Politechnika Wrocławska

DZIEDZINA: Dziedzina nauk inżynieryjno-technicznych

DYSCYPLINA: Informatyka techniczna i telekomunikacja

ROZPRAWA DOKTORSKA

Metody spersonalizowanego rozpoznawania tekstów obraźliwych

Mgr inż. Kamil Kanclerz

Promotor / Promotorzy:
dr hab. inż. Maciej Piasecki, prof. PWr

Promotor pomocniczy:
dr inż. Jan Kocoń

Słowa kluczowe: przetwarzanie języka naturalnego, personalizacja, reprezentacja użytkownika, rozpoznawanie obraźliwości, aktywne uczenie, duże modele językowe, modele generatywne

WROCŁAW 2024

Pracę tę dedykuję Julicie Bielaniewicz za bycie nieocenionym źródłem nieskończonej inspiracji oraz kochanym Rodzicom i Dziadkom za ich cierpliwość, trud i wysiłek włożone w moje wychowanie i edukację.

Streszczenie

Niniejsza rozprawa dotyczy problematyki spersonalizowanego rozpoznawania tekstów obraźliwych z wykorzystaniem technik przetwarzania języka naturalnego (ang. *natural language processing*, NLP) oraz metod głębokiego uczenia. Główne wyzwania związane z rozpoznawaniem treści obraźliwych wynikają z różnorodności percepcji użytkowników oraz złożoności semantycznej tekstu. W pracy opracowano nowe, spersonalizowane podejścia, które uwzględniają preferencje użytkowników w celu precyzyjniejszej klasyfikacji obraźliwości tekstu, w odróżnieniu od tradycyjnych metod uogólnionych, które nie dostosowują się do indywidualnych oczekiwań odbiorców. Eksperymenty obejmowały szerokie testy porównawcze z wykorzystaniem różnych metod generowania reprezentacji wektorowych tekstów, takich jak CBOW, Skip-gram, BERT, DeBERTa, MPNet, oraz XLM-RoBERTa. W badaniach zweryfikowano, że zaawansowane modele kontekstowe, takie jak DeBERTa, osiągnęły istotnie lepsze wyniki w zadaniu rozpoznawania obraźliwości w porównaniu do tradycyjnych metod opartych na modelach bezkontekstowych. Ponadto, przeanalizowano wpływ rozmiaru zbioru danych uczących na stabilność i tempo poprawy jakości predykcji, co podkreśliło dodatkową korzyść płynącą z wykorzystania modeli uwzględniających personalizację. W ramach pracy opracowano dwie, autorskie metody spersonalizowanego aktywnego uczenia wykorzystujące autorską miarę kontrowersyjności tekstu oraz stopień dyspersji anotacji. Obie metody pozwoliły na optymalizację procesu anotacji względem metody referencyjnej oraz innych metod spersonalizowanego aktywnego uczenia takich jak *Ratio Distance* oraz *Stranger Count*. Zastosowanie opracowanych metod skutkowało znacznym zwiększeniem efektywności modelu przy mniejszych zbiorach danych, co czyni te techniki szczególnie przydatnymi w sytuacjach, gdzie dostępne są ograniczone zasoby anotacji. Ponadto, praca badała zastosowanie modelu generatywnego ChatGPT-3.5, w kontekście zarówno uogólnionego, jak i spersonalizowanego rozpoznawania treści obraźliwych. Wyniki wskazują, że modele dedykowane, trenowane na specyficznych zbiorach danych, przewyższają skutecznością model generatywny w tym zadaniu, zwłaszcza w odniesieniu do personalizacji. Rozprawa przedstawia istotne wnioski dotyczące wpływu personalizacji na rozpoznawanie obraźliwych treści tekstowych oraz proponuje nowe techniki, które mogą znaleźć zastosowanie w systemach moderacji treści online.

Abstract

This dissertation addresses the problem of personalized detection of offensive text using techniques from natural language processing (NLP) and deep learning. The key challenges in detecting offensive content arise from the diversity of user perceptions and the semantic complexity of language. The study introduces novel personalized approaches that incorporate user preferences to improve the efficiency of offensive content classification, distinguishing them from traditional generalized methods, which do not adapt to individual annotator preferences. The experimental evaluation involved comparative testing across various text vectorization techniques, such as CBOW, Skipgram, BERT, DeBERTa, MPNet, and XLM-RoBERTa. The results demonstrate that advanced contextual models, such as DeBERTa, outperform traditional non-contextual approaches in identifying offensive content. Additionally, the impact of training dataset size on model stability and prediction accuracy was analyzed, highlighting the benefits of models that incorporate personalization. The dissertation also introduces two novel personalized active learning methods, based on a custom measure of text controversy and annotation dispersion. These methods optimize the annotation process compared to a reference method and other personalized active learning techniques such as Ratio Distance and Stranger Count. Their application significantly enhanced model efficiency when working with smaller datasets, making these approaches especially useful in situations with limited annotated data. Furthermore, the study explores the use of the generative model ChatGPT-3.5 for both generalized and personalized offensive content detection. The findings indicate that dedicated models trained on specific datasets outperform the generative model in this task, particularly in terms of personalization. The dissertation presents key insights into the role of personalization in offensive content detection and proposes new techniques with potential applications in online content moderation systems.

SPIS TREŚCI

SPIS RYSUNKÓW

SPIS TABEL

WYKAZ SKRÓTÓW I OZNACZEŃ

MLP MultiLayer Perceptron

NLP Natural Language Processing

BoW Bag of Words

TF-IDF Term Frequency-Inverse Document Frequency

Część I

WPROWADZENIE

1 WSTĘP

Rozwój kanałów komunikacji elektronicznej takich jak serwisy społecznościowe umożliwiła wielu ludziom komunikację z szerokim gronem odbiorców. Niestety, ułatwienie wymiany informacji umożliwiło również niekontrolowane propagowanie treści uznawanych za obraźliwe do licznej grupy użytkowników danego serwisu. W rezultacie, monitorowanie i moderowanie treści stało się kluczowym wyzwaniem dla platform internetowych. Mimo stosowania automatycznych narzędzi do wykrywania obraźliwego języka, takich jak systemy oparte na sztucznej inteligencji, istnieje wiele ograniczeń związanych z ich skutecznością. Standardowe algorytmy często opierają się na prostych regułach lub bazach danych zawierających określone słowa kluczowe, co prowadzi do problemów związanych z kontekstem, ironią czy specyficznym językiem używanym przez różne grupy użytkowników. Ponadto, treści uznawane za obraźliwe mogą mieć różne formy w zależności od kultury, społeczności lub indywidualnych doświadczeń. Stąd wynika potrzeba opracowania spersonalizowanych metod rozpoznawania treści obraźliwych, które będą uwzględniać nie tylko język i kontekst, ale również indywidualne preferencje i wrażliwość użytkowników. W tym rozdziale opisano motywację i kontekst przeprowadzonych badań. Ponadto sformułowano cel i zakres pracy oraz definicję problemu badawczego.

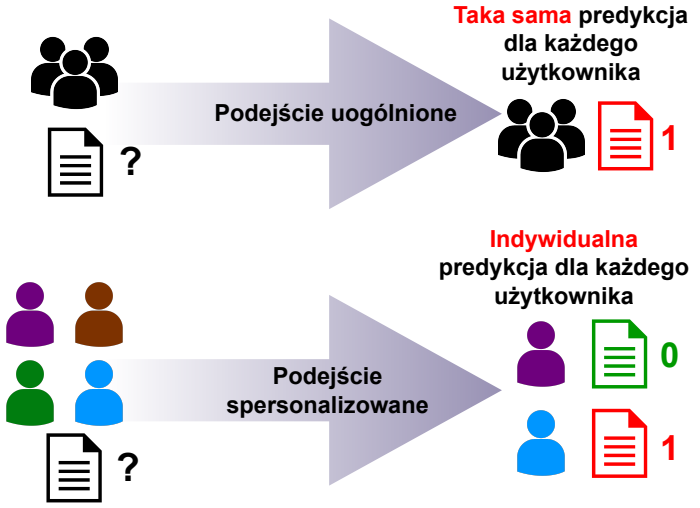
1.1 MOTYWACJA I KONTEKST BADAŃ

Konsekwencją powszechnego dostępu do kanałów komunikacyjnych, które nie wymagają bezpośredniej komunikacji „twarzą w twarz" jest wzrost liczby wiadomości zawierających treści obraźliwe (Levmore i Nussbaum, 2010). Wspomniany wzrost obserwuje się dla wielu typów obraźliwości uwzględniających zarówno mowę nienawiści (Breckheimer, 2001; Brown, 2018), społecznie nieakceptowany dyskurs (Ljubešić, Fišer i Erjavec, 2019), treści agresywne, toksyczne, jak i cyberprzemoc (Chen i in., 2012).

Brak jednoznacznej definicji treści obraźliwych prowadzi do wielu nieścisłości. Niespójność w postrzeganiu tego terminu jest zauważalna w wielu pracach na temat automatycznego rozpoznawania mowy nienawiści (Alrehili, 2019; Fortuna i Nunes, 2018; Poletto i in., 2020; Schmidt i Wiegand, 2017) lub treści agresywnych (Modha i in., 2020; Sadiq i in., 2021). Istnieje wiele definicji treści obraźliwych, które w ogólności można określić jako treści skierowane przeciwko określonym grupom społecznym w sposób dla nich szkodliwy (Jacobs, 2002). Niektóre kraje, takie jak Stany Zjednoczone,

chronią prawo do używania tego typu wypowiedzi jako dopuszczalnej formy ekspresji politycznej (Heyman, 2008). Natomiast w wielu krajach Unii Europejskiej prawo zabrania publikowania treści obraźliwych (Rosenfeld, 2002). Przepisy unijne motywowane są ochroną mniejszości, takich jak mniejszości religijne, etniczne, seksualne itp. Treści obraźliwe dotyczące mniejszości często nawołują do stosowania przemocy wobec tych grup. W związku z tym propagowanie tego typu treści jest zagrożone grzywną lub karą pozbawienia wolności. Zastosowanie się do takich przepisów często jest problematyczne dla operatorów portali społecznościowych i innych usług online w zakresie identyfikacji i moderowania niedopuszczalnych treści. Duże firmy technologiczne, takie jak Facebook i Google, są często oskarżane o to, że nie robią wystarczająco dużo, aby wykluczyć możliwość wykorzystania ich platform do propagowania treści obraźliwych (Ben-David i Fernández, 2016). Z kolei próby automatycznej kontroli treści niosą ze sobą ryzyko nieumyślnego blokowania treści nieobraźliwych.

Brak skutecznych rozwiązań uwzględniających potrzeby społeczności oraz indywidualne postrzeganie treści obraźliwych rodzi potrzebę opracowania technik spersonalizowanego rozpoznawania treści obraźliwych. Metody te uwzględniają zarówno informację na temat semantyki tekstu, jak i reprezentację użytkownika obejmującą jego preferencje określające treści, które dany użytkownik uznaje za obraźliwe. Taka reprezentacja użytkownika może występować również w wariancie skupionym na podkreślaniu różnic między percepcją danego użytkownika a reszty użytkowników traktowanej jako ogół społeczności. Opracowanie spersonalizowanych metod rozpoznawania obraźliwości umożliwi uzyskanie indywidualnych predykcji dla każdego użytkownika, uwzględniających ich osobiste preferencje. Różnicę między podejściem uogólnionym a podejściem spersonalizowanym zaprezentowano na Rysunku 1.



Rysunek 1: Różnica między podejściem uogólnionym a podejściem spersonalizowanym. [Źródło: opracowanie własne]

1.2 CEL I ZAKRES PRACY

Celem niniejszej rozprawy doktorskiej jest opracowanie spersonalizowanych metod rozpoznawania tekstowych treści obraźliwych.

1.3 HIPOTEZY BADAWCZE

W ramach niniejszej pracy sformułowano następujące hipotezy badawcze:

H1. Uwzględnienie reprezentacji użytkownika istotnie wpłynie na poprawę jakości predykcji metod rozpoznawania tekstów obraźliwych w porównaniu z podejściem wykorzystującym generalizację (Podrozdział 9.5.1).

H2. Spersonalizowane modele sieci neuronowych, w szczególności metody zaprezentowane w pracy, cechują się wyższą skutecznością w zadaniu rozpoznawania obraźliwości w porównaniu innymi metodami spersonalizowanymi oraz metodą uogólnioną (Podrozdział 9.5.1).

H3. Wykorzystanie kontekstowych metod reprezentacji tekstu (architektur typu transformer) jest lepsze niż bezkontekstowe metody reprezentacji (CBOW, Skipgram), które łącznie są lepsze niż metody niewykorzystujące wiedzy o tekście (Podrozdział 9.5.1).

H4. Język oraz rodzaj treści tekstowych (komentarze, tweety, dłuższe wypowiedzi) wpływa na jakość predykcji metod w zadaniu spersonalizowanego rozpoznawania tekstów obraźliwych (Podrozdział 9.5.1).

H5. Metody spersonalizowanego rozpoznawania tekstów obraźliwych charakteryzują się istotnie lepszą jakością predykcji niż metody zgeneralizowane niezależnie od rozmiaru zbioru uczącego (Podrozdział 9.5.2).

H6. Spersonalizowane metody aktywnego uczenia, w szczególności metody zaprezentowane w pracy, pozwalają na osiągnięcie zbliżonej jakości predykcji modelu przy istotnie mniejszym zbiorze danych w porównaniu do metody nieuwzględniającej wiedzy na temat indywidualnych anotacji użytkowników (Podrozdział 9.5.3).

H7. W zadaniu spersonalizowanego oraz uogólnionego rozpoznawania treści obraźliwych generatywne modele ogólnego przeznaczenia (np. ChatGPT-3.5) mogą mieć mniejszą skuteczność niż modele dedykowane nauczone na danym zbiorze danych (Podrozdział 9.5.4).

1.4 DEFINICJA PROBLEMU

Zakładając, że preferencje użytkownika są reprezentowane przez N-wymiarowy wektor cech $U \in R^N$, a tekst jest reprezentowany przez M-wymiarowy wektor cech $T \in R^M$ metoda $F(T, U)$ przyjmuje na wejściu zarówno reprezentację tekstu T, jak i reprezentację użytkownika U. Wyjściem metody F jest ocena obraźliwości $\hat{y}$ dla tekstu T uwzględniająca kontekst użytkownika U. Proces uzyskiwania oceny spersonalizowanej oceny obraźliwości tekstu został opisany we Wzorze 1.

$$\hat{y} = F(T, U) \tag{1}$$

Ocena obraźliwości $\hat{y}$ może przyjmować różne wartości w zależności od sposobu anotacji danych w zbiorze uczącym. W przypadku danych anotowanych binarnie ocena $\hat{y}$ może przyjmować wyłącznie wartości 0 (tekst nieobraźliwy) lub 1 (tekst obraźliwy). Dla danych anotowanych wartościami całkowitymi, należącymi do przedziału, np. $\{0, 1, \ldots, k-1, k\}$, gdzie k oznacza maksymalną wartość obraźliwości możliwą do przypisania danemu tekstowi, ocena $\hat{y}$ może uzyskać dowolną wartość rzeczywistą należącą do przedziału $[0, k]$, którą można zapisać w zastosowanym typie danych zmiennoprzecinkowych, np. *float32*. Natomiast w przypadku danych anotowanych wieloetykietowo, gdzie pojedynczy tekst może być równocześnie przypisany do wielu typów obraźliwości, ocena $\hat{y}$ przyjmuje formę wektora $\{l_0, l_1, \ldots, l_{c-1}\}$ o długości c określającej liczbę możliwych typów obraźliwości, gdzie l_{c-1} oznacza etykietę pojedynczego typu obraźliwości, przyjmującą wartość ze zbioru $\{0, 1\}$. Typy możliwych wartości wyników w zależności od rodzaju anotacji danych zostały przedstawione we Wzorze 2

$$\hat{y} = \begin{cases} \{0, 1\} & \text{, dla danych anotowanych binarnie} \\ [0, k] & \text{, dla danych anotowanych wartościami całkowitymi} \\ \{l_0, l_1, \ldots, l_{c-1}\} & \text{, dla danych anotowanych wieloetykietowo} \end{cases} \tag{2}$$

W celu zbadania skuteczności każdej opracowanej metody zostaną zastosowane różnorodne miary $\mathcal{M}(y_{test}, \hat{y}_{test})$ porównujące rzeczywiste anotacje y_{test} dla tekstów znajdujących się w zbiorze testowym z ocenami $\hat{y}_{test}$ otrzymanymi na wyjściu metody F. Zgodnie ze Wzorem 1 poprawnym jest również zapis $\mathcal{M}(y_{test}, F(T, U))$. Metoda F_1 jest uważana za lepszą od metody F_2 w zależności od wartości miary $\mathcal{M}$, przy czym interpretacja zależy od rodzaju miary. Jeśli miara $\mathcal{M}$ jest miarą skuteczności, wtedy metoda F_1 jest uznawana za lepszą od metody F_2, gdy wartość miary dla F_1 jest większa niż dla F_2, czyli $\mathcal{M}(y_{test}, F_1(T, U)) > \mathcal{M}(y_{test}, F_2(T, U))$. Z drugiej strony, jeśli

miara $\mathcal{M}$ jest miarą błędu, wtedy metoda F_1 jest uznawana za lepszą, gdy wartość miary dla F_1 jest mniejsza niż dla F_2, czyli $\mathcal{M}(y_{test}, F_1(T, U)) < \mathcal{M}(y_{test}, F_2(T, U))$.

1.5 ORYGINALNY WKŁAD PRACY

Oryginalny wkład pracy stanowią następujące elementy składowe:

1. Opracowanie dwóch metod spersonalizowanego rozpoznawania tekstowych treści obraźliwych. Zaprojektowane architektury uwzględniają indywidualne preferencje użytkowników poprzez autorskie, opracowane miary konformizmu oraz agregowane, semantyczne reprezentacje wektorowe ocenianych tekstów. (opisano w Rozdziale 5)

2. Przeprowadzenie badań na modelu generatywnym ogólnego przeznaczenia ChatGPT-3.5 mających na celu porównanie jego skuteczności względem modeli wyspecjalizowanych. Zostały one zrealizowane dla dwóch zadań spersonalizowanego i uogólnionego rozpoznawania tekstów obraźliwych (opisano w Rozdziale 6 oraz Podrozdziałach 9.4.4 i 9.5.4).

3. Opracowanie dwóch metod aktywnego uczenia pozwalających na optymalizację procesu pozyskiwania subiektywnych anotacji dotyczących obraźliwości treści tekstowych. Zaproponowane metody uwzględniają indywidualne anotacje użytkownika za pomocą autorskich miar kontrowersyjności dokumentu oraz dyspersji anotacji w celu selekcji tekstów do dalszej anotacji, które umożliwiłyby ekstrakcję maksymalnej ilości wiedzy na temat indywidualnej percepcji obraźliwości tekstów. (opisano w Rozdziale 8)

4. Przeprowadzenie ewaluacji opracowanych metod uwzględniającej wpływ różnorodnych metod generowania reprezentacji wektorowych tekstów, uwzględniające zarówno metodę referencyjną, osadzenia bezkontekstowe oraz reprezentacje kontekstowe. (opisano w Podrozdziałach 9.2.1, 9.4.1 oraz 9.5.1)

5. Przeprowadzenie badań dotyczących wpływu rozmiaru zbioru uczącego na jakość predykcji opracowanych metod w zadaniu spersonalizowanego rozpoznawania obraźliwości. Podczas eksperymentów położono nacisk na ewaluację modelu na ograniczonym zbiorze danych oraz na charakterystykę przyrostu wiedzy wraz ze wzrostem liczby anotacji w zbiorze uczącym. (opisano w Podrozdziałach 9.2.2, 9.4.2 oraz 9.5.2)

6. Przeprowadzono ewaluację opracowanych metod aktywnego uczenia, mającą na celu zbadanie wpływu strategii selekcji tekstów na tempo wzrostu jakości

wnioskowania modelu spersonalizowanego rozpoznawania obraźliwości zakładając scenariusz równomiernego anotowania tekstów przez wszystkich anotatorów w celu maksymalizacji różnorodności perspektyw użytkownika wykorzystywanych przez model w procesie wnioskowania. (opisano w Podrozdziałach 9.2.3, 9.4.3 oraz 9.5.3)

7. W efekcie pracy w projekcie CLARIN zaplanowano i przeprowadzono dwie procedury anotacji, w wyniku których opracowano dwa zbiory danych w języku polskim: Doccano 1 oeaz Doccano 2. Każdy ze zbiorów zawiera teksty anotowane na przestrzeni 26 wymiarów, w tym wymiaru obraźliwości. Zbiory te posłużyły do ewaluacji autorskich metod spersonalizowanego rozpoznawania obraźliwości, metod znanych z literatury oraz referencyjnej metody niespersonalizowanej (Podrozdziały 9.1.5 oraz 9.1.6, badania zostały opublikowane w pracy (Kanclerz i in., 2023a)).

Wyniki przeprowadzonych badań zostały wykorzystane w procesie przygotowywania następujących artykułów opublikowanych na międzynarodowych konferencjach naukowych oraz recenzowanych czasopismach naukowych. Na dzień 27.09.2024r. liczba cytowań poniższych artykułów wynosi 413 (h-index = 9) w bazie Scopus (Elsevier, 2004) oraz 60 w bazie Web of Science (Clarivate, 1997).

1. **Kamil Kanclerz**, Piotr D. Miłkowski, Jan Kocoń
   Cross-lingual deep neural transfer learning in sentiment analysis. W: Knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 24th International Conference KES 2020 / eds. M. Cristani, C. Toro, C. Zanni-Merk, R. J. Howlett and L. C. Jain. Amsterdam : Elsevier, cop. 2020. s. 1211-1220. (Kanclerz, Miłkowski i Kocoń, 2020)
   Ranga w Bazie CORE: B, Lista ministerialna: 70 pkt.

2. Jan Kocoń, Piotr D. Miłkowski, **Kamil Kanclerz**
   Multiemo: multilingual, multilevel, multidomain sentiment analysis corpus of consumer reviews. W: Computational Science - ICCS 2021 : 21st International Conference Krakow, Poland, June 16-18, 2021 : proceedings. Pt. 2 / eds. Maciej Paszynski [i in.]. Cham : Springer, cop. 2021. s. 297-312. (Kocoń, Miłkowski i Kanclerz, 2021)
   Ranga w Bazie CORE: A, Lista ministerialna: 140 pkt.

3. Piotr D. Miłkowski, Marcin Gruza, **Kamil Kanclerz**, Przemysław Kazienko, Damian Grimling, Jan Kocoń
   Personal bias in prediction of emotions elicited by textual opinions. W: The 59th Annual Meeting of the Association for Computational Linguistics and the 11th

International Joint Conference on Natural Language Processing, ACL-IJCNLP 2021, August 5-6, 2021, Bangkok, Thailand (online) : Proceedings of the Student Research Workshop / eds. Jad Kabbara [i in.]. Stroudsburg : Association for Computational Linguistics, cop. 2021. s. 248-259. (Milkowski i in., 2021)
Ranga w Bazie CORE: A*, Lista ministerialna: 200 pkt.

4. **Kamil Kanclerz**, Alicja Figas, Marcin Gruza, Tomasz Kajdanowicz, Jan Kocoń, Daria Puchalska, Przemysław Kazienko
Controversy and conformity: from generalized to personalized aggressiveness detection. W: The 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL-IJCNLP 2021, August 1-6, 2021 : Proceedings of the Conference, Vol. 1 (Long Papers) / eds. Chengqing Zong [i in.]. Stroudsburg : Association for Computational Linguistics, cop. 2021. s. 5915-5926. (Kanclerz i in., 2021)
Ranga w Bazie CORE: A*, Lista ministerialna: 200 pkt.

5. Jan Kocoń, Marcin Gruza, Julita Bielaniewicz, Damian Grimling, **Kamil Kanclerz**, Piotr D. Miłkowski, Przemysław Kazienko
Learning personal human biases and representations for subjective tasks in natural language processing. W: 21st IEEE International Conference on Data Mining ICDM 2021, 7-10 December 2021, Virtual Conference : proceedings / eds. James Bailey [i in.]. Piscataway, NJ : Institute of Electrical and Electronics Engineers, cop. 2021. s. 1168-1173. (Kocoń i in., 2021)
Ranga w Bazie CORE: A*, Lista ministerialna: 200 pkt.

6. **Kamil Kanclerz**, Maciej Piasecki
Deep neural representations for multiword expressions detection. W: The 60th Annual Meeting of the Association for Computational Linguistics : Proceedings of the Student Research Workshop, May 22-27, 2022. Stroudsburg : Association for Computational Linguistics, cop. 2022. s. 444-453. (Kanclerz i Piasecki, 2022)
Ranga w Bazie CORE: A*, Lista ministerialna: 200 pkt.

7. Maciej Piasecki, **Kamil Kanclerz**
Is context all you need? Non-contextual vs contextual multiword expressions detection. W: Computational Science - ICCS 2022 : 22nd International Conference London, UK, June 21-23, 2022 : proceedings. Pt. 1 / eds. Derek Groen [i in.]. Cham : Springer, cop. 2022. s. 248-261. (Piasecki i Kanclerz, 2022a)
Ranga w Bazie CORE: A, Lista ministerialna: 140 pkt.

8. Maciej Piasecki, **Kamil Kanclerz**
Non-contextual vs contextual word embeddings in multiword expressions de-

tection. W: Computational Collective Intelligence : 14th International Conference, ICCCI 2022, Hammamet, Tunisia, September 28-30, 2022 : proceedings / eds. Ngoc Thanh Nguyen [i in.]. Cham : Springer, cop. 2022. s. 193-206. (Piasecki i Kanclerz, 2022b)
Ranga w Bazie CORE: C, Lista ministerialna: 20 pkt.

9. **Kamil Kanclerz**, Marcin Gruza, Konrad Karanowski, Julita Bielaniewicz, Piotr D. Miłkowski, Jan Kocoń, Przemysław Kazienko
What if ground truth is subjective? Personalized deep neural hate speech detection. W: LREC 2022 : Workshop Language Resources and Evaluation Conference : 20th June 2022, 1st Workshop on Perspectivist Approaches to NLP (NLPerspectives) : proceedings / eds. Gavin Abercrombie [i in.]. Paris : European Language Resources Association, cop. 2022. s. 37-45. (Kanclerz i in., 2022)
Ranga w Bazie CORE: C, Lista ministerialna: 20 pkt.

10. Julita Bielaniewicz, **Kamil Kanclerz**, Piotr D. Miłkowski, Marcin Gruza, Konrad Karanowski, Przemysław Kazienko, Jan Kocoń
Deep-SHEEP: sense of humor extraction from embeddings in the personalized context. W: 22nd IEEE International Conference on Data Mining Workshops (ICDMW), 28 November - 1 December 2022, Orlando, Florida : proceedings / eds. K. Selçuk Candan [i in.]. Piscataway, NJ : Institute of Electrical and Electronics Engineers, cop. 2022. s. 967-974. (Bielaniewicz i in., 2022)
Ranga w Bazie CORE: A*, Lista ministerialna: 200 pkt.

11. Przemysław Kazienko, Julita Bielaniewicz, Marcin Gruza, **Kamil Kanclerz**, Konrad Karanowski, Piotr D. Miłkowski, Jan Kocoń
Human-centered neural reasoning for subjective content processing: hate speech, emotions, and humor. Information Fusion. 2023, vol. 94, s. 43-65. (Kazienko i in., 2023)
Ranga w Bazie CORE: A*, Lista ministerialna: 200 pkt., Impact Factor: 14,7

12. Jan Kocoń, Igor Cichecki, Oliwier Kaszyca, Mateusz Kochanek, Dominika Szydło, Joanna Baran, Julita Bielaniewicz, Marcin Gruza, Arkadiusz Janz, **Kamil Kanclerz**, Anna A. Kocoń, Bartłomiej Koptyra, Wiktoria Mieleszczenko-Kowszewicz, Piotr D. Miłkowski, Marcin Ł. Oleksy, Maciej Piasecki, Łukasz Radliński, Konrad D. Wojtasik, Stanisław J. Woźniak, Przemysław Kazienko
ChatGPT: Jack of all trades, master of none. Information Fusion. 2023, vol. 99, nr 101861, s. 1-37. (Kocoń i in., 2023b)
Ranga w Bazie CORE: A*, Lista ministerialna: 200 pkt., Impact Factor: 14,7

13. Jan Kocoń, Joanna Baran, **Kamil Kanclerz**, Michał Kajstura, Przemysław Kazienko

Differential dataset cartography: explainable artificial intelligence in comparative personalized sentiment analysis. W: Computational Science - ICCS 2023 : 23rd International Conference Prague, Czech Republic, July 3-5, 2023 : proceedings. Pt. 1 / eds. Jiří Mikyška [i in.]. Cham : Springer, cop. 2023. s. 148-162. (Kocoń i in., 2023a)
Ranga w Bazie CORE: A, Lista ministerialna: 140 pkt.

