7.0

ZBIORY DANYCH I MODELE JĘZYKOWE STOSOWANE W JĘZYKU POLSKIM DO IDENTYFIKACJI STWIERDZEŃ WYMAGAJĄCYCH WERYFIKACJI

Krzysztof Węcel
Uniwersytet Ekonomiczny w Poznaniu
Marcin Sawiński
Uniwersytet Ekonomiczny w Poznaniu
Ewelina Księżniak
Uniwersytet Ekonomiczny w Poznaniu
Milena Stróżyna
Uniwersytet Ekonomiczny w Poznaniu

Streszczenie

W rozdziale opisano metodykę tworzenia zbioru tekstów w języku polskim oraz trenowania modeli do wsparcia procesu weryfikacji informacji pod kątem ich prawdziwości i dokładności. Opracowany został protokół oznaczania tekstów jako „check-worthy". Posłużył on do opracowania własnego zbioru w języku polskim, wykorzystanego następnie do trenowania i ewaluacji wybranych modeli z różnymi wariantami zbiorów danych, uwzględniających zasoby w języku polskim oraz mieszane. Przeprowadzone badania wskazują, że modele przeznaczone dla języka polskiego oraz modele wielojęzyczne miały zbliżoną dokładność. Ważną rolę odgrywa sposób doboru przykładów do treningu. Uzasadnione jest wykorzystywanie modeli wielojęzycznych, umożliwiających trenowanie na większej liczbie zasobów obcojęzycznych, i dotrenowanie

Rozdział z monografii: Doligalski T., Kaszyński D. (red. nauk.), Sztuczna inteligencja w przedsiębiorstwach i gospodarce (AI Spring 2024), Warszawa 2025.

Monografia oferowana jest w swobodnym dostępie do pobrania na stronie AI Lab SGH.

na wysokiej jakości przykładach w języku polskim. Wynikiem badań jest wysokiej jakości zbiór sklasyfikowanych tekstów w języku polskim do trenowania modeli oraz najlepszy model dla języka polskiego służący do określania, czy dane stwierdzenie wymaga weryfikacji.

Słowa kluczowe: jakość informacji, wiarygodność informacji, modele językowe, fact-checking

Wprowadzenie

Jednym z kluczowych wymogów dla skutecznego zarządzania w firmie jest dostęp do prawdziwej, wiarygodnej, aktualnej i relewantnej informacji, którą określamy również jako informację wysokiej jakości. We współczesnym świecie, mimo nadmiaru informacji, obserwujemy coraz większy problem z dostępem do jakościowej informacji. Z jednej strony jest to wynikiem demokratyzacji dostępu do tworzenia treści, gdzie każdy może opublikować dowolną informację. Z różnych pobudek może być ona jednak nieprawdziwa, a celem jej tworzenia może okazać się chęć uzyskania rozgłosu, osiągnięcia konkretnych korzyści czy wywołania chaosu wśród konkurentów. Z drugiej strony mamy do czynienia z intensywnym rozwojem metod sztucznej inteligencji, wykorzystanej do automatycznego tworzenia treści. W zależności od intencji one również mogą być nieprawdziwe. Znane jest także zjawisko spontanicznego tworzenia nieprawdziwych treści przez sztuczną inteligencję, określane jako halucynacje.

Zalew fałszywych informacji w Internecie stanowi wyzwanie zarówno dla działania firm (np. nieuczciwe opinie o podmiocie), podejmowania decyzji przez konsumentów (np. nadmiernie optymistyczne opisy produktów), jak i dla bezpieczeństwa publicznego (np. ruchy antyszczepionkowe). Coraz bardziej zwraca się uwagę na systemową walkę z tym zjawiskiem, czego dowodem są niedawne regulacje unijne (Digital Services Act, DSA), które wprowadzają szereg obowiązków w zakresie przeciwdziałania rozprzestrzenianiu się dezinformacji. Na przykład bardzo duże platformy internetowe i wyszukiwarki (gdy liczba użytkowników przekracza 10% ludności UE) są zobowiązywane do usuwania dezinformujących treści.

W zapobieganiu rozprzestrzeniania dezinformacji istotną rolę odgrywają tzw. agencje fact-checkingowe, które zawodowo zajmują się monitorowaniem i weryfikacją prawdziwości informacji. Fact-checking oznacza rzetelny proces weryfikacji informacji pod kątem ich prawdziwości i dokładności. Obejmuje on krytyczną analizę poszczególnych stwierdzeń lub danych prezentowanych w różnych formach i mediach. Ele-

mentem tej oceny jest wcześniejszy wybór, z pojedynczego dokumentu lub całego zbioru, stwierdzeń wymagających weryfikacji (*check-worthy*). Są to takie stwierdzenia, które faktycznie powinny zostać poddane procesowi fact-checkingu ze względu na swoje potencjalne oddziaływanie na decyzje innych oraz prawdopodobieństwo tego, że są to stwierdzenia fałszywe.

7.1. Zagadnienie klasyfikacji treści tekstowych

Fact-checking pierwotnie rozumiany był jako potwierdzenie dokładności informacji, zamieszczonych w artykule przed samą jego publikacją. A zatem był to proces wewnętrzny, związany z odpowiedzialnością dziennikarzy za swoją pracę. Z czasem znaczenie tego pojęcia zostało rozszerzone o analizę wiadomości już po ich opublikowaniu, przede wszystkim w Internecie, wykonywaną przez osoby, które nie były autorami pierwotnego tekstu [Cazalens i in., 2018].

Proces fact-checkingu składa się z kilku etapów. Pierwszym jest wybór stwierdzeń do weryfikacji. Po nim następują kontekstualizacja i analiza, konfrontacja z zewnętrznymi bazami danych oraz wiedzą ekspertów. Całość kończy przygotowanie opracowania z przeprowadzonych badań wraz z podjęciem decyzji co do oceny informacji, np. prawda, fałsz, manipulacja, brak kontekstu [Micallef i in., 2022].

Głównym wyzwaniem obecnie jest to, że większość zadań w ramach powyższego procesu wykonywana jest manualnie. Ze względu na ilość informacji, które wymagają weryfikacji, wskazana jest jego automatyzacja. Jako lukę badawczą należy zatem wskazać brak narzędzi wspierających fact-checkerów, szczególnie w języku polskim, które ułatwiłyby im pracę, zwiększyły dokładność, przyspieszyły podejmowanie decyzji, a także pozwoliłyby na skuteczniejsze wykrywanie fake newsów.

Etapem procesu, któremu poświęcono ten rozdział, jest identyfikacja stwierdzeń do weryfikacji. Należy ona do ogólniejszego zagadnienia klasyfikacji treści tekstowych, tj. przypisywania określonej etykiety do podanego tekstu. Jeśli weźmie się przykład z analizy wydźwięku (*sentiment analysis*), tekst może być klasyfikowany jako „pozytywny", „negatywny" lub „neutralny". W naszym konkretnym przypadku będą to dwie klasy (klasyfikacja binarna): „wymagające weryfikacji" oraz „nieistotne".

Klasyfikacja tekstu odbywa się z wykorzystaniem odpowiedniego modelu, który jest wcześniej uczony maszynowo. Polega to na pokazywaniu przykładu tekstu oraz oczekiwanej odpowiedzi modelu. W przeszłości jako cechy wykorzystywane były poszczególne słowa w tekście. Obecnie do takiej klasyfikacji tekstu używane są modele językowe, które uwzględniają dodatkowo zależności między wyrazami, co określamy

potocznie jako lepsze rozumienie tekstu. Są one również traktowane jako część rozwoju sztucznej inteligencji. W zakresie klasyfikacji treści tekstowych istotne są zatem dwa obszary, które zostaną opisane w tej publikacji: zbiory danych, które mogą służyć do treningu, oraz modele, które po wyuczeniu będą służyć do klasyfikacji tekstu.

