazuje to ilustracja nr 5: zdecydowanie największym z uwzględnionych zestawów jest Pile-CC, czyli dane tekstowe zebrane w ramach Common Crawl, z których wyciągnięto tekst za pomocą biblioteki jusText. Następny, co do wielkości, jest zestaw o nazwie Books3, czyli [dane literackie udostępnione przez Shawna Pressera w 2020 roku](#)<sup>28</sup>. Składa się on z mieszanki dzieł literackich, beletrystycznych i niebeletrystycznych. Trzeci z kolei duży zestaw to Github. Zawiera on kod źródłowy z repozytoriów zebranych na danym portalu oraz oznaczonych przez autorów jako otwarte. Autorzy The Pile, zmotywowani i zachęceni wynikami osiąganymi przez model GPT-3<sup>29</sup> autorstwa OpenAI, uwzględnili ten zestaw danych po to, aby model finalny dawał lepsze wyniki w zadaniach z kodem źródłowym.

<sup>28</sup> [Wpis na platformie X](#) [za:] [The Pile: An 800GB Dataset of Diverse Text for Language Modeling](#)

<sup>29</sup> T. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Nee- lakantan, P. Shyam, G. Sastry, A. Askell et al., (2020), Language models are few-shot learners. arXiv preprint arXiv:2005.14165 [za:] [The Pile: An 800GB Dataset of Diverse Text for Language Modeling](#)

Wymienione powyżej zestawy pochodzą odpowiednio z kategorii Internet, Prose oraz Misc. Oznacza to, że każdy z trzech największych podzbiorów jest reprezentantem innej kategorii, co pozwala stwierdzić, że zbiór ten może być traktowany jako zróżnicowany i wielotematyczny.

Podzbiór	Rozmiar pierwotny	Waga	Liczba przejść (epok)	Rozmiar końcowy	Średni rozmiar dokumentu
<b>Pile-CC</b>	227.12 GiB	18.11%	1.0	227.12 GiB	4.33 KiB
<b>PubMed Central</b>	90.27 GiB	14.40%	2.0	180.55 GiB	30.55 KiB
<b>Books3*</b>	100.96 GiB	12.07%	1.5	151.44 GiB	538.36 KiB
<b>OpenWebText2</b>	62.77 GiB	10.01%	2.0	125.54 GiB	3.85 KiB
<b>ArXiv</b>	56.21 GiB	8.96%	2.0	112.42 GiB	46.61 KiB
<b>Github</b>	95.16 GiB	7.59%	1.0	95.16 GiB	5.25 KiB
<b>FreeLaw</b>	51.15 GiB	6.12%	1.5	76.73 GiB	15.06 KiB
<b>Stack Exchange</b>	32.20 GiB	5.13%	2.0	64.39 GiB	2.16 KiB
<b>USPTO Backgrounds</b>	22.90 GiB	3.65%	2.0	45.81 GiB	4.08 KiB
<b>PubMed Abstracts</b>	19.26 GiB	3.07%	2.0	38.53 GiB	1.30 KiB
<b>Gutenberg (PG-19)*</b>	10.88 GiB	2.17%	2.5	27.19 GiB	398.73 KiB
<b>OpenSubtitles*</b>	12.98 GiB	1.55%	1.5	19.47 GiB	30.48 KiB
<b>Wikipedia (en)*</b>	6.38 GiB	1.53%	3.0	19.13 GiB	1.11 KiB
<b>DM Mathematics*</b>	7.75 GiB	1.28%	2.0	15.49 GiB	8.00 KiB
<b>Ubuntu IRC</b>	5.52 GiB	0.88%	2.0	11.03 GiB	545.48 KiB
<b>BookCorpus2</b>	6.30 GiB	0.75%	1.5	9.45 GiB	369.87 KiB
<b>EuroParl*</b>	4.59 GiB	0.73%	2.0	9.17 GiB	68.87 KiB
<b>HackerNews</b>	3.90 GiB	0.62%	2.0	7.80 GiB	4.92 KiB
<b>YoutubeSubtitles</b>	3.73 GiB	0.60%	2.0	7.47 GiB	22.55 KiB
<b>PhilPapers</b>	2.38 GiB	0.38%	2.0	4.76 GiB	73.37 KiB
<b>NIH ExPorter</b>	1.89 GiB	0.30%	2.0	3.79 GiB	2.11 KiB
<b>Enron Emails*</b>	0.88 GiB	0.14%	2.0	1.76 GiB	1.78 KiB
<b>The Pile</b>	<b>825.18 GiB</b>			<b>1254.20 GiB</b>	<b>5.91 KiB</b>

Tabela nr 1. Zestawienie datasetów zebranych w ramach The Pile.

Źródło: [The Pile: An 800GB Dataset of Diverse Text for Language Modeling](#)

Zbiory danych oznaczone symbolem \* są używane z minimalnym przetworzeniem wstępnym. Dla porównania przyjrzyjmy się podobnej formie opisu zebranych danych w projekcie Spichlerz. Tabela nr 1 (pochodząca z dashboardu) prezentuje 10 największych zestawów danych na dzień 05.11.2024. Największy – *web\_artykuly\_inne\_raw* obejmuje treści stron WWW, jednakże oprócz surowych danych tekstowych są tam także nagłówki czy stopki. Drugi w kolejności zestaw to *forum\_kafeteria\_pl\_corpus*, czyli dyskusje użytkowników i użytkowników [forum Kafeteria.pl](https://forum.kafeteria.pl). Jest to największe forum, jakie do tej pory zasiliło zbiory Spichlerza. Trzecim, zamykającym czołówkę zestawem są dane pobrane z Centralnej Bazy Orzeczeń Sądów Administracyjnych. Warto też zwrócić uwagę na korpus *europiana\_eu\_pl\_corpus*. To zbiór zawierający dokumenty zagregowane w ramach [portalu europiana](https://portalu.europiana) – te materiały mają dużą wartość kulturową i dydaktyczną.

Podobnie jak w The Pile na te zestawy danych w Spichlerzu składają się teksty z odmiennych źródeł i są to odpowiednio: dane internetowe, dyskusje z forum, dane administracyjne oraz akademickie. To pozwala uznać ten zbiór za zróżnicowany.

☰ ✓	Dataset	↓ Size [MB]	Category	Tags [%]	Documents	Characters	Avg Docs Length	High Quality Docs	Update Date	Words
<input type="checkbox"/>	web_artykuly_inne_raw	36353	Internet	Uroda - 15	3,590,362	38,118,616,609	1516	10.00%	2023-11-20	5,443,688,663
<input type="checkbox"/>	forum_kafeteria_pl_corpus	19943	Forum	Zdrowie -	4,678,495	20,911,799,643	715	34.00%	2023-07-08	3,345,470,149
<input type="checkbox"/>	CBOSA_corpus	18395	Law	Różne - 10	1,610,542	19,288,978,721	1681	4.00%	2022-01-01	2,708,293,349
<input type="checkbox"/>	europiana_eu_pl_corpus	16747	Academic	Różne - 10	203,547	17,561,005,818	13653	8.00%	2023-03-23	2,779,202,638
<input type="checkbox"/>	forum_gazeta_pl_corpus	15609	Forum	Zdrowie -	2,500,970	16,367,411,171	1033	34.00%	2023-04-21	2,583,921,274
<input type="checkbox"/>	forum_wizaz_pl_corpus	15604	Forum	Uroda - 16	1,994,102	16,362,070,273	1318	58.00%	2023-04-14	2,628,423,511
<input type="checkbox"/>	web_tekst_inne	14864	Internet	Różne - 10	623,888	15,586,319,207	3352	9.00%	2023-07-07	2,091,623,140
<input type="checkbox"/>	web_artykuly_inne_147	10891	Internet	Kulinaria -	1,000,000	11,419,610,361	1666	67.00%	2023-06-27	1,666,933,715
<input type="checkbox"/>	shopping_1_general_corpus	10545	Internet	Elektronik	2,105,419	11,057,570,300	737	78.00%	2023-05-03	1,551,865,826
<input type="checkbox"/>	biblioteka_nauki_pl_corpus	9165	Academic	Różne - 10	265,089	9,610,178,304	5916	7.00%	2023-01-27	1,568,420,534