14. Wiktoria Mieleszczenko-Kowszewicz, **Kamil Kanclerz**, Julita Bielaniewicz, Marcin Ł. Oleksy, Marcin Gruza, Stanisław J. Woźniak, Ewa Dziecioł, Przemysław Kazienko, Jan Kocoń
Capturing human perspectives in NLP: questionnaires, annotations, and biases. W: Proceedings of the 2nd Workshop on Perspectivist Approaches to NLP co-located with 26th European Conference on Artificial Intelligence (ECAI 2023), Kraków, Poland, September 30th, 2023 / Gavin Abercrombie [i in.]. [B.m. : b.w.], 2023. s. 1-21. (Mieleszczenko-Kowszewicz i in., 2023)
Ranga w Bazie CORE: A, Lista ministerialna: 140 pkt.

15. Jan Kocoń, Joanna Baran, **Kamil Kanclerz**
Multi-modal personalized hate speech analysis using differenial dataset cartography. W: Proceedings of De-Factify 2: 2nd Workshop on Multimodal Fact Checking and Hate Speech Detection, co-located with AAAI 2023, Washington DC, USA, February 14, 2023 / eds. Amitava Das, Amit Sheth, Asif Ekbal. RWTH: Aachen, 2023. s. 1-13. (Kocoń, Baran i Kanclerz, 2023)
Ranga w Bazie CORE: A*, Lista ministerialna: 200 pkt.

16. **Kamil Kanclerz**, Julita Bielaniewicz, Marcin Gruza, Jan Kocoń, Stanisław J. Woźniak, Przemysław Kazienko
Towards model-based data acquisition for subjective multi-task NLP problems. W: 23nd IEEE International Conference on Data Mining Workshops (ICDMW), 1–4 December 2023 Shanghai, China : proceedings / eds. Jihe Wang [i in.]. Piscataway, NJ : Institute of Electrical and Electronics Engineers, cop. 2023. s. 726-735. (Kanclerz i in., 2023a)
Ranga w Bazie CORE: A*, Lista ministerialna: 200 pkt.

17. Aarohi Srivastava, Jan Kocoń, **Kamil Kanclerz**, Piotr D. Miłkowski i in.
Beyond the Imitation Game: quantifying and extrapolating the capabilities of language models. Transactions on Machine Learning Research. 2023, vol. 5, s. 1-95. (Srivastava i in., 2023)

18. **Kamil Kanclerz**, Konrad Karanowski, Julita Bielaniewicz, Marcin Gruza, Piotr D. Miłkowski, Jan Kocoń, Przemysław Kazienko

PALS: Personalized Active Learning for Subjective Tasks in NLP. W: The 2023 Conference on Empirical Methods in Natural Language Processing : Proceedings of the Conference : EMNLP 2023, December 6-10, 2023 / eds. Houda Bouamor, Juan Pino, Kalika Bali. Stroudsburg : Association for Computational Linguistics, cop. 2023. s. 13326-13341. (Kanclerz i in., 2023b)
Ranga w Bazie CORE: A*, Lista ministerialna: 140 pkt.

1.6 UKŁAD PRACY

Niniejsza praca została podzielona na trzy części. Pierwsza część pracy o tytule **Wprowadzenie** skupia się na wprowadzeniu do zagadnienia spersonalizowanego rozpoznawania obraźliwości.

W **Rozdziale 1 Wstęp** opisano motywację i kontekst badań (Podrozdział 1.1), cel i zakres pracy (Podrozdział 1.2), hipotezy badawcze (Podrozdział 1.3), definicję problemu (Podrozdział 1.4), oryginalny wkład pracy (Podrozdział 1.5) oraz układ pracy (Podrozdział 1.6).

Rozdział 2 Kontekst użytkownika w percepcji obraźliwości tekstu jest rozdziałem, w którym przedstawiono wyniki analizy literaturowej opisującej stan wiedzy w dziedzinie rozpoznawania tekstowych treści obraźliwych (Podrozdział 2.1), roli kontekstu użytkownika w zadaniach przetwarzania języka naturalnego (Podrozdział 2.2) oraz spersonalizowanych technik aktywnego uczenia (Podrozdział 2.3).

Następne sześć rozdziałów zostało umieszczonych w części drugiej zatytułowanej **Metody**.

W **Rozdziale 3 Referencyjna metoda uogólnionego rozpoznawania obraźliwych tekstów** opisano referencyjną metodę rozpoznawania obraźliwych tekstów nieuwzględniającą kontekstu użytkownika.

Rozdział 4 Metody spersonalizowanego rozpoznawania tekstów obraźliwych opisuje istniejące metody rozpoznawania tekstów obraźliwych uwzględniające preferencje użytkowników. Zostały w nim opisane metody HuBi-Simple (Podrozdział 4.1) oraz HuBi-Medium (Podrozdział 4.2).

W **Rozdziale 5 Nowe metody spersonalizowanego rozpoznawania tekstów obraźliwych** opisano nowe, spersonalizowane metody rozpoznawania obraźliwości. Opracowane, autorskie miary konformizmu ogólnego i ważonego zostały wyjaśnione w Podrozdziale 5.1. W Podrozdziale 5.2 opisano metodę spersonalizowanego rozpoznawania obraźliwości opierającą się na miarach konformizmu użytkownika. W Podrozdziale 5.3 zaproponowano metodę spersonalizowanego rozpoznawania obraźliwości wykorzystującą uśrednione reprezentacje semantyczne.

Rozdział 6 Wykorzystanie modelu generatywnego w zadaniu spersonalizowanego rozpoznawania tekstów obraźliwych przedstawiono metodę wykorzystania

generatywnego, dużego modelu językowego ogólnego przeznaczenia ChatGPT-3.5 w zadaniu niespersonalizowanego i spersonalizowanego rozpoznawania tekstów obraźliwych przy użyciu technik inżynierii podpowiedzi, uwzględniających uczenie o zerowej próbie treningowej (ang. *zero-shot prompting*) oraz uczenie w kontekście (ang. *in-context learning*).

W **Rozdziale 7 Metody aktywnego uczenia w zadaniu spersonalizowanego rozpoznawania tekstów obraźliwych** opisano istniejące metody aktywnego uczenia uwzględniające indywidualne anotacje w zadaniu rozpoznawania tekstów obraźliwych. Przedstawiono w nim metodę Ratio Distance (Podrozdział 7.1) oraz Stranger Count (Podrozdział 7.2).

Rozdział 8 Nowe metody aktywnego uczenia w zadaniu spersonalizowanego rozpoznawania tekstów obraźliwych została opisana opracowana miara kontrowersyjności tekstu (Podrozdział 8.1) oraz dwie metody spersonalizowanego aktywnego uczenia w zadaniu rozpoznawania tekstów obraźliwych, uwzględniające kontrowersyjność tekstu (Podrozdział 8.2) oraz stopień dyspersji anotacji (Podrozdział 8.3).

W części trzeciej zatytułowanej **Badania** umieszczono trzy ostatnie rozdziały.

W **Rozdziale 9 Badania eksperymentalne** skupiono się na przedstawieniu wykorzystanych zbiorów danych (Podrozdział 9.1), zdefiniowaniu scenariuszy eksperymentalnych (Podrozdział 9.2), omówieniu wykorzystanych testów statystycznych (Podrozdział 9.3), opisie uzyskanych wyników (Podrozdział 9.4) oraz ich analizie (Podrozdział 9.5). W Podrozdziale 9.6 przedstawiono dyskusję nad uzyskanymi wynikami.

Rozdział 10 Podsumowanie stanowi ogólny przegląd wniosków opracowanych na podstawie pracy, skupiając się na jej najważniejszych wynikach i osiągnięciach.

W **Rozdziale 11 Kierunki dalszych badań** zaproponowano dalsze kierunki prac związanych z rozwojem spersonalizowanych metod rozpoznawania obraźliwych tekstów.

2 KONTEKST UŻYTKOWNIKA W PERCEPCJI OBRAŹLIWOŚCI TEKSTU

W tym rozdziale opisano wyniki analizy literaturowej skupionej na pracach dotyczących zadania rozpoznawania treści obraźliwych oraz modelowania preferencji użytkownika.

2.1 ROZPOZNAWANIE TEKSTOWYCH TREŚCI OBRAŹLIWYCH

Metody automatycznego rozpoznawania treści obraźliwych są tematem wielu warsztatów z dziedziny przetwarzania języka naturalnego (ang. *natural language processing, NLP*), takich jak SemEval 2019 (Zampieri i in., 2019b), GermEval 2018 (Wiegand, Siegel i Ruppenhofer, 2018), FIRE/HASOC 2019 (Mandl i in., 2019) oraz PolEval 2019 (Ptaszyński, Pieciukiewicz i Dybala, 2019).

Tradycyjne metody klasyfikacji tekstu nie uwzględniały informacji na temat kontekstu i kolejności słów, np. metoda *bag of words* (*BoW*) (Harris, 1954) lub metoda ważenia częstością termów (ang. *term frequency–inverse document frequency, TF-IDF*) (Sahlgren, Isbister i Olsson, 2018). Reprezentacja tekstu może być również wzbogacona o cechy uzyskane z ontologii (Bloehdorn i Hotho, 2004) lub wordnetów (Janz i in., 2017; Kocoń i in., 2019a; Misiaszek i in., 2014; Piasecki, Broda i Szpakowicz, 2009; Scott i Matwin, 1998). Jako modele klasyfikacyjne tradycyjnie wykorzystuje się maszyny wektorów nośnych (ang. *support vector machine, SVM*) (Razavi i in., 2010) lub regresję logistyczną (Kocoń, Janz i Piasecki, 2018; Kocoń i Maziarz, 2021; Sahlgren, Isbister i Olsson, 2018; Waseem i Hovy, 2016).

Nowsze metody najczęściej wykorzystywały reprezentacje wektorowe słów (Bojanowski i in., 2017a; Wiegand, Siegel i Ruppenhofer, 2018) połączone z reprezentacjami wektorowymi znaków (Augustyniak, Kajdanowicz i Kazienko, 2019). Natomiast jako model klasyfikacyjny stosowało się konwolucyjną sieć neuronową (ang. *convolutional neural network, CNN*) (Zampieri i in., 2019a) lub sieć z długą pamięcią krótkotrwałą (ang. *long short-term memory network, LSTM*) (Yenala i in., 2017).

Obecnie najczęściej stosuje się metody oparte o sieci typu transformer, takie jak BERT (ang. *bidirectional encoder representations from transformers*) (Devlin i in., 2019a), RoBERTa (ang. *robustly optimized BERT approach*) (Liu i in., 2019), ALBERT (ang. *a lite BERT*) (Lan i in., 2019) lub XLNet (Yang i in., 2019).

Niestety większość dotychczas opublikowanych metod opiera się na generowaniu predykcji wyłącznie na podstawie reprezentacji tekstu. Wykorzystanie jakiejkolwiek

dodatkowej wiedzy dotyczy najczęściej dostarczenia informacji na temat kontekstu czasu, hierarchii wątków na forum internetowym lub cech charakterystycznych sieci społecznościowej autora (Ziems, Vigfusson i Morstatter, 2020).

W przypadku prac dotyczących rozpoznawania treści obraźliwych (Modha, Majumder i Mandl, 2018; Risch i Krestel, 2018; Safi Samghabadi i in., 2020) najczęściej wykorzystuje się zbiory opublikowane podczas cyklicznego warsztatu *Workshop on Trolling, Aggression, and Cyberbullying (TRAC)* (Kumar i in., 2020, 2018) organizowanego podczas konferencji *Language Resources and Evaluation Conference (LREC)*. Część badaczy podczas eksperymentów wykorzystywała również zbiór *Wikipedia Talk Labels: Aggression* (Wulczyn, Thain i Dixon, 2017), który zawiera nie tylko oceny obraźliwości tekstów uzyskane poprzez głosowanie większościowe, ale również indywidualne anotacje poszczególnych anotatorów. Część opublikowanych prac dotyczy również wielojęzykowego rozpoznawania treści obraźliwych (Modha, Majumder i Mandl, 2018; Risch i Krestel, 2018; Safi Samghabadi i in., 2020).

Istnieje wiele prac na temat problemu niskiej zgodności anotacji. Prace te dostarczają informacji nie tylko na temat samych anotatorów, ale również tekstów, których dotyczy proces anotacji (Aroyo i Welty, 2013). W tych pracach dopuszcza się istnienie więcej niż jednej prawidłowej oceny dla pojedynczego tekstu. Niska zgodność anotacji może zostać wykorzystana przy podziale anotatorów na grupy spolaryzowane pod kątem poglądów (Akhtar, Basile i Patti, 2020) lub w celu wykrycia osób oszukujących podczas anotacji (Raykar i Yu, 2012; Soberón i in., 2013). Część prac zwraca również uwagę na problem szkodliwego konformizmu, który objawia się poprzez zwiększoną tendencję wśród anotatorów do wystawiania podobnych anotacji. W znacznie mniejszej liczbie prac niska zgodność jest analizowana na poziomie pojedynczych tekstów, w celu pomiaru ich kontrowersyjności lub niejednoznaczności (Aroyo i Welty, 2013). W większości przeanalizowanych prac zauważa się ograniczenie wykorzystania indywidualnych ocen anotatorów wyłącznie do celów eksploracyjnych, skupiających się głównie na obliczeniu wartości współczynnika Kappa Cohena (Cohen, 1960) lub współczynnika Alfa Krippendorffa (Krippendorff, 1970).

Ponadto, podobnie jak w wielu innych zadaniach przetwarzania języka naturalnego, zauważa się preferencję w kierunku wysokiej zgodności anotacji, która poprawia stabilność modelu i skuteczność predykcji. Niestety wymienione korzyści występują wyłącznie w przypadku scenariusza badawczego zakładającego całkowite pominięcie indywidualnych cech poszczególnych anotatorów. Konsekwencją jest opracowanie metod działających skutecznie wyłącznie w przypadku tekstów ocenianych jednoznacznie przez większość użytkowników lub w przypadku użytkowników cechujących się najpopularniejszymi preferencjami. Równocześnie takie metody cechują się znacząco niższą skutecznością w przypadku tekstów kontrowersyjnych, budzących zróżnicowane wrażenia oraz użytkowników o mniej popularnych poglądach.

2.2 ROLA KONTEKSTU UŻYTKOWNIKA W PRZETWARZANIU JĘZYKA NATURALNEGO

Wiele badań skupiających się na opracowaniu metod analizy tekstu wskazuje na potrzebę uwzględniania ocen poszczególnych anotatorów na temat pojedynczego tekstu. Tendencja ta jest szczególnie zauważalna w przypadku zadania rozpoznawania emocji (Chou i Lee, 2019; Kocoń i in., 2019b; Neviarouskaya, Prendinger i Ishizuka, 2009). Wiele prac wskazuje pierwszy język, wiek, poziom wykształcenia jako czynniki mogące mieć wpływ na odbiór treści tekstowych przez użytkownika (Al Kuwatly, Wich i Groh, 2020; Wich, Al Kuwatly i Groh, 2020). Ponadto zwraca się uwagę na kraj pochodzenia (Salminen i in., 2018), płeć (Binns i in., 2017; Bolukbasi i in., 2016; Tatman, 2017; Wojatzki i in., 2018) oraz przynależność rasową (Blodgett i O'Connor, 2017; Davidson, Bhattacharya i Weber, 2019; Sap i in., 2019; Xia, Field i Tsvetkov, 2020). Część prac doszukuje się również źródła różnic ludzkiej percepcji w poglądach politycznych anotatorów (Wich, Bauer i Groh, 2020).

Jednym z podejść zakładających wykorzystanie wielu anotacji pojedynczego tekstu przy równoczesnej kontroli wpływu poglądów użytkownika na ocenę tekstu jest uwzględnienie wyłącznie ekspertów w procesie anotacji (Waseem, 2016). Tak sprecyzowana reguła rodzi ryzyko wyznaczenia zbyt małej liczby anotatorów do procesu anotacji zbioru o określonym rozmiarze (Wiegand, Ruppenhofer i Kleinbauer, 2019) lub opracowania modeli przejawiających skłonności do dyskryminacji mniejszości (Dixon i in., 2018). Trudność w pozyskaniu ekspertów o wysokich kwalifikacjach do zadań anotacji dotyczących specjalistycznych dziedzin była jedną z motywacji badań nad wykorzystaniem ocen osób nie będących ekspertami. Wyniki przeprowadzonych przez badaczy eksperymentów wykazały, że wykorzystanie uśrednionych anotacji osób nie posiadających wiedzy eksperckiej pozwala osiągnąć ekspercki poziom anotacji (Snow i in., 2008). Inne prace koncentrują się na wykorzystaniu Bayesowskich modeli probabilistycznych w celu określenia poziomu zgodności, który następnie może zostać przekształcony w wartość dyskretną, na przykład poprzez określenie wartości progowych (Kara i in., 2015).

2.3 SPERSONALIZOWANE TECHNIKI AKTYWNEGO UCZENIA

Techniki aktywnego uczenia (ang. *active learning*) są stosowane w celu optymalizacji procesu anotacji. Potencjalny zysk wynikający z zastosowania tego typu technik jest szczególnie zauważalny w przypadku zadań subiektywnych, w których pojedynczy tekst powinien zostać zaanotowany przez wielu anotatorów. Klasyczne techniki aktywnego uczenia opierają się na selekcji próbek do anotacji na podstawie stopnia pewno-

ści modelu. W tym celu wykorzystuje się selekcję na podstawie wyznaczonego progu pewności modelu (Prabhu, Dognin i Singh, 2019; Zhou i Sun, 2014) lub najniższego stopnia pewności modelu Lowell, Lipton i Wallace, 2018. Na przestrzeni lat opracowano również techniki wykorzystujące głębokie modele bayesowskie (ang. *deep bayesian models*), które pozwoliły na poprawę skuteczności selekcji próbek głównie dzięki wykorzystaniu normalizacji reprezentacji danych wejściowych w obrębie pakietu próbek przetwarzanych równolegle (ang. *batch normalization*), blokowaniu części sygnału wyjściowego z danej warstwy (ang. *dropout*) oraz zastosowaniu komitetów modeli (ang. *ensemble models*) (Ren i in., 2021; Tan, Du i Buntine, 2021; Wang i in., 2016). Najnowsze metody aktywnego uczenia w dziedzinie przetwarzania języka naturalnego skupiają się na wykorzystaniu modeli opartych o architekturę typu transformer, takich jak BERT (Devlin i in., 2019b), które okazały się skuteczniejsze niż dedykowane modele uczone wyłącznie na zbiorze anotowanym zadaniem docelowym (Lu, Henchion i Mac Namee, 2019; Prabhu, Mohamed i Misra, 2021).

Metody wizualnego aktywnego uczenia (ang. *visual active learning*) są powszechnie używane w zadaniach związanych z klasyfikacją obrazów (Harpale i Yang, 2008; Wang i in., 2016). Najczęściej skupiają się one na wykorzystaniu głębokich sieci konwolucyjnych. Proces wizualnego aktywnego uczenia rozpoczyna się od przedstawienia anotatorowi obrazów-kandydatów do anotacji wraz z predykcjami modelu wyuczonego na dotychczas zaanotowanych danych. Anotator wybiera kolejne próbki kierując się zarówno poprawą błędnych predykcji oraz dbałością o jakość i różnorodność rozszerzanego zbioru danych. Spełnienie pierwszego celu można osiągnąć poprzez wybór tych próbek, dla których model zwrócił najwyższy stopień pewności dla błędnej klasy. Natomiast spełnienie drugiego celu wymaga dogłębnej znajomości dziedziny w celu rzetelnego określenia reprezentatywności pojedynczej próbki jak i całego zbioru jako wiernego odwzorowania rzeczywistego zbioru obrazów, które będą w przyszłości oceniane za pomocą modelu uczonego na opracowywanym zbiorze danych. Takie podejście zazwyczaj pozwala na uzyskanie lepszych rezultatów jednak cechuje się znacznie wyższym kosztem i czasem anotacji. Jest to spowodowane potrzebą analizy każdego obrazu-kandydata przez wykwalifikowanego anotatora.

Większość prac dotyczących metod aktywnego uczenia skupia się na podejściu generalizującym, nieuwzględniającym żadnej formy personalizacji. Jedne z pierwszych prób uwzględnienia indywidualnego kontekstu użytkownika zostały zrealizowane poprzez rozwinięcie metod opierających się na głębokich modelach bayesowskich (Harpale i Yang, 2008; Seifert i Granitzer, 2010). Polegały one na predykcji uzyskania oceny tekstu od danego anotatora niezależnie od tego jaka była wartość samej oceny. Same metody skupiały się na selekcji tekstów, które mają największe szanse na uzyskanie oceny od poszczególnych anotatorów. Wyniki zaprezentowane przez autorów ukazują zysk wynikający z zastosowania opracowanych metod, na przykład w

procesie harmonogramowania procesu anotacji i przyporządkowywania tekstów do oceny przez poszczególnych anotatorów posiadających zróżnicowane doświadczenie oraz wiedzę dziedzinową. Niestety proces ewaluacji zakładał wyłącznie jedną prawidłową ocenę danego tekstu. W obliczu konieczności umożliwienia ekstrakcji rzetelnej wiedzy dziedzinowej ze zbioru danych powstało wiele prac mających na celu zmniejszenie wpływu anotacji indywidualnych użytkowników poprzez częściowe uwzględnienie anotacji zagregowanych. Rozszerzenie zbioru danych o uogólnione anotacje może zostać wykorzystane jako mechanizm stabilizacji wiedzy możliwej do uzyskania ze zbioru. Takie podejście połączone z równoczesnym naciskiem na wykorzystanie indywidualnych anotacji użytkowników pozwala na opracowanie zbioru umożliwiającego skuteczne modelowanie perspektywy użytkownika z zachowaniem kluczowej wiedzy dziedzinowej na temat samego zjawiska, takiego jak obraźliwość w przypadku treści tekstowych (Bernard i in., 2018; Ferdinan i Kocoń, 2023; Kocoń i in., 2023a; Mieleszczenko-Kowszewicz i in., 2023).

Część II

METODY

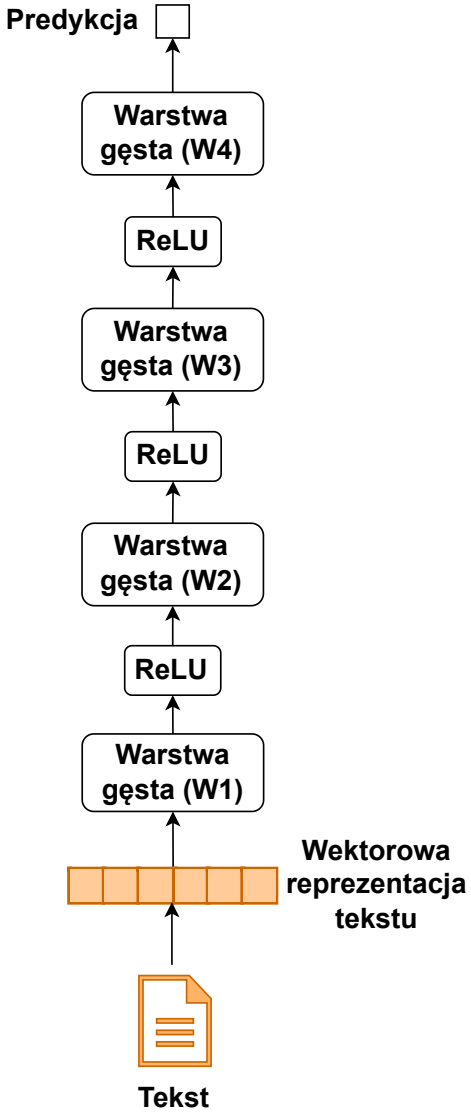
3 REFERENCYJNA METODA UOGÓLNIONEGO ROZPOZNAWANIA OBRAŹLIWYCH TEKSTÓW

W celu zbadania wpływu personalizacji na jakość predykcji modelu wykorzystano metodę referencyjną nazywaną również modelem *Baseline*. Cechą charakterystyczną tej architektury jest wykorzystanie koncepcji uogólnionającej, czyli większościowej. Metoda zwraca taką samą predykcję dla każdego użytkownika. Model ten jest efektem wspólnej pracy w zespole badawczym, której wyniki zostały opublikowane w artykule (Kocoń i in., 2021). Architektura modelu została przedstawiona na Rysunku 2. Można ją opisać za pomocą Wzoru 3:

$$\hat{y}(t) = W_4(W_3(W_2(W_1(t)))) \tag{3}$$

gdzie:

- t oznacza reprezentację wektorową ocenianego tekstu,
- W_1 oznacza wektor wag warstwy gęstej, która otrzymuje na wejściu reprezentację wektorową t ocenianego tekstu,
- W_2 oznacza wektor wag warstwy gęstej, która otrzymuje na wejściu wyjście z warstwy W_1 przekształcone funkcją ReLU (ang. *rectified linear unit*) (Fukushima, 1969),
- W_3 oznacza wektor wag warstwy gęstej, która otrzymuje na wejściu wyjście z warstwy W_2 przekształcone funkcją ReLU,
- W_4 oznacza wektor wag warstwy gęstej, która otrzymuje na wejściu wyjście z warstwy W_3 przekształcone funkcją ReLU.



Rysunek 2: Architektura modelu referencyjnego (*Baseline*). [Źródło: opracowanie własne]

4 METODY SPERSONALIZOWANEGO ROZPOZNAWANIA TEKSTÓW OBRAŹLIWYCH

Ze względu na zauważalny brak spersonalizowanych metod rozpoznawania tekstów obraźliwych, modele omawiane w niniejszym rozdziale zostały opracowane w wyniku współpracy w grupie badawczej.

4.1 METODA HUBI-SIMPLE

Metoda ta skupia się na reprezentacji użytkownika, nazywanej *human bias*(HB), za pomocą jednej liczby, która jest generowana na podstawie unikalnego identyfikatora użytkownika. Liczba ta ma za zadanie określenie stopnia, w jakim preferencje użytkownika różnią się od preferencji innych użytkowników. Metodę obliczania wartości *human bias* opisano za pomocą Wzoru 4:

$$HB(u) = \frac{\sum_{d \in D_u^{past}} \frac{v_{d,u} - \mu_d}{\sigma_d}}{|D_u^{past}|} \tag{4}$$

gdzie:

- u oznacza użytkownika, dla którego obliczamy wartość miary HB,
- $d \in D_u^{past}$ oznacza pojedynczy tekst ze zbioru tekstów zaanotowanych przez użytkownika u,
- $v_{d,u}$ oznacza wartośći anotacji przypisaną przez użytkownika u do tekstu d,
- μ_d oznacza średnią wartość anotacji przypisanych do dokumentu d,
- σ_d oznacza odchylenie standardowe anotacji przypisanych do dokumentu d,
- D_u^{past} oznacza zbiór tekstów zaanotowanych przez użytkownika u.