Modele, które są obecnie stosowane do klasyfikacji treści, wywodzą się z dwóch modeli bazowych: BERT (*bidirectional encoder representations from transformers*) [Devlin i in., 2018] oraz GPT (*generative pre-trained transformer*) [Radford i in., 2018]. Oba wykorzystują architekturę transformatorów [Vaswani i in., 2017], która stała się punktem zwrotnym w zakresie rozwoju możliwości rozumienia tekstu. Stosują ją również duże modele językowe (*large language models*, LLM), utożsamiane często ze sztuczną inteligencją.

Zagadnienie klasyfikacji tekstu ze względu na potrzebę weryfikacji (*check-worthiness*) jest obszarem szczególnego zainteresowania jednej z konferencji poświęconej przetwarzaniu języka naturalnego – Conference and Labs of the Evaluation Forum [CLEFF, 2025]. Modele oparte na transformatorach, takie jak BERT oraz GPT i ich pochodne, były bardzo często wykorzystywane przez badaczy [Barrón-Cedeño i in., 2023; Nakov i in., 2022]. Polega to na tym, że podstawowy model jest dostrajany na odpowiednio dobranym zbiorze danych.

W 2022 r. najlepsze rozwiązanie zostało opracowane z wykorzystaniem modelu RoBERTa large [Liu i in., 2019], gdzie po dodatkowym procesie wzbogacania danych (dwukrotne tłumaczenie) uzyskano miarę F1 na poziomie 0,698 [Savchev, 2022]. F1 to średnia harmoniczna pomiędzy precyzją (*precision*) i czułością (*recall*), która jest często wykorzystywana do oceny modeli klasyfikacyjnych – są one tym lepsze, im wartość F1 jest bliższa 1,0. Kolejne rozwiązanie wykorzystywało złączenie 10 modeli (*ensemble*) bazujących na BERT oraz RoBERTa, osiągając F1 = 0,667 [Buliga Nicu, 2022]. Dopiero na trzecim miejscu z F1 = 0,626 był zespół, który wykorzystywał dostrojony model GPT-3 [Agrestia, Hashemianb, Carmanc, 2022]. Jest to istotne o tyle, że modele BERT ze względu na mniejszy rozmiar oraz otwartość mogą być uruchamiane lokalnie, natomiast GPT traktowane są jako wielkie modele i mogą być wykorzystywane jedynie w chmurze poprzez API, udostępnione przez OpenAI. Niemniej w następnym roku akurat rozwiązanie wykorzystujące GPT-3 curie, dotrenowane na ograniczonym, dobranym jakościowo zbiorze, okazało się najlepsze, osiągając F1 = 0,898 [Sawiński i in., 2023]. Ten sam zespół przedstawił również inne minimalnie słabsze rozwiązanie (F1 = 0,894), które z kolei wykorzystywało model lokalny DeBERTa v3 [He, Gao, Chen, 2021].

Istotnym zagadnieniem jest również przygotowanie zbioru danych. Stosowane są różne techniki, od wzbogacenia danych (*data augmentation*) celem zwiększenia

liczebności zbioru (np. tłumaczenie zbiorów z innych języków, dwukrotne tłumaczenie celem uzyskania tożsamej semantycznie i różnej leksykalnie treści), przez ekstrakcję dodatkowych cech (np. metadane z portalu-źródła informacji), uwzględnienie wektorowych osadzeń przygotowanych przy pomocy innych modeli, do dołączenia dodatkowych nieetykietowanych danych. Zadania w ramach CLEF nie ograniczają się do języka angielskiego, a dostępne zasoby w innych językach stwarzają dodatkową szansę na wzbogacenie zbiorów uczących, np. poprzez tłumaczenie maszynowe [Eyuboglu i in., 2022] lub wręcz wykorzystanie modeli wielojęzycznych [Mingzhe Du, Gollapalli, 2022]. Z innych badań wynika, że równie dobrze wyniki może poprawić celowane ograniczenie zbioru (tzw. *undersampling*) np. poprzez usunięcie przykładów niejednoznacznych i trudnych do nauczenia [Sawiński, Węcel, Księżniak, 2024b].

Co do zasady jednak, im więcej zasobów w danym języku i im bardziej są one różnorodne, tym większa szansa na wytrenowanie dobrego modelu. Różne badania wskazują, że mniej popularne języki nie znajdują się tutaj na straconej pozycji. Jak wskazuje Jiang i Zubiaga [2024], wiedza dziedzinowa z języków bogatszych w treść może być przeniesiona do tych mniej zasobnych, co określane jest jako Cross-Lingual Transfer Learning (CLTL). Szczególnie atrakcyjne jest wykorzystanie wielojęzycznych modeli językowych, dla których przykłady można podawać w dowolnym języku, niekoniecznie tym docelowym. Pomija się w ten sposób konieczność tłumaczenia tekstu – sam model wielojęzyczny zachowuje się jak translator. Możliwe jest „przeniesienie korzyści" nawet między tak odległymi językami jak arabski i niderlandzki [Sawiński, Węcel, Księżniak, 2024a]. Wspomniana publikacja jest o tyle interesująca, że testuje 15 wariantów zbiorów, utworzonych z wykorzystaniem różnych języków oraz czterech metod doboru próbek z populacji (*undersampling*).

7.2. Metodyka

Przeprowadzone wcześniej analizy literatury oraz bieżąca obserwacja dostępności zasobów do przetwarzania języka polskiego wskazują, że brakuje metod pozwalających na identyfikację stwierdzeń wymagających weryfikacji. Hipoteza wstępna jest taka, że luka ta wynika przede wszystkim z braku odpowiednio oznaczonych zasobów w języku polskim, na którym można by wytrenować modele. Nasze doświadczenia wskazują jednak, że dobre modele można dostroić nawet na zasobach w innych językach.

Celem naszych badań jest zatem opracowanie zbiorów danych oraz modeli językowych, które pozwoliłyby na identyfikację stwierdzeń w języku polskim, wymagających weryfikacji przez fact-checkerów. Z tego wynikają cele szczegółowe. Przede wszystkim

należy określić, jakie są dostępne korpusy tekstu z klasyfikacją według *check-worthy*. Jeśli nie ma wśród nich języka polskiego, należy taki zbiór opracować poprzez wybór odpowiednich stwierdzeń i ich ręczną anotację. Sposób anotacji zbioru powinien być określony w protokole anotacji. Wcześniej konieczne jest zdefiniowanie rozumienia *check-worthy*, tak aby każda z osób anotujących była świadoma, do jakich kryteriów należy się odwołać. W obszarze podcelów dotyczących modeli w pierwszej kolejności należy zidentyfikować modele dla języka polskiego oraz modele wielojęzyczne obsługujące nasz język. Kolejny krok to przygotowanie różnych wariantów zbiorów do uczenia. Po przeprowadzonych treningach możliwe będzie określenie, który wariant daje najdokładniejsze wyniki.

W toku prowadzonych prac zamierzamy odpowiedzieć na następujące pytania badawcze:

- P1. Jak opracować wysokiej jakości zbiór tekstów w języku polskim, służący do trenowania modeli klasyfikacji według konieczności weryfikacji?
- P2. W jaki sposób wykorzystać bogate zasoby w językach obcych do poprawy modelu dla języka polskiego?
- P3. Czy lepiej wykorzystywać dedykowane modele dla jednego języka (tutaj: polskiego), czy wystarczające są modele wielojęzyczne?
- P4. Jaka kombinacja zbioru danych oraz modelu daje najlepsze wyniki?
- P5. Czy racjonalne jest podejmowanie wysiłku w kierunku opracowania większego zbioru danych dla języka polskiego?

7.3. Opracowanie zbioru danych

W momencie prowadzenia badań nie był dostępny publicznie żaden zbiór tekstów w języku polskim z etykietami dotyczącymi konieczności weryfikacji (*check-worthiness*). Na potrzeby eksperymentów przygotowaliśmy własny zbiór, w skład którego weszły wiadomości w języku polskim, pochodzące z platformy społecznościowej X (dawniej Twitter), a także wykorzystaliśmy dwa istniejące angielskojęzyczne zbiory: ClaimBuster [Arslan i in., 2020] oraz zbiór testowy z konkursu CLEF2023 CheckThat! Lab Task 1b English [CheckThat, 2025]. Oba zbiory angielskojęzyczne zostały wykorzystane zarówno w wersji oryginalnej, jak i przetłumaczonej na język polski z wykorzystaniem translatora Google [Google Cloud, 2025].