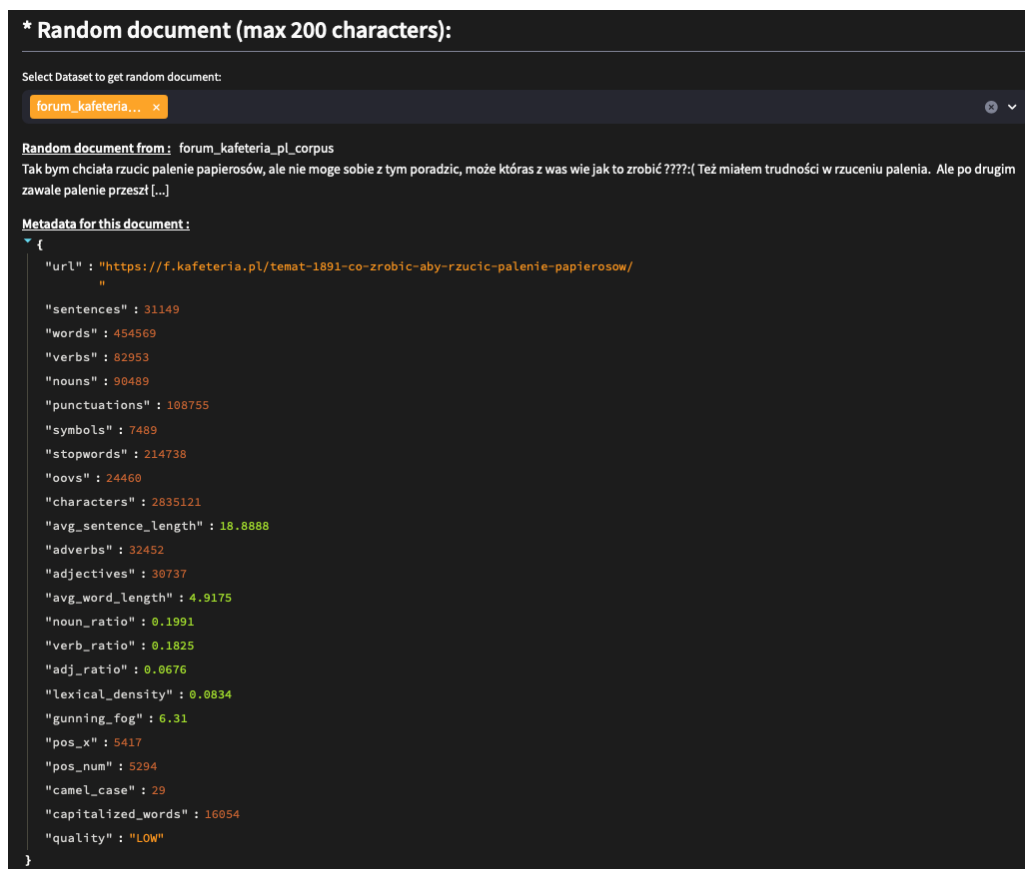
**Ilustracja nr 7.** Zestawienie 10 największych zestawów danych w projekcie Spichlerz (stan na 05.11.2024)

Źródło: [SpeakLeash / 'spix.lsg/ a.k.a. Spichlerz Dashboard](https://speakleash.github.io/spix.lsg/) [dostęp 19.11.2024]

Kolejną kluczową kwestią, pod względem której można porównać omawiane zbiory, jest zawartość metadanych, które w sposób wymierny wzbogacają surowe dane tekstowe. Jest to szczególnie ważne wobec wejścia w życie Rozporządzenia Parlamentu Europejskiego i Rady (UE) 2024/1689 z dnia 13 czerwca 2024, znanego szerszej publiczności pod nazwą [AI Act](#). Jak zostało wspomniane wcześniej, AI Act regulować będzie między innymi zarządzanie danymi i wykorzystywanie ich do budowy i trenowania modelu sztucznej inteligencji, z naciskiem na transparentność całego procesu gromadzenia, opisywania, adnotacji oraz źródła pochodzenia. The Pile już spełnia część tych warunków. w repozytorium na [platformie Hugging Face](#) oraz [w artykule stanowiącym techniczny opis](#)

[zbioru danych](#) autorzy podają źródła pochodzenia danych (o ile nazwa konkretnego podzbioru o tym nie mówi wystarczająco jasno) oraz konkretne zestawy metadanych dostępne dla podzbiorów. Przywołując ponownie trzy największe podzbiory w The Pile, tj. Pile-CC, Books3 oraz Github, dostępnymi dla nich metadanymi są:

1. **Pile-CC**: metadane dostępne w Common Crawl,
2. **Books3**: brak metadanych,
3. **Github**: nazwa repozytorium, liczba gwiazdek, język repozytorium, nazwy plików, formaty plików w repozytorium.



```
* Random document (max 200 characters):

Select Dataset to get random document:
forum_kafeteria... x

Random document from: forum_kafeteria_pl_corpus
Tak bym chciała rzucić palenie papierosów, ale nie mogę sobie z tym poradzić, może któras z was wie jak to zrobić????:( Też miałem trudności w rzuceniu palenia. Ale po drugim zawale palenie przeszł [...]

Metadata for this document:
{
  "url": "https://f.kafeteria.pl/temat-1891-co-zrobic-aby-rzucic-palenie-papierosow/"
  "sentences": 31149
  "words": 454569
  "verbs": 82953
  "nouns": 90489
  "punctuations": 108755
  "symbols": 7489
  "stopwords": 214738
  "oovs": 24460
  "characters": 2835121
  "avg_sentence_length": 18.8888
  "adverbs": 32452
  "adjectives": 30737
  "avg_word_length": 4.9175
  "noun_ratio": 0.1991
  "verb_ratio": 0.1825
  "adj_ratio": 0.0676
  "lexical_density": 0.0834
  "gunning_fog": 6.31
  "pos_x": 5417
  "pos_num": 5294
  "camel_case": 29
  "capitalized_words": 16054
  "quality": "LOW"
}
```

**Ilustracja nr 8.** Podgląd struktury i typu metadanych dla dokumentów w projekcie Spichlerz  
Źródło: [SpeakLeash / 'spix.lęs/ a.k.a. Spichlerz Dashboard](#) [dostęp 19.11.2024]

Autorzy projektu Spichlerz podeszli systematycznie i spójnie do zbierania metadanych dla swoich zestawów, tj. podają te same metryki i zbiory wartości dla każdego posiadanego podzbioru. Plik metadanych składa się z informacji ściśle powiązanych ze źródłem pochodzenia danego dokumentu (najczęściej jest to adres URL) oraz danych statystyczno-lingwistycznych, które są obliczane indywidualnie dla każdego pliku oraz zestawu danych. Widać to na ilustracji nr 7: przedstawia ona przykładowy zestaw metadanych z podzestawu *forum\_kafeteria\_pl\_corpus* dla losowego dokumentu – dostępny do podglądu z poziomu

dashboardu projektu. Oprócz wspomnianego adresu URL, metadane zawierają kilkanaście parametrów opisujących cechy językowe tekstu, takie jak liczba poszczególnych części mowy, znaków interpunkcyjnych czy symboli, współczynniki poszczególnych [części mowy](#) czy [uniwersalnie stosowane w lingwistyce miary – gęstość leksykalna](#) oraz [parametr \(Gunning\) FOG](#). Całość zamykają ocena jakości dokumentu w trzystopniowej skali (jakość niska, średnia bądź wysoka, w języku angielskim odpowiednio *low*, *medium*, *high*) obliczana na podstawie opisanych powyżej metryk oraz przypisanie do jednej z prawie 120 kategorii tematycznych (na przykład „Sport”, „Elektronika”, „Religia”, „Finanse” czy „Motoryzacja”). Wartości graniczne, opisujące przynależność do poszczególnych kategorii oraz opisujące jakość danego dokumentu, zostały wyznaczone heurystycznie na podstawie dostępnej wiedzy oraz doświadczenia uczestników projektu.

