Analiza ilościowa zestawów danych i schematów ich wykorzystania w modelach językowych

Adam Jachimczyk
ORCID: 0000-0003-2917-6926
Wydział Dziennikarstwa, Informacji i Bibliologii
Uniwersytet Warszawski

Abstrakt

Cel/Teza: Celem niniejszej analizy była ilościowa charakterystyka zestawów danych wykorzystywanych do szkolenia modeli językowych. Objęła ona: 1) analizę częstości użycia poszczególnych zestawów danych oraz schematów ich zastosowania w różnych typach modeli, 2) ocenę stopnia zróżnicowania danych wykorzystywanych przez autorów modeli, 3) analizę różnorodności wykorzystania danych w kontekście konkretnych zastosowań modeli językowych.

Koncepcja/Metody badań: Przeanalizowano metadane modeli językowych dostępnych w repozytorium Hugging Face. Uwzględniono tylko te modele, które zawierały informację o nazwie użytego zestawu danych, rozumianego jako zbiór wykorzystywany do trenowania, dostrajania lub ewaluacji modelu. Łącznie przeanalizowano 56 762 modele (13,8% ogółu), których metadane pobrano 6 marca 2025 roku za pomocą biblioteki huggingface-hub. W analizie obliczono m.in. liczbę modeli i autorów, częstotliwość występowania zestawów danych w zależności od kategorii modelu, a także korelację między liczbą unikalnych zestawów danych stosowanych przez poszczególnych autorów a łączną liczbą udostępnionych przez nich modeli. Analiza została przeprowadzona w języku Python z wykorzystaniem bibliotek pandas, matplotlib, seaborn, scipy.

Wyniki i wnioski: Łącznie zidentyfikowano 13 376 zestawów danych. Większość modeli była szkolona na podstawie tylko jednego zestawu danych. Odnotowano, że wśród autorów publikujących większą liczbę modeli językowych widoczna jest pewna tendencja do wykorzystywania bardziej zróżnicowanych zestawów danych. Zaobserwowano także, że w kategoriach z mniejszą liczbą modeli występuje większe zróżnicowanie wykorzystywanych do ich szkolenia zestawów.

Oryginalność/Wartość poznawcza: Analiza ujawnia pewne schematy dotyczące stopnia zróżnicowania wykorzystania zestawów danych w zależności od autora i kategorii modelu językowego.

Słowa kluczowe:

Hugging Face. Model językowy. Sztuczna inteligencja. Zestaw danych.

Tekst wpłynął do Redakcji: 22 kwietnia 2025 r.

1. Wstęp

W dziedzinie sztucznej inteligencji w ostatnich kilku latach rośnie zainteresowanie modelami językowymi – modelami statystycznymi zdolnymi do przewidywania słów w sekwencji słów języka naturalnego. Te, które zostały wyszkolone na dużych zestawach różnego typu danych, nazywa się dużymi modelami językowymi (Auffarth, 2024). Ich popularność ilustruje liczba pobrań z repozytorium Hugging Face (HF) (*Models – Hugging Face*, 2025); np. sentence-transformers/all-MiniLM-L6-v2 i google-bert/bert-base-uncased pobrano ponad 80 milionów razy (stan na 6 marca 2025 roku).

W repozytorium HF wszystkie modele wykonujące określone zadania z zakresu przetwarzania języka naturalnego zostały podzielone na 12 kategorii (stan: 6 marca 2025 roku). Największa liczebnie grupa (text generation, text2text generation) została wyszkolona w celu generowania nowego wyrazu lub sekwencji wyrazów na podstawie dostarczonego tekstu wejściowego (*What is Text Generation?*, 2024). Drugą pod względem liczebności stanowią modele wytrenowane dla potrzeb klasyfikowania tekstu (*text classification*) i klasyfikowania tokenów (*token classification*). Te pierwsze znajdują zastosowanie w zadaniach związanych z analizą sentymentu, a druga kategoria modeli sprawdza się m.in. w operacjach związanych z rozpoznawaniem encji czy części mowy (*What is Text Classification?*, 2023; *What is Token Classification?*, 2024). Najmniej modeli zawiera kategoria zero-shot classification, obejmująca modele klasyfikujące dane wcześniej nieznane (*What is Zero-Shot Classification?*, 2023).

Po kilkanaście tysięcy modeli zarejestrowano w takich kategoriach, jak: *feature extraction*, *fill-mask* czy *question answering*. W przypadku *feature extraction* model przekształca tekst wejściowy w reprezentację numeryczną, wykorzystywaną m.in. do wyszukiwania semantycznego czy porównywania tekstów (*Feature Extraction with LLMs, Hugging Face and MinIO*, b.d.; W*What is Feature Extraction?*, 2024). Natomiast modele typu *fill-mask* przewidują brakujące słowa w tekście, zastępując zakryte fragmenty odpowiednimi wyrazami (*What is Fill-Mask?*, 2025).

W systemach wyszukiwania informacji wykorzystuje się modele *question answering*, które odpowiadają na pytania w języku naturalnym, ekstrahując istotne informacje z dostarczonych danych (*What is Question Answering?*, 2025). W tych systemach stosuje się także modele należące do kategorii *sentence similarity*, służące do oceny semantycznego podobieństwa między tekstami (*What is Sentence Similarity?*, 2025). Odmienną funkcję pełnią modele *table question answering*, przystosowane do odpowiadania na pytania na podstawie danych tabelarycznych (*What is Table Question Answering?*, 2023). Relatywnie niewielką grupę stanowią modele przeznaczone do automatycznego tłumaczenia oraz streszczania tekstów.

W praktyce część modeli ma uniwersalny charakter i jest w stanie wykonywać zróżnicowane zadania z zakresu przetwarzania języka naturalnego (tłumaczenie, streszczanie, analiza sentymentu) (Arif, 2023).

Szkolenie modeli językowych zazwyczaj przebiega w dwóch etapach. Najpierw model jest trenowany na bardzo dużym zbiorze tekstów. Następnie jest on udoskonalany za pomocą algorytmu RLHF (*Reinforcement Learning from Human Feedback* – uczenie ze wzmocnieniem z opinii ludzi), żeby wygenerować odpowiedzi zgodne z oczekiwaniami ewaluatorów (OpenAI *et al.*, 2024). Wspomniane etapy wyznaczają też pewne kryterium podziału zestawów danych, które są wykorzystywane do szkolenia modeli.

Zhao *et al.* (2025) wyróżnili trzy główne kategorie zestawów danych. Pierwszą z nich stanowią dane wykorzystywane do wstępnego treningu modeli. Korpusy tego typu mogą pochodzić z jednego lub wielu źródeł, takich jak Wikipedia (b.d.) czy arXiv (*arXiv.org e-Print archive*, b.d.). Podstawą korpusu The Colossal Clean Crawled Corpus (C4) (Dodge *et al.*, 2021) jest baza danych Common Crawl archiwizująca zawartość stron WWW (*Common Crawl – Open Repository of Web Crawl Data*, b.d.). Modele językowe wspomagające pisanie komputerowego kodu wykorzystują publicznie dostępne zasoby takich serwisów, jak np. GitHub (*GitHub · Build and ship software on a single, collaborative platform*, b.d.).

Drugą kategorię wyróżnioną przez Zhao *et al.* (2025) stanowią zestawy danych wykorzystywane do dostrajania modelu (*fine-tuning*), które ma poprawić jego możliwości i dostosować do określonych zastosowań. Do szkolenia modeli w zakresie streszczania dialogów używany jest Samsum liczący ponad 16 tys. spisanych przez lingwistów przykładów codziennych rozmów wraz ze streszczeniami (*Samsung/samsum · Datasets at Hugging Face*, 2022; Eigenbrode *et al.*, 2024). Specyficzną grupę zestawów danych stanowią dane syntetyczne. Według Jordon *et al.* (2022), „to dane wygenerowane przy użyciu specjalnie zaprojektowanego modelu matematycznego lub algorytmu, mające na celu rozwiązanie określonego zadania lub zestawu zadań z zakresu nauki o danych". Przykładem takich danych wygenerowanych przez model text-davinci-003 jest Alpaca (*tatsu-lab/alpaca · Datasets at Hugging Face*, 2024).

Do trzeciej grupy Zhao *et al.* (2025) zaliczyli zestawy danych, których zadaniem jest takie dostrojenie modelu językowego, żeby nie generował informacji nieprzydatnych, szkodliwych lub wprowadzających w błąd (Eigenbrode *et al.*, 2024). Ponad 80 tys. wpisów zawierających pytanie, dwie odpowiedzi oraz ocenę ich szkodliwości i użyteczności zawiera np. PKU-SafeRLHF (*PKU-Alignment/PKU-SafeRLHF · Datasets at Hugging Face*, 2024). Bardziej specjalistyczny charakter ma H4 Stack Exchange Preferences (*HuggingFaceH4/stack-exchange-preferences · Datasets at Hugging Face*, b.d.), który zawiera blisko 11 milionów pytań i adnotowanych pod kątem przydatności odpowiedzi pobranych z serwisu Stack Overflow (*Newest Questions*, b.d.).