7.4. Pozyskanie i wybór danych do procesu etykietowania

Teksty do polskiego zbioru danych zostały pozyskane z platformy X w 2023 r. przez dostępny wówczas interfejs API na licencji dla badań naukowych. Wybór odbył się na podstawie dwóch zestawów kryteriów, które w zamierzeniu miały wygenerować dwa odrębne zbiory, zawierające różny stosunek wiadomości wymagających sprawdzenia do wiadomości niewymagających sprawdzenia.

Pierwszy zbiór danych został pobrany przy użyciu zestawu kryteriów obejmujących listę 100 najpopularniejszych słów w języku polskim oraz ograniczenie dotyczące czasu publikacji wiadomości (lata 2019–2021), liczby interakcji z wiadomością (liczba wszystkich interakcji, tj. suma liczb odpowiedzi, przekazań, polubień i cytowań większa niż 100). Dodatkowo wybrane zostały tylko wiadomości, które nie zawierały dołączonych mediów audio lub wideo, nie cytowały, nie przekazywały ani nie odpowiadały na inną wiadomość. Intencją zastosowania takiego zestawu kryteriów było wybranie wiadomości tekstowych, które są kompletne (tzn. nie wymagają analizy kontekstu poprzednich wiadomości lub dołączonych plików medialnych) oraz posiadają znaczący potencjał wiralności, a jednocześnie stanowią reprezentatywną próbkę dla ogółu wiadomości z X w języku polskim. Ten zbiór danych będzie dalej nazywany „zbiorem bazowym".

Drugi zbiór danych został pobrany przy użyciu zestawu kryteriów, obejmujących ograniczenie dotyczące czasu publikacji wiadomości (lata 2022–2023), liczby interakcji z wiadomością (liczba wszystkich interakcji większa niż 100) oraz braku dołączonych mediów audio lub wideo. Dodatkowo wybrane zostały tylko wiadomości, które w odpowiedziach zawierały przynajmniej jedno słowo kluczowe: 'nieprawda', 'fałsz', 'fake', 'fake-news', 'fakenews', 'fejk'. Podobnie jak w przypadku zbioru bazowego wybrane zostały jedynie wiadomości, które nie cytowały, nie przekazywały ani nie odpowiadały na inną wiadomość, a także te, których autorzy w momencie ich pobierania mieli więcej niż 5000 obserwujących użytkowników. Intencją zastosowania takiego zestawu kryteriów było wybranie wiadomości tekstowych, które są kompletne oraz posiadają najwyższy potencjał wiralności (poprzez zastosowanie filtra popularności wiadomości oraz autora), a jednocześnie reakcje użytkowników wskazują na możliwość wystąpienia w niej dezinformacji. Ten zbiór danych będzie dalej nazywany „zbiorem filtrowanym reakcjami użytkowników".

Do etykietowania zostało wybranych 1201 przykładów ze „zbioru bazowego" oraz 1500 przykładów ze „zbioru filtrowanego reakcjami użytkowników".

7.5. Etykietowanie danych

Zadanie związane z etykietowaniem (anotacją) danych miało na celu zaklasyfikowanie własnego zbioru danych ze względu na zawieranie stwierdzeń, które powinny podlegać sprawdzeniu (*check-worthiness*). Zaanotowane dane miały zostać następnie wykorzystane w procesie trenowania i testowania modeli, opisanych w sekcji czwartej.

Proces anotacji składał się z kilku etapów, omówionych w kolejnych paragrafach:

1. Przegląd literatury związanej z procesem identyfikacji stwierdzeń wymagających sprawdzenia.
2. Przygotowanie przewodnika anotacji, uwzględniającego rodzaje klas wykorzystanych w procesie anotacji zbioru oraz zdefiniowanie procedur i zasad anotacji.
3. Przygotowanie środowiska do przeprowadzenia procesu anotacji.
4. Przeprowadzenie procesu anotacji, w tym uzgodnienie ostatecznych klasyfikacji w przypadku niezgodności między anotatorami.

Pierwszym krokiem procesu był przegląd literatury w zakresie definicji pojęcia „checkworthiness", a także kryteriów i dobrych praktyk stosowanych przez organizacje fact-checkingowe do identyfikacji stwierdzeń wymagających sprawdzenia. W tym celu zapoznano się z dokumentacją opisującą metodologie stosowane przez polskie i zagraniczne agencje fact-checkingowe, m.in. Stowarzyszenie Demagog [DEMAGOG, 2025], Fakehunter [FakeHunter, 2025], Politifact [POLITIFACT, 2025], AFP [AFP, 2025], International Fact-checking Network [IFCN Code of Principles, 2025]. Analiza dokumentacji wykazała, że pojęcie „check-worthy" nie jest jednoznacznie opisane w literaturze. Agencje fact-checkingowe często używają różnych kryteriów do określenia, czy dane stwierdzenie wymaga weryfikacji. Jednak punktem wspólnym jest tzw. „factual and verifiable claim", czyli takie stwierdzenie, które może zostać sprawdzone jako prawdziwe lub fałszywe na podstawie dostępnych dowodów, gdyż opiera się na rzeczywistych informacjach, danych, zdarzeniach i nie stanowi opinii, przekonania, prognozy czy interpretacji. Na przykład, według Demagoga, sprawdzeniu powinny podlegać treści merytoryczne (tj. zawierające odwołania do faktów, np. liczb, wydarzeń z przeszłości), które wychodzą poza zakres tzw. wiedzy powszechnej i są istotne z punktu widzenia debaty publicznej, mają znaczenie dla społeczeństwa. Fakehunter przy wyborze stwierdzeń do fact-checkingu bierze również pod uwagę to, czy stwierdzenie wydaje się mylące i nasuwa przypuszczenie niezgodności ze stanem faktycznym, a także czy jest szeroko udostępniane (ocena zasięgu internetowego oraz popularności źródła udostępniania informacji). W przypadku, gdy stwierdzenie dotyczy wypowiedzi znanych osób (np. w formie mowy zależnej), analizowane mogą być z kolei dwa aspekty: 1) czy wypowiedź osoby publicznej zawiera błędne lub niezgodne z rzeczywistością

stwierdzenia; 2) przypisanie danego oświadczenia osobie publicznej, które zostało sfabrykowane lub stanowi fałszywe przedstawienie jej wypowiedzi (stwierdzenia typu „Pan X powiedział, że…").

Przeprowadzona analiza dokumentacji umożliwiła identyfikację pewnych ogólnych zasad określających, czy dane stwierdzenie powinno (lub nie) podlegać sprawdzeniu (rozróżnienie między check-worthy i non check-worthy), a także zdefiniowanie dodatkowych klas do anotacji.

Klasy (etykiety), które zostały wykorzystane w procesie etykietowania zbioru, to:

- *Absent* – tekst nie zawiera stwierdzenia (*claim*) lub nie jest możliwa jego weryfikacja.
- *Not Check-Worthy* – tekst zawiera weryfikowalne stwierdzenie, ale nie jest ono warte
- weryfikacji.
- *General Public* – tekst zawiera stwierdzenie, które powinno zostać zweryfikowane, ponieważ jest istotne z punktu widzenia debaty publicznej. Do tej kategorii zaliczane są stwierdzenia związane z polityką (np. wypowiedzi polityków, wyniki sondaży), technologią, sportem, wydarzeniami społecznymi.
- *Harmful* – tekst zawiera stwierdzenie, które powinno zostać zweryfikowane, ponieważ może bezpośrednio narazić czyjeś zdrowie lub życie. Do tej kategorii nie są jednak zaliczane stwierdzenia, które nie wpisują się w ogólną narrację (tzn. nie dotyczą nośnych tematów, np. COVID-19).
- *General Public and Harmful* – tekst zawiera stwierdzenie, które powinno zostać zweryfikowane, ponieważ może narazić czyjeś zdrowie i dotyczy szerokiego grona odbiorców (np. stwierdzenie związane z nośnymi tematami, takimi jak COVID-19, czy dotyczące ataków na mniejszości narodowe).