Na koniec przyjrzyjmy się liczbie podzbiorów dostępnych w każdym z opisywanych datasetów. dla The Pile są to 22 indywidualne zestawy danych, gdzie trzy największe, które zostały opisane i przywołane powyżej, odpowiadają za niemal połowę wolumenu całego datasetu.

Projekt Spichlerz składa się z większej liczby mniejszych zestawów danych (488 indywidualnych paczek, stan na dzień 05.11.2024).

Podsumowując, zbiory The Pile oraz Spichlerz łączy niemalże tyle podobieństw, ile różnic je dzieli. z jednej strony bazują na podobnych źródłach danych (dane akademickie, dzieła literackie czy zasoby dostępne w Internecie), ale w odmiennych proporcjach. Jedną z przyczyn jest popularność języka angielskiego, który jest dominującym językiem używanym w świecie nauki, badań czy na portalach internetowych. Stąd chociażby dysproporcja pomiędzy danymi akademickimi, ponieważ przyjętym standardem jest publikowanie badań i ich wyników właśnie w języku angielskim (nawet w przypadku polskich badaczy).

Projekt The Pile można uznać za bardziej kompaktowy (22 źródła danych), jednakże charakteryzuje go duże zróżnicowanie tematyki. Spichlerz z kolei poszczycić się może ponad 20-krotnie większą liczbą podzbiorów, co pozwala na łatwiejsze i szybsze operowanie zbiorami i wybór tylko tych, które są w danym zadaniu niezbędne – co ma swoje mocne i słabe strony. Co więcej, trzy największe zbiory The Pile stanowią około 50% objętości całego zestawu. Projekt Spichlerz z kolei jest mocno osadzony na danych z sieci Internet, które stanowią ponad 80% całości. Należy o tym pamiętać przy tworzeniu i szkoleniu modelu w oparciu o dane z tych zbiorów.

Kontekst nadchodzącego AI Act zmusza również do spojrzenia na opisywane zbiory pod kątem ich transparentności oraz metadanych zebranych obok właściwego, surowego tekstu.

Tutaj wyraźnym liderem wydaje się projekt Spichlerz. Przyjęte w nim standardy mogą niedługo stać się wymogami prawnymi w świetle zmian i regulacji zapowiadanych przez Unię Europejską.

Niezależnie od tego obydwie opisywane zbiory, czyli The Pile oraz Spichlerz, są liderami w swoich kategoriach i mogą być stawiane za wzór dla innych tego typu opracowań. Ich wykorzystanie może w przyszłości przyczynić się do powstania kolejnych uniwersalnych, dużych i wydajnych modeli językowych.



## Komentarz eksperta



### **Maria Filipkowska**

Data Scientist, Deviniti

Co-founder of [Speakleash.org](https://speakleash.org)

### Czy w Spichlerzu mamy już wszystko?

Chociaż Spichlerz to bogaty i różnorodny zestaw danych, nadal brakuje w nim niektórych ważnych typów tekstów. Nie mamy na przykład jeszcze transkrypcji treści audio – a te dane byłyby niezwykle przydatne np. do szkolenia modeli konwersacyjnych i asystentów głosowych. Trzeba bowiem pamiętać, że język mówiony rządzi się nieco innymi prawami niż język pisany, a zatem jeśli chcemy, by LLM w pełni oddawał specyfikę języka polskiego, musimy pamiętać także o polszczyźnie mówionej.

Brakuje też zapisów czatów (których styl potoczny nieco różni się od stylu forów internetowych) oraz np. e-maili prywatnych i biznesowych czy SMS-ów – zbieranie i przetwarzanie takich tekstów jest utrudnione, ponieważ mogą one umożliwić identyfikację osoby lub zawierać inne dane wrażliwe. Z podobnych względów nie udało się nam zebrać satysfakcjonującej liczby tekstów medycznych, np. zanonimizowanych epikryz, ankiet i kwestionariuszy, opisów badań, a także literatury naukowej z tej dziedziny. Takie dane wzbogaciłyby dostępny zbiór np. o nową terminologię i chociażby z tego względu warto byłoby je zgromadzić w przyszłości.



# BAN-PL

---

Zbiór BAN-PL jest efektem współpracy popularnego serwisu internetowego Wykop.pl (często nazywanego także polskim Redditem) oraz zespołu naukowców z Państwowego Instytutu Badawczego NASK pod przewodnictwem dr Inez Okulskiej. Dzięki temu projektowi powstał pierwszy publicznie dostępny zbiór treści obraźliwych i szkodliwych, które wcześniej zostały uznane za takie przez moderatorów (zbanowane). Pierwsza udostępniona wersja składa się z 24 tysięcy próbek podzielonych na dwa równe podzbiory po 12 tysięcy – odpowiednio opisane jako „szkodliwe” oraz „neutralne”<sup>30</sup>.

W swojej pracy autorzy podkreślają wpływ rozwoju platform oraz mediów społecznościowych na wzrost zachowań agresywnych, homofobicznych czy ukierunkowanych na mowę nienawiści w sieci. Biorąc pod uwagę olbrzymią ilość treści generowanych każdego dnia, nawet najsprawniejsze grupy edytorów i moderatorów nie są w stanie zapanować na przykład nad komentarzami pod materiałem wideo czy pod artykułem na portalu informacyjnym. Co istotne, często wiadomości te mogą być generowane przez zautomatyzowane boty, również oparte o modele sztucznej inteligencji, których celem może być na przykład sianie dezinformacji pod tekstem na istotne dla kraju tematy bądź celowe zwrócenie przeciwko sobie grup o różnym pochodzeniu etnicznym czy odmiennych poglądach religijnych.

Twórcy pracy opisują również dostępne do tej pory zasoby i zestawy danych skupiające się na tematyce treści obraźliwych i szkodliwych, jednakże w znakomitej większości są to dane zebrane w języku angielskim. Przywoływane są prace na nad zbiorami z platform, takich jak Twitter (aktualnie X), Wikipedia, Facebook czy YouTube. Ciekawym przykładem jest model BERT wytrenowany na bazie zbanowanych komentarzy z forum Reddit, który finalnie przyjął nazwę HateBERT<sup>31</sup>.