Podobną klasyfikację zaproponowali Liu *et al.* (2024). Badacze wyróżnili jeszcze zestawy danych służące ewaluacji modeli oraz opracowane specjalnie do trenowania, optymalizacji i testowania tradycyjnych modeli językowych znajdujących

zastosowanie w takich operacjach, jak: ekstrakcja informacji, klasyfikacja czy podsumowywanie tekstu. W ewaluacji stosuje się m.in. FACTSCORE (Min *et al.*, 2023), którego zadaniem jest ocena poprawności faktograficznej twierdzeń generowanych przez modele językowe. W innych zadaniach zastosowanie znajdują takie zestawy danych, jak Stanford Question Answering Dataset (SQuAD) (*The Stanford Question Answering Dataset*, b.d.) czy Natural Questions (*google-research-datasets/natural-questions*, 2019/2025).

Badacze kategoryzują także zestawy według rodzaju pytania, które jest wprowadzane do systemu. Wyróżniane są m.in. zbiory danych, które obejmują pytania typu faktograficznego (Kto odkrył...), niefaktograficznego (Dlaczego model językowy halucynuje?) oraz pytania typu tak/nie (Baradaran *et al.*, 2022). Przykładem faktograficznego zestawu danych jest wspomniany SQuAD (*The Stanford Question Answering Dataset*, b.d.), który składa się z zestawu pytań i odpowiedzi opracowanych na podstawie artykułów z Wikipedii. Do zestawów zawierających pytania niefaktograficzne należy np. WikiHowQA (*Lurunchik/WikiHowNFQA · Datasets at Hugging Face*, b.d.) zawierający blisko 12 tys. pytań i odpowiedzi opartych na prawie 75 tysiącach dokumentów.

Innym kryterium podziału może być rodzaj odpowiedzi generowanej przez model. Baradaran *et al.* (2022) wyróżnili dwa rodzaje zestawów danych: generatywne i ekstrakcyjne. W modelu generatywnym system samodzielnie generuje odpowiedź na podstawie dostarczonego kontekstu. Ten rodzaj zestawu danych reprezentuje np. DuoRC (*DuoRC*, b.d.), który zawiera blisko 187 tysięcy zestawów pytań i odpowiedzi dotyczących fabuł filmowych opracowanych na podstawie Wikipedii i bazy IMDb. W przypadku modelu ekstrakcyjnego odpowiedzią modelu jest fragment tekstu pobrany z kontekstu. Do większych zestawów tego typu należy TriviaQA (*TriviaQA*, b.d.) licząca ponad 650 tysięcy pytań i odpowiedzi popartych materiałami źródłowymi.

Zestawy danych można także dzielić według metody ich tworzenia. Baradaran *et al.* (2022) wyróżnili zbiory na: generowane automatycznie, tworzone przez ludzi oraz hybrydowe. Przykładem zbioru opracowanego metodą crowdsourcingu jest SQuAD – zestaw zawierający ponad 100 tys. pytań i odpowiedzi (*The Stanford Question Answering Dataset*, b.d.). Z kolei podejście hybrydowe reprezentuje zestaw Natural Questions, który zawiera rzeczywiste pytania wpisywane przez użytkowników do wyszukiwarki Google oraz odpowiedzi zaczerpnięte z Wikipedii (*google-research-datasets/natural-questions*, 2019/2025). Na uwagę w kontekście metod gromadzenia i opracowywania danych do trenowania modeli językowych zasługuje głos Goetzego i Abramsona (2021), którzy zaakcentowali m.in. problemy związane z niedostateczną ochroną praw pracowniczych osób zatrudnianych do realizacji takich zadań.

Zaproponowana z kolei przez Rogers *et al.* (2023) klasyfikacja opiera się na kluczowych umiejętnościach wymaganych do rozwiązywania zadań typu *question*

answering i *reading comprehension*, takich jak wnioskowanie, wyszukiwanie informacji czy modelowanie rzeczywistości.

Przedstawiona w artykule analiza ilościowa wpisuje się w coraz szerszy nurt badań poświęconych korpusom danych stosowanym w treningu modeli językowych. Ich zawartość i aktualność mogą bowiem istotnie wpływać na jakość generowanej odpowiedzi (Rejeleene *et al.*, 2024). Jej celem jest zbadanie częstości wykorzystywania poszczególnych zestawów oraz schematów ich użycia przez różne typy modeli. Wyniki zaprezentowanej analizy mogą być pomocne przy podejmowaniu decyzji dotyczących doboru zestawów danych służących do szkolenia modeli językowych. Mogą również stanowić element charakterystyki samych modeli, ukazując stopień użycia nowych zasobów lub częstotliwość ponownego użycia tych samych zbiorów danych.

2. Stan badań

Badacze niejednokrotnie podejmowali analizy zestawów danych wykorzystywanych w szkoleniu modeli językowych. Rozszerzoną charakterystykę 60 anglojęzycznych zbiorów danych przeprowadzili Dzendzik *et al.* (2021). Obecność Wikipedii w ponad 38% z nich skłoniła autorów do postulatu rezygnacji z tego źródła przy tworzeniu przyszłych korpusów. Zwrócili również uwagę, że ponad 60% zbiorów powstało przy udziale crowdworkerów – osób o określonej biegłości językowej, lecz niebędących ekspertami w danej dziedzinie.

Również Zeng *et al.* (2020), analizując 57 zestawów danych, zwrócili uwagę na znaczenie Wikipedii jako źródła danych treningowych i fakt, że prawie 64% zestawów danych zostało opracowanych metodą crowdsourcingu.

Przeglądu artykułów naukowych poświęconych sztucznej inteligencji dokonali Baradaran *et al.* (2022). W analizowanych tekstach z lat 2016–2020 zidentyfikowali 53 zestawy danych. Tylko 10 z nich miało zamknięty charakter, ograniczony do określonej dziedziny, a większość (36 zestawów) obejmowała wyłącznie dane w języku angielskim.

156 wielojęzycznych publicznie dostępnych zestawów danych zbadali Yu *et al.* (2022). Zidentyfikowali w nich obecność 222 języków. Tylko dla 18% z nich został przygotowany specjalnie tekst wejściowy, w pozostałych przypadkach materiałem źródłowym były wiadomości prasowe, zasoby Wikipedii czy różnego typu zasoby WWW. Odnotowali ponadto, że najwięcej zbiorów dotyczyło zadań związanych z klasyfikacją tekstu.

Systematyczne analizy zestawów prowadzą badacze z Data Provenance Initiative (DPI) (*Data Provenance Initiative*, b.d.). W obszernym opracowaniu, obejmującym ponad 1800 tekstowych zbiorów danych, Longpre *et al.* (2024) wskazali na kryzys przejrzystości danych spowodowany brakiem odpowiedniej dokumentacji źródeł

ich pochodzenia. Utrudnia to zrozumienie specyfiki danych wykorzystywanych w treningu modeli językowych.

W kolejnym badaniu Longpre *et al.* (2025) przeanalizowali 3916 zestawów obejmujących tekst, mowę i dane wideo. Zauważyli, że coraz częściej dane te pochodzą z internetu lub są generowane syntetycznie, natomiast coraz rzadziej gromadzone są ręcznie. Zwrócili także uwagę na dużą liczbę zbiorów pozbawionych jasno określonej licencji.

Problemy dotyczące jakości danych podsumowali Paullada *et al.* (2021). Chodzi np. o brak dostatecznej reprezentacji w zbiorach danych określonych kategorii socjodemograficznych czy stronniczość w etykietowaniu danych. Kontrowersje budzą także naruszające prywatność użytkowników praktyki gromadzenia i przetwarzania danych pochodzących z wyszukiwarek internetowych czy mediów społecznościowych.

Problemy z jakością danych odnotowali m.in. Kreutzer *et al.* (2022), którzy zbadali pięć dużych zestawów (CCAligned, ParaCrawl, WikiMatrix, OSCAR, mC4). W 87 z 205 korpusów językowych zawartych we wspomnianych zestawach użyteczne było mniej niż 50% tekstu. Co więcej, w 15 korpusach nie odnotowano ani jednego poprawnego zdania.