Wspomnianą wyżej narrację ogólną należy rozumieć jako zbiór powiązanych semantycznie stwierdzeń, pojawiających się w wiadomościach publikowanych przez serwisy internetowe ze względnie dużą częstością, o dużej wiralności, tzn. szybko rozprzestrzeniające się. Przykładowo, po wpisaniu do wyszukiwarki Google stwierdzenia: „szczepienia na covid wywołują autyzm" można bez problemu znaleźć wiadomości, w których rozważana jest jego prawdziwość – wokół tego tematu powstało wiele artykułów i opublikowano wiele postów w mediach społecznościowych, w których podjęto w związku z nim dyskusje. Kontrprzykładem jest natomiast stwierdzenie: „jedzenie proszku do prania pomaga na trawienie" – próba wyszukania powiązanych z nim artykułów czy postów w mediach społecznościowych nie przynosi rezultatów, niemniej stwierdzenie bezsprzecznie nawołuje do działań zagrażających życiu i zdrowiu.

Do najważniejszych wytycznych w kwestii identyfikacji stwierdzeń, które nie powinny podlegać weryfikacji, zaliczyliśmy te: 1) będące opiniami, 2) związane z osobistym

doświadczeniem, 3) odnoszące się do powszechnej wiedzy, 4) zawierające predykcje na przyszłość, 5) związane z nieistniejącymi rzeczami/zjawiskami, 6) odnoszące się do wartości estetycznych, moralnych lub etycznych, 7) dotyczące wiary lub zdolności nadprzyrodzonych, 8) wyrażające silne emocje, reakcje lub stany bez formalnego deklarowania faktów, intencji czy zobowiązań, 9) będące tautologiami w sensie językowym, np. kawalerowie nie mają żon. Informacje te zostały następnie zebrane w formie przewodnika dla anotatorów.

W procesie etykietowania brały udział trzy osoby – członkowie zespołu projektowego OpenFact, mające kilkuletnie doświadczenie badawcze w zakresie analizy tekstów zawierających fałszywe informacje. Dwie z nich pełniły rolę anotatorów, których zadaniem była niezależna klasyfikacja każdego rekordu w zbiorze, natomiast trzecia osoba pełniła rolę recenzenta oceniającego te przypadki, dla których nie było zgodności między dwoma anotatorami. W przypadku tekstów granicznych (co do których recenzent miał wątpliwości dotyczące właściwej klasyfikacji) odbywała się dodatkowa dyskusja w zespole trzyosobowym. Przykładami takich tekstów są: „W 2021 r., jako nauczyciele zarabialiśmy tak mało, że nawet uczniowie się z nas śmiali. Jeden nauczyciel wykonuje pracę za trzech (jeżeli tylko ma papier), bo tyle chętnych do pracy w szkole…" oraz „Im więcej rozmawiam i czytam na ten temat, tym bardziej utwierdzam się w przekonaniu, że za drakońskie podwyżki cen gazu odpowiada PiS i PGNiG, który stracił miliardy zł z powodu złych decyzji dotyczących formuły cenowej".

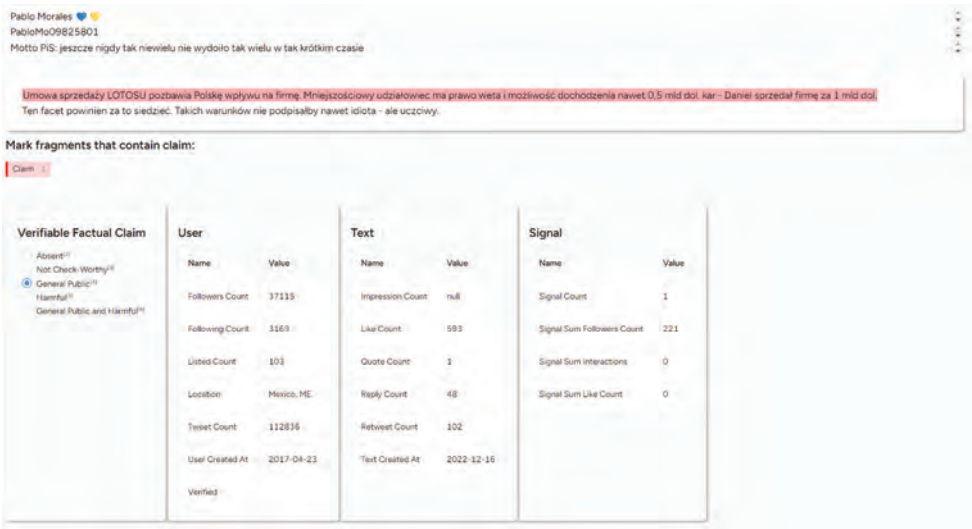
Zakres danych udostępnionych anotatorom, oprócz samej treści posta, który miał zostać zaklasyfikowany, obejmował również informacje o użytkowniku, który daną treść opublikował (nazwa konta, liczba osób śledzących, liczba profili śledzonych przez dane konto, lokalizacja, liczba opublikowanych tweetów, data utworzenia profilu, a także informacja, czy konto zostało zweryfikowane przez platformę X), statystyki dotyczące posta (liczba wyświetleń, polubień, cytowań, odpowiedzi, retweetów, data stworzenia posta), statystyki dotyczące tzw. sygnałów[1] (liczba sygnałów, suma liczb śledzących dla kont, od których sygnał pochodzi, liczba wyświetleń i polubień sygnału).

Do procesu anotacji wykorzystano oprogramowanie LabelStudio [Label Studio, 2025] w wersji Enterprise, które oprócz udostępnienia anotatorom opisanych wyżej danych oraz możliwości przeprowadzenia procesu etykietowania (rysunek 7.1) umożliwiło również zarządzanie całym procesem anotacji (publikacja przewodnika dla anotatorów, przypisanie anotatorów do zadań, analiza wyników anotacji

[1] Sygnał to odpowiedź na oryginalny post, np. w formie komentarza, sugerująca, że dana treść jest nieprawdziwa. Sygnały są identyfikowane na podstawie zdefiniowanej listy słów kluczowych, takich jak: 'nieprawda', 'kłamstwo', 'fake news'.

i identyfikacja braku zgodności, dodawanie komentarzy do przykładów wątpliwych). W przypadku dłuższych tekstów narzędzie pozwalało również na zaznaczenie jego fragmentu, stanowiącego ocenione stwierdzenie (czerwony fragment tekstu widoczny na rysunku 7.1).

Rysunek 7.1. Narzędzie LabelStudio wykorzystane w procesie anotacji danych



Źródło: opracowanie własne.

7.6. Przygotowanie zbiorów do treningu modeli językowych

Dane pochodzące ze zbiorów ClaimBuster, CLEF2023 CheckThat! Lab oraz wiadomości z X zostały podzielone na podzbiory, których następnie użyto do skonstruowania zbiorów treningowych, ewaluacyjnych i testowych dla treningu modeli językowych.

Podstawę treningu stanowiły dane ze zbioru ClaimBuster, który zawiera 23 533 przykłady z etykietami, przy czym liczebność stwierdzeń z etykietą *check-worthy* (w dalszej części określanej jako klasa pozytywna) wynosi 5485. Poprzednie badania [Sawiński, Węcel, Księżniak, 2024b] pokazały, że redukcja liczebności próby klasy większościowej zapewnia efektywny trening modelu. Dysponując informacją o jakości etykiet, do zbioru treningowego włączono wszystkie przykłady pozytywne oraz taką samą liczbę przykładów negatywnych, wylosowanych spośród tekstów o etykietach wysokiej jakości. Zbiór będzie dalej nazywany *ClaimBuster-en*, a jego przetłumaczona na polski wersja będzie dalej nazywana *ClaimBuster-pl*.