W języku polskim jedynym dostępnym datasetem do niedawna był zestaw opublikowany w ramach zadania nr 6 w konkursie PolEval w edycji z roku 2019. Aktualnie jest to część benchmarku KLEJ, opisywanego wcześniej w tym raporcie. Podzestaw ten nosi nazwę CBD (ang. Cyberbullying Detection, wykrywanie cyberprześladowania) i został opisany w pracy

---

<sup>30</sup> I. Okulska, K. Głabińska, A. Kołos, A. Karlińska et. al., (2023), [BAN-PL: a Novel Polish Dataset of Banned Harmful and Offensive Content from Wykop.pl web service](#)

<sup>31</sup> por. artykuł T. Caselli, V. Basile, J. Mitrović i M. Granitzer, (2021), [HateBERT: Retraining BERT for Abusive Language Detection in English](#) [za]: I. Okulska et al. (2023)

autorstwa P. Rybaka, R. Mroczkowskiego, J. Tracza, I. Gawlika<sup>32</sup>. Niedawno, bo w 2023 roku, autorzy pracy „TrelBERT: A pre-trained encoder for Polish Twitter”<sup>33</sup> opublikowali dataset o nazwie *harmful\_tweets\_1k9*. Jest to zbiór 1000 losowych tweetów z lat 2019-2022 adnotowany przez część autorów według zasad opisanych w pracy dotyczącej analizy zadania nr 6 we wspomnianym powyżej konkursie PolEval 2019<sup>34</sup>.

Serwis [Wykop.pl](http://Wykop.pl), z którego to pochodzą dane tworzące opisywany dataset, powstał w grudniu 2005 roku. Aktualnie jest jedną z dziesięciu najpopularniejszych witryn społecznościowych w Polsce. Autorzy pracy opisują strukturę wiekową użytkowników, udział poszczególnych płci oraz systemy promocji wiadomości publikowanych w ramach portalu, aby dotrzeć do systemu oceny i klasyfikacji komentarzy. Te są pod ciągłym nadzorem zespołu moderatorów. Oprócz tego każdy użytkownik może samodzielnie oznaczyć daną wiadomość jako naruszającą regulamin (czyli jako potencjalny materiał do zbanowania) na podstawie wewnętrznej polityki portalu, która to wyróżnia ponad 20 różnych kategorii. Następnie zgłoszony w ten sposób komentarz jest sprawdzany przez moderującego użytkownika i oznaczony jako niegroźny lub jako groźny i usunięty z witryny.

Zbieranie danych i budowa zestawu były nieco uproszczone w stosunku do polityki portalu, dlatego też stworzono klasy „zawartości obraźliwej” (która uwzględnia akty nienawiści lub agresji oraz akty ataków personalnych, niezależnie od ich sposobu i tła, na jakim są realizowane) oraz „zawartości niestosownej” (pod tę kategorię podpadają na przykład wulgarny język, zachęcanie do przestępstw, promowanie używek czy materiały zawierające drastyczne treści). Zdecydowana większość analizowanych danych datowana była na lata 2013-2023. Ponadto, aby uzyskać zrównoważony zbiór, oprócz danych obraźliwych i nieodpowiednich potrzebne były również dane normalne, w tym wypadku uzyskane metodami [web scrapingu](#) w trzech okresach w latach 2021–2023. Następnie po eksperymentach z automatyczną klasyfikacją przy użyciu modelu RoBERTa large stworzony został dataset o wielkości ponad 345 tysięcy komentarzy oznaczonych jako „nieodpowiednie” (również „szkodliwe”). dla wyrównania dodano drugie tyle komentarzy neutralnych, co w sumie pozwoliło na uzyskanie około 691 662 wystąpień do analizy.

---

<sup>32</sup> P. Rybak, R. Mroczkowski, J. Tracz, I. Gawlik, (2020), [KLEJ: Comprehensive Benchmark for Polish Language Understanding](#) [za]: I. Okulska et al. (2023)

<sup>33</sup> W. Szmyd, A. Kotyla, M. Zobniów, (2023), [TrelBERT: A pre-trained encoder for Polish Twitter](#) [za]: I. Okulska et al. (2023)

<sup>34</sup> M. Ptaszyński, A. Pieciukiewicz, P. Dybała, (2019), [Results of the PolEval 2019 Shared Task 6: First Dataset and Open Shared Task for Automatic Cyberbullying Detection in Polish Twitter](#). [za]: I. Okulska et al. (2023)

Kolejnym krokiem w pracy nad datasetem było przeprowadzenie anonimizacji danych. Zawierały one w związku ze swoją różnorodnością informacje, takie jak numery PESEL, numery telefonów czy adresy, które można potraktować jako dane wrażliwe, umożliwiające identyfikację użytkownika. Zaliczały się do nich także dane osobowe osób publicznych (na przykład polityków, aktorów, celebrytów, którzy często stanowią cel internetowego hejtu). Zostały one kompletnie zanonimizowane poprzez usunięcie dodatkowego kontekstu, który mógłby wskazywać na daną postać pomimo braku jasnego wskazania.

Przeprowadzona została analiza lingwistyczna stworzonego datasetu, której jednym z wielu efektów jest chmura słów, pokazana na ilustracji nr 8. Przedstawiono w niej najczęściej pojawiające się słowa w analizowanych zbiorach, co pozwoliło wskazać główne grupy, które wyróżniają się na tle pozostałych komentarzy.



**Ilustracja nr 9.** Chmura słów szkodliwych z datasetu BAN-PL

Źródło: I. Okulska et al. (2023)

Analizowany tekst został też wyczyszczony z niechcianych elementów, takich jak hashtagi, emotikony (w tym bardzo [popularny na Wykopie lennyface](#)) oraz anomaliami w tekście pisany (jak zamiana litery „o” na cyfrę „0” czy używanie znaków specjalnych w miejscu liter podobnych optycznie). Stworzono także bazę funkcji, która dostępna jest w repozytorium projektu [w serwisie Github](#).

W przyszłości planowane jest powiększenie dostępnego dzisiaj datasetu (który liczy 24 tysiące komentarzy) oraz dalsze badania nad poprawą możliwości analizy mowy nienawiści i sprawnym jej demaskowaniu. Wszystkie aktualizacje będą publicznie dostępne we wspomnianym repozytorium.

# PolQA

---

OpenQA, czyli [Open-domain Question Answering](#), to rodzaj systemu przetwarzania języka naturalnego (NLP), który jest zaprojektowany do odpowiadania na pytania sformułowane w sposób otwarty, bez ograniczania się do konkretnej dziedziny wiedzy. Systemy te wymagają dużej ilości danych treningowych, aby osiągnąć jak najwyższą możliwą wydajność oraz skuteczność. Krajobraz dostępnych zbiorów danych przeznaczonych do OpenQA zdominowany jest przez zbiory anglojęzyczne lub chińskie<sup>35</sup>, podczas gdy zbiory w języku polskim występowały dotychczas w postaci niewielkich zasobów istniejących już datasetów. Aby zmienić taki stan rzeczy, zespół z Instytutu Podstaw Informatyki Polskiej Akademii Nauk (IPI PAN) w składzie Piotr Rybak, Piotr Przybyła oraz Maciej Ogrodniczuk podjął się wyzwania stworzenia PolQA, pierwszego polskiego zbioru danych OpenQA.

Głównym źródłem pytań dla zbioru PolQA były istniejące zasoby, w tym 6000 pytań opublikowanych w ramach konkursu PolEval 2021 dotyczącego QA<sup>36</sup>, a także dodatkowe 1000 pytań zebranych przez jednego z uczestników tego konkursu<sup>37</sup>. Pytania te wywodziły się z różnych kolekcji, w tym z oficjalnie opublikowanej książki M. Karzewskiego „Jeden z dziesięciu – pytania i odpowiedzi”, wydanej przez Muza SA w 1997 roku, jak również z materiałów online, zebranych przez fanów teleturniejów oraz z pytań wykorzystywanych w rzeczywistych konkursach quizowych, zarówno tych telewizyjnych, jak i internetowych.