Ponadto w procesie wnioskowania model ten wykorzystuje reprezentację wektorową tekstu oraz uśrednione wartości obraźliwości dla każdego słowa znajdującego się w tekście nazywane *word bias*. Ostatnia warstwa modelu zwraca predykcję na podstawie konkatenacji wartości *human bias* użytkownika, wyjścia z warstwy gęstej T2 oraz sumy wartości *word bias* dla słów znajdujących się w ocenianym tekście. Model

ten jest efektem wspólnej pracy w zespole badawczym, której wyniki zostały opublikowane w artykule (Kocoń i in., 2021). Architektura modelu została przedstawiona na Rysunku 3. Można ją opisać za pomocą Wzoru 5:

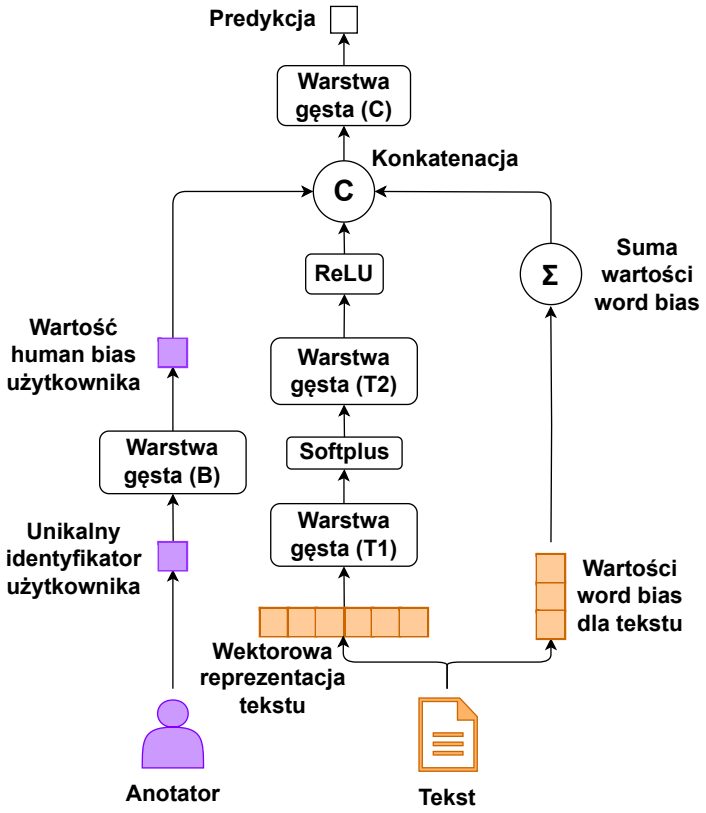
$$\hat{y}(a,t) = W_C(B(a) \frown ReLU(W_{T2}(Softplus(W_{T1}(t)))) \frown \sum_{b \in B_t} b) \tag{5}$$

gdzie:

- a oznacza unikalny identyfikator anotatora,
- t oznacza reprezentację wektorową ocenianego tekstu,
- W_C oznacza oznacza wektor wag warstwy gęstej C otrzymującej na wejściu wynik konkatenacji wyjścia z warstw gęstych B oraz T2 zmodyfikowanej za pomocą funkcji ReLU oraz sumę średnich wartości obraźliwości dla słów znajdujących się w tekście,
- $B(a)$ oznacza wektor wag warstwy E otrzymującej na wejściu unikalny identyfikator anotatora a,
- $\frown$ oznacza operację konkatenacji wektorów,
- $ReLU$ oznacza funkcję aktywacji ReLU,
- W_{T2} oznacza wektor wag warstwy gęstej T2, otrzymującej na wejściu wyjście z warstwy gęstej T1 zmodyfikowanej funkcją aktywacji Softplus,
- $Softplus$ oznacza funkcję aktywacji Softplus,
- W_{T1} oznacza wektor wag warstwy gęstej T1 otrzymującej na wejściu wektorową reprezentację ocenianego tekstu t,
- $\sum_{b \in B_t} b$ oznacza sumę średnich wartości obraźliwości dla słów znajdujących się w tekście,
- b oznacza pojedyncze słowo znajdujące się w tekście t,
- B_t oznacza zbiór słów znajdujących się w tekście t.

4.2 METODA HUBI-MEDIUM

Metoda ta jest rozwinięciem metody HuBi-Simple opisanej w Podrozdziale 4.1. Cechą szczególną jest rozbudowanie reprezentacji użytkownika, którego odrębność preferencji jest w tej architekturze reprezentowana jest poprzez wieloelementowy wektor.



Rysunek 3: Architektura modelu HuBi-Simple. [Źródło: opracowanie własne]

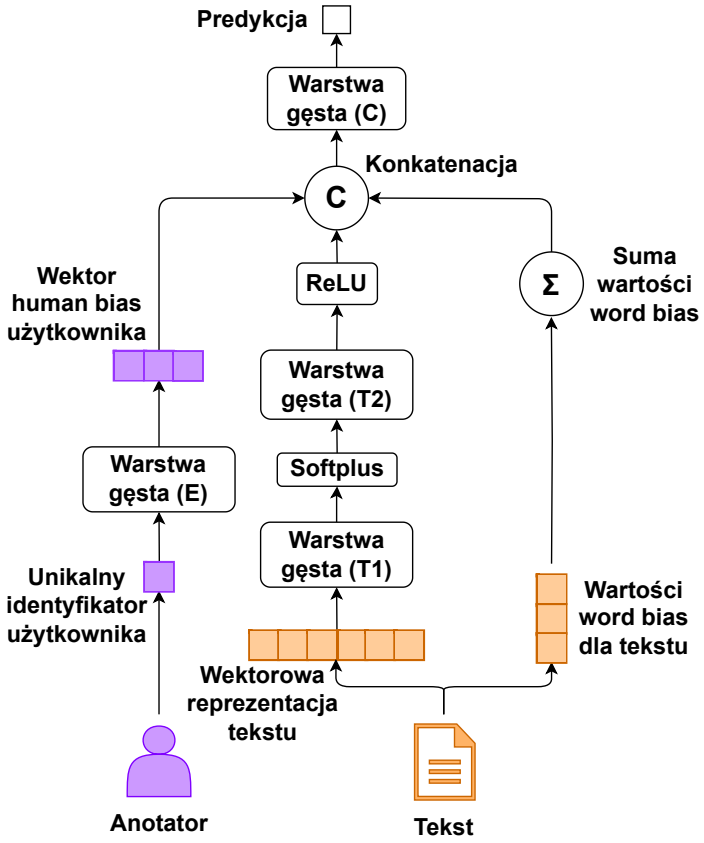
Stanowi to rozszerzenie koncepcji kontekstu użytkownika znanej z modelu HuBi-Simple, gdzie anotator był reprezentowany za pomocą pojedynczej wartości. Architektura ta, podobnie jak HuBi-Simple, wykorzystuje reprezentację użytkownika, osadzenie tekstu i sumę średnich obraźliwości słów zawartych w ocenianym tekście w celu zwrócenia finalnej predykcji. Metoda ta powstała jako wynik współpracy w zespole badawczym, której efekty zostały zaprezentowane w artykule (Kocoń i in., 2021). Architekturę modelu przedstawiono na Rysunku 4. Została również opisana za pomocą Wzoru 6:

$$\hat{y}(a,t) = W_C(E(a) \frown ReLU(W_{T2}(Softplus(W_{T1}(t)))) \frown \sum_{b \in B_t} b) \tag{6}$$

gdzie:

- a oznacza identyfikator anotatora,
- t oznacza osadzenie ocenianego tekstu,
- W_C oznacza oznacza wektor wag warstwy gęstej C przyjmującej jako dane wejściowe wynik konkatenacji wyjścia z warstw gęstych E oraz T2 przetworzonej za pomocą funkcji aktywacji ReLU oraz sumę uśrednionych wartości obraźliwości dla słów z tekstu t,

- $E(a)$ oznacza wektor wag warstwy E przyjmującej jako dane wejściowe identyfikator anotatora a,
- $\frown$ oznacza operację konkatenacji wektorów,
- $ReLU$ oznacza funkcję aktywacji ReLU,
- W_{T2} oznacza wektor wag warstwy gęstej T2, przetwarzającej wyjście z warstwy gęstej T1 uprzednio przetworzone funkcją Softplus,
- $Softplus$ oznacza funkcję Softplus,
- W_{T1} oznacza wektor wag warstwy gęstej T1 przetwarzającej osadzenie ocenianego tekstu t,
- $\sum_{b \in B_t} b$ oznacza operację sumy uogólnionej na średnich wartościach obraźliwości dla wszystkich słów znajdujących się w tekście t,
- b oznacza słowo znajdujące się w tekście t,
- B_t oznacza zbiór słów znajdujących się w tekście t.



Rysunek 4: Architektura modelu HuBi-Medium. [Źródło: opracowanie własne]

5 NOWE METODY SPERSONALIZOWANEGO ROZPOZNAWANIA TEKSTÓW OBRAŹLIWYCH

W celu umożliwienia uzyskiwania spersonalizowanej predykcji obraźliwości dla każdego użytkownika, należy uwzględnić jego kontekst w procesie wnioskowania. W niniejszym rozdziale zaproponowano nowe architektury modeli służących do spersonalizowanego rozpoznawania tekstów obraźliwych.

5.1 MIARY KONFORMIZMU UŻYTKOWNIKA

W celu zmierzenia odrębności perspektywy danego użytkownika od reszty zbiorowości opracowano miary ogólnego i ważonego konformizmu. Miara ogólnego konformizmu $GConf(a, c)$ określa częstość anotacji tekstu przez danego użytkownika a zgodnie z głosem większości. Jej wartość oblicza się dla każdej możliwej klasy osobno. Miarę tę można opisać za pomocą Wzoru 7:

$$GConf(a, c) = \frac{\sum_{d \in A_a} \mathbb{1}_{\{l_d = c \wedge l_d = l_{d,a}\}}}{\sum_{d \in A_a} \mathbb{1}_{\{l_d = c\}}} \tag{7}$$

gdzie:

- a oznacza użytkownika, dla którego obliczamy wartość ogólnego konformizmu,
- c oznacza klasę, dla której obliczana jest wartość miary ogólnego konformizmu,
- d oznacza pojedynczy tekst oceniony przez danego użytkownika,
- A_a oznacza zbiór dokumentów zaanotowanych przez użytkownika a,
- l_d jest klasą wybraną przez większość anotatorów dla dokumentu d,
- $l_{d,a}$ oznacza anotację przypisaną przez użytkownika a do tekstu d.

Miara $GConf(a, c)$ osiąga wartości z przedziału $[0, 1]$, gdzie wartość $GConf(a, c) = 1$ oznacza, że użytkownik a zaanotował wszystkie teksty zgodnie z głosem większości, a wartość $GConf(a, c) = 0$ oznacza, że wszystkie anotacje użytkownika a są niezgodne z anotacjami większości dla poszczególnych tekstów $d \in A_a$.

Miara ważonego konformizmu $WConf(a, c)$ jest swoistym rozwinięciem miary ogólnego konformizmu. Kładzie ona nacisk na stosunek rozmiaru grupy anotatorów n_d^c, którzy zaanotowali tekst d tą samą klasą c, którą zaanotował ten tekst użytkownik

a do rozmiaru grupy wszystkich użytkowników n_d, którzy zaanotowali ten tekst. Miara $WConf(a,c)$ jest wyznaczana dla wszystkich tekstów, które anotator a zaanotował klasą c, a następnie uśredniana przez liczbę tekstów, dla której została obliczona. Miarę tę określa się osobno dla każdej możliwej klasy obraźliwości $c \in C$. Można ją opisać za pomocą Wzoru 8:

$$WConf(a,c) = \frac{\sum_{d \in A_a} \sum_{c \in C} \frac{n_d^c}{n_d} \mathbb{1}_{\{l_{d,a}=c\}}}{\sum_{d \in A_a} \mathbb{1}_{\{l_{d,a}=c\}}} \tag{8}$$

gdzie:

- a oznacza użytkownika, dla którego obliczamy wartość ważonego konformizmu,
- c oznacza klasę, dla której obliczana jest wartość miary ważonego konformizmu,
- d oznacza pojedynczy tekst oceniony przez danego użytkownika,
- A_a oznacza zbiór dokumentów zaanotowanych przez użytkownika a,
- n_d^c oznacza liczbę użytkowników, którzy zaanotowali tekst d klasą obraźliwości c,
- n_d oznacza liczbę wszystkich użytkowników, którzy zaanotowali tekst d,
- $l_{d,a}$ oznacza anotację przypisaną przez użytkownika a do tekstu d.

5.2 METODA USERCONF

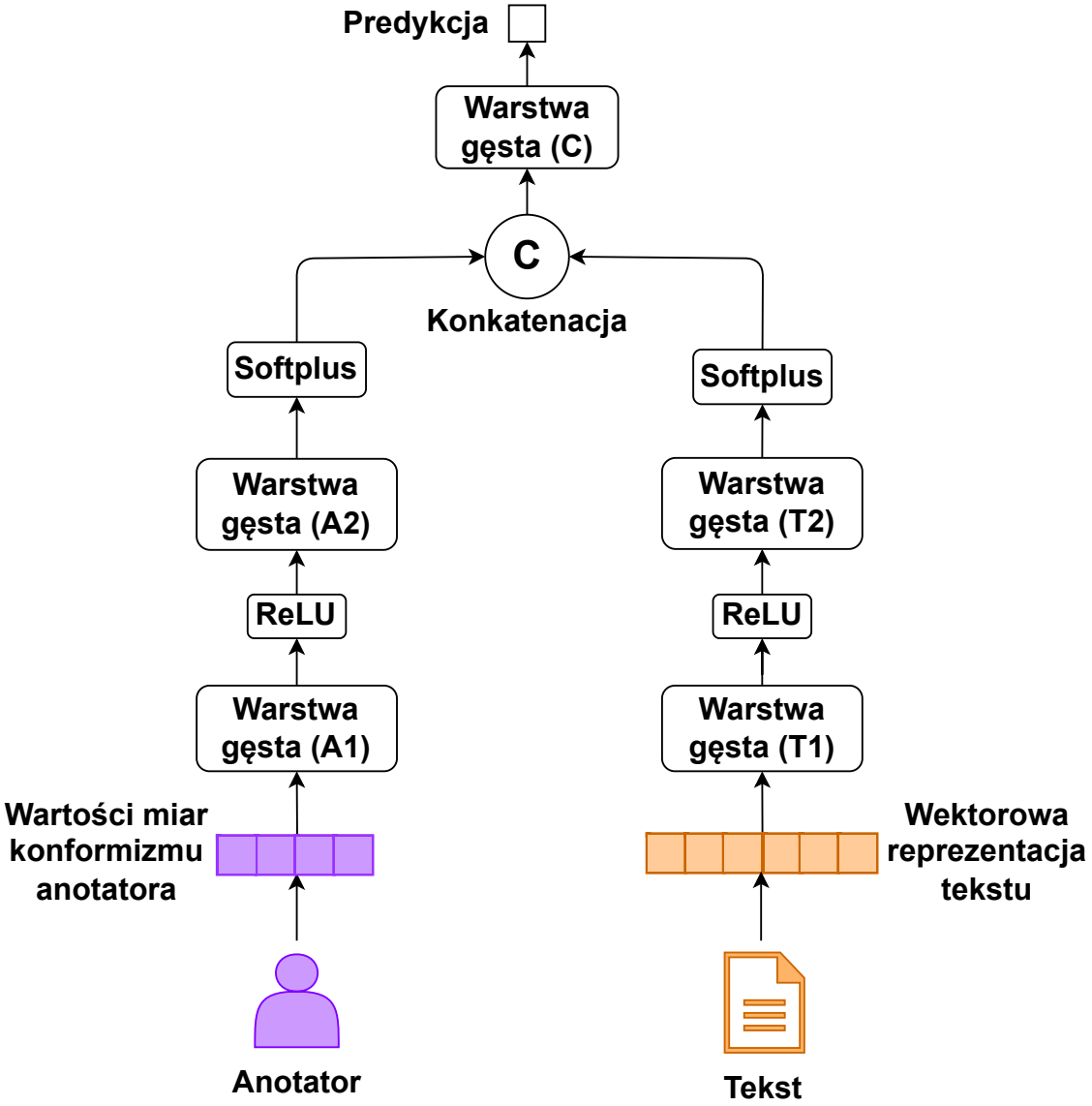
Pierwsza metoda spersonalizowanego rozpoznawania tekstów obraźliwych skupia się na wykorzystaniu miary konformizmu. Miara ta określa w jakim stopniu dany anotator odróżnia się swoją percepcją obraźliwości od innych anotatorów. W celu wykorzystania miary konformizmu użytkownika w zadaniu spersonalizowanego rozpoznawania tekstów obraźliwych opracowano architekturę modelu głębokiej sieci neuronowej o nazwie UserConf. Na wejściu model przyjmuje reprezentację wektorową ocenianego tekstu oraz wartości miar konformizmu dla danego użytkownika. Wektor wartości miar konformizmu anotatora zostaje umieszczony na wejściu warstwy gęstej A1 (ang. *fully connected layer (FC)*). Wyjście z warstwy gęstej A1 zostaje zmodyfikowane za pomocą funkcji aktywacji ReLU (Fukushima, 1969). Wynik funkcji aktywacji zostaje przekazany do drugiej warstwy gęstej A2. Wyjście z warstwy gęstej A2 zostaje zmodyfikowane za pomocą funkcji Softplus (Dugas i in., 2000). Dla tekstu, którego obraźliwość będzie predykowana zostaje wygenerowana reprezentacja wektorowa. Następnie reprezentacja wektorowa zostaje umieszczona na wejściu

warstwy gęstej T1. Wyjście z warstwy T1 zostaje zmodyfikowane za pomocą funkcji aktywacji ReLU. Zmodyfikowane wyjście z warstwy T1 zostaje umieszczone na wejściu warstwy gęstej T2. Wyjście z warstwy gęstej T2 zostaje zmodyfikowane za pomocą funkcji Softplus. Następnie zmodyfikowane wyjścia z warstw A2 oraz T2 zostają poddane operacji konkatenacji. Wynik konkratenacji zostaje umiesczony na wejściu warstwy gęstej C. Wyjście z warstwy gęstej C zostaje zmodyfikowane za pomocą funkcji aktywacji Softmax (Bridle, 1989) w celu uzyskania finalnej predykcji. Architekturę modelu zaprezentowano na Rysunku 5. Metodę wykorzystującą miary konformizmu użytkownika można również opisać za pomocą Wzoru 9:

$$\hat{y}(a,t) = W_C(Softplus(W_{A2}(ReLU(W_{A1}(a)))) \frown Softplus(W_{T2}(ReLU(W_{T1}(t))))) \quad (9)$$

gdzie:

- a oznacza wektor zawierający wartości miar konformizmu użytkownika, dla którego predykujemy obraźliwość tekstu,
- t oznacza reprezentację wektorową ocenianego tekstu,
- W_C oznacza wektor wag warstwy gęstej C otrzymującej na wejściu wynik konkatenacji wyjść z warstw gęstych A2 oraz T2 uprzednio zmodyfikowanych funkcją aktywacji Softplus,
- $Softplus$ oznacza funkcję aktywacji Softplus,
- W_{A2} oznacza wektor wag warstwy gęstej A2, która otrzymuje na wejściu wyjście z warstwy gęstej A1 uprzednio zmodyfikowane funkcją aktywacji ReLU,
- $ReLU$ oznacza funkcję aktywacji ReLU,
- W_{A1} oznacza wektor wag warstwy gęstej A1, która otrzymuje na wejściu wektor a zawierający wartości miar konformizmu użytkownika,
- $\frown$ oznacza operację konkatenacji dwóch wektorów,
- W_{T2} oznacza wektor wag warstwy gęstej T2, która otrzymuje na wejściu wyjście z warstwy gęstej T1 uprzednio zmodyfikowane funkcją aktywacji ReLU,
- W_{T1} oznacza wektor wag warstwy gęstej T1, która otrzymuje na wejściu reprezentację wektorową t ocenianego tekstu.



Rysunek 5: Architektura modelu UserConf wykorzystującego miary konformizmu użytkownika. [Źródło: opracowanie własne]

5.3 METODA USEREMB

Metoda UserEmb uwzględniająca kontekst użytkownika w procesie spersonalizowanego rozpoznawania tekstów obraźliwych zakłada reprezentację indywidualnej perspektywy użytkownika poprzez uśrednione reprezentacje semantyczne zaanotowanych przez niego tekstów. W tym wypadku wszystkie teksty ocenione przez anotatora zostają podzielone na K podzbiorów, gdzie K oznacza liczbę możliwych klas w danym zadaniu anotacji obraźliwości tekstu. W przypadku klasyfikacji binarnej ($K = \{p, n\}$) wszystkie teksty zaanotowane przez użytkownika a zostają podzielone na dwa podzbiory: (1) podzbiór tekstów ocenionych przez użytkownika a klasą pozytywną p oraz (2) podzbiór tekstów ocenionych przez użytkownika klasą negatywną n. Wydzielone podzbiory zostają następnie wykorzystane do wygenerowania wektorowych reprezentacji tekstów znajdujących się w każdym podzbiorze. Następnie

dla każdego podzbioru zostaje obliczona uśredniona reprezentacja wektorowa znajdujących się w nim tekstów. Otrzymane uśrednione osadzenia tekstów t_a^p oraz t_a^n mają za zadanie reprezentować semantykę tekstów, które anotator a uważa za obraźliwe lub nieobraźliwe. Uśrednione reprezentacje wektorowe t_a^p oraz t_a^n zostają wykorzystane w procesie konkatenacji do uzyskania reprezentacji perspektywy użytkownika a. Skonkatenowany wektor zostaje umieszczony na wejściu warstwy gęstej A1. Wyjście z warstwy gęstej A1 zostaje zmodyfikowane za pomocą funkcji aktywacji ReLU. Wektor wartości funkcji aktywacji ReLU zostaje umieszczona na wejściu warstwy gęstej A2. Wyjście z warstwy gęstej A2 zostaje zmodyfikowane za pomocą funkcji aktywacji Softplus. Dla tekstu, którego obraźliwość jest predykowana, zostaje wygenerowana reprezentacja wektorowa. Następnie uzyskana reprezentacja wektorowa zostaje umieszczona na wejściu warstwy gęstej T1. Wyjście z warstwy gęstej T1 zostaje zmodyfikowane za pomocą funkcji aktywacji ReLU. Wynik funkcji aktywacji ReLU zostaje umieszczony na wejściu warstwy gęstej T2. Wyjście warstwy gęstej T2 zostaje zmodyfikowane funkcją aktywacji Softplus. W kolejnym kroku wyjścia z warstw gęstych A2 oraz T2 osobno zmodyfikowane funkcją aktywacji Softmax zostają poddane operacji konkatenacji. Wynik tej operacji zostaje umieszczony na wejściu warstwy gęstej C. Wektor otrzymany na wyjściu warstwy gęstej C zostaje zmodyfikowany za pomocą funkcji aktywacji Softmax w celu uzyskania finalnej predykcji.

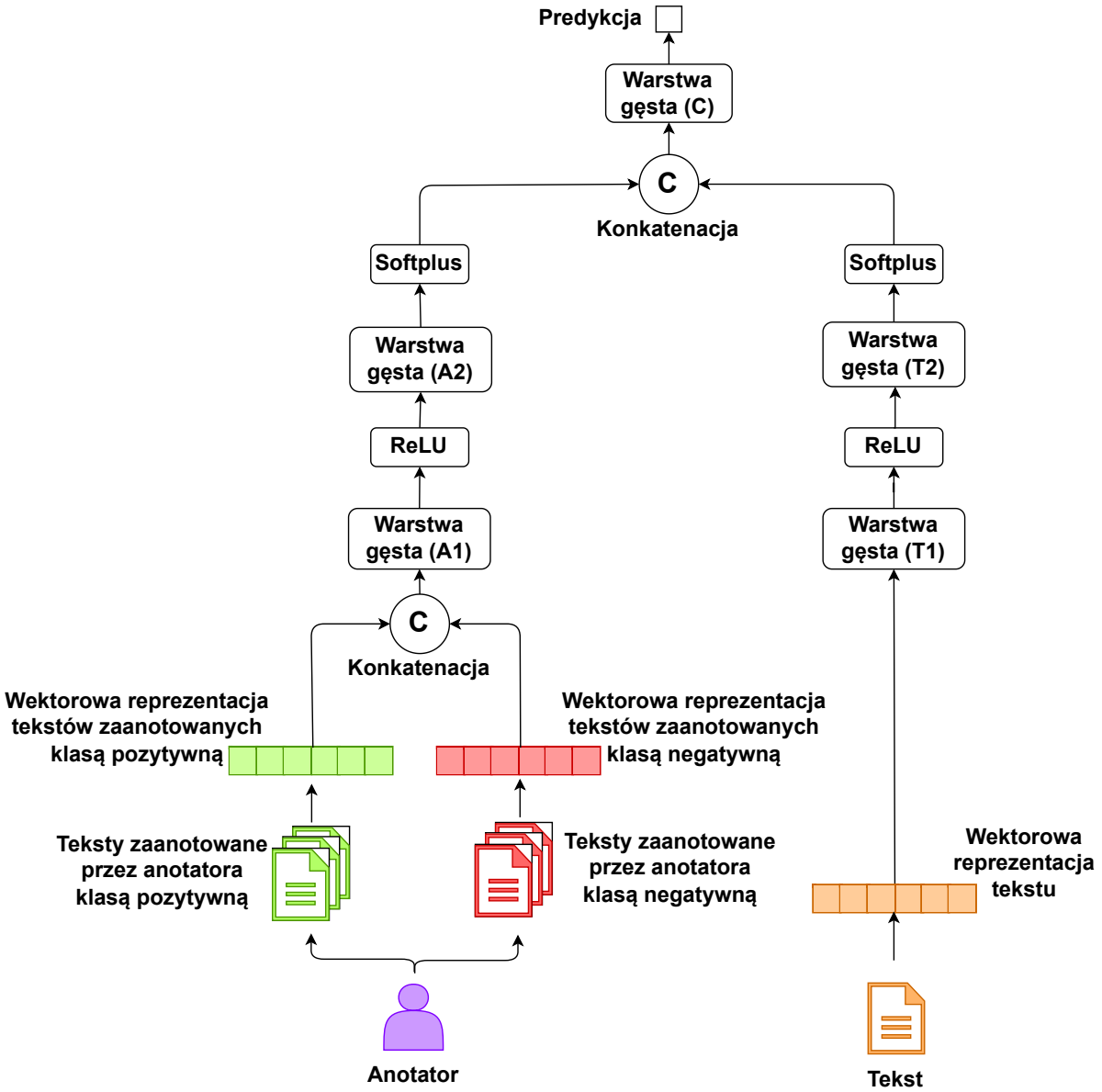
Architekturę modelu dla problemu klasyfikacji binarnej przedstawiono na Rysunku 6. Metoda wykorzystująca uśrednione reprezentacje semantyczne w zadaniu klasyfikacji binarnej może zostać opisana za pomocą Wzoru 10:

$$\hat{y}(a,t) = W_C(Softplus(W_{A2}(ReLU(W_{A1}(t_a^p \frown t_a^n)))) \frown Softplus(W_{T2}(ReLU(W_{T1}(t))))) \quad (10)$$

gdzie:

- a oznacza anotatora, dla którego predykujemy obraźliwość określonego tekstu,
- t oznacza reprezentację wektorową tekstu, którego obraźliwość jest predykowana,
- W_C oznacza wektor wag warstwy gęstej C otrzymującej na wejściu konkatenację wektorów wartości funkcji aktywacji Softplus obliczonej osobno dla wyjść z warstw gęstych A2 oraz T2,
- $Softplus$ oznacza funkcję aktywacji Softplus,
- W_{A2} oznacza wektor wag warstwy gęstej A2, która otrzymuje na wejściu wektor wartości funkcji aktywacji ReLU obliczonej dla wyjścia z warstwy gęstej A1,
- $ReLU$ oznacza funkcję aktywacji ReLU,
- W_{A1} oznacza wektor wag warstwy gęstej A1, która otrzymuje na wejściu konkatenację uśrednionej reprezentacji wektorowej t_a^p tekstów ocenionych przez anotatora a klasą pozytywną oraz uśrednionej reprezentacji wektorowej t_a^n, które anotator a zaanotował klasą negatywną,

- t_a^p oznacza uśrednioną reprezentację wektorową tekstów, które anotator a ocenił klasą pozytywną,
- $\frown$ oznacza operację konkatenacji dwóch wektorów,
- t_a^n oznacza uśrednioną reprezentację wektorową tekstów, które anotator a zaanotował klasą negatywną,
- W_{T2} oznacza wektor wag warstwy gęstej T2, która otrzymuje na wejściu wektor wartości funkcji aktywacji ReLU obliczoną dla wyjścia z warstwy gęstej T1,
- W_{T1} oznacza wektor wag warstwy gęstej T1, która otrzymuje na wejściu reprezentację wektorową t ocenianego tekstu.