Zbiór testowy z konkursu CLEF2023 CheckThat! Lab Task 1b English, zawierający 315 przykładów, został dołączony jako dodatkowy zbiór testowy, aby móc odnieść wyniki eksperymentów do wyników konkursu. Oryginalne teksty w języku angielskim zostały użyte do stworzenia zbioru nazywanego dalej *CheckThat23-en*. Przetłumaczone teksty posłużyły do stworzenia analogicznego zbioru w języku polskim, nazywanego dalej *CheckThat23-pl*. Wszystkie 1201 przykładów z ręcznie przygotowanymi etykietami ze zbioru bazowego zostały włączone do podzbioru treningowego, nazywanego dalej BAZA. Przykłady ze zbioru filtrowanego reakcjami użytkowników podzielono losowo na 3 równe grupy o liczebności 500, z których pierwsza została podzbiorem zbioru treningowego, nazywanym dalej FILTR. Druga grupa przykładów została zbiorem ewaluacyjnym, nazywanym dalej EWALUACJA, a trzecia – zbiorem testowym, nazywanym dalej TEST.

Złączenia różnych podzbiorów posłużyły do skonstruowania ośmiu wariantów zbioru treningowego. Cztery z nich składały się wyłącznie z przykładów w języku polskim i były użyte do treningu wszystkich modeli: 1) ClaimBuster-pl, 2) ClaimBuster-pl + BAZA, 3) ClaimBuster-pl + FILTR, 4) ClaimBuster-pl + BAZA + FILTR.

Pozostałe warianty składały się w głównej mierze z przykładów w języku angielskim, do których były dokładane mniejsze zbiory w języku polskim. Warianty te zostały użyte wyłącznie do treningu modeli wielojęzycznych: 1) ClaimBuster-en, 2) ClaimBuster-en+BAZA, 3) ClaimBuster-en + FILTR, 4) ClaimBuster-en + BAZA + FILTR.

W każdym treningu do ewaluacji użyto tego samego zbioru, oznaczonego jako *EWALUACJA*. W przypadku treningu wszystkich modeli do testów wykorzystano dwa zbiory polskojęzyczne, oznaczone jako *TEST* i *CheckThat2023-pl*. Dodatkowo dla modeli wielojęzycznych miary jakości klasyfikacji były obliczane dla angielskojęzycznego zbioru *CheckThat2023-en*.

Szczegóły zostały zaprezentowane w tabeli 7.1.

Tabela 7.1. Zbiory danych treningowych, ewaluacyjnych i testowych

Nazwa podzbioru	Język	Liczba przykładów	Stosunek przykładów pozytywnych do negatywnych	Źródło	Anotacja
BAZA	PL	1201	1:3.5	X, *zbiór bazowy*	własna
FILTR	PL	500	1:1.7	X, *zbiór filtrowany reakcjami użytkowników*	własna
EWALUACJA	PL	500	1:1.6		
TEST	PL	500	1:1.8		
ClaimBuster-en	EN	10 970	1:1	debaty prezydenckie w USA	zbiorowa
ClaimBuster-pl	PL				

Nazwa podzbioru	Język	Liczba przykładów	Stosunek przykładów pozytywnych do negatywnych	Źródło	Anotacja
CheckThat23-en	EN	315	1:1.9	brak danych	brak danych
CheckThat23-pl	PL				

Źródło: opracowanie własne.

7.7. Opracowanie modeli

7.7.1. Wybór modeli bazowych

Wyboru modeli wykorzystanych do treningu dokonano na podstawie trzech kryteriów:

- **Wyniki osiągnięte w benchmarku KLEJ** – benchmark KLEJ (Kompleksowa Lista Ewaluacji Językowych) jest zestawem testów oceniających zdolności modeli przetwarzania języka naturalnego w zakresie zadań obejmujących m.in. klasyfikację tekstu czy analizę wydźwięku [Rybak i in., 2020] (Kryterium 1: K1).
- **Zasoby sprzętowe** – eksperymenty przeprowadzano na serwerze wyposażonym w cztery karty NVIDIA GeForce RTX 2080 Ti z pamięcią 11 GB na kartę; zasoby te nie pozwoliły na uruchomienie niektórych większych modeli językowych (Kryterium 2: K2).
- **Specyfika modeli** – ze względu na ogólny charakter zadania klasyfikacji (szeroki zakres tematyczny twierdzeń, takich jak zdrowie, polityka, nastroje społeczne) nie wybierano modeli przystosowanych do specyficznych zadań (np. dotrenowanych do detekcji cyberagresji) (Kryterium 3: K3).

Tabela 7.2 przedstawia wyniki osiągnięte przez różne modele według poszczególnych kryteriów. Na tej podstawie zdecydowano się na przeprowadzenie treningu z wykorzystaniem modelu wielojęzycznego XLM-RoBERTa base [Conneau i in., 2019] i dwóch modeli jednojęzycznych: HerBERT base [Mroczkowski i in., 2021], Polish RoBERTa v2 base [Dadas, Perełkiewicz, Poświata, 2020]. Dodatkowo podjęto decyzję o przeprowadzeniu treningu z wykorzystaniem – nieujętego w benchmarku KLEJ – wielojęzycznego modelu mDeBERTa v3 base [He i in., 2020], który dzięki zastosowanej architekturze poprawia wyniki testu w zadaniu XNLI dla 15 języków, osiągając średni wynik 79,8 w porównaniu z 76,2 osiągniętym przez wielojęzyczny model RoBERTa [Dadas, Perełkiewicz, Poświata, 2020].

Tabela 7.2. Wyniki benchmarku KLEJ dla różnych modeli

Nazwa modelu	Wynik benchmarku KLEJ (K1)	Zasoby sprzętowe (K2)	Charakterystyka zastosowań (K3)
Polish RoBERTa v2 large	88,9	nie spełnia	spełnia
HerBERT large	88,4	nie spełnia	spełnia
XLM-RoBERTa large + NKJP	87,8	nie spełnia	spełnia
Polish RoBERTa large	87,8	nie spełnia	spełnia
XLM-RoBERTa large	87,5	nie spełnia	spełnia
Polish RoBERTa-v2 base	**86,8**	**spełnia**	**spełnia**
XLM-RoBERTa large	86,6	nie spełnia	spełnia
HerBERT base	**86,3**	**spełnia**	**spełnia**
TrelBERT	86,1	spełnia	nie spełnia
Polish RoBERTa base	85,3	spełnia	spełnia
XLM-RoBERTa large	84,7	nie spełnia	spełnia
PoLitBert v32k linear 50k (kilka rodzajów)	83,5	spełnia	nie spełnia
Sup-SimCSE SNLI XLM-R base	81,7	spełnia	nie spełnia
Polbert (cased)	81,7	spełnia	spełnia
XLM-RoBERTa base	**81,5**	**spełnia**	**spełnia**

Źródło: opracowanie własne.

7.7.2. Trenowanie modeli

Wskazane wyżej modele zostały dostrojone na podstawie opracowanych zbiorów danych z uwzględnieniem następujących wartości hiperparametrów: learning rate (testowano wartości: 2e-6, 3e-6, 5e-6, 1e-5) oraz optimizer (adamw torch dla wszystkich wybranych modeli i dodatkowo rmsprop oraz nadam dla modeli: mDeBERTa v3 base, HerBERT base).

Wykonano optymalizację hiperparametrów metodą bayesowską z wykorzystaniem funkcji sweep, dostępnej w narzędziu Weights & Biases [2025]. Metoda ta polega na tworzeniu probabilistycznego modelu funkcji celu, którą chcemy zminimalizować lub zmaksymalizować (przyjęto konfigurację: maksymalizacja wartości funkcji F1 dla klasy pozytywnej) podczas trenowania modelu. Proces ten polega na iteracyjnym wyborze kombinacji hiperparametrów, trenowaniu modelu, ocenie wyników i aktualizacji przekonań na temat najlepszych wartości hiperparametrów. Łącznie wykonano 192 próby w poszukiwaniu optymalnych wartości hiperparametrów. Najlepsze wyni-

ki osiągnięto z zastosowaniem optymalizatora „adamw torch" oraz wartości learning rate: 1e-5 (model: herbert base oraz Polish RoBERTa base), 2e-6 (model XLM-RoBERTa base), 5e-6 (model mDeBERTa v3 base). Wskazane hiperparametry zostały użyte do przeprowadzenia ostatecznych eksperymentów.