Każda para pytanie-odpowiedź została ręcznie sklasyfikowana według trzech głównych kryteriów: sposobu sformułowania pytania, rodzaju zadawanego pytania oraz charakteru podmiotu, którego dotyczy pytanie. Sposób sformułowania odnosi się do konstrukcji językowej, która została wykorzystana do postawienia pytania, przy czym można wyróżnić trzy kategorie: standardowe pytania, polecenia oraz wyrażenia złożone. Rodzaj pytania określa rodzaj poszukiwanej informacji, mogą to być pytania o pojedynczą jednostkę, ich zbiór, czy pytania zamknięte typu tak/nie. w kontekście charakteru podmiotu, pytania mogą

---

<sup>35</sup> [Improving Question Answering Performance through Manual Annotation: Costs, Benefits and Strategies](#) Piotr Rybak, Piotr Przybyła, and Maciej Ogrodniczuk.

<sup>36</sup> Ogrodniczuk, M., Przybyła, P.: PolEval 2021 Task 4: Question Answering Challenge. In: Ogrodniczuk, M., Kobyliński, Ł. (eds.) Proceedings of the PolEval 2021 Workshop. pp. 123–136. Institute of Computer Science, Polish Academy of Sciences, Warsaw (2021)

<sup>37</sup> Rybak, P.: Retrieve and Refine System for Polish Question Answering. Proceedings of the PolEval 2021 Workshop pp. 151–157 (2021)

odnosić się zarówno do podmiotów zdefiniowanych jak i tych bardziej ogólnych czy niezdefiniowanych.

Odpowiedzi były formułowane w języku naturalnym w taki sposób, w jaki odpowiadałby na nie polskojęzyczny użytkownik. Oznacza to, że mogą mieć różne formy gramatyczne, przymyki czy różne znaki interpunkcyjne. w wielu przypadkach odpowiedź może mieć kilka poprawnych wariantów ze względu na występowanie synonimów, skrótów i ich rozwinięć bądź liczb zapisanych za pomocą cyfr lub słownie. w procesie tworzenia zbioru uwzględniono wszelkie możliwe warianty odpowiedzi.

Jako główne źródło fragmentów (z ang. *passages*), czyli tekstów zawierających potencjalne informacje do zadanego pytania, wybrano Wikipedię. Odpowiada ona za ponad 93% wszystkich pytań, ze względu na wysokie zróżnicowanie wszelkich dziedzin wiedzy. Pozostałe pytania – dotyczące podstawowej arytmetyki, przysłów czy treści książek lub filmów – są w formie tak/nie, nie mają odpowiedzi lub są pytaniami porównawczymi, które wymagają znacznej ilości treści. Artykuły z Wikipedii zostały przeanalizowane za pomocą [WikiExtractora](#)<sup>38</sup>, a następnie podzielone na fragmenty nie dłuższe niż 500 znaków, z cięciami wykonanymi na granicach zdań. w ten sposób uzyskane zostało obszerne źródło wiedzy złożone 7097322 fragmentów tekstowych.

W procesie adnotacji fragmentów dowodowych dla poszczególnych pytań autorzy zastosowali trzy różne metody. Każda z nich składała się z dwóch etapów: odnalezienia pięciu najbardziej obiecujących fragmentów, które mogły zawierać potencjalne odpowiedzi, a następnie ich ręcznej weryfikacji pod kątem trafności na zadane pytanie.

Pierwszym podejściem było *Unconstrained Retrieval* (wyszukiwanie bez ograniczeń). w tej metodzie annotatory wykorzystywali zarówno wewnętrzne zasoby, takich jak Wikipedia, jak i zewnętrzne, typu Google, aby znaleźć odpowiednie fragmenty w oparciu o dowolne słowa kluczowe, które uważali za użyteczne. Założeniem takiego podejścia było uzyskanie bardziej obiektywnych oraz różnorodnych przykładów poprzez swobodny wybór fragmentów. Annotatory zostali również poinstruowani do znajdowania do pięciu różnych fragmentów, najlepiej z różnych artykułów, co miało na celu zwiększenia potencjalnej różnorodności. Każdy wybrany fragment był następnie poddawany podwójnej weryfikacji, aby ocenić jego istotność w kontekście zadanego pytania.

---

<sup>38</sup> [Attardi, G.: Wikiextractor](#). (2015)

Drugim podejściem było *Zero-shot Verification* (weryfikacja zero-shot), które polegało na przedstawianiu annotatorom par pytanie – fragment i uzyskania nich oceny, czy dany fragment jest adekwatny do zadanego pytania. Metoda ta okazała znacznie szybsza niż Unconstrained Retrieval, choć wymagała użycia funkcji próbkowania do wyboru fragmentów. Autorzy zastosowali w tym celu próbkowanie zero-shot. Najpierw [przeprowadzono lematyzację pytań](#) i fragmentów przy użyciu narzędzia Spacy, następnie wybrano 100 najlepszych fragmentów za pomocą algorytmu BM-25, dokonano klasyfikacji przy użyciu metody krzyżowej (ang. [cross-encoder](#)) i ostatecznie wybrano pięć najlepszych fragmentów do ręcznego etykietowania. Unikano specyficznych dla języka polskiego zasobów, z wyjątkiem lematyzatora przeszkolonego na polskich tekstach.

Trzecim podejściem było *Hard Negatives*, czyli wybór nieistotnych (negatywnych) fragmentów, kluczowych dla szkolenia modułu wyszukiwania (ang. [neural retriever](#)) w tym przypadku wykorzystano fragmenty znalezione w ramach wyszukiwania bez ograniczeń (ang. Unconstrained Retrieval), a następnie każdy z nich był weryfikowany ręcznie.

Stosując różne podejścia przeprowadzania adnotacji podczas tworzenia zbioru danych PolQA, autorzy doszli do wniosku, że wykorzystanie bezstronnych fragmentów dowodowych nie przyczynia się do zwiększenia wydajności modeli OpenQA. Metoda zero-shot dała najlepsze wyniki poprzez selekcję kandydatów na fragmenty oraz ich ręczną weryfikację. Pozwoliła ona na obniżenie kosztów adnotacji ponad pięciokrotnie, jednocześnie poprawiając wydajność modeli OpenQA.

Ostatecznym efektem pracy zespołu z Instytutu Podstaw Informatyki Polskiej Akademii Nauk nad danymi i ich adnotacją, było stworzenie zbioru danych PolQA, który ma następującą strukturę:

1. 7000 pytań,
2. 8713 wariantów odpowiedzi,
3. 87525 fragmentów dowodowych,
4. korpus składający się z 7097322 nieoznakowanych kandydatów na fragmenty/odpowiedzi,
5. 97,7% pytań stanowią pytania proste,
6. 79,8% pytań dotyczy pojedynczego podmiotu,
7. 55,1 % odpowiedzi stanowią typy określone,
8. 37,3 % odpowiedzi stanowią typy nieokreślone.

Tak rozbudowany dataset PolQA znajduje zastosowanie w ocenie jakości i wydajności modeli OpenQA oraz trenowaniu modeli różnego typu: odpowiadających na pytania z dziedziny ogólnej, przeznaczonych do przeszukiwania dokumentów (ang. *document-retrieval*) czy służących do odpowiadania na pytania abstrakcyjne.



## Komentarz eksperta



### **Sebastian Kondracki**

Chief Innovation Officer, Deviniti  
Lider SoDA AI Research Group  
Twórca projektu SpeakLeash



### **Dr inż. Paweł Tadejko**

Ekspert AI, Data Scientist, wykładowca akademicki  
Wydział Informatyki Politechniki Białostockiej,  
Polski Klaster Budowlany

## Polskie modele językowe a źródła encyklopedyczne

W kontekście rozwoju polskich modeli językowych, takich jak Polbert, HerBert czy APT-3, ważnym aspektem jest ich źródło danych treningowych. Większość z nich została wytrenowana na podstawie zawartości polskiej Wikipedii. Podobnie, zestawy pytań i odpowiedzi takie jak poqaud, będące cennym źródłem instrukcji, również bazują na treściach z Wikipedii. Zasadnicze pytanie, które nasuwa się w tym kontekście, dotyczy ryzyka wynikającego z opierania wielu modeli i zestawów danych na jednym źródle, jakim jest Wikipedia.