Coraz poważniejszym problemem staje się także zalewająca internet fałszywa treść, którą przetwarzają systemy sztucznej inteligencji i która w konsekwencji zanieczyszcza wygenerowane przez nie odpowiedzi. Dziennikarze serwisu NewsGuard udowodnili na przykład, że dziesięć czołowych czatbotów (m.in. ChatGPT-4o, Copilot, Meta AI, Claude, Gemini) powielało tezy rosyjskiej propagandy publikowanej przez sieć powiązanych ze sobą stron internetowych (NewsGuard, 2024).

Wysokie koszty związane z tworzeniem treningowych zestawów danych, obawy dotyczące zachowania poufności czy niedostępność specjalistycznych danych skłaniają badaczy do wykorzystania dużych modeli językowych w procesie tworzenia danych syntetycznych (Rossi *et al.*, 2024; Guo & Chen, 2024).

Sun *et al.* (2024) wykorzystali opracowany przez OpenAI model text-davinci-003 do wygenerowania zapytań, na podstawie których wytrenowano konwersacyjny model MOSS. Tang *et al.* (2023) wykazali relatywnie wysoką skuteczność lokalnego modelu językowego wytrenowanego na syntetycznych danych wygenerowanych przez ChatGPT w rozpoznawaniu jednostek nazewniczych i relacji w tekstach medycznych. Na duży potencjał generatywnych modeli językowych (w tym przypadku ChatGPT) w tworzeniu danych treningowych na potrzeby szkolenia sieci neuronowych zwrócili także uwagę Askari *et al.* (2023). Dla odmiany Li *et al.* (2023) wykazali zróżnicowaną skuteczność modeli językowych wytrenowanych na danych stworzonych przez model GPT-3.5-Turbo do różnych zadań związanych z klasyfikacją tekstu.

Dane syntetyczne mogą być też wykorzystane do korygowania niepożądanych zachowań dużych modeli językowych. Wei *et al.* (2024) użyli ich, żeby ograniczyć

sytuację, kiedy model językowy dostosowuje swoje odpowiedzi do poglądów użytkownika. Hämäläinen *et al.* (2023) wykorzystali z kolei model GPT-3 do wygenerowania odpowiedzi użytkowników na temat postrzegania gier video jako sztuki i w konkluzji stwierdzili, że z uwagi na niski koszt i wysoką szybkość generowania danych, syntetyczne dane mogą być przydatne na etapie opracowywania pomysłów i wstępnych eksperymentów.

3. Cel analizy

Celem analizy było zbadanie częstości wykorzystania poszczególnych zestawów danych oraz schematów ich użycia w różnych typach modeli językowych. Zasadnicza część badania została poprzedzona ogólną charakterystyką ilościową modeli, obejmującą identyfikację ich autorów oraz przypisanie do kategorii zadań z zakresu przetwarzania języka naturalnego.

Celem głównej części analizy była odpowiedź na następujące pytania badawcze:

- Pytanie 1. Jakie zestawy danych i jak często były wykorzystywane w szkoleniu modeli językowych?
- Pytanie 2. Czy istnieje korelacja między liczbą unikalnych zestawów danych stosowanych przez poszczególnych autorów a łączną liczbą modeli, które udostępnili? Pytanie to miało na celu ocenę zróżnicowania wykorzystanych danych przez autorów modeli.
- Pytanie 3. Czy istnieje zależność między zestawami danych a kategoriami modeli, w których je wykorzystano? Zbadano, jak często poszczególne zestawy danych były stosowane w kontekście różnych kategorii zadań realizowanych przez modele. Przeanalizowano również stosunek liczby unikalnych zestawów danych do liczby modeli przypisanych do danej kategorii. Pozwoliło to ocenić różnorodność źródeł danych wykorzystywanych w konkretnych zastosowaniach.

4. Metoda badawcza

Analiza ma charakter ilościowy i polega na statystycznym opracowaniu metadanych modeli językowych dostępnych w repozytorium HF. Uwzględniono wyłącznie metadane tych modeli, które w polu tags zawierały informację o nazwie wykorzystanego zestawu danych.

Na potrzeby niniejszej analizy zestaw danych definiowany jest jako zbiór przeznaczony do szkolenia, dostrajania lub ewaluacji modelu językowego. Może obejmować jeden lub wiele zbiorów danych zawierających różne rodzaje informacji wykorzystywanych w procesie trenowania i oceny modelu. Przyjęto, że każda

nazwa zestawu danych występująca w opisie modelu stanowi unikalny identyfikator, odnoszący się do konkretnej wersji zbioru danych użytej przy jego tworzeniu.

Metadane modeli językowych zostały pobrane 6 marca 2025 roku za pomocą biblioteki huggingface-hub (wersja 0.29.3) służącej do wyszukiwania i zarządzania modelami w repozytorium HF.

Proces wyszukiwania modeli oraz kryteria selekcji przedstawiono na rysunku 1. Uwzględniono wyłącznie metadane modeli językowych zaprojektowanych do wykonywania zadań z zakresu przetwarzania języka naturalnego. Do charakterystyki zasobu wykorzystano następujące metadane: modelId (nazwa modelu), author (autor modelu), pipeline_tag (kategoria zadań NLP wykonywanych przez model), tags (słowa kluczowe zawierające charakterystykę modelu, w tym informację o nazwie zestawu danych).

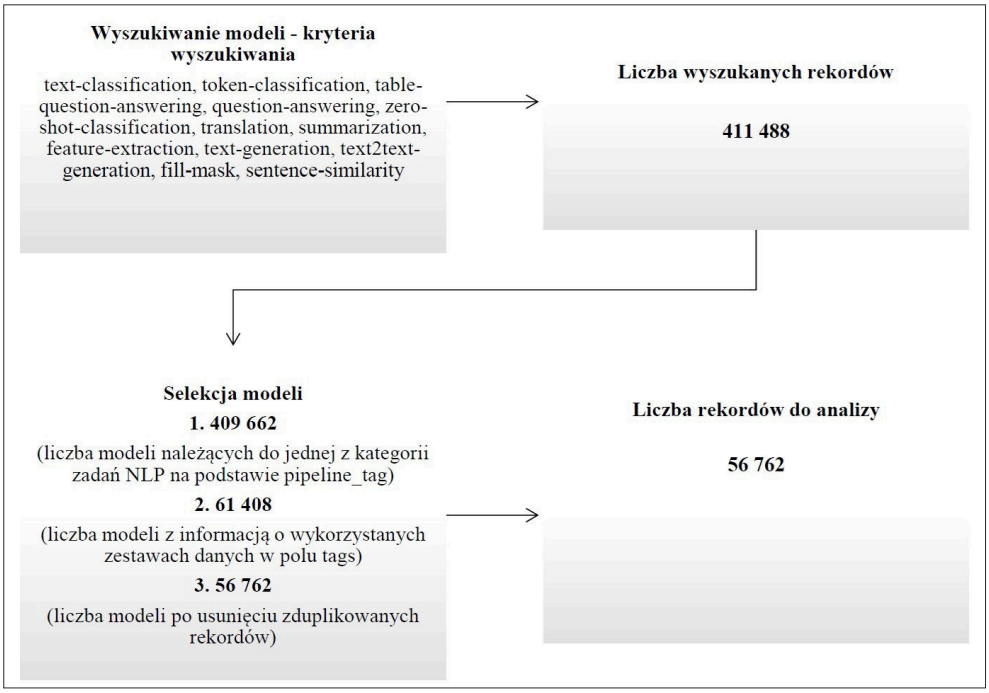
W analizie uwzględniono łącznie 56 762 modele (13,80% ogółu), w których podano informację o wykorzystanym zestawie danych.

W celu oceny reprezentatywności analizowanego zestawu (n=56 762) porównano rozkład kategorii modeli (pipeline_tag) z pełnym zbiorem odszukanych modeli (n=409 662). Test chi-kwadrat ($p < 0{,}0001$) wykazał istotne statystycznie różnice między rozkładami, co wskazuje na brak reprezentatywności próbki względem populacji. W szczególności stwierdzono nadmierną reprezentację kategorii token-classification (+4,17 punktu procentowego) i question-answering (+2,60) oraz niedostateczną reprezentację kategorii text-generation (–6,45). Należy to wziąć pod uwagę interpretując wyniki przedstawionej analizy.

W ramach analizy obliczono:

(1) Liczbę autorów udostępnionych modeli oraz średnią liczbę modeli przypadających na autora.
(2) Liczbę modeli należących do poszczególnych kategorii.
(3) Liczbę kategorii modeli językowych przypadających na jednego autora.
(4) Liczbę zestawów danych wykorzystanych do szkolenia modeli językowych oraz średnią liczbę zestawów przypadających na jeden model językowy.
(5) Korelację (współczynnik Pearsona) między liczbą zestawów danych stosowanych przez poszczególnych autorów a łączną liczbą udostępnionych przez nich modeli językowych.
(6) Częstość występowania zestawów danych w poszczególnych kategoriach modeli językowych.
(7) Stosunek liczby zestawów do ogólnej liczby modeli w danej kategorii.
(8) Średnią liczbę zestawów danych przypadających na modele z danej kategorii.