7.8. Doświadczenia z przeprowadzonych eksperymentów

Zaprezentowane wyniki zostały zebrane w trakcie 240 treningów modeli językowych, przy 24 wariantach kombinacji rodzaju modelu i zbioru treningowego oraz 10 wariantach losowej inicjalizacji modelu.

We wszystkich przypadkach treningu modeli jednojęzycznych model HerBERT base osiągnął wyższe wyniki F1 dla pozytywnej klasy niż model Polish RoBERTa v2 base podczas porównania zarówno wartości maksymalnych, jak i średnich osiągniętych w kolejnych treningach z różną losową inicjalizacją. Podobnie we wszystkich przypadkach treningów modeli wielojęzycznych model mDeBERTa v3 base osiągnął wyższe wyniki F1 dla pozytywnej klasy niż model XLMRoBERTa base podczas porównania zarówno wartości maksymalnych, jak i średnich osiągniętych w kolejnych treningach z różną losową inicjalizacją (tabela 7.3).

Tabela 7.3. Maksymalne i średnie wartości F1 dla klasy pozytywnej, wyliczone dla zbioru testowego TEST według wybranych czterech modeli trenowanych z dziesięcioma różnymi losowymi inicjalizacjami parametrów (CB = ClaimBuster)

Model	HerBERT base		Polish RoBETa v2 base		mDeBERTa v3 base		XLM-RoBERTa base	
Zbiór treningowy	maks.	śr.	maks	śr.	maks.	śr.	maks.	śr.
CB-pl	72,73	70,80	72,59	69,91	72,35	70,77	69,87	67,21
CB-pl + BAZA	75,42	73,42	73,76	71,86	74,44	72,13	72,11	70,80
CB-pl + FILTR	76,68	75,13	74,36	72,15	75,53	72,89	73,49	71,81
CB-pl + BAZA + FILTR	79,56	76,25	76,57	74,29	75,74	74,70	74,81	72,39
CB-en	–	–	–	–	73,26	69,32	71,03	68,84
CB-en + BAZA	–	–	–	–	77,11	73,91	72,90	71,50
CB-en + FILTR	–	–	–	–	77,30	75,96	75,44	74,21
CB-en + BAZA + FILTR	–	–	–	–	79,78	77,30	77,60	75,34

Źródło: opracowanie własne.

Zastosowanie do treningu wyłącznie danych ze zbioru ClaimBuster pozwalało osiągnąć na polskim testowym zbiorze *TEST* najlepszy wynik F1 na poziomie 73,26 dla zbioru treningowego w języku angielskim oraz 72,73 dla zbioru w języku polskim.

Wyniki osiągnięte na angielskojęzycznym zbiorze testowym *CheckThat23-en* wyniosły 91,08, czyli o 2% więcej niż zwycięski model zgłoszony na ten konkurs, gdy do treningu została użyta angielska wersja zbioru ClaimBuster, a tylko 88,12, gdy model był trenowany na przetłumaczonej na polski wersji zbioru. Wyniki osiągnięte na tym samym zbiorze testowym przetłumaczonym na polski *CheckThat23-pl* wyniosły 88,12, gdy do treningu została użyta angielska wersja zbioru ClaimBuster i aż 91,35, gdy model był trenowany na wersji zbioru przetłumaczonej na polski. Podobna, około trzyprocentowa, różnica w obu przypadkach przemawiała na korzyść stosowania zbioru treningowego i testowego w tym samym języku.

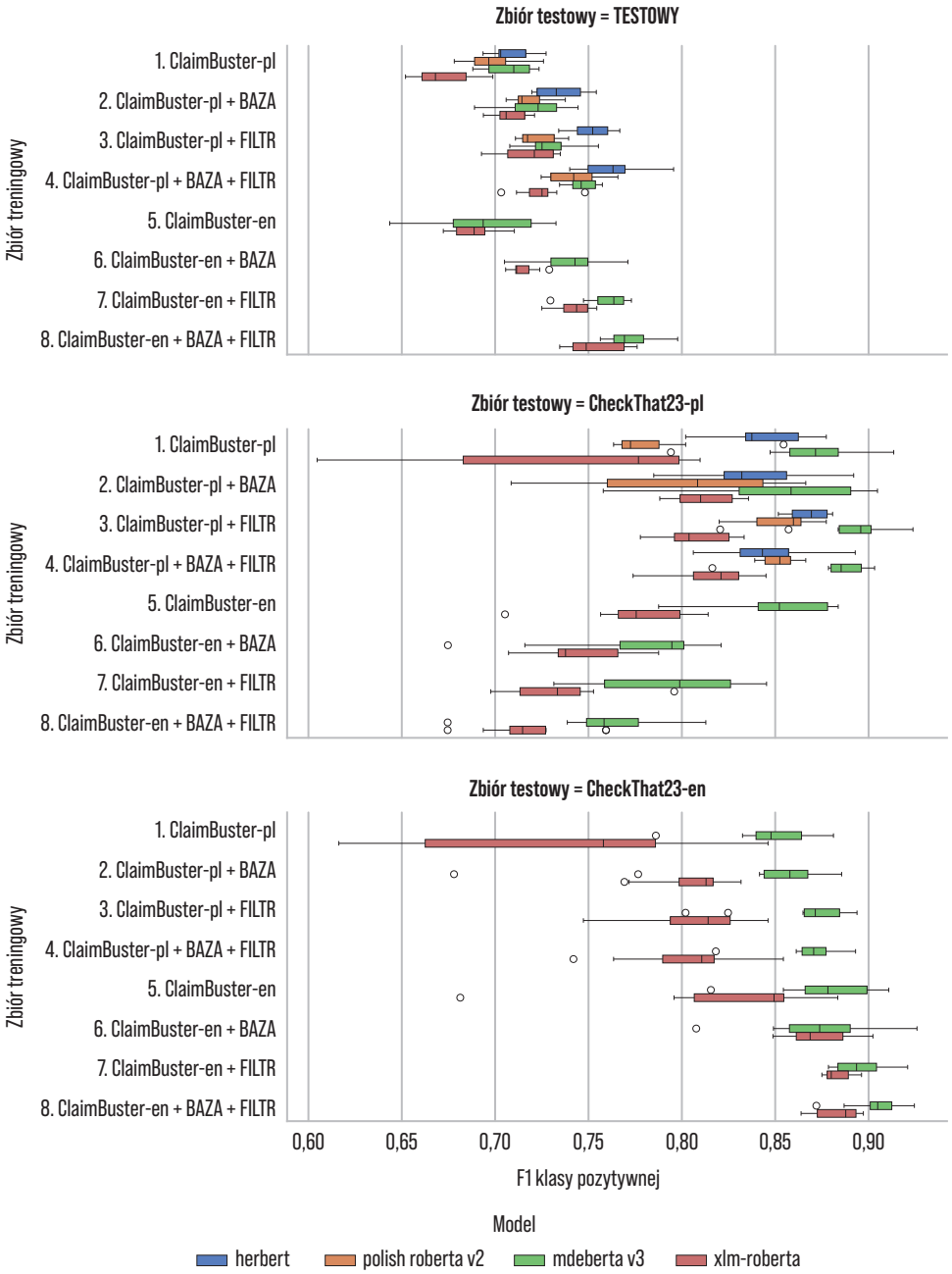
Rozszerzenie zbioru treningowego o dodatkowe przykłady w języku polskim poprawiało wynik na polskim testowym zbiorze TEST. Przy treningu na angielskojęzycznym ClaimBusteren dodanie zbioru BAZA poprawiało średnio wynik o 3,62%, dodanie zbioru FILTR poprawiało średnio wynik o 6,0%, a dodanie obu zbiorów BAZA i FILTR poprawiało średnio wynik o 7,24%. Podobne wyniki na zbiorze testowym CheckThat23-en zostały osiągnięte przy treningu na obu zbiorach ClaimBuster-pl oraz ClaimBuster-en. Jednocześnie rozszerzenie zbioru treningowego o dodatkowe przykłady w języku polskim pogarszało wynik na polskim testowym zbiorze CheckThat23-pl przy treningu na angielskojęzycznym ClaimBuster-en. Spadki wynosiły średnio 5,24% po dodaniu, dodanie zbioru FILTR pogarszało średnio wynik o 4,96%, a dodanie obu zbiorów BAZA i FILTR pogarszało średnio wynik o 7,49%. Rozszerzenie zbioru treningowego ClaimBuster-en poprawiało wyniki dla CheckThat23-pl jak w poprzednich przypadkach. Uśrednione wyniki przedstawione zostały w tabeli 7.4, a wyniki w formie wizualnej na rysunku 7.2.