Rysunek 6: Architektura modelu UserEmb wykorzystującego uśrednione reprezentacje semantyczne dla problemu klasyfikacji binarnej. [Źródło: opracowanie własne]

6 WYKORZYSTANIE MODELU GENERATYWNEGO W ZADANIU SPERSONALIZOWANEGO ROZPOZNAWANIA TEKSTÓW OBRAŹLIWYCH

Nowoczesne, duże generatywne modele językowe (ang. *Large Language Models, LLM*) również mogą zostać wykorzystane w zadaniach uogólnionego oraz spersonalizowanego rozpoznawania treści obraźliwych. Generatywny charakter modeli LLM kładzie dodatkowy nacisk na zrozumienie semantyki tekstów wykorzystanych w procesie uczenia. Pomaga to skupić się modelowi na kluczowych aspektach lingwistycznych i tym samym wykryć niestandardowe kolokacje oraz części wyrazów (ang. *word tokens*), które rzadko występują obok siebie. Czułość na zaburzenia kontekstu połączona z dogłębną procesem uczenia nakierowanym na estymację rozkładu słownictwa w dużym zbiorze uczącym zawierającym najczęściej teksty z wielu dziedzin może umożliwić modelowi bardziej szczegółową analizę ocenianego tekstu w przeciwieństwie do modeli dyskryminatywnych. Ponadto, mnogość tekstów, które były wykorzystane w procesie uczenia oraz kalibracji (ang. *fine-tuning*) takiego modelu pozwala na skuteczne wykorzystanie go w wielu zróżnicowanych zadaniach bez konieczności douczania na zbiorze danych dedykowanym danemu zadaniu (Kaplan i in., 2020). W celu ewaluacji jakości predykcji dużego modelu językowego ogólnego przeznaczenia, wykorzystano model GPT-3.5 Turbo (OpenAI, 2023). W przypadku ewaluacji zakładającej uogólnione rozpoznawanie tekstów obraźliwych w treści tekstu wejściowego (ang. *prompt*) umieszczono opis zadania oceny obraźliwości wraz z treścią tekstu do oceny. Natomiast dla zadania spersonalizowanego rozpoznawania tekstów obraźliwych, tekst wejściowy został wzbogacony o trzy przykłady tekstów, z których dwa zostały przez użytkownika ocenione jako obraźliwe, a jeden jako nieobraźliwy. Miało to na celu wykorzystanie zdolności modelu do zmiany generowanej sekwencji na podstawie dostarczonych przykładów (ang. *few-shot prompting*) (Brown i in., 2020). W celu ułatwienia procesu ewaluacji każdy tekst wejściowy został rozszerzony o informację, aby odpowiedź modelu miała formę listy w języku Python.

7 METODY AKTYWNEGO UCZENIA W ZADANIU SPERSONALIZOWANEGO ROZPOZNAWANIA TEKSTÓW OBRAŹLIWYCH

W konsekwencji zauważalnego braku spersonalizowanych metod aktywnego uczenia dla zadania rozpoznawania tekstów obraźliwych, metody opisane w niniejszym rozdziale zostały przygotowane w efekcie wspólnej pracy w grupie badawczej.

7.1 METODA RATIO DISTANCE

Metoda Ratio Distance (Kanclerz i in., 2023b), nazywana również *RatioDist*, skupia się na różnicy pomiędzy stosunkiem $ratio(d)$ anotatorów, którzy zaanotowali dany tekst jako obraźliwy ($l_{d,u} = 1$) a uśrednionym stosunkiem $avg\_ratio(u)$ liczby takich anotatorów dla wszystkich tekstów d zaanotowanych przez użytkownika u ($d \in D_u$). Przykładowe zastosowanie metody Ratio Distance przedstawiono na Rysunku 7, Sposób obliczania stosunku $ratio(d)$ opisano we Wzorze 11

$$ratio(d) = \frac{\sum_{u \in U_d} \mathbb{1}_{\{l_{d,u}=1\}}}{\sum_{u \in U_d} \mathbb{1}_{\{l_d \in C\}}} \tag{11}$$

gdzie:

- d oznacza tekst, dla którego obliczana jest wartość miary $ratio(d)$,
- U_d oznacza zbiór wszystkich anotatorów, którzy zaanotowali tekst d,
- $l_{d,u}$ oznacza wartość anotacji przypisaną przez anotatora u do tekstu d,
- l_d oznacza pojedynczą anotację dokumentu d,
- C oznacza zbiór wszystkich możliwych klas obraźliwości w danym procesie anotacji.

Metodę obliczania $avg\_ratio(u)$ określono we Wzorze 12:

$$avg\_ratio(u) = \frac{\sum_{d \in D_u} ratio(d)}{|D_u|} \tag{12}$$

gdzie:

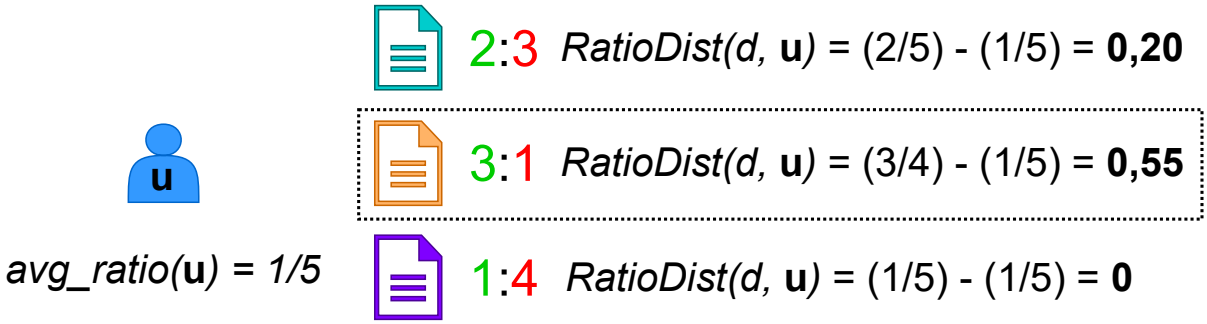
- u oznacza pojedynczego anotatora,
- d oznacza pojedynczy tekst,
- D_u oznacza zbiór tekstów zaanotowanych przez anotatora u.

Finalnie, w celu uzyskania wartości $RatioDist(a, t)$ należy obliczyć różnicę pomiędzy $ratio(d)$ a $avg\_ratio(u)$ zgodnie ze Wzorem 13:

$$RatioDist(d, u) = ratio(d) - avg\_ratio(u) \tag{13}$$

gdzie:

- d oznacza pojedynczy tekst,
- u oznacza pojedynczego anotatora.



Rysunek 7: Przykładowe zastosowanie metody Ratio Distance. [Źródło: opracowanie własne]

7.2 METODA STRANGER COUNT

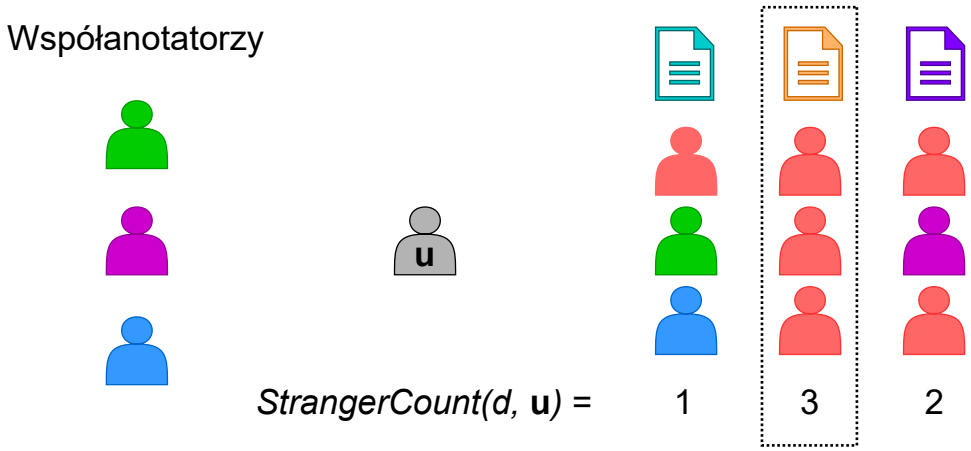
Metoda Stranger Count (Kanclerz i in., 2023b) kładzie nacisk na maksymalizację liczby unikalnych anotatorów, z którymi dany anotator anotuje te same teksty. Głównym założeniem jest umożliwienie uzyskania bardziej relewantnej reprezentacji użytkownika poprzez zestawienie jego anotacji z jak największą liczbą innych anotatorów. Powinno to umożliwić bardziej precyzyjne uwydatnienie odrębności jego preferencji względem reszty anotatorów. Wartość miary określającej jak bardzo dany tekst d powinien być zaanotowany przez użytkownika u określa się poprzez obliczenie różnicy między liczbą anotatorów $|U_d|$, którzy zaanotowali tekst d a liczbą współanotatorów $a \in U_d$, którzy zaanotowali przynajmniej jeden tekst uprzednio zaanotowany przez anoatatora u ($|D_u \cap D_a| > 0$). Przykładowe zastosowanie metody Stranger Count zaprezentowano na Rysunku 8. Sposób obliczania wartości $StrangerCount(d, u)$ został opisany we Wzorze 14:

$$StrangerCount(d, u) = |U_d| - |\sum_{a \in U_d} \mathbb{1}_{\{|D_u \cap D_a| > 0\}}| \tag{14}$$

gdzie:

- d oznacza pojedynczy tekst, dla którego obliczamy wartość $StrangerCount(d, u)$,
- u oznacza użytkownika, dla którego obliczamy wartość $StrangerCount(d, u)$,
- U_d oznacza zbiór użytkowników, którzy zaanotowali tekst d,

- a oznacza pojedynczego użytkownika, który już zaanotował tekst d,
- D_u oznacza zbiór tekstów zaanotowanych przez użytkownika u,
- D_a oznacza zbiór tekstów zaanotowanych przez użytkownika a.



Rysunek 8: Przykładowe zastosowanie metody Stranger Count. Kolorem czerwonym zostali oznaczeni anotatorzy, którzy nie zaanotowali ani jednego tekstu spośród tekstów zaanotowanych przez anotatora szarego. [Źródło: opracowanie własne]

8 NOWE METODY AKTYWNEGO UCZENIA W ZADANIU SPERSONALIZOWANEGO ROZPOZNAWANIA TEKSTÓW OBRAŹLIWYCH

Istotnym czynnikiem wpływającym na jakość predykcji modeli uczenia maszynowego jest jakość tekstów wybranych do anotacji w procesie przygotowywania zbioru uczącego. Opracowanie zbiorów danych niezbędnych w procesie uczenia spersonalizowanych architektur służących do predykcji obraźliwości tekstu wymaga pozyskania anotacji dla każdego tekstu dokonanych przez wielu anotatorów. Tak zdefiniowany proces anotacji jest znacznie bardziej kosztowny i czasochłonny w porównaniu z konwencjonalnymi scenariuszami anotacji, zakładającymi anotowanie każdej próbki przez 2 ekspertów połączone z późniejszą korektą ze strony kierownika anotacji. W celu optymalizacji kosztów i czasu anotacji opracowano dwie nowe metody spersonalizowanego aktywnego uczenia dla zadania rozpoznawania tekstów obraźliwych. Pierwsza metoda wykorzystuje wiedzę na temat kontrowersyjności tekstu, natomiast druga skupia się na stopniu dyspersji anotacji poszczególnych tekstów.

8.1 MIARA KONTROWERSYJNOŚCI TEKSTU

Efektywne modelowanie kontekstu użytkownika wymaga wykorzystania zbioru tekstów, który cechuje się zarówno akceptowalnym stopniem reprezentowania przestrzeni semantycznej, jak i zawieraniem przykładów tekstów, których obraźliwość jest silnie zależna od perspektywy użytkownika. Poprawnie przeprowadzona selekcja tekstów umożliwia wyuczenie modelu, który będzie w stanie wykorzystać informację na temat preferencji użytkownika w zależności od semantyczno-pragmatycznych właściwości ocenianego tekstu. W procesie selekcji tekstów warto wykorzystać przykłady, które skutecznie dzielą anotatorów, co można zaobserwować na podstawie wysokiego stopnia zróżnicowania anotacji danego tekstu.

W celu określenia rozbieżności anotacji w obrębie pojedynczego tekstu d opracowano miarę kontrowersyjności $Contr(d)$. Przyjmuje ona wartości z przedziału $[0, 1]$, gdzie wartość $Contr(d) = 1$ jest osiągana wyłącznie w sytuacji, gdy tekst d został zaanotowany każdą możliwą klasą obraźliwości przez taką samą liczbę anotatorów. Natomiast wartość $Contr(d) = 0$ występuje wyłącznie w sytuacji, gdy wszyscy anotatorzy oceniający tekst d przypisali do niego tę samą klasę obraźliwości. Miarę kontrowersyjności tekstu opisano za pomocą Wzoru 15:

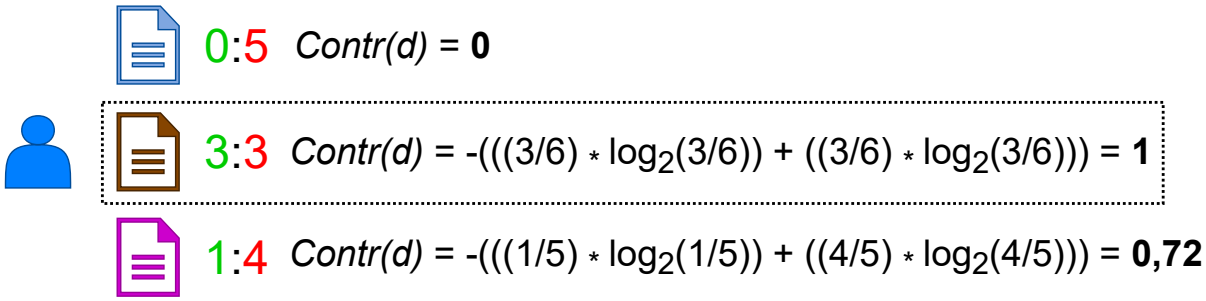
$$Contr(d) = \begin{cases} 0 & \text{, jeżeli } \bigvee_{c \in C} n_d^c = n_d \\ -\sum_{c \in C} \frac{n_d^c}{n_d} \log_2 \left(\frac{n_d^c}{n_d} \right) & \text{, w pozostałych przypadkach} \end{cases} \tag{15}$$

gdzie:

- d oznacza tekst, dla którego obliczana jest wartość miary kontrowersyjności,
- n_d^c oznacza liczbę anotatorów, którzy zaanotowali tekst d klasą obraźliwości c,
- n_d oznacza liczbę anotatorów, którzy zaanotowali tekst d,
- c oznacza pojedynczą klasę obraźliwości,
- C oznacza zbiór wszystkich klas obraźliwości możliwych do przypisania danemu tekstowi d w procesie anotacji.

8.2 METODA OPARTA O KONTROWERSYJNOŚĆ TEKSTU

Pierwsza opracowana metoda spersonalizowanego aktywnego uczenia skupia się na wykorzystaniu miary kontrowersyjności opisanej w Podrozdziale 8.1. Metoda nazywana jest również Kontrowersyjnością w celu skrócenia zapisu nazwy. W przypadku tego podejścia premiowane są teksty, które cechują się rozkładem anotacji najbardziej zbliżonym do rozkładu jednostajnego. Selekcja tekstów o maksymalnie zbalansowanym rozkładzie anotacji pozwala na zbudowanie zbioru danych umożliwiającego modelowi skupienie się na kluczowych cechach preferencji użytkownika. Kładzie to nacisk na zdolność modelu do maksymalnej ekstrakcji wiedzy z reprezentacji anotatora w celu predykcji obraźliwości niejednoznacznych tekstów. Przykładowe zastosowanie metody opartej o miarę kontrowersyjności tekstu dla klasyfikacji binarnej przedstawiono na Rysunku 9.



Rysunek 9: Przykładowe zastosowanie metody aktywnego uczenia opartej o kontrowersyjność tekstu. [Źródło: opracowanie własne]

8.3 METODA OPARTA O STOPIEŃ DYSPERSJI ANOTACJI

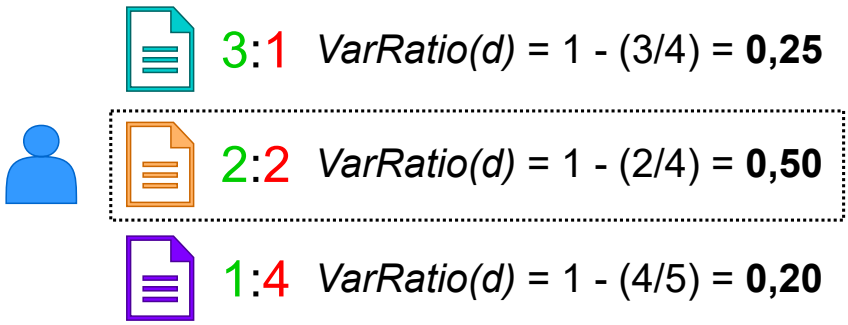
Drugą opracowaną metodą spersonalizowanego aktywnego uczenia jest *VarRatio*. Metoda ta skupia się na pomiarze zróżnicowania anotacji poprzez obliczenie stosunku liczby anotatorów n_d^{mode}, którzy zaanotowali tekst d klasą najczęściej przypisywaną temu tekstowi do liczby wszystkich anotatorów n_d, którzy zaanotowali ten tekst. Miara $VarRatio(d)$ przyjmuje wartości z przedziału $[0, \frac{|C|-1}{|C|}]$, gdzie $|C|$ oznacza liczbę wszystkich możliwych do wyboru klas obraźliwości w danym zbiorze danych. Wartość $VarRatio(d) = 0$ jest osiągana w sytuacji,

gdy wszyscy anotatorzy zaanotowali dany tekst tą samą klasą obraźliwości. Natomiast wartość $VarRatio(d) = \frac{|C|-1}{|C|}$ występuje, gdy liczba anotatorów tekstu d jest taka sama dla każdej z możliwych klas obraźliwości. Przykład zastosowania metody *VarRatio* dla zadania binarnej klasyfikacji obraźliwości przedstawiono na Rysunku 10. Miarę $VarRatio(d)$ można opisać za pomocą Wzoru 16:

$$VarRatio(d) = 1 - \frac{n_d^{mode}}{n_d} \tag{16}$$

gdzie:

- d oznacza tekst, dla którego obliczamy wartość miary $VarRatio(d)$,
- n_d^{mode} oznacza liczbę anotatorów, którzy zaanotowali tekst d, która została przypisana do tego tekstu największą liczbę razy,
- n_d oznacza liczbę wszystkich anotatorów, którzy zaanotowali tekst d.



Rysunek 10: Przykładowe zastosowanie metody VarRatio w zadaniu binarnej klasyfikacji obraźliwości. [Źródło: opracowanie własne]

Część III

BADANIA

9 BADANIA EKSPERYMENTALNE

Zrealizowane badania skupiały się na analizie wpływu rozmiaru zbioru uczącego, metody generowania reprezentacji wektorowej tekstu oraz zastosowanej metody aktywnego uczenia w zadaniu spersonalizowanego rozpoznawania tekstów obraźliwych. W procesie realizacji tych badań wykonano szereg eksperymentów mających na celu określenie w jakim stopniu opracowane metody mogą pomóc w spersonalizowanej ocenie obraźliwości tekstu.

9.1 ZBIORY DANYCH

W celu zbadania jakości predykcji opracowanych metod przeprowadzono ewaluację wykorzystującą różnorodne zbiory zawierające indywidualne anotacje dla każdego tekstu. Oznacza to, że każdy zbiór zawierał zarówno teksty jak i wartości anotacji powiązane z jednoznacznymi identyfikatorami anotatorów, którzy byli autorami poszczególnych anotacji. Ponadto każdy tekst był powiązany z wieloma anotacjami, gdzie każda anotacja została dokonana przez inną osobę. Szczegółowe informacje na temat wykorzystanych zbiorów danych znajdują się w Tabeli 1 oraz w Tabeli 2.

Zbiory Doccano 1 (Podrozdział 9.1.5) oraz Doccano 2 (Podrozdział 9.1.6) zawierające teksty w języku polskim zostały opracowane w ramach pracy w zespole badawczym. W procesie anotacji wzięli udział przeszkoleni pracownicy Politechniki Wrocławskiej. Procedury opracowania zbiorów danych uwzględniały zebranie, analizę i selekcję tekstów do anotacji, przygotowanie listy wymiarów anotacji, implementację oraz konfigurację środowiska dla anotatorów oraz analizę zebranych danych. Wyniki badań na zbiorach Doccano 1 oraz Doccano 2 zostały opublikowane w pracach (Bielaniewicz i Kazienko, 2023; Kanclerz i in., 2023a; Mieleszczenko-Kowszewicz i in., 2023).

Tabela 1: Szczegóły zbiorów WikiDetox: Aggression, WikiDetox: Toxicity oraz Measuring Hate Speech. Dodatkowe informacje na temat poszczególnych zbiorów umieszczono w Podrozdziale 9.1.

Właściwości \ Zbiór danych	WikiDetox: Aggression	WikiDetox: Toxicity	Measuring Hate Speech
Źródło danych tekstowych	Wikipedia	Wikipedia	YouTube, Twitter, Reddit
Liczba tekstów	115864	159686	39565
Liczba anotacji	1365217	1598289	135556
Liczba anotatorów	4053	4301	7912
Średnia liczba anotacji na tekst	11,78	10,01	3,43
Średnia liczba anotacji na anotatora	336,84	371,61	17,13
Język	Angielski	Angielski	Angielski

Tabela 2: Szczegóły zbiorów Unhealthy Conversations, Doccano 1 oraz Doccano 2. Dodatkowe informacje na temat poszczególnych zbiorów umieszczono w Podrozdziale 9.1.

Właściwości \ Zbiór danych	Unhealthy Conversations	Doccano 1	Doccano 2
Źródło danych tekstowych	Globe, Mail	ZnanyLekarz, Onet, Wirtualna Polska	ZnanyLekarz, Onet, Wirtualna Polska
Liczba tekstów	44355	880	8891
Liczba anotacji	227975	31521	17533
Liczba anotatorów	558	39	49
Średnia liczba anotacji na tekst	5,14	35,82	1,97
Średnia liczba anotacji na anotatora	408,56	808,23	357,81
Język	Angielski	Polski	Polski

9.1.1 WikiDetox: Aggression

Cechą charakterystyczną serwisu Wikipedia (Wikipedia, 2001) jest dbałość o jakość zamieszczanych artykułów realizowana między innymi poprzez dyskusje między ich autorami. Niestety, prowadzone dyskusje nieraz cechują się wysokim nasyceniem agresji i ataków personalnych. W celu przeciwdziałania szerzącej się agresji słownej opracowano zbiór WikiDetox: Aggression (Wulczyn, Thain i Dixon, 2017). W procesie anotacji użytkownicy skupili się na anotacji tekstów określających ich agresywność. W efekcie 4053 anotatorów zaanotowało łącznie 115864 teksty.

9.1.2 WikiDetox: Toxicity

Drugim zbiorem opracowanym w celu analizy i automatycznej detekcji negatywnych komentarzy na forum autorów Wikipedii jest WikiDetox: Toxicity (Wulczyn, Thain i Dixon, 2017). W tym wypadku procedura anotacji skupiała się na anotacji tekstów toksycznych. Udział w procesie wzięło 4301 użytkowników, którzy zaanotowali 159686 tekstów.

9.1.3 Measuring Hate Speech

Zbiór Measuring Hate Speech (Kennedy i in., 2020) zawiera 39565 komentarzy pozyskanych z serwisów YouTube (Karim i in., 2005), Twitter (Dorsey i in., 2006) oraz Reddit (Huffman, Swartz i Ohanian, 2005). Zostały one zaanotowane przez 7912 anotatorów będących pracownikami usługi Amazon Mechanical Turk (Harinarayan, 2001) mieszkającymi w Stanach Zjednoczonych. Autorzy zbioru wykorzystywali go głównie w celach eksploracyjnych, gdyż głównym celem projektu było określenie kluczowych czynników wpływających na występowanie obraźliwych tekstów w sekcjach komentarzy popularnych serwisów internetowych.

9.1.4 Unhealthy Conversations

Unhealthy Conversations (Price i in., 2020) jest zbiorem 44355 komentarzy o długości nieprzekraczającej 250 znaków. Teksty zostały pozyskane z serwisów Globe oraz Mail, które były częścią korpusu Simon Fraser University Opinion and Comments Corpus (Kolhatkar i in., 2020). Każdy komentarz był indywidualnie zaanotowany przez minimum 3 anotatorów. Komentarze były przedstawiane do oceny bez dostarczania artykułów, których dotyczyły. Miało to na celu minimalizację wpływu tematyki i słownictwa używanego w artykułach na ocenę obraźliwości komentarzy.

9.1.5 Doccano 1

Zbiór Doccano 1 jest efektem projektu Doccano 1.0 mającego na celu zbadanie wielowymiarowej subiektywnej percepcji tekstu. Teksty w języku polskim są komentarzami z forów internetowych, takich jak znanylekarz.pl, onet.pl, wp.pl, a ich długość nie przekracza 132 słów ($\mu = 24,5$, $\sigma = 16,2$). Średnio pojedyncza osoba anotowała około 808 tekstów. Statystycznie każdy tekst otrzymał anotacje od około 36 anotatorów. Cały zbiór składa się z około 31521 anotacji, z których każda dotyczy 26 różnych wymiarów: (1) *spokój*, (2) *współczucie*, (3) *zadowolenie*, (4) *inspiracja*, (5) *radość*, (6) *pozytywność*, (7) *zdziwienie*, (8) *złość*, (9) *obrzydzenie*, (10) *strach*, (11) *negatywność*, (12) *smutek*, (13) *zgoda*, (14) *zakłopotanie*, (15) *śmieszne dla mnie*, (16) *śmieszne dla kogoś*, (17) *niezrozumiałe*, (18) *interesujące*, (19) *ironia*, **(20) *obraźliwe dla mnie***, (21) *obraźliwe dla kogoś*, (22) *polityczne*, (23) *sympatia*, (24) *zrozumiałe*, (25) *zaufanie*, i (26) *wulgarne*. Dla każdego tekstu anotatorzy określali indywidualnie obecność każdego z wymiarów. Ano-

tator mógł wstrzymać się od oceny danego wymiaru dla danego tekstu jeżeli w jego ocenie dany wymiar był zupełnie niepowiązany z tematyką ocenianego tekstu. Ze względu na tematykę niniejszej rozprawy podczas eksperymentów skupiono się na wymiarze (20) *obraźliwe dla mnie*.

9.1.6 Doccano 2

Doccano 2 jest kolejnym efektem projektu Doccano 1.0. W ramach procedury anotacji 49 anotatorów zaanotowało 8891 tekstów. Pojedynczy tekst otrzymał anotacje od około 2 anotatorów, w wyniku czego łączna liczba anotacji wynosi 17533. W przypadku tego zbioru procedura anotacji obejmowała 26 niezależnych wymiarów, tożsamych z wymiarami uwzględnionymi w procedurze anotacji dla zbioru Doccano 1 opisanego w Podrozdziale 9.1.5. W tym wypadku również zadaniem anotatorów było określenie nacechowania w tekście każdego z 26 wymiarów z możliwością wstrzymania się od odpowiedzi dla pojedynczych wymiarów jeżeli anotator uznał, że anotacja dla danego wymiaru nie ma zastosowania dla ocenianego tekstu. Z uwagi na badania dotyczące spersonalizowanego rozpoznawania tekstów obraźliwych, w przypadku tego zbioru, eksperymenty ograniczono również do wymiaru (20) *obraźliwe dla mnie*.

9.2 SCENARIUSZE EKSPERYMENTALNE

W celu realizacji badań nad spersonalizowaną percepcją obraźliwości tekstów, sporządzono następujące scenariusze eksperymentalne, mające na celu umożliwienie analizy różnicy jakości predykcji opracowanych metod spersonalizowanych w porównaniu z podejściem niespersonalizowanym zakładające zbadanie:

- Wpływu metody generowania reprezentacji wektorowej tekstu oraz modelu głębokiej sieci neuronowej służącej do predykcji obraźliwości tekstu,
- Wpływu rozmiaru zbioru wykorzystanego w procesie uczenia na jakość predykcji obraźliwości tekstu,
- Wpływu wykorzystania metody aktywnego uczenia na jakość predykcji obraźliwości modeli uczonych na ograniczonym zbiorze danych,
- Jakości spersonalizowanej i niespersonalizowanej predykcji obraźliwości tekstu przez generatywny model językowy ogólnego przeznaczenia ChatGPT-3.5.

Dla każdego scenariusza jakość modeli badano przy pomocy miary F1 macro (Rijsbergen, 1979). W celu zmierzenia różnicy między jakością predykcji opracowanej metody, a metody referencyjnej zastosowano miarę $F1_{roznica}$ opisaną Wzorem 17:

$$F1_{roznica} = F1_{model} - F1_{baseline} * 100 \tag{17}$$

gdzie:

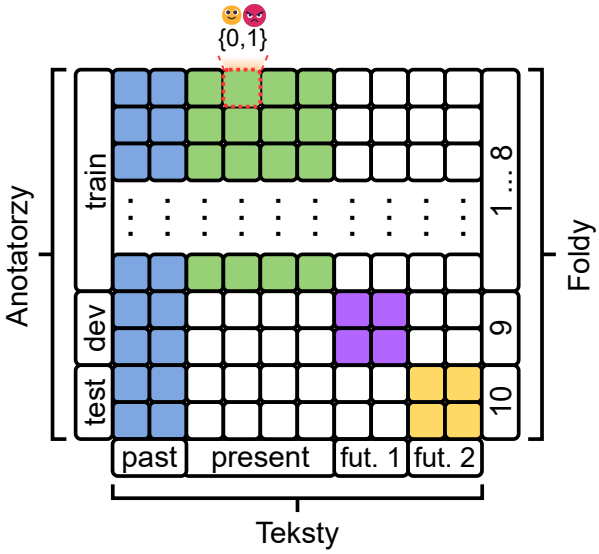
- $F1_{model}$ oznacza wartość miary F1 macro modelu, dla którego obliczamy wartość miary $F1_{roznica}$,
- $F1_{baseline}$ oznacza wartość miary F1 macro modelu referencyjnego, z którym porównujemy model poddawany ewaluacji.