Analiza została przeprowadzona w języku Python (wersja 3.12.7) z wykorzystaniem biblioteki pandas (wersja 2.2.2), matplotlib (wersja 3.9.2), seaborn (wersja 0.13.2) oraz scipy (wersja 1.13.1).



Rys. 1. Selekcja modeli w repozytorium Hugging Face.

Źródło: opracowanie własne

5. Wyniki analizy

5.1. Ogólna charakterystyka modeli językowych

Ogółem 56 762 modele udostępniło w repozytorium 17 477 podmiotów. W Tabeli 1 przedstawiono wykaz 20 autorów z największą liczbą opublikowanych modeli. Maksymalna liczba modeli udostępnionych przez jeden podmiot to 1136, a średnia liczba modeli przypadających na jednego autora wynosi 3,25. Ponad 65% ogółu autorów opublikowało tylko jeden model, a odsetek tych, którzy udostępnili nie więcej niż 5, wynosi 92%.

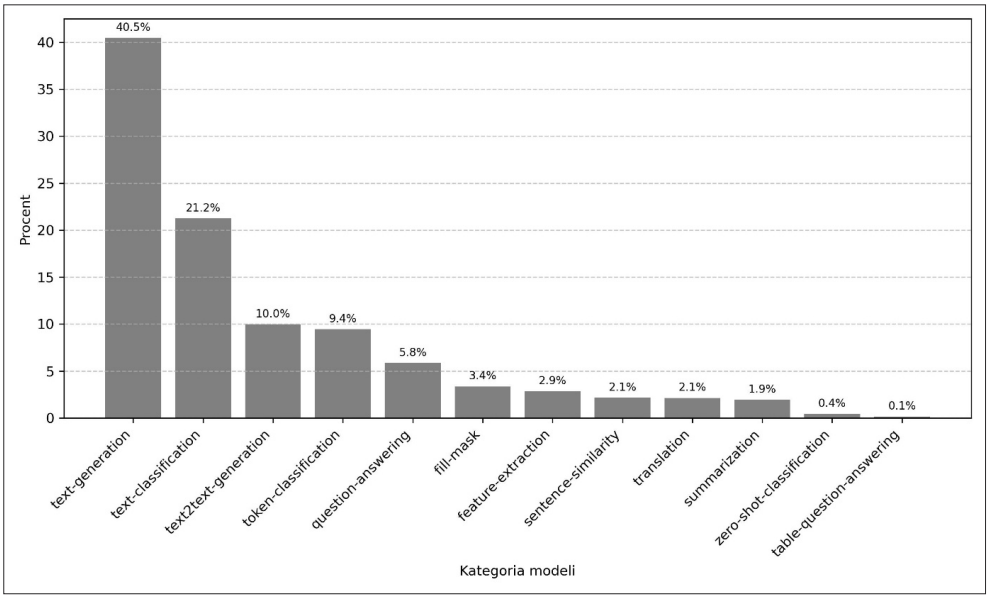
Modele przeznaczone do zadań związanych z klasyfikacją tekstu/tokenu/generowaniem tekstu stanowiły ponad 80% wszystkich analizowanych modeli językowych (Wykres 1). Znaczącą mniejszość (4%) stanowiły modele przeznaczone do zadań związanych z tłumaczeniami lub streszczeniami.

Tab. 1. Wykaz 20 autorów z największą liczbą udostępnionych modeli językowych

Nazwa autora	Liczba modeli	Procent ogółu modeli (n=56 762)
goldfish-models	1136	2,0
gokuls	1003	1,77
LoneStriker	953	1,68
TheBloke	849	1,5
research-backup	709	1,25
TransferGraph	597	1,05
fine-tuned	508	0,89
bartowski	507	0,89
gokulsrinivasagan	484	0,85
tensorblock	390	0,69
NasimB	381	0,67
anas-awadalla	373	0,66
ShenaoZ	312	0,55
blockblockblock	308	0,54
fbaldassarri	271	0,48
CharlesLi	269	0,47
google	264	0,47
KoichiYasuoka	264	0,47
MaziyarPanahi	225	0,4
Triangle104	210	0,37
Inni	46 749	82,36

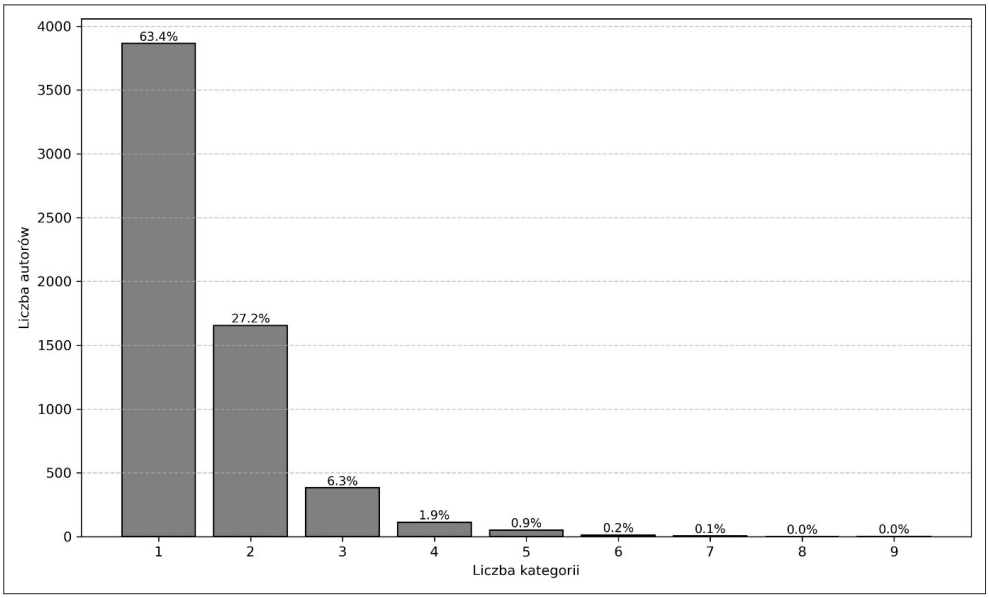
Źródło: opracowanie własne

Po wykluczeniu podmiotów, które opublikowały tylko jeden model, największa grupa autorów (niemal 64%) (Wykres 2) udostępniła modele należące wyłącznie do jednej kategorii. Nieliczni autorzy zdecydowali się opracować modele obejmujące trzy lub więcej kategorii.



Wykres 1. Podział modeli według kategorii

Źródło: opracowanie własne



Wykres 2. Rozkład autorów według liczby kategorii modeli językowych (tylko autorzy z ≥2 modelami)

Źródło: opracowanie własne

5.2. Jakie zestawy danych i jak często były wykorzystywane w szkoleniu modeli językowych?

Ogółem w analizowanym zbiorze zidentyfikowano 13 376 unikalnych zestawów danych. Tabela 2 zawiera wykaz 20 najczęściej cytowanych zbiorów. Każdy model średnio wykorzystywał mniej niż dwa zestawy (1,74).