Istotne jest zwrócenie uwagi na faktograficzne ograniczenia Wikipedii, szczególnie w kontekście nowych technologii oraz na pewne skrzywienie w opisach subiektywnych faktów. Przykładem może być różnica w narodowości Kopernika w polskiej i niemieckiej wersji encyklopedii. Ponadto, czasami Wikipedia jest podatna na manipulacje, takie jak promowanie fałszywych informacji, nieistotnych wynalazków i artystów czy wpisów sponsorowanych, co może prowadzić do skrzywień w modelach językowych. z kolei zakres haseł dotyczących nowych technologii jest bardzo wąski a ich jakość pozostawia bardzo wiele do życzenia. Widać to nie tylko porównując treść haseł z ich odpowiednikami w angielskiej Wikipedii, ale też z materiałami źródłowymi. i nie dotyczy to tylko haseł z ostatnich lat, ale generalnie haseł związanych np. z branżą ICT, czy technologiami tzw. przełomowymi, tj. AI, cyberbezpieczeństwo, obliczenia kwantowe, etc.



W związku z tym, podczas trenowania i rozwoju polskich modeli językowych powinno się rozważyć wykorzystanie różnorodnych źródeł encyklopedycznych, aby zapewnić większą obiektywność i różnorodność danych. Zróżnicowanie źródeł danych może przyczynić się do zwiększenia rzetelności i obiektywizmu oraz podniesienia jakości językowej generowanych przez AI treści. Istnieje również pytanie o dostępność encyklopedycznych źródeł danych: czy nie powinny one stać się otwartym dobrem narodowym, aby ułatwić rozwój nowych, bardziej zrównoważonych modeli językowych?

Ostatecznie, dla rozwoju efektywnych i rzetelnych modeli językowych, kluczowe jest zapewnienie zróżnicowanego i wiarygodnego zbioru danych treningowych, co może wymagać szerszego spojrzenia na dostępne źródła wiedzy.

## Podsumowanie

# Ocena zaspokojenia potrzeb biznesowych

---

Jak już wspomniano we wstępie niniejszego dokumentu, przetwarzanie języka naturalnego (NLP) jest jednym z najważniejszych obszarów AI wspierających trwającą transformację cyfrową w biznesie, edukacji i sektorze publicznym. Wynika to głównie z ogromnej liczby dokumentów cyfrowych, które są wymieniane między firmami, organizacjami i urzędami. Streszczanie dokumentów, klasyfikacja, strukturyzacja, korekta i tłumaczenie maszynowe to jedne z największych potrzeb nowoczesnych organizacji. Zatem przyjrzyjmy się obszarowi NLP, poszczególnym technikom i modelom w służbie firmy, klienta, obywatela, pacjenta czy petenta.

Dane językowe w formie zestawów mogą być generalnie przydatne w dwóch obszarach:

1. specjalizowanych modelach uczenia maszynowego, takich jak: klasyfikatory dokumentów, modele NER (rozpoznawanie nazwanych bytów), analiza emocji w tekście, klasyfikacja tekstu pod kątem treści ofensywnych itp.,
2. budowaniu i dostrajaniu modeli językowych.

Pierwsza grupa modeli wymaga zwłaszcza danych adnotowanych i może być rozwijana również na komputerach (serwerach) wyposażonych w CPU. Natomiast druga grupa wymaga tekstów bez żadnych adnotacji (np. książek, artykułów, stron WWW) oraz instrukcji i ogromnej mocy obliczeniowej (GPU). Dzięki temu dostrojone modele językowe mogą być agnostyczne wobec zadań, co oznacza, że są zdolne do wykonywania większości zadań z obszaru NLP.

## Dane językowe do specjalizowanych modeli uczenia maszynowego trenowanych pod kątem zadań NLP

W tej grupie największa potrzeba dotyczy zestawów danych **etykietowanych tekstów** (np. treści e-maili z etykietami SPAM/NIE-SPAM) różnych typów, takich jak wiadomości, artykuły, opinie, dokumenty oraz listy czy drzewa kategorii (emocje, kategorie tematyczne, typy dokumentów itp.). Istnieje kilka zestawów danych o wymaganej wolumetrii i jakości gotowych do trenowania modeli. Można tutaj wskazać na przykład zbiory: PolEmo (analiza emocji), Ban-PL (wykrywanie tekstów ofensywnych) oraz Polish GEC (korekta językowa). Jest również grupa **specjalizowanych danych**, z których po niewielkiej obróbce można stworzyć pożądaną zestaw, np. do klasyfikacji tekstu. Innym sposobem pozyskania danych jest korzystanie z **obszernych repozytoriów**, takich jak polska Wikipedia, ISAP (internetowa baza aktów prawnych) czy EUR-Lex (baza aktów prawnych Unii Europejskiej), a także przetwarzanie tekstów (np. streszczanie) i łączenie ich z metadanymi (kategoria, słowa kluczowe), aby stworzyć odpowiedni zestaw danych. Jednak wymaga to znacznej mocy obliczeniowej i wykorzystania gotowych modeli językowych.

Trzeba też pamiętać, że ta grupa danych ma istotne znaczenie edukacyjne. Większość początkujących naukowców ds. danych (z ang. *data scientists*) czy programistów zajmujących się przetwarzaniem języka naturalnego potrzebuje takich zbiorów do budowania swoich pierwszych modeli ML (tj. modeli uczenia maszynowego – ang. *machine learning*). Często realizują oni swoje projekty na podstawie zagranicznych zestawów danych, np. w języku angielskim, co może prowadzić do braków kompetencyjnych związanych z pracą na polskich tekstach (dla których charakterystyczna jest między innymi fleksja).

## Zestawy danych językowych wymagane do budowania i dostrajania modeli językowych

Polskie zasoby danych językowych wymagane do budowania i dostrajania dużych modeli językowych, chociaż stale powiększane, wciąż są jeszcze niekompletne. Jak wspomniano w raporcie, brakuje m.in. zapisów polszczyzny mówionej, warto byłoby je więc wzbogacić o **dane dialogowe i konwersacyjne**. Należą do nich m.in. ogólnodostępne transkrypcje audio – zapisy rozmów telefonicznych, a także napisy do filmów (tworzących wraz z materiałem wideo **dane multimodalne**). Na razie polscy badacze i innowatorzy mają

do dyspozycji głównie transkrypcje automatyczne (będące często tłumaczeniami z języka angielskiego), których jakość często jest niezadowalająca.

Brakuje również ręcznie tworzonych **zestawów instrukcji** do różnych zadań, takich jak odpowiedzi na pytania otwarte i zamknięte, streszczenia, parafrazy itp. Te nieliczne dostępne na ogół są zbiorami syntetycznymi – generowanymi z szablonu lub tłumaczonymi z języka angielskiego.

Instrukcje można tworzyć również na wysokiej jakości **danych tekstowych z adnotacjami**. Najbardziej wiarygodne i wartościowe są zbiory adnotowane ręcznie – jednak ich pozyskanie jest trudniejsze ze względu na duży nakład kosztów i długi czas realizacji tego typu zadań. Warto zwiększyć wysiłki, aby poszerzyć istniejące już zbiory o kolejne adnotowane ręcznie dane (np. pozyskane w ramach konkursów lub projektów naukowych).