Każda trójka $\langle$*zbiór danych*, *model językowy*, *model predykcyjny*$\rangle$ została ewaluowana przy użyciu 10-krotnej walidacji krzyżowej. W celu uniknięcia przecieku nadmiarowych danych zbiór anotacji został podzielony zarówno pod kątem anotatorów jak i tekstów. Metodę podziału zbioru podczas walidacji krzyżowej przedstawiono na Rysunku 11. W pojedynczej iteracji walidacji krzyżowej model jest uczony na (1) anotacjach wszystkich anotatorów dotyczących tekstów wybranych do zbioru *past* (niebieskie bloki na Rysunku 11) oraz (2) anotacjach anotatorów na tekstach ze zbioru *present* (zielone bloki na Rysunku 11). Jako zbiór walidacyjny wykorzystywany jest zbiór anotacji na tekstach ze zbioru *fut.1* dokonane przez anotatorów ze zbioru *dev* (fioletowe bloki na Rysunku 11), a jako zbiór testowy są używane anotacje tekstów ze zbioru *fut.1* wykonane przez anotatorów ze zbioru *test* (żółte bloki na Rysunku 11). Podział na podstawie anotatorów zapobiegł dostarczeniu zbyt wielu próbek uczących na temat anotatorów, których anotacje znajdowały się w zbiorze walidacyjnym *dev* oraz testowym *test*. Natomiast podział zbioru pod kątem tekstów pozwolił na uniknięcie przecieku danych na temat tekstów, które zostały wykorzystywane podczas walidacji *fut.1* oraz testowania *test* modelu. Dostarczenie niewielkiej liczby anotacji dla anotatorów ze zbiorów *dev* oraz *test* w procesie uczenia było kluczowe w celu dostarczenia niezbędnej ilości wiedzy na temat preferencji tych anotatorów. Była ona niezbędna w celu predykcji obraźliwości tekstów w zbiorze walidacyjnym i testowym z uwzględnieniem perspektywy anotatorów, dla których dostarczono ograniczoną liczbę anotacji w zbiorze treningowym.

Zróżnicowanie językowe tekstów znajdujących się w wykorzystanych zbiorach danych (teksty anglojęzyczne w WikiDetox: Aggression, WikiDetox: Toxicity, Measuring Hate Speech, Unhealthy Conversations oraz teksty polskojęzyczne w Doccano 1 i Doccano 2) umożliwiły weryfikację skuteczności opracowanych metod na tekstach w języku angielskim jak i w języku polskim.

9.2.1 Badania wpływu zróżnicowanych modeli językowych i głębokich metod predykcyjnych na jakość predykcji obraźliwości tekstu

W głównym scenariuszu eksperymentalnym skupiono się na wpływie metody generowania reprezentacji wektorowej tekstu jak i samego modelu predykcyjnego. W szczególności uwzględniono wyniki metod zaproponowanych w niniejszej pracy: metody opartej o miary konformizmu użytkownika (Podrozdział 5.2) oraz metody opartej o uśrednione reprezentacje wektorowe (Podrozdział 5.3). Ich wyniki zostały porównane z innymi spersonalizowanymi architekturami opracowanymi w ramach pracy w zespole badawczym: HuBi-Simple oraz



Rysunek 11: Podział danych zastosowany podczas walidacji krzyżowej w celu zapobiegnięcia przeciekowi nadmiarowych danych na temat anotatorów i tekstów spoza zbioru uczącego. [Źródło: opracowanie własne]

HuBi-Medium (Rozdział 4). Wyniki metod spersonalizowanych posłużyły do ich zestawienia z metodą referencyjną (Rozdział 3). Ostatnia wspomniana metoda była uruchamiana wyłącznie na tekście. Jednak w celu zachowania spójności danych wejściowych pojedynczy tekst był przepuszczany przez model *Baseline* tyle razy, ilu użytkowników go anotowało. Przy każdym przepuszczeniu tekstu przez model otrzymywał on klasę obraźliwości przypisaną przez kolejnego anotatora.

W celu ewaluacji wszystkich architektur na wszystkich zbiorach wykorzystano różnorodne metody generowania reprezentacji wektorowych tekstów: Random, CBOW, Skipgram, XLM-R, LaBSE, BERT, DeBERTa, MPNet. Zdecydowano się na wybór szerokiej listy metod generowania osadzeń tekstu, aby zbadać ich wpływ na skuteczność metod predykcyjnych:

- Random – metoda generująca losowe reprezentacje wektorowe tekstu, wykorzystana jako punkt odniesienia dla innych metod.
- CBOW (Bojanowski i in., 2017b) – metoda generowania bezkontekstowych reprezentacji oparta o technikę *Continuous Bag of Words*, za pomocą której w procesie uczenia model przewiduje brakujące słowo w tekście na podstawie kontekstu słów w jego otoczeniu.
- Skipgram (Bojanowski i in., 2017b) – metoda generowania bezkontekstowych reprezentacji wykorzystująca technikę *Skipgram*, dzięki której model podczas uczenia na podstawie słowa przewiduje słowa w jego otoczeniu.
- BERT (Devlin i in., 2019c) – model *Bidirectional Encoder Representations from Transformers* wprowadził dwukierunkowe przetwarzanie tekstu, umożliwiając lepsze zrozumienie kontekstu zarówno z lewej, jak i prawej strony słowa.

- LaBSE (Feng i in., 2022) – *Language-agnostic BERT Sentence Embeddings* jest modelem opartym na architekturze BERT, zaprojektowanym do generowania jednorodnych osadzeń zdań dla ponad 100 języków.
- DeBERTa (He i in., 2020) — model wprowadzający mechanizm rozdzielonej uwagi (ang. *disentangled attention*) jest ulepszoną wersją modelu BERT.
- MPNet (Song i in., 2020) – model ten wprowadził nową strategię maskowania i permutowania cząstek słów (ang. *tokens*).
- XLM-RoBERTa (XLM-R) (Conneau i in., 2020) – wielojęzykowa wersja modelu RoBERTa (Liu, 2019), oparta na architekturze typu transformer, dostarczająca reprezentacje wektorowe dla 100 języków.

Podczas badań wykorzystano implementacje modeli wykorzystujących architekturę typu transformer dostępne w bibliotece HuggingFace (Wolf, 2019). W celu uzyskania osadzenia tekstu obliczono uśrednioną reprezentację wszystkich jego tokenów. Jest to podejście różniące się od standardowo wykorzystywanej reprezentacji specjalnego tokenu [CLS], którego rolą jest przechowywanie reprezentacji wektorowej całego tekstu. Wstępne eksperymenty wykazały, że wykorzystanie uśrednionej reprezentacji wszystkich tokenów prowadzi do uzyskania wyższej jakości predykcji niż podejście zakładające wykorzystanie reprezentacji tekstu uzyskanej wyłącznie z tokenu [CLS].

9.2.2 Badania wpływu rozmiaru zbioru danych wykorzystanego w procesie uczenia na jakość predykcji obraźliwości tekstu

Scenariusz ten kładzie nacisk na zbadanie tempa uczenia się zarówno preferencji użytkownika jak i wpływu semantyki tekstu na jego obraźliwość. W ramach przeprowadzonych eksperymentów rozmiar zbioru uczącego został początkowo ustalony na 1 fold uczący utworzony w wyniku dziesięciokrotnej walidacji krzyżowej. Następnie w każdym kolejnym etapie eksperymentu, zbiór uczący był rozszerzany o 1 fold. W ostatniej iteracji zbiór treningowy zawierał 8 foldów, gdyż na zbiór walidacyjny i testowy wykorzystano po 1 foldzie. Dla każdego modelu predykcyjnego wykorzystano metodę generowania reprezentacji tekstu, dla której zaobserwowano najlepsze wyniki w scenariuszu eksperymentalnym opisanym w Podrozdziale 9.2.1.

9.2.3 Badania wpływu wykorzystania metody aktywnego uczenia na jakość predykcji obraźliwości modeli uczonych na ograniczonym zbiorze danych

W ramach tego scenariusza jako zbiór uczący wykorzystano początkowo teksty ze zbioru *past* zaanotowane przez wszystkich anotatorów (niebieskie bloki na Rysunku 11). Następnie w każdej iteracji rozszerzano zbiór uczący o jedną anotację ze zbioru tekstów *present* dobieraną osobno dla każdego anotatora ze zbioru *train* (zielone bloki na Rysunku 11) na podstawie

wybranej metody aktywnego uczenia lub metody losowej, nazywanej Random. W metodzie Random dla każdego użytkownika ze zbioru *train* w kolejnych iteracjach eksperymentu do zbioru uczącego była dodawana losowo wybrana anotacja tekstu dokonana przez tego użytkownika spośród anotacji tego użytkownika dotyczących tekstów ze zbioru *present*, które nie zostały jeszcze dodane do zbioru uczącego. Dla każdego zbioru eksperymenty wykonywano dla 14 iteracji. Każdą sekwencję 14 iteracji powtórzono dziesięciokrotnie w ramach walidacji krzyżowej opisanej na Rysunku 11. W celu zbadania jakości zbioru przygotowanego za pomocą badanej metody podczas każdej iteracji wyuczano model UserConf na zbiorze składającym się z tekstów wybranych w danej iteracji oraz poprzednich. Przykładowo, w iteracji numer 6 zbiór uczący zawierał 6 anotacji dla każdego anotatora. Procedurę selekcji tekstów do anotacji dla poszczególnych użytkowników opisano w Algorytmie 1.

Algorytm 1 Procedura selekcji tekstów $a_j \in A_u$ możliwych do anotacji dla anotatorów $u_i \in U$. Wartość m_j miary wykorzystywanej w danej metodzie aktywnego uczenia ψ jest obliczana dla każdego tekstu, którego anotacja przez użytkownika u_i nie została jeszcze dodana do zbioru uczącego. Wartości m_j są umieszczane w słowniku *text_measures*, gdzie kluczami są teksty a_j. Następnie teksty znajdujące się w *text_measures* są sortowane na podstawie wartości miary m_j i umieszczane w słowniku K, gdzie kluczem jest a_j, a wartością m_j. W kolejnym kroku uzyskiwana jest lista kluczy słownika K, która jest dodawana jako osobny wektor do listy wektorów *orders* opisującej kolejność anotacji tekstów opracowaną osobno dla każdego użytkownika u_i.

```
orders ← [[]]
for u_i ∈ U do
    text_measures ← {}
    for a_j ∈ A_u do
        m_j ← ψ(a_j)
        text_measures[a_j] ← m_j
    end for
    K = text_measures.sort_by_values()
    orders[i] ← K.keys()
end for
return orders
```

9.2.4 Badania jakości predykcji obraźliwości tekstu przez generatywny model językowy ogólnego przeznaczenia ChatGPT-3.5

W celu zbadania możliwości spersonalizowanego i niespersonalizowanego rozpoznawania tekstów obraźliwych za pomocą generatywnego, dużego modelu językowego ogólnego przeznaczenia ChatGPT-3.5 wykorzystano dwie metody inżynierii podpowiedzi (ang. *prompt engineering*): *zero-shot prompting* (Wei i in., 2022) oraz *in-context learning* (Brown i in., 2020). W przypadku pierwszej techniki w podpowiedzi umieszczanej na wejściu modelu dostarczono opis zadania niespersonalizowanego rozpoznawania obraźliwości wraz z tekstem, którego

obraźliwość powinna zostać oceniona. Natomiast drugą metodę zastosowano w przypadku spersonalizowanego rozpoznawania obraźliwości, gdzie do modelu dostarczono dodatkowo trzy przykłady tekstów zaanotowanych przuz użytkownika wraz z ich oceną obraźliwości w celu dostarczenia reprezentacji preferencji anotatora. W przypadku obu zastosowanych metod inżynierii podpowiedzi do instrukcji dodano informację, aby odpowiedź modelu miała formę listy znanej z języka programowania Python (Van Rossum i Drake, 2009). Umożliwiło to skuteczną ekstrakcję predykcji modelu z jego odpowiedzi.

W celu ewaluacji dużego modelu językowego względem opracowanych spersonalizowanych metod rozpoznawania obraźliwych tekstów posłużono się dwiema miarami. Miara *Loss* (Kocoń i in., 2023b) skupia się na określeniu różnicy między jakością predykcji modelu ChatGPT-3.5 $F1_{chatgpt}$ a wartością miary F1 macro dla najlepszego modelu spersonalizowanego rozpoznawania tekstów obraźliwych $F1_{SOTA}$. Sposób obliczania miary *Loss* opisano we Wzorze 18:

$$Loss = \frac{100\% * (F1_{SOTA} - F1_{chatgpt})}{F1_{SOTA}} \tag{18}$$

gdzie:

- $F1_{SOTA}$ oznacza wartość miary F1 macro najlepszego modelu dla danego zbioru danych (ang. *state of the art*, *SOTA*),
- $F1_{chatgpt}$ oznacza wartość miary F1 macro dla modelu ChatGPT-3.5.

Miara *Gain* (Kocoń i in., 2023b) kładzie nacisk na określenie zysku płynącego z wzbogacenia podpowiedzi umieszczanej na wejściu modelu o przykłady tekstów zaanotowanych przez użytkownika w celu uzyskania spersonalizowanej predykcji obraźliwości tekstu. Można ją zdefiniować za pomocą Wzoru 19:

$$Gain = \frac{100\% * (F1_{Per} - F1_{NonPer})}{100\% - F1_{NonPer}} \tag{19}$$

gdzie:

- $F1_{Per}$ oznacza wartość miary F1 macro dla modelu ChatGPT-3.5 przy wykorzystaniu metody *in-context learning*,
- $F1_{NonPer}$ oznacza wartość miary F1 macro dla modelu ChatGPT-3.5 przy wykorzystaniu metody *zero-shot prompting*.

9.3 TESTY STATYSTYCZNE

Każdy przeprowadzony eksperyment został powtórzony 10-krotnie, aby umożliwić przeprowadzenie testów statystycznych. Wstępnie skupiono się na sprawdzeniu założeń testu statystycznego t-Studenta (Student, 1908). W tym celu zweryfikowano następujące założenia:

- normalność rozkładów – wykorzystano test Shapiro-Wilka (Shapiro i Wilk, 1965) w celu zbadania normalności rozkładów dla obu porównywanych metod,
- homogeniczność wariancji – posłużono się testem Bartletta (Bartlett, 1937) weryfikującym homogeniczność wariancji wyników obu metod,
- niezależność obserwacji – upewniono się, że poszczególne wyniki w ramach każdej z prób są od siebie niezależne.

W przypadku spełnienia wszystkich założeń testu t-Studenta, przeprowadzono test t-Studenta dla dwóch prób zależnych lub niezależnych w zależności od charakteru danych. W przeciwnym przypadku, gdy założenia testu t-Studenta nie zostały spełnione, wykorzystano nieparametryczne testy Wilcoxona (Wilcoxon, 1945) dla dwóch prób zależnych lub Manna–Whitneya (Mann i Whitney, 1947) w przypadku dwóch prób niezależnych.

Mając na względzie kontrolę występowania błędu typu 1, zastosowano poprawkę Bonferroniego (Dunn, 1961). W przypadku porównań wielu par prób, ryzyko błędnego uznania istotności różnic między próbami zostaje znacząco zwiększone. Jako środek zapobiegawczy zastosowano korektę poziomu istotności ($\alpha = 0,05$) na podstawie listy przeprowadzonych testów.

9.4 WYNIKI

Niniejszy rozdział przedstawia szczegółowe wyniki przeprowadzonych badań skupiających się na wpływie różnych metod generowania reprezentacji wektorowych tekstów na jakość predykcji obraźliwości, wpływie rozmiaru zbioru uczącego na jakość predykcji obraźliwości, wpływie zastosowanych technik aktywnego uczenia na tempo przyrostu jakości predykcji obraźliwości oraz jakość predykcji obraźliwości dla dużego, generatywnego modelu językowego. Opisane analizy opisują ocenę efektywności metod ewaluowanych na różnych zbiorach danych. Każdy podrozdział szczegółowo omawia wyniki przeprowadzonych eksperymentów i przedstawia wnioski na temat jakości predykcji modeli.

9.4.1 Badania wpływu metody generowania reprezentacji wektorowej tekstu oraz modelu głębokiej sieci neuronowej służącej do predykcji obraźliwości tekstu

Tabela 3 przedstawia wyniki dla zbioru WikiDetox: Aggression. Modele DeBERTa oraz MPNet systematycznie osiągają najwyższe wartości miary F1 macro, niezależnie od zastosowanej metody predykcyjnej. Najwyższy wynik uzyskano w kombinacji metody UserConf z osadzeniem wygenerowanym modelem MPNet, gdzie F1 macro wyniosło 82.14 z odchyleniem standardowym wynoszącym 0.66. Warto także podkreślić, że osadzenie DeBERTa osiągnęło bardzo zbliżone wyniki, zwłaszcza w metodzie UserConf (81.87, odchylenie standardowe 0.40), co świadczy o jego wysokiej efektywności w relacji do innych metod generowania reprezentacji wektorowych tekstów. Wyniki te wyraźnie sugerują, że modele osadzeń kontekstowych, oparte na architekturze transformerów, takie jak DeBERTa i MPNet, dostarczają

bogatszych reprezentacji tekstu, co przekłada się na wyższą skuteczność w zadaniu predykcji obraźliwości tekstu.

Z drugiej strony, najgorsze wyniki uzyskało osadzenie Random, którego wartość F1 macro waha się od 45.00 z odchyleniem standardowym wynoszącym 0.25 dla Metody referencyjnej do 74.62 z odchyleniem 0.46 w przypadku metody UserEmb. Wyniki te wskazują, że brak relewantnej reprezentacji tekstu znacząco obniża jakość predykcji, co jest szczególnie widoczne w przypadku Metody referencyjnej.

W odniesieniu do metod predykcyjnych, zauważa się istotną przewagę modelu UserConf, który osiągnął najwyższe wyniki dla większości osadzeń. Przykładowo, dla osadzenia BERT wartość F1 macro wyniosła 80.55 przy odchyleniu standardowym wynoszącym 0.74, co czyni tę metodę najbardziej skuteczną w porównaniu z pozostałymi technikami. Także UserEmb wykazuje wysoką efektywność, szczególnie w połączeniu z osadzeniem DeBERTa (81.02, odchylenie standardowe 0.24). Warto jednak zauważyć, że różnice między wynikami uzyskanymi przez UserConf i UserEmb są minimalne, co wskazuje na ich porównywalną skuteczność w zadaniu rozpoznawania obraźliwości tekstu.

Metoda referencyjna generuje najniższe wartości F1 macro we wszystkich osadzeniach, co świadczy o jej istotnie niższej skuteczności. Na przykład, dla osadzenia BERT wartość F1 macro wyniosła jedynie 73.11 przy odchyleniu standardowym równym 0.75, co jest wyraźnie niższe niż w przypadku metod spersonalizowanych, takich jak UserConf czy UserEmb. Podobną tendencję można zaobserwować w przypadku metod HuBi-Simple oraz HuBi-Medium, które choć uzyskują zadowalające wyniki, zazwyczaj wypadają gorzej niż metody UserConf i UserEmb.

Tabela 3: Wartości miary F1 macro w zależności od wykorzystanej metody generowania wektorowej reprezentacji tekstu dla zbioru WikiDetox: Aggression. Wartości **pogrubione** oznaczają najlepszy wynik dla danej reprezentacji wektorowej. Natomiast wartości podkreślone oznaczają najlepszy wynik dla metody rozpoznawania tekstów obraźliwych.

Metoda \ Osadzenia	Random	CBOW	Skipgram	BERT	LaBSE	DeBERTa	MPNet	XLM-R
Metoda referencyjna	45.00 ± 0.25	65.39 ± 0.86	67.24 ± 0.45	73.11 ± 0.75	73.86 ± 0.47	74.68 ± 0.89	73.62 ± 0.71	72.92 ± 0.86
UserConf	74.02 ± 0.53	**77.21** ± 0.36	77.32 ± 0.57	**80.55** ± 0.74	**79.91** ± 0.51	**81.87** ± 0.40	**82.14** ± 0.66	**81.70** ± 0.85
UserEmb	**74.62** ± 0.46	76.71 ± 0.51	**77.53** ± 0.50	80.26 ± 0.61	79.84 ± 0.45	81.02 ± 0.24	80.15 ± 0.64	80.77 ± 0.52
HuBi-Simple	73.90 ± 0.24	76.19 ± 0.75	77.21 ± 0.49	79.13 ± 0.53	79.46 ± 0.27	80.09 ± 0.51	79.87 ± 0.63	78.93 ± 0.44
HuBi-Medium	72.37 ± 0.94	76.50 ± 0.60	77.13 ± 0.43	79.48 ± 0.53	79.57 ± 0.26	80.61 ± 0.52	79.98 ± 0.65	79.15 ± 0.58

Tabela 4 opisuje wartości miary precyzji (Fawcett, 2006) dla zbioru WikiDetox: Aggression. Najlepsze rezultaty precyzji osiągnęły reprezentacje wektorowe wygenerowane modelami DeBERTa oraz MPNet, co potwierdza ich przewagę nad innymi modelami. Najwyższa wartość precyzji została uzyskana przez metodę UserConf w połączeniu z osadzeniem MPNet, gdzie precyzja wyniosła 81.80 z odchyleniem standardowym wynoszącym 0.74. Osadzenia modelu DeBERTa również pozwoliły na uzyskanie bardzo dobrych wyników, szczególnie w przypadku metody UserConf (81.53, odchylenie standardowe 0.63). Wyniki te dowodzą, że osadzenia oparte na architekturze transformerów, takie jak DeBERTa i MPNet, dostarczają

bardziej dokładnych reprezentacji tekstu, co przekłada się na wyższą precyzję w zadaniu spersonalizowanej predykcji obraźliwości tekstu.

Z kolei najniższe wartości precyzji uzyskało osadzenie Random, co jest szczególnie widoczne w przypadku Metody referencyjnej, gdzie precyzja wyniosła 44.89 z odchyleniem standardowym wynoszącym 0.70. To wskazuje na niską jakość predykcji tej metody w rozpoznawaniu obraźliwych treści tekstowych.

Pod względem metod predykcyjnych, model UserConf ponownie osiągnął najwyższe wyniki precyzji dla większości osadzeń. Dla przykładu, w połączeniu z osadzeniem BERT precyzja wyniosła 80.30 przy odchyleniu standardowym wynoszącym 0.82. UserEmb uzyskało podobnie wysokie wartości, zwłaszcza z osadzeniem DeBERTa (80.91, odchylenie standardowe 0.44). Choć różnice między UserConf a UserEmb są stosunkowo niewielkie, UserConf zazwyczaj uzyskuje nieznacznie lepsze rezultaty.

Metoda referencyjna konsekwentnie generuje najniższe wyniki precyzji, niezależnie od zastosowanego osadzenia. Na przykład, dla osadzenia BERT precyzja wyniosła 72.85 przy odchyleniu standardowym wynoszącym 0.41, co jest istotnie niższym wynikiem w porównaniu do bardziej zaawansowanych metod, takich jak UserConf czy UserEmb.

Podobnie jak w przypadku miary F1 macro, wyniki precyzji jednoznacznie wskazują na wyższość osadzeń kontekstowych (BERT, LaBSE, DeBERTa, MPNet, XLM-R) nad osadzeniami bezkontekstowymi (CBOW, Skipgram) oraz osadzeniami uzyskanymi metodą Random. Najwyższe wartości precyzji uzyskano dla osadzeń DeBERTa i MPNet, co potwierdza ich znaczącą efektywność w zadaniu predykcji obraźliwości tekstu.

Tabela 4: Wartości miary precyzji w zależności od wykorzystanej metody generowania wektorowej reprezentacji tekstu dla zbioru WikiDetox: Aggression. Wartości **pogrubione** oznaczają najlepszy wynik dla danej reprezentacji wektorowej. Natomiast wartości podkreślone oznaczają najlepszy wynik dla metody rozpoznawania tekstów obraźliwych.

Metoda \ Osadzenia	Random	CBOW	Skipgram	BERT	LaBSE	DeBERTa	MPNet	XLM-R
Metoda referencyjna	44.89 ± 0.70	65.26 ± 0.05	67.11 ± 0.88	72.85 ± 0.41	73.65 ± 0.24	74.54 ± 0.14	73.34 ± 0.70	72.70 ± 0.19
UserConf	73.77 ± 0.75	**77.01** ± 0.09	77.00 ± 0.18	**80.30** ± 0.82	**79.77** ± 0.67	**81.53** ± 0.63	**81.80** ± 0.74	**81.39** ± 0.72
UserEmb	**74.29** ± 0.68	76.60 ± 0.32	**77.29** ± 0.07	79.92 ± 0.21	79.52 ± 0.17	80.91 ± 0.44	79.95 ± 0.40	80.38 ± 0.46
HuBi-Simple	73.57 ± 0.90	76.04 ± 0.06	76.95 ± 0.87	78.86 ± 0.49	79.09 ± 0.29	79.88 ± 0.29	79.69 ± 0.24	78.75 ± 0.81
HuBi-Medium	72.02 ± 0.86	76.21 ± 0.87	76.73 ± 0.80	79.36 ± 0.62	79.41 ± 0.41	80.30 ± 0.95	79.60 ± 0.95	78.86 ± 0.52

W tabeli 5 przedstawiono wyniki miary kompletności (ang. *recall*) (Fawcett, 2006) dla zbioru WikiDetox: Aggression. Najwyższe wartości miary uzyskano przy zastosowaniu osadzeń DeBERTa oraz MPNet, szczególnie w metodach UserConf oraz UserEmb. Najlepszy wynik osiągnęła metoda UserConf w połączeniu z MPNet, uzyskując wartość recall równą 82.47 przy minimalnym odchyleniu standardowym wynoszącym 0.04. Równie imponujące rezultaty osiągnęło osadzenie DeBERTa, które dla metody UserConf zanotowało wartość 82.20 przy odchyleniu standardowym wynoszącym 0.97. Wyniki te wskazują na wyższość osadzeń opartych na architekturze transformerów, które oferują bogatsze reprezentacje semantyczne, co przekłada się na większą skuteczność w detekcji obraźliwych treści.

Osadzenie Random wykazuje najniższe wartości kompletności, szczególnie w przypadku Metody referencyjnej, gdzie wartość kompletności wyniosła jedynie 45.10 przy odchyleniu standardowym równym 0.80. Wyniki te podkreślają ograniczoną skuteczność tej metody w wychwytywaniu obraźliwych treści, co kontrastuje z wyraźnie lepszymi wynikami osiągniętymi przez osadzenia kontekstowe.

W kontekście metod predykcyjnych, metoda UserConf konsekwentnie osiąga najlepsze wyniki w większości osadzeń. Dla osadzenia BERT wartość kompletności wyniosła 80.78 przy odchyleniu standardowym wynoszącym 0.49, co czyni tę metodę wysoce skuteczną w rozpoznawaniu obraźliwych treści. Metoda UserEmb osiągnęła podobnie wysokie wyniki, zwłaszcza w połączeniu z osadzeniem DeBERTa (81.11, odchylenie standardowe 0.12), co wskazuje na jej wysoką efektywność.

Metoda referencyjna, jak w poprzednich miarach, uzyskuje najniższe wartości kompletności, niezależnie od zastosowanego osadzenia. Przykładowo, dla osadzenia BERT wartość kompletności wyniosła 73.35 przy odchyleniu standardowym 0.45, co jest wyraźnie niższą wartością niż w przypadku bardziej zaawansowanych metod, takich jak UserConf i UserEmb. Podobne tendencje zaobserwowano w metodach HuBi-Simple oraz HuBi-Medium, które chociaż generują stosunkowo dobre wyniki, zazwyczaj ustępują skutecznością metodom UserConf i UserEmb.