Tab 2. Wykaz 20 najczęściej wykorzystywanych zestawów danych

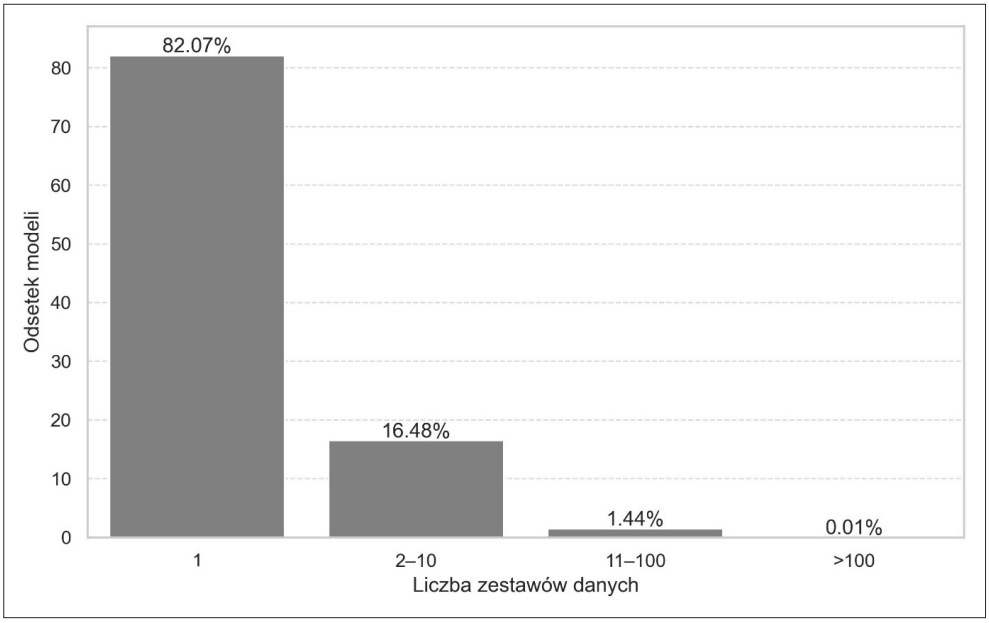
Nazwa zestawu	Liczba wystąpień
glue**	2458
emotion	1750
squad	1736
imdb	1298
generator*	1233
xtreme	1126
conll2003	1077
wikipedia	866
cis-lmu/Glot500	783
allenai/MADLAD-400	683
allenai/c4	670
tweet_eval	622
samsum	609
allenai/nllb	584
HuggingFaceH4/ultrafeedback_binarized	581
oscar-corpus/OSCAR-2109	574
teknium/OpenHermes-2.5	566
squad_v2	538
kde4	509

* Taką nazwą oznaczono zestaw danych użyty do wyszkolenia modelu, ale nie identyfikuje on żadnego konkretnego zbioru. W charakterystyce modelu zawarto np. określenie: „This model is a fine-tuned version of ShenaoZhang/0.001_idpo_iter_1 on the updated and the original datasets".

** Zestaw bez znaku ukośnika w nazwie nie został udostępniony w HF. Autorzy mogli w tym przypadku wykorzystać zasoby spoza HF.

Źródło: opracowanie własne.

Ponad 80% modeli (Wykres 3) opierało się tylko na jednym zestawie danych. Niecałe 17% wykorzystywało od 2 do 10 zbiorów. Znaczącą mniejszość stanowiły modele trenowane na więcej niż 10 zestawach. W przypadku 6 modeli wykorzystano 200 i więcej zestawów danych. Najwięcej zbiorów, 289, zostało użytych do wytrenowania modelu tasksource/deberta-small-long-nli (*tasksource/deberta-small-long-nli · Hugging Face*, 2023).

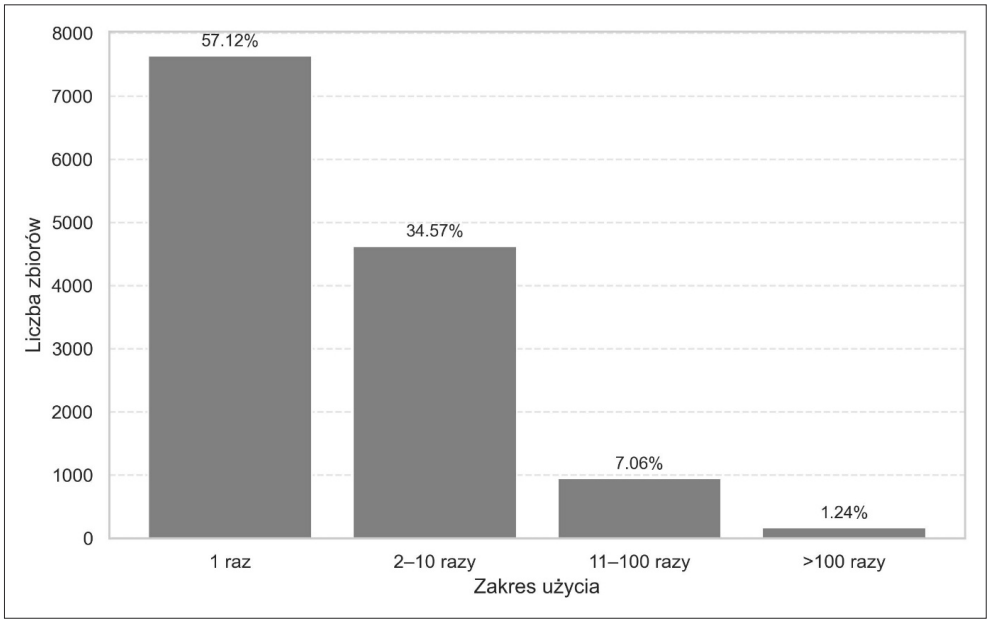


Wykres 3. Liczba zestawów danych wykorzystywanych przez modele językowe

Źródło: opracowanie własne

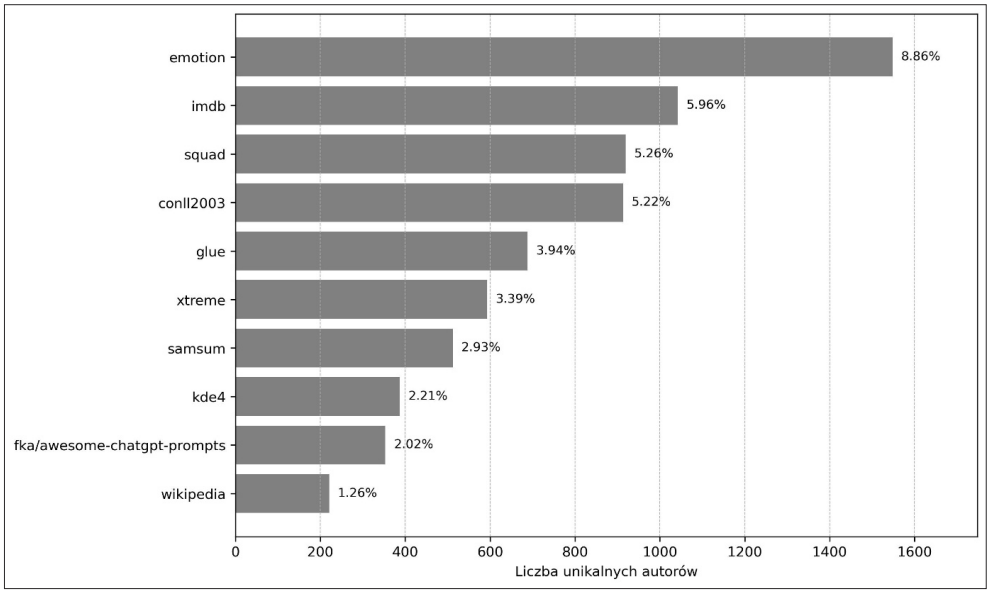
Blisko 58% zestawów danych zostało wykorzystanych tylko raz w szkoleniu modeli językowych (Wykres 4). Generalnie, znaczącą (prawie 92%) grupę stanowią zestawy, które wykorzystano najwyżej dziesięciokrotnie. Niewielki odsetek, nieprzekraczający 1,3%, to zestawy, które zostały wykorzystane ponad 100 razy.

Znaczenie pewnych zestawów danych w szkoleniu modeli językowych ilustruje Wykres 5 z odsetkiem autorów wykorzystujących dany zestaw. Blisko 9% autorów wykorzystało zestaw emotion, siedmiokrotnie więcej niż popularną Wikipedię. Powyżej 5% podmiotów wyszkoliło model na bazie zestawów imdb i squad. Do najczęściej wymienianego zbioru glue odwołało się mniej niż 4% autorów.



Wykres 4. Częstość wykorzystania zestawów danych przez modele językowe

Źródło: opracowanie własne



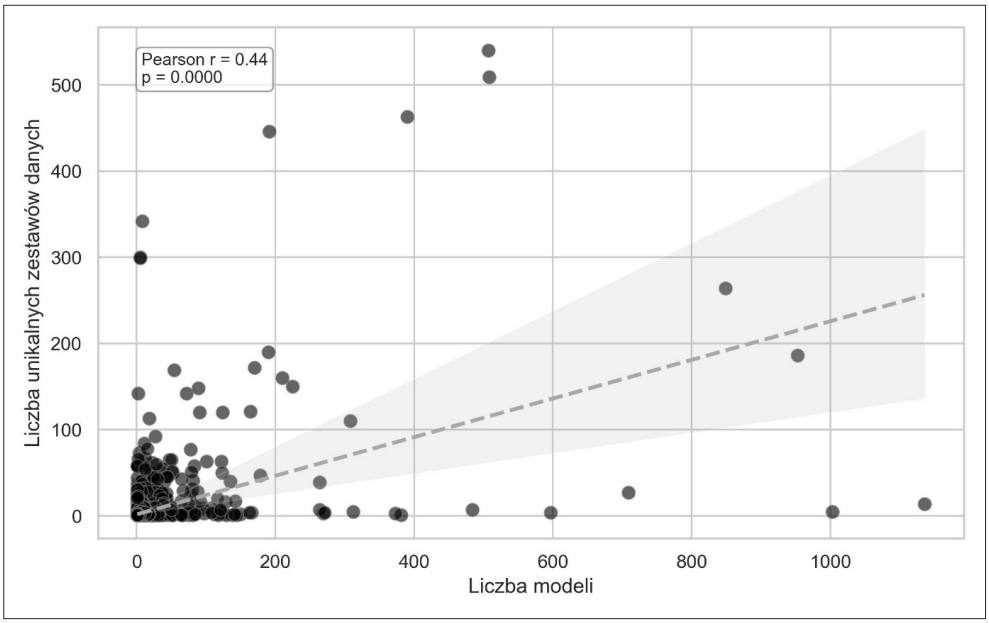
Wykres 5. Dziesięć zestawów danych wykorzystywanych przez największą liczbę autorów modeli

Źródło: opracowanie własne

5.3. Czy istnieje korelacja między liczbą unikalnych zestawów danych stosowanych przez poszczególnych autorów a łączną liczbą modeli, które udostępnili?