Tabela 7.4. Maksymalne i średnie wartości F1 dla klasy pozytywnej, wyliczone dla trzech zbiorów testowych według wybranych czterech modeli trenowanych z dziesięcioma różnymi losowymi inicjalizacjami parametrów (CB = ClaimBuster)

Zbiór testowy	Średni wynik F1			Maksymalny wynik F1		
Zbiór treningowy	CheckThat 23-en	CheckThat 23-pl	TEST	CheckThat 23-en	CheckThat 23-pl	TEST
CB-pl	78,96	81,10	69,67	88,12	91,35	72,73
CB-pl + BAZA	82,06	82,83	72,05	88,56	90,48	75,42
CB-pl + FILTR	83,85	85,39	73,00	89,38	92,38	76,68
CB-pl + BAZA + FILTR	83,62	84,87	74,41	89,30	90,32	79,56
CB-en	85,19	**81,28**	69,08	91,08	88,37	73,26
CB-en + BAZA	87,32	**76,04**	72,70	92,59	82,11	77,11
CB-en + FILTR	88,94	**76,32**	75,08	92,09	84,54	77,30
CB-en + BAZA + FILTR	89,32	**73,79**	76,32	92,45	81,28	79,78

Źródło: opracowanie własne.

Rysunek 7.2. **Wykresy pudełkowe miar F1 dla klasy pozytywnej, wyliczonych dla trzech różnych zbiorów testowych (poszczególne wykresy) według czterech różnych modeli (kolor) trenowanych na ośmiu różnych zbiorach treningowych (wiersze w grupie)**



Źródło: opracowanie własne.

Najlepsze wyniki F1 na polskim testowym zbiorze *TEST*, czyli 79,78%, osiągnął model mDeBERTa v3 base z wykorzystaniem do treningu połączonych zbiorów *ClaimBuster-en* oraz *BAZA* i *FILTR*. Tylko nieznacznie gorszy wynik, 79,56%, osiągnął model HerBERT base trenowany z użyciem *ClaimBuster-pl* oraz *BAZA* i *FILTR*.

Podsumowanie

Przeprowadzona analiza literatury wskazała na niejednoznaczne rozumienie pojęcia „stwierdzenie wymagające weryfikacji" (*check-worthy*). Zebrane wskazówki oraz wcześniejsze doświadczenia pozwoliły jednak na opracowanie spójnego przewodnika anotacji i utworzenie własnego zbioru tekstów wysokiej jakości, co odpowiada na pierwsze pytanie badawcze (P1).

Istniejące zbiory w językach obcych mogą być wykorzystane na dwa sposoby, których skuteczność została potwierdzona w toku prowadzonych badań. Pierwszy z nich to maszynowe tłumaczenie zbioru i wykorzystanie do trenowania dedykowanego modelu w języku polskim – potwierdzone dla zbioru ClaimBuster-pl i modelu HerBERT. Drugi zaś to bezpośrednie wykorzystanie dla trenowania modelu wielojęzycznego, co potwierdziliśmy dla ClaimBuster-en i modelu mDeBERTa v3 base. Wspomniane sposoby odpowiadają na pytanie badawcze P2. Przeprowadzone badania wskazują, że oba modele osiągnęły porównywalną skuteczność, mierzoną wynikiem F1 dla klasy pozytywnej i wynoszącą w zaokrągleniu 80%. Odpowiadając na trzecie pytanie badawcze P3, należy wskazać, że korzystniejsze jest użycie modeli wielojęzycznych: unikamy konieczności tłumaczenia zasobów na język polski oraz pozostawiamy sobie możliwość wykorzystania większych i bardziej różnorodnych zasobów w językach innych niż polski.

Aby uzyskać odpowiedź na czwarte pytanie badawcze P4, analizując kombinacje zbioru danych oraz modelu pod kątem najlepszych rezultatów uczenia maszynowego, należy w pierwszej kolejności odnieść się do składu zbiorów. Niezależnie od tego, czy były trenowane modele jedno- czy wielojęzyczne, najlepsze wyniki osiągnęliśmy podczas wykorzystania kombinacji trzech zbiorów: ClaimBuster oraz własnych etykietowanych ręcznie danych z X, zarówno tych wyszukanych na podstawie popularnych słów kluczowych, jak i filtrowanych reakcjami użytkowników. Ciekawa jest również obserwacja, że modele wielojęzyczne uczyły się lepiej, jeśli były trenowane na danych w dwóch językach, w porównaniu z zasobami jednojęzycznymi. Ponownie należy zauważyć, że różnorodność dała przewagę. Jeśli weźmiemy pod uwagę same mode-

le wielojęzyczne, okazuje się, że praktycznie we wszystkich eksperymentach model mDeBERTa v3 base osiągnął lepsze rezultaty niż model XLM-RoBERTa base.

W toku prowadzonych badań stwierdziliśmy, że zwiększanie liczebności i różnorodności zbioru jest zasadne. Z danych zgromadzonych w tabeli 7.3 wynika, że zwiększanie liczebności zbioru treningowego prowadzi do zwiększenia miary F1 na różnych zbiorach testowych, praktycznie niezależnie od użytego modelu. Należy również zauważyć, że modele trenowane ze zbiorem FILTR, tj. wybrane z uwzględnieniem reakcji użytkowników, osiągają lepsze wyniki niż BAZA, wybrane losowo. A zatem, odpowiadając na pytanie badawcze P5, należy stwierdzić, że warto opracować większy zbiór dla języka polskiego z bardzo dobrymi jakościowo anotacjami. Modele wielojęzyczne mają tę dodatkową przewagę, że mogą korzystać z większych zasobów, bez konieczności tłumaczenia. Wynika z tego, że powinno się poszukiwać podobnych zasobów w innych językach, gdyż również one przyczyniają się do lepszych wyników modeli wielojęzycznych do stosowania dla języka polskiego. Rozbudowa zbioru dla języka polskiego nie musi ograniczać się do zasobów wyłącznie polskich.

Wnioski z przeprowadzonych badań ograniczają się do języka polskiego oraz angielskiego. Z innych naszych prac wynika jednak, że wzbogacać mogą się tak odległe języki jak niderlandzki i arabski. Zauważamy również pewne różnice w zbiorach testowych używanych w ramach ClaimBuster oraz opracowanych przez nas (TEST) – ten drugi wydaje się większym wyzwaniem dla modeli. Kolejnym ograniczeniem jest liczebność zbiorów wysokiej jakości, gdzie ręcznie ocenione zostało 1500 dokumentów. W najbliższym czasie liczbę tę zwiększymy docelowo do 15 000, z wykorzystaniem metodyki opracowanej w ramach niniejszych badań. Modele będą również trenowane z większą liczbą losowych inicjalizacji parametrów (tzw. *seed*), gdyż zauważyliśmy, że wartość F1 może znacznie się wahać ze względu na wpadanie w różne minima lokalne.

Bibliografia

AFP (2025). *AFP Fact-Checking Stylebook*, https://sprawdzam.afp.com/node/54477 (dostęp: 30.03.2025.