Tabela 5: Wartości miary kompletności w zależności od wykorzystanej metody generowania wektorowej reprezentacji tekstu dla zbioru WikiDetox: Aggression. Wartości **pogrubione** oznaczają najlepszy wynik dla danej reprezentacji wektorowej. Natomiast wartości podkreślone oznaczają najlepszy wynik dla metody rozpoznawania tekstów obraźliwych.

Osadzenia / Metoda	Random	CBOW	Skipgram	BERT	LaBSE	DeBERTa	MPNet	XLM-R
Metoda referencyjna	45.10 ± 0.80	65.50 ± 0.35	67.36 ± 0.84	73.35 ± 0.45	74.06 ± 0.26	74.81 ± 0.20	73.88 ± 0.38	73.12 ± 0.31
UserConf	74.25 ± 0.75	**77.39** ± 0.53	77.63 ± 0.71	**80.78** ± 0.49	80.03 ± 0.79	**82.20** ± 0.97	**82.47** ± 0.04	**82.00** ± 0.07
UserEmb	**74.94** ± 0.70	76.80 ± 0.29	**77.76** ± 0.68	80.60 ± 0.72	**80.15** ± 0.51	81.11 ± 0.12	80.33 ± 0.68	81.14 ± 0.13
HuBi-Simple	74.21 ± 0.07	76.32 ± 0.83	77.46 ± 0.42	79.39 ± 0.12	79.81 ± 0.54	80.30 ± 0.58	80.03 ± 0.84	79.10 ± 0.59
HuBi-Medium	72.71 ± 0.65	76.77 ± 0.05	77.51 ± 0.91	79.58 ± 0.51	79.71 ± 0.87	80.90 ± 0.13	80.34 ± 0.31	79.42 ± 0.38

W tabeli 6 zaprezentowano wyniki dla zbioru WikiDetox: Toxicity. Modele DeBERTa oraz BERT konsekwentnie uzyskują wysokie wartości miary F1 macro. Najwyższy wynik osiągnęła metoda UserEmb z osadzeniem DeBERTa, gdzie wartość F1 macro wyniosła 81.68 z odchyleniem standardowym 0.78. UserConf także uzyskała wysokie wyniki, szczególnie z osadzeniem BERT (81.22, odchylenie standardowe 0.59), co potwierdza skuteczność osadzeń opartych na modelach typu transformer w zadaniu predykcji obraźliwości tekstu. Osadzenia LaBSE oraz MPNet również osiągnęły dobre wyniki, zwłaszcza w metodach UserEmb (LaBSE: 81.32, odchylenie 0.37) oraz UserConf (MPNet: 79.70, odchylenie 0.86).

Osadzenie Random generuje najniższe wyniki, co jest szczególnie widoczne w Metodzie referencyjnej, gdzie wartość F1 macro wyniosła 46.07 z odchyleniem standardowym 0.14. Wynik ten wskazuje na niską efektywność tego podejścia, co kontrastuje z bardziej zaawansowanymi osadzeniami kontekstowymi.

W odniesieniu do metod predykcyjnych, UserEmb okazuje się być najbardziej efektywnym modelem, osiągając najlepsze wyniki w większości osadzeń. Dla przykładu, w połączeniu z LaBSE wartość F1 macro wyniosła 81.32 z odchyleniem standardowym 0.37. UserConf także uzyskuje konkurencyjne wyniki, szczególnie dla osadzeń BERT i DeBERTa. Metody HuBi-Simple oraz HuBi-Medium generują stabilne, lecz niższe wyniki w porównaniu do metod UserConf i UserEmb, co wskazuje na ich mniejszą skuteczność w zadaniu spersonalizowanego rozpoznawania tekstów obraźliwych.

Tabela 6: Wartości miary F1 macro w zależności od wykorzystanej metody generowania wektorowej reprezentacji tekstu dla zbioru WikiDetox: Toxicity. Wartości **pogrubione** oznaczają najlepszy wynik dla danej reprezentacji wektorowej. Natomiast wartości podkreślone oznaczają najlepszy wynik dla metody rozpoznawania tekstów obraźliwych.

Metoda \ Osadzenia	Random	CBOW	Skipgram	BERT	LaBSE	DeBERTa	MPNet	XLM-R
Metoda referencyjna	46.07 ± 0.14	70.56 ± 0.76	70.89 ± 0.49	74.64 ± 1.00	75.66 ± 0.47	75.88 ± 0.98	74.71 ± 0.39	73.83 ± 0.88
UserConf	74.89 ± 0.34	78.05 ± 0.11	**78.24** ± 0.36	**81.22** ± 0.59	79.57 ± 0.79	80.47 ± 0.85	79.70 ± 0.86	79.61 ± 0.56
UserEmb	**74.98** ± 0.43	**78.12** ± 0.61	78.17 ± 0.55	80.40 ± 0.68	**81.32** ± 0.37	**81.68** ± 0.78	**80.58** ± 0.50	**81.17** ± 0.45
HuBi-Simple	74.36 ± 0.45	77.51 ± 0.42	77.92 ± 0.16	78.56 ± 0.87	79.14 ± 0.81	79.50 ± 0.60	78.42 ± 0.74	78.17 ± 0.91
HuBi-Medium	73.54 ± 0.85	77.68 ± 0.88	77.24 ± 0.69	79.01 ± 0.81	78.88 ± 0.48	79.96 ± 0.61	79.21 ± 0.36	78.74 ± 0.79

Tabela 7 opisuje wyniki dla zbioru Measuring Hate Speech. Osadzenia BERT, LaBSE oraz DeBERTa konsekwentnie osiągają najwyższe wartości miary F1 macro w porównaniu z innymi osadzeniami. Najlepszy wynik uzyskano dla osadzenia LaBSE w połączeniu z metodą UserConf, gdzie F1 macro wyniosło 42.34 z odchyleniem standardowym 3.02. Równie wysokie wartości osiągnięto w metodzie UserEmb z osadzeniem DeBERTa, gdzie wartość F1 macro wyniosła 41.64 z odchyleniem standardowym 3.82. Wyniki te wskazują na wyższą efektywność osadzeń kontekstowych opartych na modelach transformerowych w rozpoznawaniu tekstów obraźliwych.

Z kolei najniższe wartości F1 macro osiągnęło osadzenie Random, gdzie w Metodzie referencyjnej uzyskano wartość 28.43 z dużym odchyleniem standardowym wynoszącym 3.89. Wyniki te wyraźnie pokazują, że osadzenia losowe nie dostarczają wystarczająco bogatych reprezentacji tekstu, co prowadzi do niższej skuteczności w zadaniu klasyfikacji tekstów obraźliwych.

W odniesieniu do metod predykcyjnych, UserConf uzyskuje najlepsze wyniki dla większości osadzeń. Przykładowo, dla osadzenia BERT wartość F1 macro wyniosła 40.79 przy odchyleniu standardowym wynoszącym 2.91, co czyni tę metodę jedną z najbardziej skutecznych w rozważanym zadaniu. UserEmb także wykazuje wysoką efektywność, szczególnie dla osadzeń LaBSE (40.55, odchylenie 3.80) i DeBERTa (41.64, odchylenie 3.82). Metody HuBi-Simple oraz HuBi-Medium osiągają stabilne, ale niższe wyniki, co wskazuje na ich relatywnie mniejszą skuteczność w porównaniu z metodami UserConf i UserEmb.

W tabeli 8 przedstawione są wyniki dla zbioru Unhealthy Conversations. Osadzenia kontekstowe, takie jak BERT, LaBSE, DeBERTa, MPNet oraz XLM-R, osiągają wyższe wartości miary F1 macro w porównaniu z osadzeniami bezkontekstowymi, jak CBOW i Skipgram oraz metodą Random. Najwyższe wyniki uzyskano dla osadzenia BERT w połączeniu z metodą

Tabela 7: Wartości miary F1 macro w zależności od wykorzystanej metody generowania wektorowej reprezentacji tekstu dla zbioru Measuring Hate Speech. Wartości **pogrubione** oznaczają najlepszy wynik dla danej reprezentacji wektorowej. Natomiast wartości podkreślone oznaczają najlepszy wynik dla metody rozpoznawania tekstów obraźliwych.

Metoda \ Osadzenia	Random	CBOW	Skipgram	BERT	LaBSE	DeBERTa	MPNet	XLM-R
Metoda referencyjna	28.43 ± 3.89	29.63 ± 4.05	30.29 ± 4.16	33.77 ± 4.16	31.58 ± 4.47	34.54 ± 4.39	31.36 ± 4.23	32.69 ± 4.45
UserConf	**33.83** ± 3.58	35.36 ± 3.62	**36.71** ± 2.91	**40.79** ± 2.91	**42.34** ± 3.02	37.72 ± 3.38	37.80 ± 2.92	**39.45** ± 3.07
UserEmb	33.24 ± 3.83	**35.78** ± 3.62	36.57 ± 4.37	39.56 ± 4.09	40.55 ± 3.80	**41.64** ± 3.82	**39.91** ± 4.35	38.23 ± 3.91
HuBi-Simple	32.12 ± 3.69	33.60 ± 4.25	34.56 ± 4.11	39.46 ± 3.92	39.86 ± 4.68	35.12 ± 4.10	36.34 ± 4.13	38.52 ± 4.70
HuBi-Medium	31.43 ± 3.26	33.27 ± 3.56	36.05 ± 3.52	37.98 ± 3.72	40.44 ± 3.91	36.43 ± 3.48	36.75 ± 3.21	37.71 ± 2.89

UserConf, gdzie wartość F1 macro wyniosła 53.12 przy odchyleniu standardowym wynoszącym 0.52. Równie wysokie rezultaty osiągnęła metoda UserEmb z osadzeniem DeBERTa, gdzie F1 macro wyniosło 51.89 przy odchyleniu 0.29. Wyniki te wskazują na wysoką skuteczność osadzeń opartych na architekturach transformerowych w zadaniu rozpoznawania tekstów obraźliwych.

Z kolei osadzenia wygenerowane metodą Random osiągnęły najniższe wyniki we wszystkich metodach predykcyjnych, zwłaszcza w Metodzie referencyjnej, gdzie wartość F1 macro wyniosła 32.17 z odchyleniem standardowym równym 0.38. Wynik ten jednoznacznie sugeruje, że losowe reprezentacje tekstu nie są w stanie zapewnić efektywnego rozpoznawania tekstów obraźliwych. Podkreśla to istotną rolę metod generowania reprezentacji kontekstowych oraz bezkontekstowych.

W odniesieniu do metod predykcyjnych, UserConf konsekwentnie osiąga najlepsze wyniki dla większości osadzeń. Na przykład, dla osadzenia LaBSE F1 macro wyniosło 45.85 przy odchyleniu standardowym 0.19. Metoda UserEmb także uzyskała konkurencyjne wyniki, szczególnie dla osadzenia DeBERTa (51.89, odchylenie 0.29). Warto zauważyć, że HuBi-Simple oraz HuBi-Medium również osiągają dobre wyniki, szczególnie z osadzeniami BERT i LaBSE, lecz nie dorównują jakością predykcji metodom UserConf i UserEmb.

Tabela 8: Wartości miary F1 macro w zależności od wykorzystanej metody generowania wektorowej reprezentacji tekstu dla zbioru Unhealthy Conversations. Wartości **pogrubione** oznaczają najlepszy wynik dla danej reprezentacji wektorowej. Natomiast wartości podkreślone oznaczają najlepszy wynik dla metody rozpoznawania tekstów obraźliwych.

Metoda \ Osadzenia	Random	CBOW	Skipgram	BERT	LaBSE	DeBERTa	MPNet	XLM-R
Metoda referencyjna	32.17 ± 0.38	35.21 ± 0.42	37.60 ± 0.22	41.27 ± 0.40	42.81 ± 0.25	41.05 ± 0.45	43.27 ± 0.38	42.42 ± 0.18
UserConf	**39.57** ± 0.40	**40.92** ± 0.40	41.25 ± 0.26	**53.12** ± 0.52	45.85 ± 0.19	42.72 ± 0.54	44.01 ± 0.27	**49.48** ± 0.34
UserEmb	38.23 ± 0.45	40.27 ± 0.47	**43.08** ± 0.47	50.37 ± 0.38	45.37 ± 0.36	**51.89** ± 0.29	44.39 ± 0.23	47.99 ± 0.33
HuBi-Simple	36.23 ± 0.33	38.74 ± 0.32	40.64 ± 0.21	47.12 ± 0.47	**49.89** ± 0.31	44.18 ± 0.42	44.02 ± 0.44	42.72 ± 0.46
HuBi-Medium	35.17 ± 0.44	37.28 ± 0.53	39.87 ± 0.57	47.34 ± 0.38	48.89 ± 0.51	41.47 ± 0.60	**45.96** ± 0.40	46.70 ± 0.59

Tabela 9 opisuje wyniki dla zbioru Doccano 1. Najlepsze wyniki osiągnięto przy zastosowaniu osadzeń kontekstowych, szczególnie modelem MPNet, które w połączeniu z metodą UserEmb uzyskały wartość F1 macro równą 53.76 z odchyleniem standardowym wynoszącym

1.13. DeBERTa także uzyskała wysoki wynik w metodzie UserConf, osiągając 51.70 z odchyleniem 1.42. Warto zwrócić uwagę, że metoda UserConf również osiągnęła dobre rezultaty w połączeniu z MPNet, uzyskując wynik 53.23 z odchyleniem standardowym wynoszącym 1.79. Wyniki te potwierdzają skuteczność osadzeń opartych na architekturach transformerowych w zadaniu predykcji obraźliwych treści.

Z kolei osadzenie Random uzyskało najniższe wyniki we wszystkich metodach, zwłaszcza w Metodzie referencyjnej, gdzie wartość F1 macro wyniosła 32.47 z odchyleniem standardowym równym 1.36. Wynik ten potwierdza ograniczoną użyteczność losowych reprezentacji tekstu w zadaniu rozpoznawania obraźliwych treści.

Metody UserConf i UserEmb uzyskały najwyższe wyniki, szczególnie w połączeniu z osadzeniami MPNet i DeBERTa. Przykładowo, dla osadzenia LaBSE metoda UserEmb osiągnęła wartość 49.47 z odchyleniem 2.31, a metoda UserConf uzyskała 46.15 z odchyleniem 1.42. HuBi-Simple oraz HuBi-Medium osiągnęły stabilne wyniki, ale w większości przypadków nie dorównują skutecznością metodom UserConf i UserEmb.

Tabela 9: Wartości miary F1 macro w zależności od wykorzystanej metody generowania wektorowej reprezentacji tekstu dla zbioru Doccano 1. Wartości **pogrubione** oznaczają najlepszy wynik dla danej reprezentacji wektorowej. Natomiast wartości podkreślone oznaczają najlepszy wynik dla metody rozpoznawania tekstów obraźliwych.

Osadzenia / Metoda	Random	CBOW	Skipgram	BERT	LaBSE	DeBERTa	MPNet	XLM-R
Metoda referencyjna	32.47 ± 1.36	34.29 ± 1.39	36.00 ± 1.67	42.58 ± 1.45	43.12 ± 1.61	40.69 ± 1.60	44.03 ± 1.44	37.42 ± 1.61
UserConf	39.78 ± 1.28	41.46 ± 1.03	43.73 ± 1.10	**48.79** ± 2.12	46.15 ± 1.42	**51.70** ± 1.42	53.23 ± 1.13	51.15 ± 1.52
UserEmb	**39.98** ± 1.01	**43.32** ± 0.82	**43.96** ± 2.18	46.86 ± 0.87	49.47 ± 2.31	45.65 ± 0.91	**53.76** ± 1.79	50.70 ± 2.31
HuBi-Simple	37.39 ± 1.12	40.22 ± 2.52	42.38 ± 2.67	47.02 ± 1.48	51.42 ± 2.55	42.63 ± 1.58	44.59 ± 1.47	49.19 ± 1.91
HuBi-Medium	38.39 ± 2.56	39.81 ± 1.99	42.89 ± 1.88	46.39 ± 0.88	**51.98** ± 1.58	51.63 ± 1.23	48.21 ± 1.43	**51.26** ± 1.46

W tabeli 10 opisano wyniki dla zbioru danych Doccano 2. Osadzenia kontekstowe, takie jak BERT, LaBSE i DeBERTa, uzyskują najlepsze wyniki. Najwyższa wartość F1 macro została osiągnięta dla osadzenia DeBERTa w połączeniu z metodą UserConf, gdzie F1 macro wyniosło 51.78 z odchyleniem standardowym wynoszącym 1.36. BERT również osiągnęło bardzo dobre wyniki w metodzie UserConf (50.82, odchylenie 1.42), co potwierdza dominację osadzeń opartych na modelach transformerowych w tym zadaniu.

Osadzenie Random uzyskało najniższe wartości F1 macro w każdej z metod predykcyjnych. Najniższy wynik uzyskano w Metodzie referencyjnej, gdzie wartość F1 macro wyniosła 31.23 przy odchyleniu standardowym równym 2.34. Wynik ten wskazuje na ograniczoną skuteczność losowej reprezentacji tekstu w zadaniu klasyfikacji obraźliwych treści.

Metody predykcyjne UserConf oraz UserEmb generują najwyższe wyniki. W metodzie UserEmb, osadzenie LaBSE uzyskało wynik 47.92 z odchyleniem standardowym 2.17, natomiast osadzenie BERT uzyskało 48.37 z odchyleniem 2.30. HuBi-Simple oraz HuBi-Medium także osiągnęły stabilne wyniki, jednak ich skuteczność była nieco niższa w porównaniu do wyników uzyskanych w metodach UserConf i UserEmb.

Tabela 10: Wartości miary F1 macro w zależności od wykorzystanej metody generowania wektorowej reprezentacji tekstu dla zbioru Doccano 2. Wartości **pogrubione** oznaczają najlepszy wynik dla danej reprezentacji wektorowej. Natomiast wartości podkreślone oznaczają najlepszy wynik dla metody rozpoznawania tekstów obraźliwych.

Osadzenia / Metoda	Random	CBOW	Skipgram	BERT	LaBSE	DeBERTa	MPNet	XLM-R
Metoda referencyjna	31.23 ± 2.34	33.55 ± 2.47	34.79 ± 2.29	40.10 ± 0.79	41.72 ± 1.23	37.21 ± 0.78	35.80 ± 1.77	41.19 ± 2.35
UserConf	**41.17** ± 1.27	**43.16** ± 1.19	**44.05** ± 1.88	**50.82** ± 1.42	46.17 ± 1.63	**51.78** ± 1.36	**48.90** ± 1.57	47.13 ± 1.82
UserEmb	39.17 ± 2.37	40.48 ± 2.32	43.00 ± 1.74	48.37 ± 2.30	**47.92** ± 2.17	44.21 ± 1.91	45.63 ± 2.35	**49.78** ± 2.46
HuBi-Simple	36.13 ± 1.96	38.67 ± 2.47	41.05 ± 2.78	46.49 ± 2.23	42.22 ± 2.47	43.49 ± 2.89	43.98 ± 2.69	41.49 ± 1.80
HuBi-Medium	37.17 ± 2.30	38.42 ± 1.35	40.30 ± 1.33	41.13 ± 2.17	42.35 ± 2.40	43.36 ± 1.96	42.08 ± 1.60	45.78 ± 1.86

9.4.2 Badania wpływu rozmiaru zbioru treningowego w zadaniu predykcji obraźliwości tekstu

Tabela 11 opisuje wyniki dla zbioru WikiDetox: Aggression. Metoda referencyjna, podobnie jak w przypadku innych zbiorów danych, osiąga najniższe wyniki, zaczynając od wartości 45.23 przy jednym foldzie i stopniowo poprawiając się do 74.68 przy ośmiu foldach, co pozostaje najgorszym wynikiem w porównaniu do innych metod. Metoda UserConf wykazuje wyraźną przewagę nad metodą referencyjną już od samego początku, osiągając 57.27 dla jednego folda i ostatecznie 81.87 dla ośmiu foldów, co czyni ją najlepszą metodą w tym zestawieniu. Metoda UserEmb również odnotowuje dobre wyniki, jednak jej efektywność wzrasta wolniej w pierwszych etapach. Przy jednym foldzie osiąga wartość 55.19, a dopiero od pięciu foldów zaczyna dorównywać innym metodom, kończąc z wynikiem 81.02 przy ośmiu foldach, co plasuje ją tuż za UserConf. Metoda HuBi-Simple osiąga umiarkowane wyniki początkowe, zaczynając od 52.42 dla jednego folda, jednak jej skuteczność gwałtownie wzrasta wraz z liczbą foldów, by ostatecznie osiągnąć 80.09 dla pełnej liczby foldów. Podobnie HuBi-Medium charakteryzuje się początkowo umiarkowaną skutecznością, z wynikiem 53.67 przy jednym foldzie, ale kończy z wynikiem 80.61 przy ośmiu foldach, co czyni ją jedną z lepszych metod w tej analizie. Najlepszą metodą w tym zestawieniu jest UserConf, która osiąga najwyższą wartość F1 macro, wynoszącą 81.87 przy pełnej liczbie foldów, przewyższając inne metody zarówno przy mniejszych, jak i większych liczbach foldów.

Tabela 11: Wartości miary F1 macro w zależności od liczby foldów uwzględnionych w zbiorze uczącym na zbiorze danych WikiDetox: Aggression. Wartości **pogrubione** oznaczają najlepszy wynik dla danej liczby foldów uczących. Natomiast wartości podkreślone oznaczają najlepszy wynik dla danej metody rozpoznawania tekstów obraźliwych.

Foldy uczące / Metoda	1	1-2	1-3	1-4	1-5	1-6	1-7	1-8
Metoda Referencyjna	45.23 ± 0.36	48.99 ± 0.91	55.47 ± 1.02	61.37 ± 0.63	64.56 ± 1.02	65.95 ± 0.82	69.88 ± 0.51	74.68 ± 0.89
UserConf	**57.27** ± 0.63	**62.72** ± 0.46	**65.93** ± 0.50	**67.75** ± 0.71	**73.23** ± 0.77	**75.84** ± 0.57	77.06 ± 0.47	**81.87** ± 0.40
UserEmb	55.19 ± 0.52	60.54 ± 0.64	60.86 ± 0.63	62.48 ± 0.60	64.25 ± 0.57	75.16 ± 0.53	76.77 ± 0.52	81.02 ± 0.24
HuBi-Simple	52.42 ± 0.45	55.09 ± 0.54	55.17 ± 0.39	61.03 ± 0.78	66.42 ± 0.74	71.51 ± 0.42	75.90 ± 0.61	80.09 ± 0.51
HuBi-Medium	53.67 ± 0.45	56.36 ± 0.63	62.02 ± 0.53	66.45 ± 0.44	69.44 ± 0.44	71.44 ± 0.70	**77.49** ± 0.71	80.61 ± 0.52

W tabeli 12 przedstawiono wyniki dla zbioru WikiDetox: Toxicity. Metoda referencyjna cechuje się najniższymi wynikami w porównaniu do pozostałych metod, niezależnie od liczby foldów uwzględnionych w procesie uczenia. Początkowo, przy jednym foldzie uczącym, osiąga wartość 42.36, stopniowo zwiększając swój wynik do 75.88 przy pełnym zestawie foldów (1-8). Metoda UserConf od początku przewyższa metodę referencyjną, osiągając 51.43 dla jednego folda, a kończąc na 80.47 przy ośmiu foldach, co świadczy o stabilnym wzroście jakości predykcji. Z kolei metoda UserEmb charakteryzuje się jeszcze lepszą skutecznością, przewyższając UserConf już przy dwóch foldach uczących, osiągając 60.17 w porównaniu do 54.69 dla UserConf. W miarę wzrostu liczby foldów UserEmb systematycznie poprawia swoje wyniki, osiągając ostatecznie najwyższą wartość F1 macro spośród wszystkich metod — 81.68 przy ośmiu foldach, co czyni ją najlepszą metodą w tej analizie. Metoda HuBi-Simple, choć początkowo osiąga wyniki niższe niż UserConf i UserEmb, wykazuje stabilny wzrost jakości predykcji, od 47.93 dla jednego folda do 79.50 przy ośmiu foldach. Podobnie metoda HuBi-Medium, która zaczyna z wynikiem 46.18, osiąga 79.96 przy ośmiu foldach, co czyni ją drugą najskuteczniejszą metodą na poziomie pełnej liczby foldów. Ogólnie, zarówno UserEmb, jak i HuBi-Medium wykazują wyższą skuteczność w miarę wzrostu liczby foldów, natomiast metoda referencyjna, mimo poprawy, pozostaje najgorsza pod względem skuteczności predykcji. UserEmb, osiągając wartość 81.68, jest zdecydowanie najlepszą metodą na zbiorze WikiDetox: Toxicity, zarówno przy pełnym zestawie foldów, jak i przy mniejszej ich liczbie.

Tabela 12: Wartości miary F1 macro w zależności od liczby foldów uwzględnionych w zbiorze uczącym na zbiorze danych WikiDetox: Toxicity. Wartości **pogrubione** oznaczają najlepszy wynik dla danej liczby foldów uczących. Natomiast wartości podkreślone oznaczają najlepszy wynik dla danej metody rozpoznawania tekstów obraźliwych.

Foldy uczące / Metoda	1	1-2	1-3	1-4	1-5	1-6	1-7	1-8
Metoda Referencyjna	42.36 ± 0.45	46.16 ± 0.91	53.01 ± 0.61	55.35 ± 0.65	60.79 ± 0.91	64.04 ± 0.44	70.61 ± 0.94	75.88 ± 0.98
UserConf	**51.43** ± 0.39	54.69 ± 0.46	55.30 ± 0.43	61.54 ± 0.53	62.13 ± 0.39	64.24 ± 0.89	72.59 ± 0.69	80.47 ± 0.85
UserEmb	50.37 ± 0.51	**60.17** ± 0.46	**63.03** ± 0.73	**69.37** ± 0.41	**72.29** ± 0.81	**75.42** ± 0.45	76.40 ± 0.95	**81.68** ± 0.78
HuBi-Simple	47.93 ± 0.53	56.17 ± 0.36	60.56 ± 0.87	66.46 ± 0.73	70.09 ± 0.88	70.42 ± 0.96	72.54 ± 0.71	79.50 ± 0.60
HuBi-Medium	46.18 ± 0.74	52.38 ± 0.63	59.11 ± 0.66	61.14 ± 0.59	65.94 ± 0.67	72.47 ± 0.41	**76.58** ± 0.89	79.96 ± 0.61

Tabela 13 opisuje wyniki dla zbioru danych Measuring Hate Speech. Metoda referencyjna osiąga najniższe rezultaty, zaczynając od 21.07 dla jednego folda i stopniowo poprawiając się do 34.54 przy ośmiu foldach, jednak pozostaje najgorszą metodą w każdej konfiguracji liczby foldów. UserConf wykazuje wyższą skuteczność od samego początku, osiągając 34.17 dla jednego folda i kończąc na 42.34 przy ośmiu foldach, co czyni ją najlepszą metodą w tym zestawieniu. UserEmb również odznacza się stabilnym wzrostem efektywności, zaczynając od 33.47 przy jednym foldzie i kończąc na 41.64 przy ośmiu foldach, zajmując drugie miejsce w tej analizie. HuBi-Simple początkowo osiąga wyniki zbliżone do UserEmb, startując z wynikiem 30.64 przy jednym foldzie, a kończąc na 39.86, co plasuje ją tuż za UserEmb. Z kolei HuBi-Medium, choć początkowo nieco gorsza niż HuBi-Simple, poprawia swoje wyniki z 29.77 dla jednego folda do 40.44 przy pełnym zestawie foldów, kończąc na trzeciej pozycji. Z analizy wynika, że UserConf osiąga najlepsze rezultaty spośród wszystkich metod, uzyskując najwyższą wartość F1 macro, wynoszącą 42.34 przy ośmiu foldach.

Tabela 13: Wartości miary F1 macro w zależności od liczby foldów uwzględnionych w zbiorze uczącym na zbiorze danych Measuring Hate Speech. Wartości **pogrubione** oznaczają najlepszy wynik dla danej liczby foldów uczących. Natomiast wartości podkreślone oznaczają najlepszy wynik dla danej metody rozpoznawania tekstów obraźliwych.