Współczynnik korelacji Pearsona (Wykres 6) między liczbą wykorzystanych unikalnych zestawów danych a liczbą modeli językowych udostępnionych przez danego autora wynosi 0,44 (p=0,0000), co wskazuje na umiarkowanie dodatnią korelację. Oznacza to, że wśród autorów publikujących większą liczbę modeli językowych widoczna jest pewna tendencja do wykorzystywania bardziej zróżnicowanych zestawów danych. Nie jest to jednak silna zależność.

Odmienne pod względem funkcji modele językowe mogą bowiem wykorzystywać zróżnicowane zestawy danych, zarówno pod względem ich rodzaju, jak i liczby. Jednocześnie kolejne wersje tego samego modelu zwykle opierają się na identycznych zestawach danych, co może ograniczać różnorodność danych wykorzystywanych przez autora.

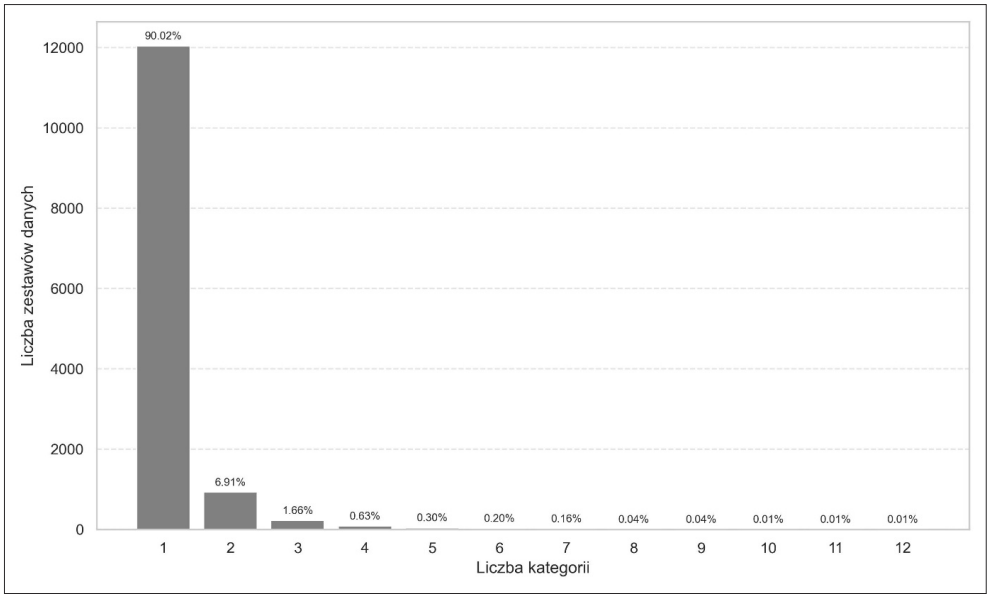


Wykres 6. Korelacja między liczbą wykorzystanych unikalnych zestawów danych a liczbą modeli językowych udostępnionych przez danego autora

Źródło: opracowanie własne

5.4. Czy istnieje zależność między zestawami danych a kategoriami modeli, w których je wykorzystano?

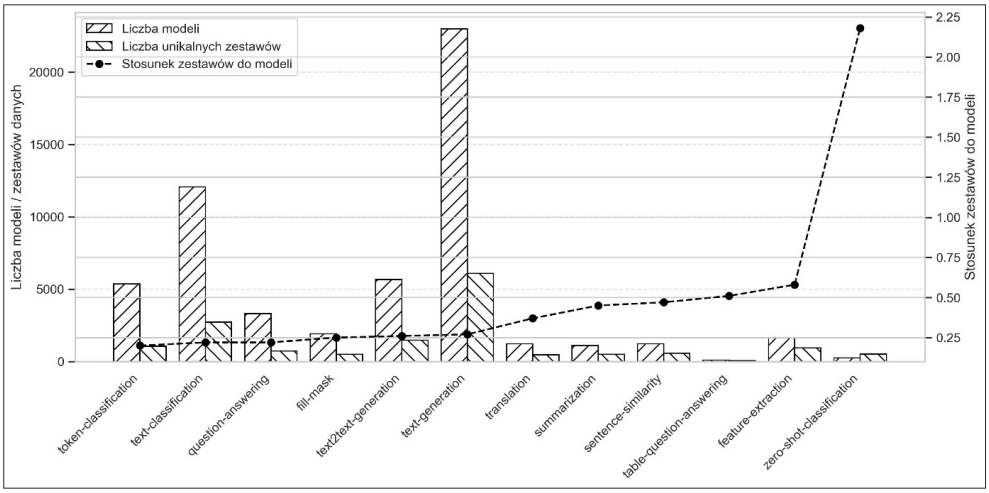
Zdecydowana większość zestawów danych, jak wynika z Wykresu 7, była wykorzystywana tylko w jednej kategorii modeli. Mniej niż pół procenta zestawów zostało przypisanych do więcej niż 5 kategorii. Zazwyczaj jeden zestaw danych występuje powszechnie tylko w jednej kategorii, a w innych bywa wykorzystywany już sporadycznie. Na przykład na 1298 wystąpień zestawu imdb prawie 74% przypada na kategorię text classification. We wszystkich 12 kategoriach znajduje się jedynie fka/awesome-chatgpt-prompts (*fka/awesome-chatgpt-prompts · Datasets at Hugging Face*, 2024). Większość z kategorii charakteryzuje użycie jednego dominującego zestawu danych. Dla modeli z grupy feature extraction jest to allenai/c4, a w modelach realizujących zadania z grupy text classification glue. Z kolei w kategorii question answering widoczna jest dominacja zestawu SQuAD.



Wykres 7. Wykorzystanie zestawów danych w kategoriach modeli językowych

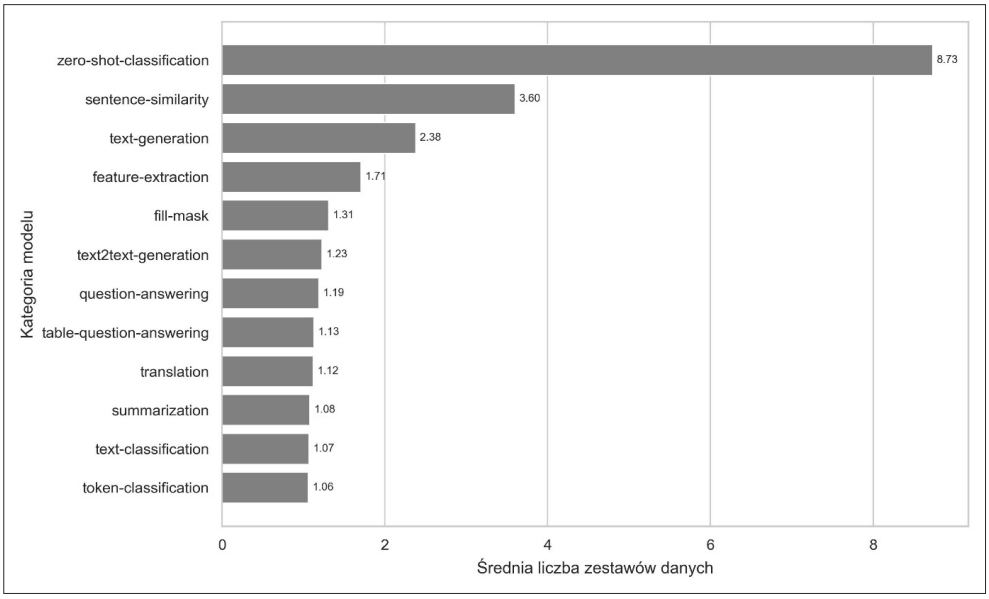
Źródło: opracowanie własne

W tych kategoriach, w których zarejestrowano 5000 i więcej modeli, stosunek liczby unikalnych zestawów do liczby modeli jest stosunkowo niski i średnio wynosi niecałe 0,24 (Wykres 8). W przypadku pozostałych kategorii (wyjąwszy zero shot classification) ten wskaźnik osiąga wartość blisko 0,41. Razem z kategorią zero shot classification będzie on jeszcze większy, ale w tym przypadku kilka modeli było szkolonych na blisko 300 zestawach danych, co znacząco podnosi wspomniany wskaźnik.