Agrestia, S., Hashemianb, A., Carmanc, M. (2022). *PoliMi-FlatEarthers at CheckThat! 2022: GPT-3 Applied to Claim Detection*, "Working Notes of CLEF 2022" – Conference and Labs of the Evaluation Forum.

Arslan, F., Hassan, N., Li, C., Tremayne, M. (2020). *ClaimBuster: A Benchmark Dataset of Check-worthy Factual Claims*. DOI: 10.5281/zenodo. 3836810.

Barrón-Cedeño, A., Alam, F., Galassi, A., Da San Martino, G., Nakov, P., Elsayed, T., Azizov, D., Caselli, T., Cheema, G.S., Haouari, F., Hasanain, M., Kutlu, M., Li, C., Ruggeri, F., Struß, J.M., Zaghouani, W. (2023). Overview of the CLEF – 2023 CheckThat! Lab on Checkworthiness, Subjectivity, Political Bias, Factuality, and Authority of News Articles and Their Source. W: *Experimental IR Meets Multilinguality, Multimodality, and Interaction* (s. 251–275), A. Arampatzis, E. Kanoulas, T. Tsikrika, S. Vrochidis, A. Giachanou, D. Li, M. Aliannejadi, M. Vlachos, G. Faggioli, N. Ferro (Eds.). Cham: Springer. DOI: 10.1007/978-3-031-42448-9.

Buliga Nicu, R. (2022). *Zorros at CheckThat! 2022: Ensemble Model for Identifying Relevant Claims in Tweets*, "Working Notes of CLEF 2022" – Conference and Labs of the Evaluation Forum.

Cazalens, S., Leblay, J., Lamarre, P., Manolescu, I., Tannier, X. (2018). Computational fact checking: a content management perspective, *Proceedings of the VLDB Endowment*, 11(12), s. 2110–2113.

CheckThat (2025). *CT23_1B_checkworthy_english_test.zip*, https://gitlab.com/checkthat_lab/clef2023-checkthat-lab/-/blob/main/task1/data/CT23_1B_checkworthy_english_test.zip (dostęp: 30.03.2025).

CLEF (2025). *Conference and Labs of the Evaluation Forum*, http://www.clef-initiative.eu/ (dostęp: 30.03.2025).

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V. (2019). *Unsupervised Cross-lingual Representation Learning at Scale*. DOI: 10.48550/arXiv.1911.02116.

Dadas, S., Perełkiewicz, M., Poświata, R. (2020). Pre-Training Polish Transformer-Based Language Models at Scale. W: *Artificial Intelligence and Soft Computing* (s. 301–314), L. Rutkowski, R. Scherer, M. Korytkowski, W. Pedrycz, R. Tadeusiewicz, J.M. Zurada (Eds.). Cham: Springer. DOI: 0.1007/978-3-030-61534-5_27.

DEMAGOG (2025). *Metodologia*, https://demagog.org.pl/metodologia/ (dostęp: 30.03.2025).

Devlin, J., Chang, M., Lee, K., Toutanova, K. (2018). *BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding*. DOI: 10.48550/arXiv.1810.04805.

Eyuboglu, A., Arslan, M., Sonmezer, E., Kutlu, M. (2022). *TOBB ETU at CheckThat! 2022: Detecting Attention-Worthy and Harmful Tweets and Check-Worthy Claims*, "Working Notes of CLEF 2022" – Conference and Labs of the Evaluation Forum.

FakeHunter (2025). *Reguły weryfikacji faktów*, https://fake-hunter.pap.pl/sites/default/files/2022-07/reguly_weryfikacji_faktow.pdf (dostęp: 30.03.2025).

Google Cloud (2025). *Translate Documents, Websites, Apps, Audio Files, Videos with Support for More than 135 Languages*, https://cloud.google.com/translate (dostęp: 30.03.2025).

He, P., Gao, J., Chen, W. (2021). *DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing*. DOI: 10.48550/arXiv.2111.09543.

He, P., Liu, X., Gao, J., Chen, W. (2020). *DeBERTa: Decoding-Enhanced BERT with Disentangled Attention*. DOI: 10.48550/arXiv.2006.03654.

IFCN Code of Principles (2025). *The Commitments of the Code of Principles*, https://ifcncodeofprinciples.poynter.org/the-commitments (dostęp: 29.04.2025).

Jiang, A., Zubiaga, A. (2024). *Cross-Lingual Offensive Language Detection: A Systematic Review of Datasets, Transfer Approaches and Challenges*. DOI: 10.48550/arXiv.2401.09244.

Label Studio (2025). *Open Source Data Labeling Platform*, https://labelstud.io/ (dostęp: 30.03.2025).

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*. DOI: 10.48550/arXiv.1907.11692.

Micallef, N., Armacost, V., Memon, N., Patil, S. (2022). True or False: Studying the Work Practices of Professional Fact-Checkers, *Proceedings of the ACM on Human-Computer Interaction*, 6 (CSCW1), s. 1–44.

Mingzhe Du, S., Gollapalli, S.D. (2022). *NUS-IDS at CheckThat! 2022: Identifying Checkworthiness of Tweets Using ChecktaT5*, "Working Notes of CLEF 2022" – Conference and Labs of the Evaluation Forum.

Mroczkowski, R., Rybak, P., Wróblewska, A., Gawlik, I. (2021). HerBERT: Efficiently Pretrained Transformer-based Language Model for Polish. W: *Proceedings of the 8th WorkshoponBalto-SlavicNaturalLanguageProcessing* (s. 1–10). Kiyv: Association for Computational Linguistics, https://www.aclweb. org/anthology/2021.bsnlp-1.1 (dostęp: 30.03.2025).

Nakov, P., Barrón-Cedeño, A., Da San Martino, G., Alam, F., Kutlu, M., Zaghouani, W., Li, C., Shaar, S., Mubarak, H., Nikolov, A. (2022). Overview of the CLEF-2022 CheckThat! Lab Task 1 on Identifying Relevant Claims in Tweets. W: *CLEF 2022. Conference and Labs of the Evaluation Forum* (s. 368–392). Bologna: CEUR Workshop Proceedings.

POLITIFACT (2025). *Here's a Fact: You Ended up in the Wrong Place!*, https://www.politifact.com/article/2018/feb/12/principles-truth-o-meter-politifacts-methodology- (dostęp: 30.03.2025).

Radford, A., Narasimhan, K., Salimans, T., Sutskever, I. (2018). *Improving Language Understanding by Generative Pre-Training*, https://api. semanticscholar.org/CorpusID:49313245 (dostęp: 30.03.2025).

Rybak, P., Mroczkowski, R., Tracz, J., Gawlik, I. (2020). KLEJ: Comprehensive Benchmark for Polish Language Understanding. W: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (s. 1191–1201). Association for Computational Linguistics. DOI: 10.18653/v1/2020.acl-main.111.

Savchev, A. (2022). *AI Rational at CheckThat! 2022: Using Transformer Models for Tweet Classification*, "Working Notes of CLEF 2022" – Conference and Labs of the Evaluation Forum.

Sawiński, M., Węcel, K., Księżniak, E. (2024a). *OpenFact at CheckThat! 2024: Cross-Lingual Transfer Learning for Check-Worthiness Detection*, "Working Notes of CLEF 2024" – Conference and Labs of the Evaluation Forum, https://www.dei.unipd.it/~faggioli/temp/clef2024/ (dostęp: 30.03.2025).

Sawiński, M., Węcel, K., Księżniak, E. (2024b). *Under-Sampling Strategies for Better Transformer-Based Classifications Models*, "Proceedings of CLEF 2024" – Conference and Labs of the Evaluation Forum.

Sawiński, M., Węcel, K., Księżniak, E., Stróżyna, M., Lewoniewski, W., Stolarski, P., Abramowicz, W. (2023*). OpenFact at CheckThat! 2023: Head-to-Head GPT vs. BERT – A Comparative Study of Transformers Language Models for the Detection of Checkworthy Claims*, "Working Notes of CLEF 2023" – Conference and Labs of the Evaluation Forum.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I. (2017). *Attention Is All You Need*. DOI: 10.48550/arXiv.1706.03762.