Foldy uczące / Metoda	1	1-2	1-3	1-4	1-5	1-6	1-7	1-8
Metoda Referencyjna	21.07 ± 3.03	23.43 ± 4.17	25.79 ± 4.95	26.48 ± 3.95	28.99 ± 4.79	30.16 ± 3.36	31.79 ± 4.29	34.54 ± 4.39
UserConf	**34.17** ± 3.47	34.61 ± 3.58	**35.74** ± 3.34	**37.20** ± 3.90	**37.28** ± 3.11	**39.08** ± 3.02	**39.82** ± 3.34	**42.34** ± 3.02
UserEmb	33.47 ± 4.11	**34.73** ± 4.09	34.73 ± 4.22	36.47 ± 4.21	37.09 ± 4.33	38.30 ± 4.05	39.39 ± 4.47	41.64 ± 3.82
HuBi-Simple	30.64 ± 4.53	31.55 ± 4.52	33.74 ± 4.11	35.75 ± 4.70	37.00 ± 3.75	37.10 ± 4.58	37.85 ± 4.82	39.86 ± 4.68
HuBi-Medium	29.77 ± 3.96	30.59 ± 3.81	31.85 ± 3.45	33.32 ± 3.07	35.60 ± 4.00	37.88 ± 3.42	39.28 ± 4.16	40.44 ± 3.91

W tabeli 14 przedstawiono wyniki dla zbioru Unhealthy Conversations. Metoda referencyjna osiąga najniższe wyniki w każdej konfiguracji liczby foldów, zaczynając od wartości 31.09 dla jednego folda i osiągając 43.27 przy ośmiu foldach, co czyni ją najmniej skuteczną metodą w tym zestawieniu. UserConf systematycznie przewyższa metodę referencyjną we wszystkich etapach, począwszy od wartości 37.89 dla jednego folda, a kończąc na 53.12 przy ośmiu foldach, co czyni ją najlepszą metodą w tej analizie. UserEmb, choć nieco gorsza od UserConf, również wykazuje stabilny wzrost skuteczności, od 37.01 dla jednego folda do 51.89 przy ośmiu foldach, co plasuje ją na drugim miejscu. HuBi-Simple zaczyna od wartości 35.19 przy jednym foldzie, a następnie osiąga 49.89 przy ośmiu foldach, co czyni ją jedną z lepszych metod, choć nie dorównuje skutecznością metodom UserConf i UserEmb. HuBi-Medium, z wynikiem startowym 34.81 dla jednego folda, poprawia swoje rezultaty do 48.89 przy ośmiu foldach, plasując się za HuBi-Simple. Ostatecznie UserConf okazuje się najlepszą metodą w tej analizie, osiągając najwyższą wartość miary F1 macro — 53.12 przy pełnej liczbie foldów, co wyraźnie odróżnia ją od pozostałych metod.

Tabela 14: Wartości miary F1 macro w zależności od liczby foldów uwzględnionych w zbiorze uczącym na zbiorze danych Unhealthy Conversations. Wartości **pogrubione** oznaczają najlepszy wynik dla danej liczby foldów uczących. Natomiast wartości podkreślone oznaczają najlepszy wynik dla danej metody rozpoznawania tekstów obraźliwych.

Foldy uczące / Metoda	1	1-2	1-3	1-4	1-5	1-6	1-7	1-8
Metoda Referencyjna	31.09 ± 0.27	33.34 ± 0.53	35.19 ± 0.63	38.24 ± 0.74	40.24 ± 0.62	42.17 ± 0.30	42.38 ± 0.77	43.27 ± 0.38
UserConf	**37.89** ± 0.73	**42.52** ± 0.67	**43.20** ± 0.37	**44.55** ± 0.55	**47.97** ± 0.57	**48.30** ± 0.28	**51.55** ± 0.32	**53.12** ± 0.52
UserEmb	37.01 ± 0.26	38.28 ± 0.70	40.23 ± 0.45	43.21 ± 0.77	44.94 ± 0.83	47.91 ± 0.40	49.14 ± 0.51	51.89 ± 0.29
HuBi-Simple	35.19 ± 0.22	37.35 ± 0.29	40.28 ± 0.59	42.81 ± 0.39	45.63 ± 0.28	46.51 ± 0.53	49.74 ± 0.63	49.89 ± 0.31
HuBi-Medium	34.81 ± 0.35	37.91 ± 0.34	38.20 ± 0.70	40.57 ± 0.58	41.63 ± 0.28	42.60 ± 0.69	45.49 ± 0.36	48.89 ± 0.51

Tabela 15 przedstawia wyniki dla zbioru danych Doccano 1. Metoda referencyjna osiąga najniższe wyniki we wszystkich konfiguracjach liczby foldów, zaczynając od 30.18 dla jednego folda i stopniowo zwiększając swoją wartość do 44.03 przy ośmiu foldach, co świadczy o pewnej poprawie, ale nadal pozostaje ona najmniej skuteczną metodą. UserConf wykazuje wyraźną przewagę nad metodą referencyjną we wszystkich etapach, począwszy od wartości 43.02 dla jednego folda, a kończąc na 53.23 przy pełnym zestawie foldów, co czyni ją jedną z

najlepszych metod w tej analizie. UserEmb osiąga również bardzo dobre wyniki, startując z poziomu 41.37 przy jednym foldzie i kończąc na 53.76 przy ośmiu foldach, co plasuje ją nieco wyżej niż UserConf. HuBi-Simple zaczyna z wynikiem 38.53 przy jednym foldzie i systematycznie poprawia się, osiągając wartość 51.42 przy ośmiu foldach. Podobnie HuBi-Medium, choć początkowo uzyskuje nieco gorsze wyniki niż HuBi-Simple (37.64 przy jednym foldzie), stopniowo poprawia swoją efektywność, kończąc z wynikiem 51.98 dla ośmiu foldów. Ostatecznie najlepszą metodą w tym zestawieniu okazuje się UserEmb, osiągając najwyższą wartość miary F1 macro, wynoszącą 53.76 przy pełnej liczbie foldów, co plasuje ją na pierwszym miejscu, przewyższając UserConf, który osiąga wartość 53.23.

Tabela 15: Wartości miary F1 macro w zależności od liczby foldów uwzględnionych w zbiorze uczącym na zbiorze danych Doccano 1. Wartości **pogrubione** oznaczają najlepszy wynik dla danej liczby foldów uczących. Natomiast wartości podkreślone oznaczają najlepszy wynik dla danej metody rozpoznawania tekstów obraźliwych.

Foldy uczące / Metoda	1	1-2	1-3	1-4	1-5	1-6	1-7	1-8
Metoda Referencyjna	30.18 ± 1.88	32.21 ± 1.65	34.45 ± 1.37	35.90 ± 1.24	37.79 ± 1.41	40.11 ± 1.83	42.78 ± 1.25	44.03 ± 1.44
UserConf	**43.02** ± 1.16	**45.37** ± 1.16	46.04 ± 1.27	**48.80** ± 1.11	**50.53** ± 1.35	50.96 ± 1.12	**52.40** ± 1.09	53.23 ± 1.13
UserEmb	41.37 ± 1.17	44.19 ± 1.98	**47.24** ± 2.20	47.29 ± 1.81	49.14 ± 2.24	**51.77** ± 2.05	52.30 ± 2.33	**53.76** ± 1.79
HuBi-Simple	38.53 ± 2.32	41.27 ± 1.34	42.19 ± 2.19	43.76 ± 1.51	45.42 ± 1.85	47.19 ± 1.29	48.59 ± 1.73	51.42 ± 2.55
HuBi-Medium	37.64 ± 1.95	38.58 ± 1.72	41.01 ± 1.79	44.47 ± 1.81	47.54 ± 1.76	48.23 ± 2.33	51.21 ± 1.91	51.98 ± 1.58

Tabela 16 opisuje wyniki dla zbioru danych Doccano 2. Metoda referencyjna odnotowuje najniższe wartości w każdej konfiguracji liczby foldów, zaczynając od 28.17 dla jednego folda i stopniowo poprawiając się do 41.19 przy ośmiu foldach. Niemniej jednak, mimo pewnej poprawy, metoda referencyjna pozostaje najgorszą metodą w zestawieniu. UserConf wykazuje wyraźną przewagę już od początku, osiągając 39.46 dla jednego folda, a kończąc na 51.78 przy pełnej liczbie foldów, co czyni ją najlepszą metodą w tej analizie. UserEmb również osiąga dobre wyniki, zaczynając od 37.62 dla jednego folda, a kończąc na 49.78 przy ośmiu foldach, co plasuje ją na drugim miejscu pod względem skuteczności. HuBi-Simple stopniowo poprawia swoje wyniki z początkowego poziomu 35.31 przy jednym foldzie, osiągając wartość 46.49 przy ośmiu foldach, co stawia ją w dalszym ciągu za UserEmb, ale przed metodą referencyjną. Z kolei HuBi-Medium, zaczynając od 34.26 dla jednego folda, poprawia swoje wyniki do 45.78 przy ośmiu foldach, co czyni ją nieco słabszą od HuBi-Simple. Ostatecznie, UserConf okazuje się najlepszą metodą, osiągając najwyższą wartość F1 macro — 51.78 przy pełnej liczbie foldów, wyprzedzając pozostałe metody w całym zakresie liczby foldów.

9.4.3 Badania wpływu wykorzystania metody aktywnego uczenia na jakość predykcji obraźliwości tekstu

W tabeli 17 opisano wyniki dla zbioru WikiDetox: Aggression. Metoda VarRatio odznacza się wyraźną przewagą nad innymi metodami, szczególnie w początkowych iteracjach. Już w pierwszej iteracji osiągnęła najlepszy wynik na poziomie 58.88, co stanowi znaczny wzrost w porównaniu do metody Random (56.50) oraz pozostałych metod. VarRatio utrzymywała

Tabela 16: Wartości miary F1 macro w zależności od liczby foldów uwzględnionych w zbiorze uczącym na zbiorze danych Doccano 2. Wartości **pogrubione** oznaczają najlepszy wynik dla danej liczby foldów uczących. Natomiast wartości podkreślone oznaczają najlepszy wynik dla danej metody rozpoznawania tekstów obraźliwych.

Metoda \ Foldy uczące	1	1-2	1-3	1-4	1-5	1-6	1-7	1-8
Metoda Referencyjna	28.17 ± 1.05	29.61 ± 2.08	29.74 ± 1.51	33.87 ± 1.40	36.52 ± 1.44	36.75 ± 2.34	39.79 ± 1.69	41.19 ± 2.35
UserConf	**39.46** ± 1.35	**39.84** ± 1.47	41.07 ± 1.33	41.20 ± 1.17	44.08 ± 1.35	**47.53** ± 1.57	**48.43** ± 1.54	**51.78** ± 1.36
UserEmb	37.62 ± 2.15	38.14 ± 2.15	**41.20** ± 2.28	**43.70** ± 2.50	**46.03** ± 2.39	47.15 ± 2.24	47.69 ± 2.08	49.78 ± 2.46
HuBi-Simple	35.31 ± 1.73	38.08 ± 2.10	39.24 ± 2.46	41.58 ± 2.04	43.21 ± 2.48	43.30 ± 2.30	45.33 ± 2.17	46.49 ± 2.23
HuBi-Medium	34.26 ± 1.97	35.71 ± 2.30	38.26 ± 2.14	38.65 ± 1.42	39.67 ± 1.56	42.34 ± 2.08	43.84 ± 2.30	45.78 ± 1.86

Tabela 17: Wartość miary F1 macro dla różnych metod aktywnego uczenia na zbiorze WikiDetox: Aggression. Numer iteracji określa liczbę anotacji dla każdego użytkownika, która została umieszczona w zbiorze uczącym. Wartości **pogrubione** oznaczają najlepszy wynik w danej iteracji. Natomiast wartości podkreślone oznaczają najlepszy wynik dla danej metody na przestrzeni wszystkich iteracji.

Metoda \ Iteracja	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Random	56.50±0.67	56.71±0.84	57.42±0.50	57.77±0.66	58.41±0.18	58.75±0.89	59.33±0.49	59.98±0.48	60.44±0.53	60.99±0.91	60.74±0.68	60.71±0.34	60.56±0.62	60.86±0.42
Kontrowersyjność	58.29±0.39	60.28±0.85	60.59±0.19	61.93±0.26	62.31±0.74	62.38±0.88	63.15±0.52	62.54±0.22	**63.13**±0.96	63.23±0.98	63.04±0.80	**63.85**±0.77	63.01±0.18	63.62±0.19
VarRatio	**58.88**±0.67	**60.73**±0.30	**61.82**±0.54	**62.42**±0.21	**62.74**±0.67	**62.61**±0.29	**63.53**±0.63	**63.05**±0.93	62.89±0.23	**63.55**±0.51	**63.41**±0.90	63.74±0.85	**63.81**±0.21	**63.68**±0.97
Ratio Distance	56.16±0.83	56.67±0.45	57.02±0.19	56.96±0.44	57.44±0.67	57.69±0.64	58.23±0.26	58.68±0.60	58.36±0.79	58.75±0.91	58.73±0.62	59.50±0.78	59.83±0.57	59.87±0.36
Stranger Count	56.59±0.62	56.55±0.80	57.04±0.18	57.74±0.93	58.33±0.64	58.81±0.92	58.95±0.66	59.68±0.25	60.09±0.50	60.30±0.80	60.44±0.96	59.82±0.50	60.67±0.77	60.88±0.43

swoją dominację niemal przez całą długość eksperymentu, z najlepszym wynikiem w 13. iteracji (63.81) oraz 10. iteracji (63.55).

Metoda Kontrowersyjność, choć początkowo nieznacznie gorsza od VarRatio, od 4. iteracji zaczyna osiągać wyniki porównywalne z tą metodą. W szczególności, w 12. iteracji osiągnęła wynik 63.85, co stanowiło najwyższą wartość spośród wszystkich uzyskanych w całym eksperymencie. Warto podkreślić, że Kontrowersyjność była jedyną metodą, która w niektórych iteracjach zbliżyła się lub nawet przewyższyła rezultaty VarRatio, co świadczy o jej wysokim potencjale, szczególnie w późniejszych etapach.

Z kolei metoda Random charakteryzowała się stabilnym wzrostem jakości predykcji, jednak jej wyniki były generalnie niższe niż dla bardziej zaawansowanych metod. Najlepszy wynik Random został uzyskany w 10. iteracji (60.99), ale już od 9. iteracji różnice między Random a metodami VarRatio i Kontrowersyjność stały się bardzo widoczne.

Metody Ratio Distance oraz Stranger Count wypadały najsłabiej, osiągając wyraźnie niższe wyniki niż VarRatio i Kontrowersyjność. Ratio Distance osiągnęła swoje maksymalne wyniki w 14. iteracji (59.87), natomiast Stranger Count w tej samej iteracji (60.88). Ich stabilność w późniejszych iteracjach jest zauważalna, jednak nadal odstają od czołowych metod.

Tabela 18 przedstawia wyniki dla zbioru danych WikiDetox: Toxicity. Już od pierwszej iteracji metoda Kontrowersyjność osiągała najlepsze wyniki, sukcesywnie zwiększając swoją przewagę w kolejnych iteracjach. Najwyższą wartość, 61.46, uzyskano w 10. iteracji, co jest najlepszym wynikiem na całej przestrzeni eksperymentu. Metoda Kontrowersyjność wykazała wysoką stabilność, utrzymując się powyżej innych metod, osiągając najlepsze wyniki w większości iteracji.

Tabela 18: Wartość miary F1 macro dla różnych metod aktywnego uczenia na zbiorze WikiDetox: Toxicity. Numer iteracji określa liczbę anotacji dla każdego użytkownika, która została umieszczona w zbiorze uczącym. Wartości **pogrubione** oznaczają najlepszy wynik w danej iteracji. Natomiast wartości podkreślone oznaczają najlepszy wynik dla danej metody na przestrzeni wszystkich iteracji.

Metoda \ Iteracja	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Random	56.62±0.88	55.89±0.57	55.73±0.76	57.05±0.85	56.65±0.49	56.57±0.57	56.97±0.71	57.69±0.34	58.08±0.54	57.54±0.66	57.77±0.95	58.00±0.63	58.11±0.80	58.56±0.61
Kontrowersyjność	**57.54**±0.45	**58.55**±0.97	**58.56**±0.19	**59.67**±0.58	**59.79**±0.71	**60.22**±0.99	**60.22**±0.56	**60.22**±0.47	**60.71**±0.71	**61.46**±0.70	**60.99**±0.53	**61.16**±0.92	**61.29**±0.66	**60.95**±0.35
VarRatio	57.04±0.50	57.55±0.28	57.74±0.94	58.62±0.22	58.36±0.93	58.69±0.60	58.95±0.61	59.21±0.19	59.72±0.28	59.76±0.55	58.83±0.53	59.18±0.26	59.30±0.46	59.74±0.80
Ratio Distance	57.15±0.39	56.62±0.44	57.04±0.86	57.58±0.61	57.34±0.62	57.50±0.19	57.92±0.69	58.22±0.89	57.27±0.95	58.44±0.71	58.23±0.26	58.23±0.51	58.28±0.46	58.61±0.79
Stranger Count	54.92±0.62	53.92±0.70	54.24±0.34	54.89±0.71	54.79±0.81	55.72±0.79	56.46±0.29	56.16±0.18	56.18±0.21	56.09±0.70	56.17±0.86	56.15±0.92	56.72±0.25	56.89±0.93

Metoda VarRatio, choć mniej skuteczna niż Kontrowersyjność, osiągała stabilne wyniki, z tendencją wzrostową w kolejnych iteracjach. W 10. iteracji VarRatio uzyskała wynik 59.76, co było jej najwyższym rezultatem, choć wyraźnie niższym niż dla Kontrowersyjności. Metoda ta zachowywała przewagę nad Random oraz Stranger Count, co sugeruje jej efektywność w zadaniach związanych z detekcją toksycznych treści, mimo że nie dorównywała Kontrowersyjności w najwyższych iteracjach.

Metoda Random charakteryzowała się dość zmiennymi wynikami, z początkowo słabszymi wynikami, ale osiągnęła maksymalny wynik 58.56 dopiero w 14. iteracji. Mimo tego nie dorównywała ona metodom Kontrowersyjność i VarRatio w większości iteracji, co sugeruje, że losowe wybieranie próbek do anotacji nie jest optymalnym podejściem w kontekście tego zadania.

Metoda Ratio Distance wykazywała nieco lepsze wyniki niż Random, choć jej najlepszy wynik 58.61 w 14. iteracji również był niższy niż Kontrowersyjność i VarRatio. Ratio Distance utrzymywała jednak względnie stabilne wyniki przez cały eksperyment, co może sugerować jej przydatność w zadaniach o mniejszej liczbie iteracji.

Najgorzej wypadała metoda Stranger Count, która od pierwszej iteracji osiągała niższe wartości niż pozostałe metody. Najlepszy wynik tej metody, uzyskany w 14. iteracji (56.89), był znacznie niższy niż wyniki osiągane przez pozostałe techniki. Wynika z tego, że Stranger Count nie jest optymalną metodą aktywnego uczenia w kontekście wykrywania toksycznych treści, a jej efektywność jest wyraźnie niższa niż bardziej zaawansowanych technik, takich jak Kontrowersyjność czy VarRatio.

Tabela 19 opisuje wyniki dla zbioru Measuring Hate Speech. Już od pierwszej iteracji metoda Kontrowersyjność wyprzedziła inne metody, uzyskując wynik 20.49, a następnie sukcesywnie poprawiała swoje wyniki, osiągając najlepszy wynik w 14. iteracji (39.98). Ta metoda wykazała się dużą stabilnością i przewagą, szczególnie w późniejszych iteracjach, kiedy różnice między nią a pozostałymi metodami stały się wyraźniejsze.

Metoda VarRatio była również dość efektywna, choć nie na poziomie Kontrowersyjności. Osiągnęła najwyższy wynik 38.74 w 14. iteracji, co plasuje ją na drugim miejscu. Choć w początkowych iteracjach wyniki tej metody były niższe od Kontrowersyjności, z czasem różnice zaczęły się zacierać, a VarRatio stopniowo zwiększała swoją skuteczność.

Random prezentowała początkowo bardzo słabe wyniki, z wynikiem 15.46 w pierwszej iteracji, ale stopniowo poprawiała swoją efektywność, osiągając swój najwyższy wynik 36.78

Tabela 19: Wartość miary F1 macro dla różnych metod aktywnego uczenia na zbiorze Measuring Hate Speech. Numer iteracji określa liczbę anotacji dla każdego użytkownika, która została umieszczona w zbiorze uczącym. Wartości **pogrubione** oznaczają najlepszy wynik w danej iteracji. Natomiast wartości podkreślone oznaczają najlepszy wynik dla danej metody na przestrzeni wszystkich iteracji.

Metoda \ Iteracja	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Random	15.46±3.62	15.55±2.96	16.64±3.25	17.72±4.10	19.16±3.02	21.54±3.65	21.60±3.29	24.15±3.45	26.67±4.07	29.57±3.81	31.10±4.05	31.85±3.97	33.78±3.24	36.78±3.44
Kontrowersyjność	**20.49**±3.38	**23.12**±3.30	**24.87**±2.87	**25.93**±3.16	**26.27**±3.32	**27.35**±3.50	**29.84**±3.70	**33.19**±2.80	**34.83**±2.56	**34.99**±3.31	**37.03**±3.07	**39.03**±2.86	**39.07**±2.91	**39.98**±3.39
VarRatio	18.63±3.72	20.36±3.88	20.78±3.92	20.93±2.80	21.13±2.68	21.28±3.43	23.18±2.77	25.59±3.70	26.77±2.79	27.56±2.75	30.61±2.75	33.56±3.92	36.08±3.31	38.74±2.54
Ratio Distance	16.29±3.94	17.09±3.51	18.66±2.82	21.34±3.02	23.83±3.78	25.05±3.13	26.20±3.17	26.89±3.75	28.83±3.40	30.90±3.41	32.93±3.33	33.94±2.75	35.42±2.50	35.36±3.96
Stranger Count	17.06±3.40	17.39±3.25	20.19±2.58	20.49±3.26	21.39±2.50	21.97±3.01	24.23±2.72	25.20±3.75	26.53±4.06	28.69±3.44	30.90±2.68	32.52±2.71	34.20±3.33	36.14±2.89

w 14. iteracji. Choć Random zdołała osiągnąć stosunkowo wysoką skuteczność w ostatnich iteracjach, jej wyniki były systematycznie gorsze od Kontrowersyjności i VarRatio.

Metoda Ratio Distance osiągnęła swoje maksymalne wyniki na poziomie 35.42 w 13. iteracji, co pokazuje, że jest mniej skuteczna niż metody Kontrowersyjność i VarRatio, choć lepsza od Random na wcześniejszych etapach eksperymentu.

Stranger Count, choć początkowo wypadała nieco lepiej niż Random i Ratio Distance, ostatecznie nie zdołała osiągnąć wyników porównywalnych z Kontrowersyjnością i VarRatio. Najlepszy wynik tej metody to 36.14, uzyskany w 14. iteracji, co pokazuje, że Stranger Count, choć konkurencyjna, jest mniej efektywna w porównaniu do pozostałych strategii.

W tabeli 20 przedstawiono wyniki dla zbioru danych Unhealthy Conversations. Kontrowersyjność od pierwszej iteracji utrzymuje najwyższe wyniki w każdej z kolejnych iteracji. Już w pierwszej osiąga wartość 19.75, przez co Kontrowersyjność ta znacznie przewyższyła pozostałe metody. W miarę postępu iteracji jej przewaga stawała się jeszcze bardziej wyraźna, a maksymalny wynik 35.10 został uzyskany w 12. iteracji, co stanowi najwyższy rezultat w całym eksperymencie.

Metoda Random wykazywała stabilny wzrost jakości predykcji, ale jej wyniki były znacznie niższe niż dla Kontrowersyjności. Random osiągnęła swój najlepszy wynik w 12. iteracji (28.38), co sugeruje, że metoda ta, mimo iż jest prostsza, nie dorównuje bardziej zaawansowanym technikom.

VarRatio odznaczała się stosunkowo niskimi wynikami w początkowych iteracjach, ale po 10. iteracji zaczęła wykazywać poprawę, osiągając najwyższy wynik 22.20 w 14. iteracji. Choć VarRatio nie była tak skuteczna jak Kontrowersyjność, pokazała potencjał w późniejszych etapach eksperymentu.

Metoda Ratio Distance była bardziej efektywna niż VarRatio w większości iteracji, a jej wyniki systematycznie rosły. W 14. iteracji osiągnęła wynik 29.89, co czyni ją drugą najlepszą metodą po Kontrowersyjności. Jest to wskazówka, że Ratio Distance może być dobrą alternatywą dla bardziej złożonych metod.

Stranger Count, choć lepsza od Random i VarRatio w początkowych iteracjach, osiągnęła stosunkowo niższe wyniki w porównaniu do Kontrowersyjności i Ratio Distance. Jej maksymalny wynik (25.62) w 14. iteracji pozostaje daleko za najlepszymi wynikami Kontrowersyjności i Ratio Distance.

Tabela 20: Wartość miary F1 macro dla różnych metod aktywnego uczenia na zbiorze Unhealthy Conversations. Numer iteracji określa liczbę anotacji dla każdego użytkownika, która została umieszczona w zbiorze uczącym. Wartości **pogrubione** oznaczają najlepszy wynik w danej iteracji. Natomiast wartości podkreślone oznaczają najlepszy wynik dla danej metody na przestrzeni wszystkich iteracji.

Metoda \ Iteracja	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Random	14.17±0.59	17.87±0.37	20.92±0.20	20.80±0.49	21.38±0.38	23.70±0.74	22.60±0.96	24.57±0.73	26.82±0.60	26.82±0.90	27.68±0.27	28.38±0.19	27.93±0.25	27.33±0.67
Kontrowersyjność	**19.75**±0.72	**21.77**±0.28	**25.66**±0.60	**27.75**±0.90	**29.44**±0.85	**29.74**±0.96	**30.44**±0.96	**33.01**±0.97	**32.45**±0.26	**33.54**±0.18	**33.97**±0.40	**35.10**±0.25	**34.34**±0.63	**34.99**±0.97
VarRatio	15.22±0.37	15.05±0.65	17.66±0.70	15.91±0.29	17.59±0.93	17.14±0.32	18.06±0.31	18.73±0.92	18.88±0.63	21.13±0.24	19.60±0.51	19.96±0.94	21.20±0.39	22.20±0.96
Ratio Distance	13.65±0.32	16.20±0.99	16.82±0.99	19.83±0.61	23.10±0.60	24.89±0.40	25.90±0.28	26.72±0.35	26.95±0.78	28.22±0.64	27.90±0.31	28.35±0.77	28.87±0.89	29.89±0.42
Stranger Count	14.17±0.63	16.30±0.32	18.97±0.59	19.96±0.58	20.92±0.89	21.66±0.74	22.05±0.77	23.17±0.51	24.30±0.73	24.83±0.69	25.04±0.20	24.95±0.82	25.33±0.82	25.62±0.71

Tabela 21: Wartość miary F1 macro dla różnych metod aktywnego uczenia na zbiorze Doccano 1. Numer iteracji określa liczbę anotacji dla każdego użytkownika, która została umieszczona w zbiorze uczącym. Wartości **pogrubione** oznaczają najlepszy wynik w danej iteracji. Natomiast wartości podkreślone oznaczają najlepszy wynik dla danej metody na przestrzeni wszystkich iteracji.

Metoda \ Iteracja	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Random	14.41±1.71	15.69±1.98	17.19±1.56	18.46±2.10	20.80±1.88	22.95±1.78	23.52±1.72	26.01±1.65	26.58±1.69	28.05±1.92	28.11±1.63	29.04±1.78	31.56±1.51	32.23±1.90
Kontrowersyjność	**19.32**±1.57	**20.03**±2.09	20.12±1.81	22.81±1.89	23.34±1.96	23.83±1.71	26.38±1.61	**28.10**±1.63	**30.67**±2.06	30.81±1.87	**34.04**±1.69	**35.16**±2.05	**37.17**±1.91	**37.54**±1.64
VarRatio	18.26±1.59	19.99±1.87	**21.66**±1.57	**23.51**±1.70	**24.56**±1.93	**25.08**±2.03	**26.46**±1.78	27.67±2.03	29.52±1.59	**31.14**±1.99	32.29±1.62	33.37±1.60	35.16±1.78	36.15±1.80
Ratio Distance	16.31±2.02	16.32±1.84	16.83±2.00	18.85±1.72	20.79±1.52	22.96±1.77	24.58±1.93	24.73±1.96	25.73±1.84	27.69±1.64	28.33±1.70	30.52±1.96	32.40±2.00	34.18±1.94
Stranger Count	17.02±1.61	19.08±1.83	19.74±1.60	19.96±1.92	21.19±1.86	22.31±1.83	23.58±1.81	25.66±2.10	27.48±1.95	28.87±1.93	31.23±1.79	31.63±1.68	33.26±1.80	34.78±1.79

W tabeli 21 opisano wyniki dla zbioru Doccano 1. Już w pierwszej iteracji metoda Kontrowersyjność osiągnęła najlepszy wynik (19.32), a jej efektywność wzrastała w kolejnych iteracjach, osiągając maksymalny wynik 37.54 w 14. iteracji. Wysoka stabilność i przewaga nad innymi metodami świadczą o skuteczności tej techniki, szczególnie w późniejszych etapach eksperymentu.