## Proces walidacji zestawów danych

---

Kluczowym aspektem podczas walidacji zestawów danych jest ocena stopnia ich dojrzałości, a także gotowości technologicznej. Ocena taka powinna obejmować kilka kluczowych aspektów:

1. **Zakres zawartości** – ocena, czy zestaw danych obejmuje szeroki zakres tematów, kontekstów i stylów językowych. w przypadku zestawów danych językowych ważne jest, aby zawierały one różnorodne formy języka, w tym język potoczny, formalny, techniczny oraz specjalistyczny.
2. **Jakość i dokładność adnotacji** – w przypadku zestawów danych adnotowanych istotna jest precyzja i spójność adnotacji. To obejmuje ocenę, czy etykietowanie jest konsekwentne, czy annotatory postępowali zgodnie z jednolitymi wytycznymi, a także jakość samego procesu adnotacji.
3. **Reprezentatywność i zróżnicowanie** – czyli analiza, czy dane są reprezentatywne dla docelowej populacji użytkowników czy zastosowań. Należy zwrócić uwagę na to, czy zestaw danych uwzględnia różnorodność lingwistyczną, kulturową i demograficzną.
4. **Rozmiar i wolumetria danych** – wielkość zestawu danych ma kluczowe znaczenie dla skuteczności modeli uczenia maszynowego. Wykorzystanie większych zestawów danych zwykle prowadzi do opracowania modeli o lepszej ogólnej wydajności i ogólności, ale wymaga to większych zasobów do przetwarzania i analizy.

5. **Aktualność danych** – ustalenie, czy dane są aktualne i czy odzwierciedlają współczesne użycie języka oraz aktualne tematy i wydarzenia.
6. **Dostępność i licencjonowanie** – ocena, czy zestaw danych jest łatwo dostępny dla naukowców i praktyków oraz czy istnieją ograniczenia licencyjne dotyczące jego użycia.
7. **Zgodność ze standardami etycznymi i prawnymi** – sprawdzenie, czy zestaw danych jest zgodny z obowiązującymi przepisami prawnymi (np. RODO) oraz czy respektuje zasady etyczne, szczególnie w kontekście prywatności i anonimowości danych.
8. **Informacja zwrotna od użytkowników i społeczności** – zbieranie informacji zwrotnych od użytkowników i społeczności naukowej, które mogą wskazywać na realną wartość i użyteczność zestawu danych.

W związku z wynikami naszych badań i ich ograniczeniami, zaleca się rozważenie możliwości przeprowadzenia podobnych analiz we współpracy z jednostkami akademickimi oraz urzędami centralnymi. Takie wspólne przedsięwzięcia umożliwiłyby nie tylko dogłębną analizę dostępnych zestawów danych, ale także umożliwiłyby przeprowadzenie ich szczegółowej oceny jakościowej. Warto w tym kontekście zastosować [metodykę unijną Open Data Maturity](#) (ODM), szczególnie jej sekcję dotyczącą jakości danych otwartych (z ang. *open data quality*), co pozwoliłoby na bardziej obiektywne i standaryzowane podejście do oceny danych.

## Rekomendacje dotyczące danych językowych

---

Kluczowym brakiem w obecnych zestawach danych językowych jest niedostatek zestawów danych o średniej wolumetrii (powyżej 25 tysięcy rekordów), adnotowanych i o wysokiej jakości, najlepiej ręcznie tworzonych lub weryfikowanych. Takie zbiory danych otworzyłyby znaczące możliwości dla centrów szkoleniowych i edukacyjnych. Ponadto umożliwiłyby one biznesowi wydajne tworzenie modeli do klasyfikacji tekstów, dokumentów, analizy emocji i innych zastosowań.

Wysokiej jakości ręcznie adnotowane zbiory danych są niezwykle cenne, ponieważ zapewniają dokładność i precyzję niezbędną do trenowania skutecznych modeli językowych. Warto zauważyć, że im większa jest jakość adnotacji, tym model lepiej radzi sobie z zadaniami NLP, co jest kluczowe w zastosowaniach komercyjnych i akademickich.

Dodatkowo duża liczba takich zestawów danych byłaby znakomitym punktem wyjścia do budowania zróżnicowanych zestawów instrukcji potrzebnych do dostrajania dużych modeli językowych pod kątem różnych domen. Takie instrukcje wspierają rozwój modeli, które mogą być dostosowane do konkretnych potrzeb i wymagań różnych branż, od finansów po opiekę zdrowotną. Taka specjalizacja jest niezbędna, by modele mogły efektywnie rozwiązywać specyficzne problemy i zadania.

Ostatecznie inwestycja w rozwój wysokiej jakości adnotowanych zestawów danych o średniej wolumetrii przyniesie korzyści zarówno światu akademickiemu, jak i sektorowi biznesowemu, przyczyni się do postępu w dziedzinie sztucznej inteligencji oraz do wdrażania jej praktycznych zastosowań.

## Kierunki rozwoju i kontynuacja prac

---

### Analiza pogłębiona obszaru medycznego i farmaceutycznego (healthcare/pharma)

W tym opracowaniu skoncentrowano się na analizie polskich zasobów językowych w otwartym dostępie, które mają potencjał zastosowania w technologiach sztucznej inteligencji (AI) w różnych sektorach biznesowych. Jednakże, ze względu na ograniczenia czasowe, analiza ta nie mogła w pełni objąć obszarów medycznego i farmaceutycznego, które są niezwykle istotne i dostarczają specyficznych rodzajów danych, wykraczających poza tradycyjne zasoby tekstowe. Te dane mają różne formaty, należą do nich np. obrazy medyczne (np. zdjęcia rentgenowskie, skany MRI), zapisy dźwiękowe (np. z badania serca), dane genetyczne i wiele innych, które mają ogromny potencjał pod względem zastosowań AI. w kolejnych opracowaniach skupimy się na tych typach danych, rozszerzając zakres badań również o zapisy w innych formatach. Analiza takich zbiorów może przynieść przełomowe wyniki w diagnozowaniu, leczeniu i badaniach nad nowymi lekami, a także w personalizacji terapii i monitoringu stanu zdrowia pacjentów. Należy też zwrócić uwagę na wyzwania związane z przetwarzaniem i analizą tych złożonych formatów danych, w tym kwestie związane z ochroną prywatności, etyką wykorzystania danych medycznych oraz integracją z istniejącymi systemami opieki zdrowotnej.

Kolejnym istotnym powodem, dla którego warto zająć się zbiorami agregującymi dane medyczne i farmaceutyczne, jest nadchodzące wdrożenie Europejskiej przestrzeni danych dotyczących zdrowia ([European Health Data Space, EHDS](#)) i związane z tym wymagania,

które w niedalekiej przyszłości zaczną obowiązywać również w Polsce. Europejska przestrzeń danych dotyczących zdrowia ma na celu ułatwienie wymiany tych danych w całej Europie, zarówno w celach świadczenia opieki zdrowotnej, jak i badań, innowacji oraz realizacji polityki zdrowotnej. EHDS stawia przed krajami członkowskimi, w tym Polską, nowe wyzwania i wymogi dotyczące standardów, bezpieczeństwa, interoperacyjności oraz ochrony danych osobowych.