Wykres 8. Stosunek liczby unikalnych zestawów danych do liczby modeli w danej kategorii

Źródło: opracowanie własne



Wykres 9. Średnia liczba zestawów danych przypadających na modele z danej kategorii

Źródło: opracowanie własne

Średnia liczba zestawów wykorzystywanych przez modele językowe (Wykres 9), nie jest wysoka i dla 9 kategorii nie przekracza 2. Jedynie w modelach z grupy sentence similarity i text generation wynosi od 2,38 do 3,6. Wyjątkiem są modele typu zero shot classification, gdzie średnia wynosi prawie 9, ale jak wspomniano wcześniej, w kilku modelach z tej grupy wykorzystano blisko 300 zestawów danych.

6. Zakończenie

Analiza abstrahuje od celu użycia określonego zestawu danych (szkolenie, dostrajanie, ewaluacja). Koncentruje się na skali wykorzystania zestawów danych w trenowaniu modeli językowych różnego rodzaju. Zdecydowana większość z nich jest szkolona tylko na podstawie jednego zestawu danych. Relatywnie duża liczba zidentyfikowanych zestawów danych, które zostały wykorzystane tylko raz, może wskazywać na ich specjalistyczny charakter. Liczne użycie określonych zbiorów można interpretować dwojako: jako potwierdzenie ich skuteczności, ale też jako oznakę braku alternatywnych jakości danych o odpowiedniej jakości. Można również zaobserwować pewną prawidłowość polegającą na tym, że w kategoriach z mniejszą liczbą modeli występuje większe zróżnicowanie wykorzystywanych do ich szkolenia zestawów danych.

Szereg ograniczeń wpływa na dokładność wyników przedstawionej analizy. Zbadano tylko niecałe 14% modeli zawierających nie zawsze kompletne metadane. Brakuje w nich niejednokrotnie nazw wszystkich wykorzystanych zestawów danych, jak w przypadku modelu RoBERTa (*FacebookAI/roberta-base · Hugging Face*, 2024). W tym przypadku kompletna charakterystyka modelu, jeśli chodzi o zestawy danych, jest dostępna dopiero w opisującej go publikacji (Liu *et al.*, 2019).

Warto zaznaczyć, że wielu autorów publikuje różne warianty jednego modelu, korzystając przy tym z tych samych zbiorów danych. Każdy wariant, będący odrębną pozycją, nieco sztucznie zawyża więc statystykę wykorzystania określonych zasobów.

Wprawdzie w repozytorium różne modele teoretycznie mogą odwoływać się do tego samego zestawu danych, jednak w praktyce często korzystają z różnych wersji tych danych. Dlatego też informację o zestawach danych należy traktować raczej jako wskazówkę dotyczącą ich pochodzenia, a nie dokładne określenie użytej wersji.

Bibliografia

Arif, A. (2023). Introduction to Large Language Models (LLMs): An overview of BERT, GPT, and other popular models. *John Snow Labs*. https://www.johnsnowlabs.com/introduction-to-large-language-models-llms-an-overview-of-bert-gpt-and-other-popular-models/

arXiv.org e-Print archive. (b.d.). https://arxiv.org/

Askari, A., Aliannejadi, M., Kanoulas, E., Verberne, S. (2023). *Generating synthetic documents for cross-encoder re-rankers: A aomparative study of ChatGPT and human experts* (No. arXiv:2305.02320). arXiv. https://doi.org/10.48550/arXiv.2305.02320

Auffarth, B. (2024). *Generatywna sztuczna inteligencja z LangChain: Budowanie aplikacji AI opartych na LLM z użyciem Pythona, ChatGPT i innych modeli językowych*. Helion.

Baradaran, R., Ghiasi, R., & Amirkhani, H. (2022). A survey on machine reading comprehension systems. *Natural Language Engineering*, *28*(6), 683–732. https://doi.org/10.1017/S1351324921000395

Common Crawl—Open Repository of Web Crawl Data. (b.d.). https://commoncrawl.org/

Data Provenance Initiative. (b.d.). https://www.dataprovenance.org/about

Dodge, J., Sap, M., Marasović, A., Agnew, W., Ilharco, G., Groeneveld, D., Mitchell, M., Gardner, M. (2021). *Documenting large webtext corpora: a case study on the Colossal Clean Crawled Corpus* (No. arXiv:2104.08758). arXiv. https://doi.org/10.48550/arXiv.2104.08758

DuoRC. (b.d.). https://duorc.github.io/

Dzendzik, D., Foster, J., Vogel, C. (2021). English machine reading comprehension datasets: A survey. In M.-F. Moens, X. Huang, L. Specia, S. W. Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (pp. 8784–8804). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.693

Eigenbrode, S., Barth, A., Fregly, C., Watrak, A. (red.). (2024). *Generatywna sztuczna inteligencja na platformie AWS: Tworzenie multimodalnych aplikacji wnioskujących kontekstowo*. Helion.

FacebookAI/roberta-base · Hugging Face. https://huggingface.co/FacebookAI/roberta-base

Feature extraction with LLMs, Hugging Face and MinIO. (b.d.). https://www.e2enetworks.com/blog/feature-extraction-with-large-language-models-hugging-face-and-minio

Fka/awesome-chatgpt-prompts · Datasets at Hugging Face. (2024). https://huggingface.co/datasets/fka/awesome-chatgpt-prompts

GitHub · Build and ship software on a single, collaborative platform. (b.d.). https://github.com/

Goetze, T. S., Abramson, D. (2021). Bigger isn't better: The ethical and scientific vices of extra-large datasets in language models. *Companion Publication of the 13th ACM Web Science Conference 2021*, 69–75. https://doi.org/10.1145/3462741.3466809

Google-research-datasets/natural-questions. (2025). [Dataset]. Google Research Datasets. https://github.com/google-research-datasets/natural-questions (Original work published 2019)

Guo, X., Chen, Y. (2024). *Generative AI for synthetic data generation: methods, challenges and the future* (No. arXiv:2403.04190). arXiv. https://doi.org/10.48550/arXiv.2403.04190

Hämäläinen, P., Tavast, M., Kunnari, A. (2023). Evaluating Large Language Models in generating synthetic HCI research data: A case study. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–19. https://doi.org/10.1145/3544548.3580688

HuggingFaceH4/stack-exchange-preferences · Datasets at Hugging Face. (b.d.). https://huggingface.co/datasets/HuggingFaceH4/stack-exchange-preferences

Jordon, J., Szpruch, L., Houssiau, F., Bottarelli, M., Cherubin, G., Maple, C., Cohen, S. N., Weller, A. (2022). *Synthetic Data—What, why and how?* (No. arXiv:2205.03257). arXiv. https://doi.org/10.48550/arXiv.2205.03257

Kreutzer, J., Caswell, I., Wang, L., Wahab, A., van Esch, D., Ulzii-Orshikh, N., Tapo, A., Subramani, N., Sokolov, A., Sikasote, C., Setyawan, M., Sarin, S., Samb, S., Sagot, B., Rivera, C., Rios, A., Papadimitriou, I., Osei, S., Suarez, P. O., ... Adeyemi, M. (2022). Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, *10*, 50–72. https://doi.org/10.1162/tacl_a_00447

Li, Z., Zhu, H., Lu, Z., Yin, M. (2023). *Synthetic data generation with Large Language Models for text classification: potential and limitations* (No. arXiv:2310.07849). arXiv. https://doi.org/10.48550/arXiv.2310.07849

Liu, Y., Cao, J., Liu, C., Ding, K., Jin, L. (2024). *Datasets for Large Language Models: A comprehensive survey* (No. arXiv:2402.18041). arXiv. https://doi.org/10.48550/arXiv.2402.18041

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V. (2019). *RoBERTa: A Robustly Optimized BERT Pretraining Approach* (No. arXiv:1907.11692). arXiv. https://doi.org/10.48550/arXiv.1907.11692