Skupienie się na zbiorach danych medycznych i farmaceutycznych w kontekście EHDS ma kilka kluczowych zalet:

1. **Zgodność z przepisami**, z którą związane jest z przygotowanie i dostosowanie krajowych systemów danych do wymogów EHDS. To będzie wymagało przeglądu i ewentualnej modyfikacji istniejących baz danych, co jest istotne dla zapewnienia zgodności z europejskimi standardami.
2. **Wspieranie innowacji**, ponieważ agregacja i standaryzacja danych medycznych i farmaceutycznych umożliwi lepsze wykorzystanie tych danych w badaniach i rozwoju, co może przyczynić się do postępu w medycynie i farmacji.
3. **Poprawa jakości opieki zdrowotnej** poprzez dostęp do zharmonizowanych i interoperacyjnych danych medycznych pozwoli na lepszą koordynację opieki zdrowotnej, personalizację terapii oraz efektywniejsze zarządzanie systemami zdrowia.
4. **Współpraca międzynarodowa** związana z integracją danych i weryfikacją ich pod kątem zgodności z EHDS.

W związku z tym praca nad zbiorami danych medycznych i farmaceutycznych w Polsce, z uwzględnieniem wymogów EHDS, jest nie tylko koniecznością wynikającą z nadchodzących regulacji, ale również szansą na rozwój sektora zdrowia i farmacji, poprawę jakości opieki zdrowotnej oraz wspieranie innowacji i badań naukowych na skalę międzynarodową. Chociaż niniejsze opracowanie dostarcza szerokiej wiedzy na temat polskich zasobów tekstowych z otwartym dostępem, istnieje potrzeba dalszych badań skoncentrowanych na danych medycznych i farmaceutycznych, które mogą znacząco przyczynić się do rozwoju tych krytycznych dla społeczeństwa sektorów.

Katalizatorem, który zmieni wiele na rynku analizy danych medycznych, a co za tym idzie medycznych zbiorów danych, nowych platform i aplikacji mobilnych, jest także

rozporządzenie UE „Data Act”. Data Act<sup>39</sup> wprowadza szczególne regulacje dla dostawców usług przetwarzania danych, a także nakłada na dostawców rozwiązań obowiązki dotyczące dostępu do danych i możliwości ich ponownego wykorzystania. Nie tylko wyżej wymienione np. dane rejestrowane jako zdjęcia rentgenowskie, skany MRI, badania EKG, ale też inne dane zbierane z urządzeń określanych mianem urządzeń noszonych (z *ang.* *wearables* – tj. urządzenia z grupy IoT), np. rozruszniki serca, monitory glukozy i inteligentne urządzenia insulinowe, urządzenia *wellness wearables* i *fitness trackery*, czy czujniki połykane.

## Eksploracja nowych dziedzin

Jako autorzy raportu podjęliśmy decyzję o skoncentrowaniu się w pierwszej kolejności na ewidencjonowaniu i opisie zestawów danych językowych o dużym potencjale w sektorze biznesowym. Jak już wspomnieliśmy, taki krok wynika głównie z rosnących potrzeb w obszarze przetwarzania języka naturalnego w biznesie, które obejmują komunikację, dokumentację, media i inne zastosowania. Szczególnie istotny okazuje się tu rozwój narzędzi wspierających efektywną i precyzyjną komunikację między przedsiębiorstwami a klientami, jak również ułatwiających zarządzanie dokumentacją biznesową. Obszarem dalszych prac (oddzielny nowy raport), który został już zarysowany, jest sektor farmacji, opieki medycznej i medycyny (patrz powyższy podrozdział).

Konieczna jest też dalsza eksploracja innych obszarów i domen, jak również identyfikacja potrzeb w zakresie zestawów danych dla AI w takich sektorach jak rolnictwo, przemysł, logistyka, a także obszary wspierające turystykę czy e-commerce. Są to sektory, w których analiza danych może przynieść znaczne korzyści, na przykład dzięki optymalizacji łańcuchów dostaw, personalizacji ofert turystycznych czy efektywnemu zarządzaniu relacjami z klientami w e-commerce.

Bardzo ważną kwestię w kontekście otwartych zbiorów danych i trenowania modeli stanowią regulacje prawne w zakresie ochrony praw autorskich. Rok 2023 obfitował w wiele pozwów o naruszenia praw autorskich, m.in. przez OpenAI<sup>40</sup>, Meta<sup>41</sup>, Midjourney<sup>42</sup>. Rozpoczęło to

---

<sup>39</sup> [Adapting to the new EU Data Act: implications for medical devices and other health devices](#)

<sup>40</sup> [New York Times Sues OpenAI and Microsoft Over Use of Copyrighted Work](#) - The New York Times

<sup>41</sup> [Erotica, Atwood, and 'For Dummies': The Books Behind Meta's Generative AI](#) - The Atlantic

<sup>42</sup> [Data Leak: Midjourney's Unauthorised Use of 16,000+ Artists' Works Sparks Legal and Ethical Showdown](#) | by Caleb | Jan, 2024 | Medium



także dyskusje na forum publicznym, jak w nowej rzeczywistości dużych modeli (nie tylko językowych, tj. LLM) należy podchodzić do ochrony praw autorskich. Praw w zakresie dostępu do danych, na których są trenowane modele, bo osobną kwestią są prawa autorskie do treści tworzonych przy pomocy narzędzi generatywnej AI.

Japonia jest chyba jedynym krajem, który podjął działania w tym kierunku już kilka lat temu. w 2018 roku kraj ten modyfikował swoją ustawę o prawie autorskim tak, aby umożliwić bezpłatne korzystanie z utworów chronionych prawem autorskim w celu trenowania modeli uczenia maszynowego, o ile celem „nie jest czerpanie przyjemności z myśli lub uczuć wyrażonych w utworze”<sup>43</sup>. Widać wyraźnie, że świadomość znaczenia transformacji cyfrowej i wykorzystania nowych technologii na potrzeby gospodarki napędza w Japonii spektakularne zmiany prawne od kilku lat. Członkowie G7 także dostrzegli ten problem i konsekwencje wynikające z różnych regulacji prawnych na poziomie UE oraz pozostałych krajów G7. w 2023 roku rozpoczęto pierwsze prace G7 w zakresie ujednoczenia międzynarodowych zasad dotyczących sztucznej inteligencji i kodeksu postępowania dla programistów AI w ramach kodowej nazwy „Hiroshima AI Process”<sup>44</sup>. Opracowanie, określane mianem kodeksu „International Code of Conduct for Organizations Developing Advanced AI Systems”<sup>45</sup>, jest dokumentem, który będzie podlegał zmianom w zależności od rozwoju technologii i innych uwarunkowań (tzw. *living document*). Kodeks, który jest de facto odpowiednikiem unijnego rozporządzenia „AI Act”, zawiera w szczególności punkt 11. „Wdrożenie odpowiednich środków wprowadzania danych oraz ochrony danych osobowych i własności intelektualnej”. w tym zakresie należy oczekiwać dalszych prac związanych z podejściem do wykorzystania w trenowaniu modeli treści objętych prawem autorskim. Całkiem możliwe, że mimo tego, że Japonia należy do G7, okaże się trudnym negocjatorem i będzie bronić wypracowanych już zasad krajowych.

Grupa robocza ds. AI zamierza eksplorować te obszary, korzystając z artykułów branżowych, badań wstępnych oraz współpracy z ekspertami z różnych dziedzin. Celem jest stworzenie kompleksowego przeglądu dostępnych danych, identyfikacja luk i proponowanie kierunków rozwoju narzędzi AI, które mogą wspierać różne branże w ich specyficznych wyzwaniach.

---

<sup>43</sup> [Japan's AI Data Laws, Explained.](#)

<sup>44</sup> [G7 Leaders' Statement on the Hiroshima AI Process](#) | Shaping Europe's digital future,

<sup>45</sup> [Hiroshima Process International Code of Conduct for Organizations Developing Advanced AI Systems](#) (30/10/2023) - G7/G20 Documents Database