Longpre, S., Mahari, R., Chen, A., Obeng-Marnu, N., Sileo, D., Brannon, W., Muennighoff, N., Khazam, N., Kabbara, J., Perisetla, K., Wu, X. (Alexis), Shippole, E., Bollacker, K., Wu, T., Villa, L., Pentland, S., Hooker, S. (2024). A large-scale audit of dataset licensing and attribution in AI. *Nature Machine Intelligence*, *6*(8), 975–987. https://doi.org/10.1038/s42256-024-00878-8

Longpre, S., Singh, N., Cherep, M., Tiwary, K., Materzynska, J., Brannon, W., Mahari, R., Obeng-Marnu, N., Dey, M., Hamdy, M., Saxena, N., Anis, A. M., Alghamdi, E. A., Chien, V. M., Yin, D., Qian, K., Li, Y., Liang, M., Dinh, A., ... Kabbara, J. (2025). *Bridging the data provenance gap across text, speech and video* (No. arXiv:2412.17847). arXiv. https://doi.org/10.48550/arXiv.2412.17847

Lurunchik/WikiHowNFQA · Datasets at Hugging Face. (b.d.). https://huggingface.co/datasets/Lurunchik/WikiHowNFQA

Min, S., Krishna, K., Lyu, X., Lewis, M., Yih, W., Koh, P., Iyyer, M., Zettlemoyer, L., Hajishirzi, H. (2023). FActScore: Fine-grained Atomic Evaluation of Factual Precision in long form text generation. In H. Bouamor, J. Pino, K. Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 12076–12100). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.741

Models—Hugging Face. (2025). https://huggingface.co/models

Newest Questions. (b.d.). Stack Overflow. https://stackoverflow.com/questions

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., ... Zoph, B. (2024). *GPT-4 Technical Report* (No. arXiv:2303.08774). arXiv. https://doi.org/10.48550/arXiv.2303.08774

Paullada, A., Raji, I. D., Bender, E. M., Denton, E., Hanna, A. (2021). Data and its (dis)contents: A survey of dataset development and use in machine learning research. *Patterns*, *2*(11). https://doi.org/10.1016/j.patter.2021.100336

PKU-Alignment/PKU-SafeRLHF · Datasets at Hugging Face. (2024). https://huggingface.co/datasets/PKU-Alignment/PKU-SafeRLHF

Rejeleene, R., Xu, X., Talburt, J. (2024). Towards trustable Language Models: Investigating information quality of Large Language Models (No. arXiv:2401.13086). arXiv. https://doi.org/10.48550/arXiv.2401.13086

Rogers, A., Gardner, M., Augenstein, I. (2023). QA dataset explosion: A taxonomy of NLP resources for question answering and reading comprehension. *ACM Comput. Surv.*, *55*(10), 197:1–197:45. https://doi.org/10.1145/3560260

Rossi, L., Harrison, K., Shklovski, I. (2024). The problems of LLM-generated data in social science research. *Sociologica*, *18*(2), Article 2. https://doi.org/10.6092/issn.1971–8853/19576

Samsung/samsum · Datasets at Hugging Face. (2022). https://huggingface.co/datasets/Samsung/samsum

Sun, T., Zhang, X., He, Z., Li, P., Cheng, Q., Liu, X., Yan, H., Shao, Y., Tang, Q., Zhang, S., Zhao, X., Chen, K., Zheng, Y., Zhou, Z., Li, R., Zhan, J., Zhou, Y., Li, L., Yang, X., ... Qiu, X. (2024). MOSS: An open conversational Large Language Model. *Machine Intelligence Research*, *21*(5), 888–905. https://doi.org/10.1007/s11633-024-1502-8

Tang, R., Han, X., Jiang, X., & Hu, X. (2023). *Does Synthetic Data Generation of LLMs Help Clinical Text Mining?* (No. arXiv:2303.04360). arXiv. https://doi.org/10.48550/arXiv.2303.04360

Tasksource/deberta-small-long-nli · Hugging Face. (2023). https://huggingface.co/tasksource/deberta-small-long-nli

Tatsu-lab/alpaca · Datasets at Hugging Face. (2024). [Dataset]. https://huggingface.co/datasets/tatsu-lab/alpaca

The Stanford Question Answering Dataset. (b.d.). https://rajpurkar.github.io/SQuAD-explorer/

TriviaQA. (b.d.). https://nlp.cs.washington.edu/triviaqa/

Wei, J., Huang, D., Lu, Y., Zhou, D., Le, Q. V. (2024). *Simple synthetic data reduces sycophancy in large language models* (No. arXiv:2308.03958). arXiv. https://doi.org/10.48550/arXiv.2308.03958

What is Feature Extraction? – Hugging Face. (2024). https://huggingface.co/tasks/feature-extraction

What is Fill-Mask? – Hugging Face. (2025). https://huggingface.co/tasks/fill-mask

What is Question Answering? | IBM. (2025). https://www.ibm.com/think/topics/question-answering

What is Sentence Similarity? – Hugging Face. (2025). https://huggingface.co/tasks/sentence-similarity

What is Table Question Answering? – Hugging Face. (2023). https://huggingface.co/tasks/table-question-answering

What is Text Classification? – Hugging Face. (2023). https://huggingface.co/tasks/text-classification

What is Text Generation? – Hugging Face. (2024). https://huggingface.co/tasks/text-generation

What is Token Classification? – Hugging Face. (2024). https://huggingface.co/tasks/token-classification

What is Zero-Shot Classification? – Hugging Face. (2023). https://huggingface.co/tasks/zero-shot-classification

Wikimedia Downloads. (b.d.). https://dumps.wikimedia.org/

Yu, X. V., Asai, A., Chatterjee, T., Hu, J., Choi, E. (2022). *Beyond counting datasets: A survey of multilingual dataset construction and necessary resources* (No. arXiv:2211.15649). arXiv. https://doi.org/10.48550/arXiv.2211.15649

Zeng, C., Li, S., Li, Q., Hu, J., Hu, J. (2020). *A survey on machine reading comprehension: tasks, evaluation metrics and benchmark datasets* (No. arXiv:2006.11880). arXiv. https://doi.org/10.48550/arXiv.2006.11880

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., ... Wen, J.-R. (2025). *A survey of Large Language Models* (No. arXiv:2303.18223). arXiv. https://doi.org/10.48550/arXiv.2303.18223

Quantitative analysis of datasets and the patterns of their use in language models

Abstract

Purpose/Thesis: The objective of this analysis was to provide a quantitative characterisation of the datasets used for training language models. Specifically, the study encompassed: 1) an analysis of the frequency of use of individual datasets and the patterns of their usage across different types of models; 2) an assessment of the diversity of data employed by model developers; and 3) an examination of the variety in dataset usage in relation to specific applications of language models.

Approach/Methods: The metadata of language models available in the Hugging Face repository was analysed. Only models that included information about the name of the dataset used – understood as a collection used for training, fine-tuning, or evaluating a language model – were considered. In total, 56 762 models (13.8% of all retrieved) were analysed, with metadata collected on March 6, 2025, using the huggingface-hub library. The analysis included, among others, the number of models and their authors, the frequency of dataset use depending on model category, and the correlation between the number of unique datasets used by each author and the total number of models they released. The analysis was conducted in Python using the pandas, matplotlib, seaborn, and scipy.

Results and Conclusions: A total of 13 376 datasets were identified. Most models were trained using only a single dataset. A tendency toward greater dataset diversity was observed among authors who released a larger number of language models. It was also observed that categories with fewer models exhibited greater diversity in the datasets used for training.

Originality/Value: The analysis reveals specific patterns concerning the degree of diversity in dataset usage, depending on the author and the category of the language model.

Keywords:

Artificial intelligence. Dataset. Hugging Face. Language model.

ADAM JACHIMCZYK jest adiunktem w Katedrze Bibliografii i Dokumentacji na Wydziale Dziennikarstwa, Informacji i Bibliologii Uniwersytetu Warszawskiego. Jego zainteresowania naukowe obejmują zastosowanie technologii informatycznej w działalności informacyjnej. Opublikował m.in.: razem z M. Chrapek i Z. Chrapkiem: Web directories: selected features and their impact on directory quality (2016). Program-Electronic Library and Information Systems 2016, vol. 50 no 3; Ochrona znaków towarowych przez biblioteki w świetle zawartości bazy danych TMview (2019). Przegląd Biblioteczny, 87(4); Patent Applications for Electronic Publishing Market (2014–2018). Selected Issues (2019). Zagadnienia Informacji Naukowej – Studia Informacyjne, 57(2a).

Kontakt do autora:
a.jachimczyk@uw.edu.pl
Uniwersytet Warszawski
Wydział Dziennikarstwa, Informacji i Bibliologii
Nowy Świat 69, 00-927 Warszawa