VarRatio była kolejną skuteczną metodą, uzyskując relatywnie wysokie wyniki w kolejnych iteracjach. Choć w początkowych etapach ustępowała Kontrowersyjności, stopniowo poprawiała swoje wyniki, osiągając w 14. iteracji wynik 36.15. Metoda ta utrzymywała stabilny wzrost efektywności, co czyni ją dobrą alternatywą, choć nie tak efektywną jak Kontrowersyjność.

Random, choć prostsza i mniej przewidywalna, również odnotowała wzrost skuteczności na przestrzeni iteracji, osiągając maksymalny wynik 32.23 w 14. iteracji. Wynik ten był jednak znacznie niższy od rezultatów Kontrowersyjności i VarRatio, co wskazuje, że losowe podejście do wyboru próbek jest mniej efektywne.

Ratio Distance miała bardziej zmienne wyniki, szczególnie w początkowych iteracjach, ale w miarę postępu eksperymentu jej skuteczność rosła, osiągając maksymalny wynik 34.18 w 14. iteracji. Choć Ratio Distance była mniej efektywna niż Kontrowersyjność i VarRatio, przewyższała metodę Random w większości iteracji.

Stranger Count osiągnęła swój najwyższy wynik 34.78 w 14. iteracji, co wskazuje na jej umiarkowaną efektywność. Choć metoda ta była mniej efektywna niż Kontrowersyjność i VarRatio, wyprzedzała Random w późniejszych iteracjach, co czyni ją solidną, choć nieoptymalną strategią.

Tabela 22 przedstawia wynuiki dla zbioru danych Doccano 2. Metoda Kontrowersyjność w 1. iteracji osiągnęła najwyższą wartość miary F1 macro wśród wszystkich badanych metod

Tabela 22: Wartość miary F1 macro dla różnych metod aktywnego uczenia na zbiorze Doccano 2. Numer iteracji określa liczbę anotacji dla każdego użytkownika, która została umieszczona w zbiorze uczącym. Wartości **pogrubione** oznaczają najlepszy wynik w danej iteracji. Natomiast wartości podkreślone oznaczają najlepszy wynik dla danej metody na przestrzeni wszystkich iteracji.

Metoda \ Iteracja	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Random	13.36±1.97	15.07±1.31	16.47±2.09	16.59±2.16	17.88±1.76	18.14±2.13	18.20±1.22	18.79±1.21	20.59±1.29	22.03±1.27	22.32±1.99	24.34±2.32	24.59±1.59	25.78±1.73
Kontrowersyjność	**18.26**±1.77	**19.37**±2.33	**22.03**±1.88	**23.40**±2.14	**24.99**±1.45	**25.05**±1.34	**26.91**±2.41	**29.44**±1.99	**31.06**±2.19	**33.55**±1.53	**35.40**±1.33	**35.58**±2.25	**37.39**±1.51	**38.58**±2.72
VarRatio	17.93±1.78	18.89±1.16	20.70±1.15	22.67±1.68	23.51±2.58	24.74±2.19	25.13±1.18	26.97±2.46	29.97±1.53	32.12±2.88	32.23±2.18	32.98±2.14	34.59±1.69	36.12±2.22
Ratio Distance	15.43±1.37	15.61±1.54	17.68±1.84	18.78±2.50	19.64±2.32	21.95±2.77	22.55±1.88	24.10±1.17	25.13±2.11	26.64±2.21	28.40±1.30	30.20±2.52	32.18±2.41	32.46±1.52
Stranger Count	15.06±2.65	16.75±1.56	18.42±1.82	19.94±2.64	21.04±1.60	22.79±1.87	22.97±2.00	24.35±2.68	26.10±1.72	27.83±1.33	29.39±1.40	31.28±1.17	32.83±2.65	33.29±2.82

(18.26), a następnie stopniowo zwiększała swoją skuteczność, osiągając maksymalny wynik 38.58 w 14. iteracji. Metoda ta wyraźnie przewyższa inne strategie we wszystkich iteracjach, a jej wyniki w ostatnich iteracjach znacznie wyprzedzają pozostałe metody.

VarRatio jest drugą najskuteczniejszą metodą w eksperymencie, osiągając swój najlepszy wynik (36.12) w 14. iteracji. Podobnie jak w przypadku Kontrowersyjności, metoda ta wykazuje stały wzrost skuteczności, a różnica między VarRatio a Kontrowersyjnością zaczyna się zmniejszać w późniejszych iteracjach. Choć metoda ta nie osiągnęła wyników Kontrowersyjności, pozostaje konkurencyjna w kontekście wykrywania obraźliwych treści.

Metoda Random, jak w poprzednich zbiorach, wypada najsłabiej, choć odnotowuje systematyczny wzrost na przestrzeni iteracji. Osiągnęła swój najwyższy wynik (25.78) w 14. iteracji, co jest znacznie niższym wynikiem niż dla Kontrowersyjności czy VarRatio. Pomimo pewnych wzrostów, Random nie jest tak skuteczna, jak bardziej zaawansowane metody aktywnego uczenia.

Ratio Distance oraz Stranger Count utrzymują się na poziomie pośrednim, osiągając swoje maksymalne wyniki odpowiednio 32.46 i 33.29 w 14. iteracji. Choć są one mniej efektywne niż Kontrowersyjność i VarRatio, wykazują stopniowy wzrost skuteczności w miarę wzrostu liczby iteracji. Obie te metody przewyższają Random, ale są mniej efektywne niż czołowe techniki.

9.4.4 Badania jakości predykcji generatywnego modelu ogólnego przeznaczenia w zadaniu predykcji obraźliwości tekstu

Tabela 23 przedstawia wyniki dla modelu generatywnego ChatGPT-3.5 oraz modelu predykcyjnego UserConf oraz wartość miar Loss i Gain na zbiorach WikiDetox: Aggression oraz Unhealthy Conversations. Model UserConf osiągnął wyraźnie lepsze wyniki w obu zbiorach danych. Dla zbioru WikiDetox: Aggression, miara F1 macro wyniosła 81.87, co znacznie przewyższa wyniki uzyskane przez ChatGPT-3.5, który w trybie zero-shot prompting uzyskał wynik 69.10, a w trybie in-context learning nieco lepszy wynik 72.57. Analogiczna sytuacja miała miejsce dla zbioru Unhealthy Conversations, gdzie UserConf uzyskał F1 macro na poziomie 53.12, przewyższając zarówno zero-shot prompting (45.21), jak i in-context learning (52.02) modelu ChatGPT-3.5.

Tabela 23: Wartość miary F1 macro dla modelu ChatGPT-3.5 oraz modelu UserConf oraz wartość miar Loss i Gain na zbiorach WikiDetox: Aggression oraz Unhealthy Conversations.

Miara \ Zbiór danych	WikiDetox: Aggression	Unhealthy Conversations
F1 macro (UserConf)	81.87	53.12
F1 macro (ChatGPT-3.5 *zero-shot prompting*)	69.10	45.21
F1 macro (ChatGPT-3.5 *in-context learning*)	72.57	52.02
Loss (*zero-shot prompting*)	15.60	14.90
Loss (*in-context learning*)	11.36	2.07
Gain	11.23	12.43

Wartości miar Loss pokazują, jak ChatGPT-3.5 radził sobie w obu trybach. W trybie zero-shot prompting, model wykazywał większe straty, wynoszące 15.60 dla WikiDetox: Aggression oraz 14.90 dla Unhealthy Conversations. Z kolei w trybie in-context learning straty były znacznie mniejsze: 11.36 i 2.07 odpowiednio dla obu zbiorów, co pokazuje, że ten tryb poprawia wyniki modelu ChatGPT-3.5, choć nadal nie dorównuje on modelowi UserConf.

Ostatecznie, wartość Gain, która odzwierciedla poprawę jakości predykcji, była nieco wyższa w przypadku zbioru Unhealthy Conversations (12.43) niż w WikiDetox: Aggression (11.23), co może sugerować, że różnice między modelami były bardziej widoczne w trudniejszych zadaniach.

9.5 ANALIZA WYNIKÓW

Niniejszy rozdział prezentuje szczegółową analizę wyników badań nad efektywnością spersonalizowanych metod rozpoznawania tekstów obraźliwych. Wyniki te wyraźnie pokazują, że personalizacja znacząco poprawiła jakość predykcji, umożliwiając lepsze rozpoznanie treści obraźliwych w różnych kontekstach. Zastosowanie zaawansowanych modeli oraz technik aktywnego uczenia pozwoliło na uzyskanie wyższej precyzji w klasyfikacji treści obraźliwych, nawet przy ograniczonych zbiorach danych. W kolejnych podrozdziałach szczegółowo omówiono wpływ takich elementów jak reprezentacja wektorowa tekstu, rozmiar zbioru treningowego, skuteczność aktywnego uczenia na jakość predykcji modeli, a także porównanie dedykowanych metod predykcji obraźliwości z generatywnym modelem ogólnego przeznaczenia.

9.5.1 Analiza wpływu metody generowania reprezentacji wektorowej tekstu oraz modelu głębokiej sieci neuronowej służącej do predykcji obraźliwości tekstu

Na podstawie otrzymanych wyników można zauważyć, że wartości F1 macro przewyższały metodę referencyjną od kilku do kilkunastu punktów, co pokazuje przewagę metod spersonalizowanych nad uogólnionymi w kontekście rozpoznawania obraźliwych treści tekstowych.

Wartości F1 macro dla tych modeli przewyższały metodę referencyjną od kilku do kilkunastu punktów, co pokazuje przewagę metod spersonalizowanych nad uogólnionymi w kontekście rozpoznawania obraźliwych treści tekstowych. Wartości F1 macro dla metody referencyjnej na poszczególnych zbiorach danych były stosunkowo niskie, wynosząc od 43 do 47 punktów. Wynika to z prostoty podejścia uogólnionego, które nie uwzględnia różnic w percepcji użytkowników i nie dostosowuje się do specyfiki każdego przypadku. Metoda referencyjna bazuje na bardziej tradycyjnych technikach, które mimo stabilności, nie osiągają takiej skuteczności jak nowoczesne, personalizowane modele. Analiza wyników pokazuje, że uwzględnienie reprezentacji użytkownika istotnie wpływa na poprawę jakości predykcji metod rozpoznawania tekstów obraźliwych w porównaniu z podejściem wykorzystującym generalizację, co potwierdza hipotezę H1.. Personalizacja pozwala na lepsze dopasowanie modelu do indywidualnej percepcji każdego użytkownika, co prowadzi do lepszej klasyfikacji.

Model UserConf, który wprowadza personalizację poprzez uwzględnienie miar konformizmu użytkownika, osiągnął najlepsze wyniki we wszystkich zbiorach danych (Rys.18). Kluczowym elementem jego skuteczności jest sposób generowania reprezentacji wektorowych tekstu, w którym model ten nie tylko wykorzystuje kontekst semantyczny, ale również wzbogaca reprezentację o kontekst użytkownika wyrażony miarami konformizmu obliczanymi na podstawie zachowań użytkownika, co zapewnia wyższą trafność predykcji w porównaniu z metodą referencyjną. W szczególności różnice w miarach F1 między UserConf a metodą referencyjną były najbardziej zauważalne w zbiorach WikiDetox: Aggression oraz WikiDetox: Toxicity, gdzie agresywne i toksyczne wypowiedzi zostały precyzyjniej rozpoznane dzięki miarom konformizmu. Również w zbiorach Measuring Hate Speech i Unhealthy Conversations, które zawierały bardziej złożone konteksty mowy nienawiści, model UserConf skutecznie przetwarzał reprezentacje wektorowe, dostosowując je do specyficznych cech użytkownika, co przełożyło się na wyższą jakość predykcji. W przypadku zbiorów Doccano 1 i Doccano 2, model wykazał podobnie wysoką skuteczność w rozpoznawaniu obraźliwości tekstu, przewyższając metodę referencyjną zarówno w generowaniu reprezentacji wektorowych, jak i w precyzji predykcji (Rys.19). Spersonalizowane modele sieci neuronowych, w szczególności UserConf i UserEmb, cechują się wyższą skutecznością w zadaniu rozpoznawania obraźliwości niż zarówno metoda referencyjna, jak i inne, prostsze metody spersonalizowane, takie jak HuBi-Simple czy HuBi-Medium, co potwierdza hipotezę H2.. Zdolność tych bardziej zaawansowanych modeli do uchwycenia kontekstu użytkownika oraz zastosowania nowoczesnych reprezentacji wektorowych, takich jak te oparte na transformatorach, wyraźnie podnosi jakość predykcji. Różnice w wynikach F1 macro pomiędzy tymi modelami a metodą referencyjną potwierdzają przewagę personalizacji w kontekście rozpoznawania obraźliwości tekstów.

Dla zbioru WikiDetox: Aggression model UserConf uzyskał wartość miary F1 macro na poziomie 53.12, co daje różnicę 9.85 punktów w stosunku do metody referencyjnej (Rys. 12). W przypadku zbioru WikiDetox: Toxicity, model UserConf osiągnął wartość F1 macro wyższą o 9.85 punktów w porównaniu do metody referencyjnej. Model UserEmb przewyższył metodę referencyjną o 8.62 punktów, natomiast HuBi-Simple i HuBi-Medium uzyskały różnice na poziomie odpowiednio 5.85 i 4.85 punktów (Rys. 13). Istotną rolę reprezentacji wektorowych

tekstów potwierdzają również wysokie wartości miary $F1_{roznica}$ dla każdego modelu spersonalizowaego. Jedną z przyczyn jest charakterystyczne słownictwo oraz teksty niosące ze sobą ładunek negatywny silnie powiązany z ich semantyką. Taki charakter zbioru danych skutkuje wysoką poprawą jakości predykcji modeli uczonych na relewantnych reprezentacjach wektorowych w przeciwieństwie do modelu, który nie ma dostępu do reprezentatywnej reprezentacji semantyki tekstu. Dla zbioru Measuring Hate Speech, różnica F1 macro między modelem UserConf a metodą referencyjną wyniosła 7.12 punktów, podczas gdy model UserEmb uzyskał 6.00 punktów więcej. HuBi-Simple i HuBi-Medium przewyższyły metodę referencyjną o odpowiednio 4.00 i 3.00 punkty procentowe (Rys. 14). Na zbiorze Unhealthy Conversations, model UserConf uzyskał wartość F1 macro wyższą o 8.62 punktów w stosunku do metody referencyjnej, a model UserEmb o 7.62 punktów. Modele HuBi-Simple i HuBi-Medium przewyższyły metodę referencyjną o odpowiednio 5.62 i 4.62 punktów (Rys. 15). Dla zbioru Doccano 1, różnica między modelem UserConf a metodą referencyjną wyniosła 8.85 punktów, natomiast dla UserEmb było to 7.85 punktów. Modele HuBi-Simple i HuBi-Medium uzyskały przewagę nad metodą referencyjną o odpowiednio 6.85 i 5.85 punktów (Rys. 16). W zbiorze Doccano 2, model UserConf przewyższył metodę referencyjną o 8.12 punktów F1 macro, a UserEmb o 7.12 punktów. Dla modeli HuBi-Simple i HuBi-Medium różnice wyniosły odpowiednio 6.12 i 5.12 punktów (Rys. 17).

Kluczową zaletą modelu UserConf jest zdolność do lepszego uchwycenia kontekstu użytkownika, co pozwala na bardziej precyzyjne dopasowanie predykcji do indywidualnych preferencji i oczekiwań. Podobnie model UserEmb, który wykorzystuje średnie reprezentacje semantyczne, również osiągnął lepsze wyniki, choć w mniejszym stopniu niż UserConf. Różnice F1 macro wynosiły od 7 do 8 punktów w zależności od zbioru danych, co pokazuje, że średnie reprezentacje semantyczne wprowadzają wartościowy, choć nieco mniej precyzyjny kontekst.

Modele HuBi-Simple i HuBi-Medium, które wprowadzają uproszczoną personalizację poprzez mniej złożone mechanizmy adaptacji do preferencji użytkownika, również przewyższały metodę referencyjną, ale różnice były mniejsze. Wyniki tych modeli są zbliżone, a różnice F1 macro wahały się od 3 do 6 punktów. Te modele są kompromisem między złożonością a skutecznością, co czyni je bardziej efektywnymi w sytuacjach, gdy dostępne są ograniczone zasoby obliczeniowe lub gdy bardziej zaawansowana personalizacja, jak w modelu UserConf, nie jest konieczna.

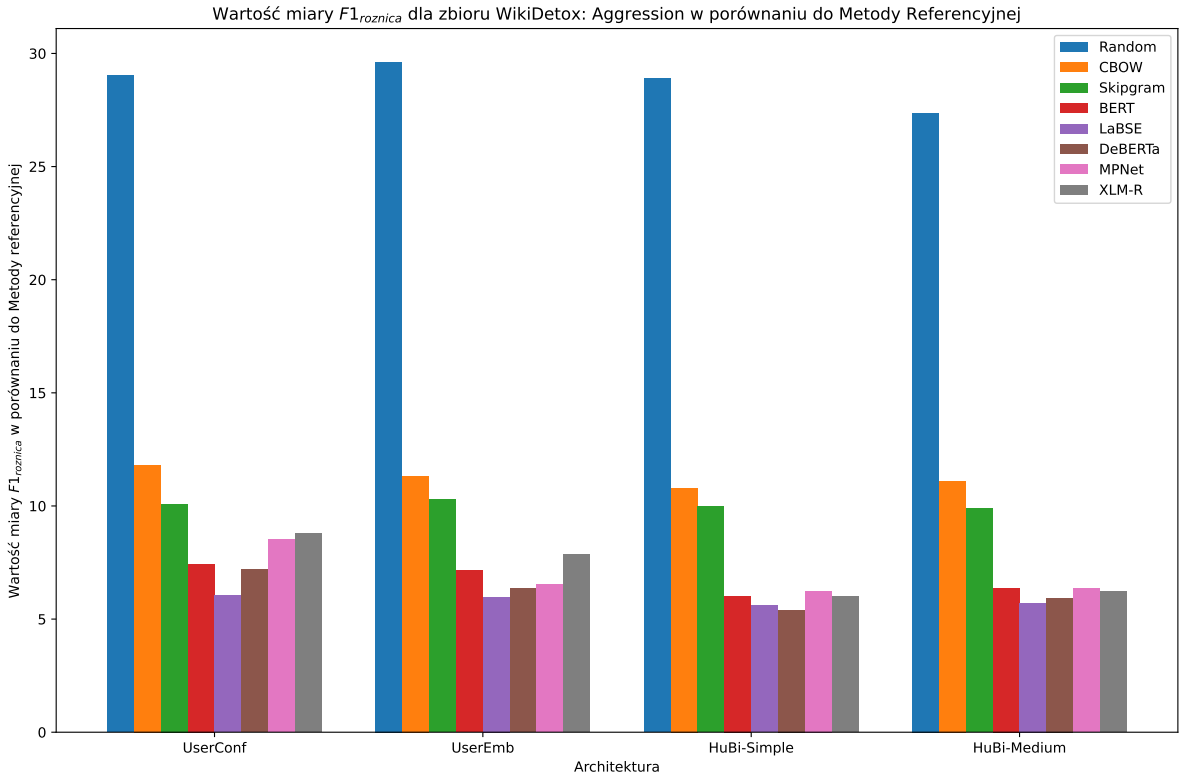
Wyniki te jednoznacznie wskazują na przewagę metod spersonalizowanych, które uwzględniają zróżnicowaną percepcję użytkowników i potrafią lepiej dostosować swoje predykcje do specyficznych przypadków. Dlaczego jednak te modele są lepsze? Głównym powodem jest ich zdolność do uchwycenia kontekstu użytkownika, co umożliwia bardziej precyzyjne rozpoznanie treści obraźliwych w zależności od tego, jak są one postrzegane przez różnych użytkowników. Modele oparte na reprezentacjach kontekstowych, takich jak DeBERTa czy MPNet, dodatkowo zwiększają skuteczność tych predykcji, gdyż potrafią lepiej analizować złożoność semantyczną tekstów.

Metoda referencyjna, mimo swojej prostoty, nie jest w stanie osiągnąć porównywalnych wyników, gdyż nie uwzględnia różnic w percepcji użytkowników i bazuje na tradycyjnych technikach generowania reprezentacji tekstu, które są mniej skuteczne w zadaniu predykcji obraźliwości tekstu. Wyniki F1 macro dla tego modelu były w każdym przypadku najniższe, co pokazuje, że bardziej złożone modele oparte na personalizacji są kluczowe dla zwiększenia skuteczności w tej dziedzinie.

W badaniach nad generowaniem reprezentacji wektorowych tekstu uwzględniono zarówno kontekstowe metody reprezentacji, jak i bezkontekstowe. Modele kontekstowe, takie jak BERT, LaBSE, DeBERTa, MPNet i XLM-R, które opierają się na architekturach typu transformer, osiągnęły najlepsze wyniki we wszystkich zbiorach danych. Na przykład, reprezentacje uzyskane za pomocą modelu MPNet uzyskały najwyższe wartości miary F1 macro w połączeniu z metodą UserConf, co świadczy o zdolności tego modelu do skutecznego przetwarzania informacji zawartej w tekście w zadaniu rozpoznawania obraźliwości tekstu. Również model DeBERTa uzyskał bardzo wysokie wyniki, co podkreśla siłę tych zaawansowanych metod w porównaniu do tradycyjnych, bezkontekstowych podejść.

Z kolei metody bezkontekstowe, takie jak CBOW i Skipgram, radziły sobie znacznie gorzej niż modele kontekstowe. Te metody przewidują słowa na podstawie otoczenia i generują reprezentacje wektorowe, ale nie uwzględniają pełnego kontekstu całego zdania. Mimo to, przewyższały one metodę Random, która generowała losowe wektory dla tekstów, uzyskując najniższe wyniki F1 macro, co pokazuje, że nawet proste, bezkontekstowe modele są bardziej skuteczne niż podejścia losowe. Wykorzystanie kontekstowych metod reprezentacji tekstu (architektur typu transformer) jest lepsze niż bezkontekstowe metody reprezentacji (CBOW, Skipgram), które łącznie są lepsze niż metody niewykorzystujące wiedzy o tekście, co potwierdza hipotezę H3..

Analiza wpływu różnych rodzajów treści tekstowych na skuteczność metod generowania reprezentacji wektorowych tekstu wykazała, że długość i styl wypowiedzi miały znaczący wpływ na jakość predykcji modeli. W krótszych tekstach, takich jak tweety i komentarze, obecnych w zbiorach WikiDetox: Aggression i Unhealthy Conversations, spersonalizowane modele, w tym UserConf, osiągnęły wyższe wyniki F1, szczególnie dzięki efektywnemu wykorzystaniu kontekstowych reprezentacji wektorowych. Modele te były w stanie lepiej wychwytywać subtelne różnice semantyczne w agresywnych i toksycznych komentarzach, co przełożyło się na lepszą identyfikację obraźliwości tekstu. Z kolei w dłuższych wypowiedziach, takich jak te obecne w zbiorach Measuring Hate Speech, Doccano 1 i Doccano 2, metody spersonalizowanego rozpoznawania obraźliwości również radziły sobie dobrze, choć wymagały bardziej zaawansowanego przetwarzania semantycznego oferowanego przez kontekstowe modele językowe, aby prawidłowo zrozumieć kontekst i złożoność tych tekstów. Model UserConf, korzystający z kontekstowych reprezentacji wektorowych, był w stanie uwzględnić indywidualne preferencje użytkowników, co poprawiało precyzję predykcji w tych bardziej złożonych przypadkach. Język oraz rodzaj treści tekstowych (komentarze, tweety, dłuższe wypowiedzi) wpływa na jakość predykcji metod w zadaniu spersonalizowanego rozpoznawania tekstów obraźliwych, co potwierdza hipotezę H4..



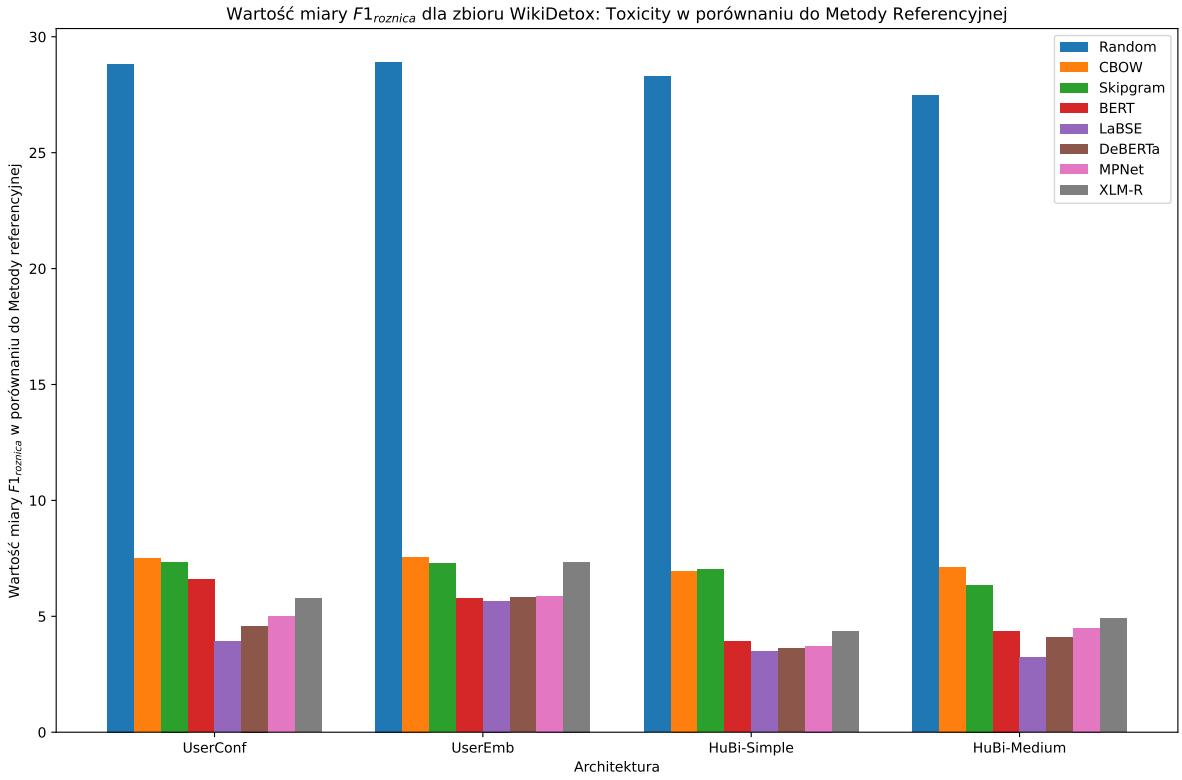
Rysunek 12: Wartość miary $F1_{roznica}$ względem Metody Referencyjnej dla zbioru WikiDetox: Aggression. [Źródło: opracowanie własne]

9.5.2 Analiza wpływu rozmiaru zbioru treningowego w zadaniu predykcji obraźliwości tekstu

Analiza wpływu rozmiaru zbioru treningowego na jakość predykcji została przeprowadzona na zbiorach danych: WikiDetox: Aggression, WikiDetox: Toxicity, Measuring Hate Speech, Unhealthy Conversations, Doccano 1 oraz Doccano 2. Dla każdego z tych zbiorów obserwowano wyraźny wzrost wartości F1 macro wraz ze zwiększaniem liczby foldów w zbiorze uczącym.

W przypadku zbioru WikiDetox: Aggression, różnice w wartościach F1 macro pomiędzy metodą referencyjną a modelami spersonalizowanymi były znaczące, szczególnie dla modelu UserConf, który przy pełnym zbiorze danych przewyższał metodę referencyjną o 9.85 punktów (Rys. 20). Podobne zależności zauważono dla WikiDetox: Toxicity, gdzie model UserConf również osiągnął najwyższe wartości, a jego przewaga nad metodą referencyjną rosła z każdym dodatkowym foldem (Rys. 21).

Zbiór Measuring Hate Speech również potwierdził te tendencje. Modele spersonalizowane, takie jak UserConf i UserEmb, znacznie przewyższały metodę referencyjną, zwłaszcza przy większej liczbie foldów (Rys. 22). Wyniki dla Unhealthy Conversations pokazały podobny trend – modele personalizowane osiągały coraz lepsze wyniki w miarę zwiększania liczby danych treningowych, przy czym przewaga UserConf nad metodą referencyjną była najbardziej widoczna (Rys. 23).

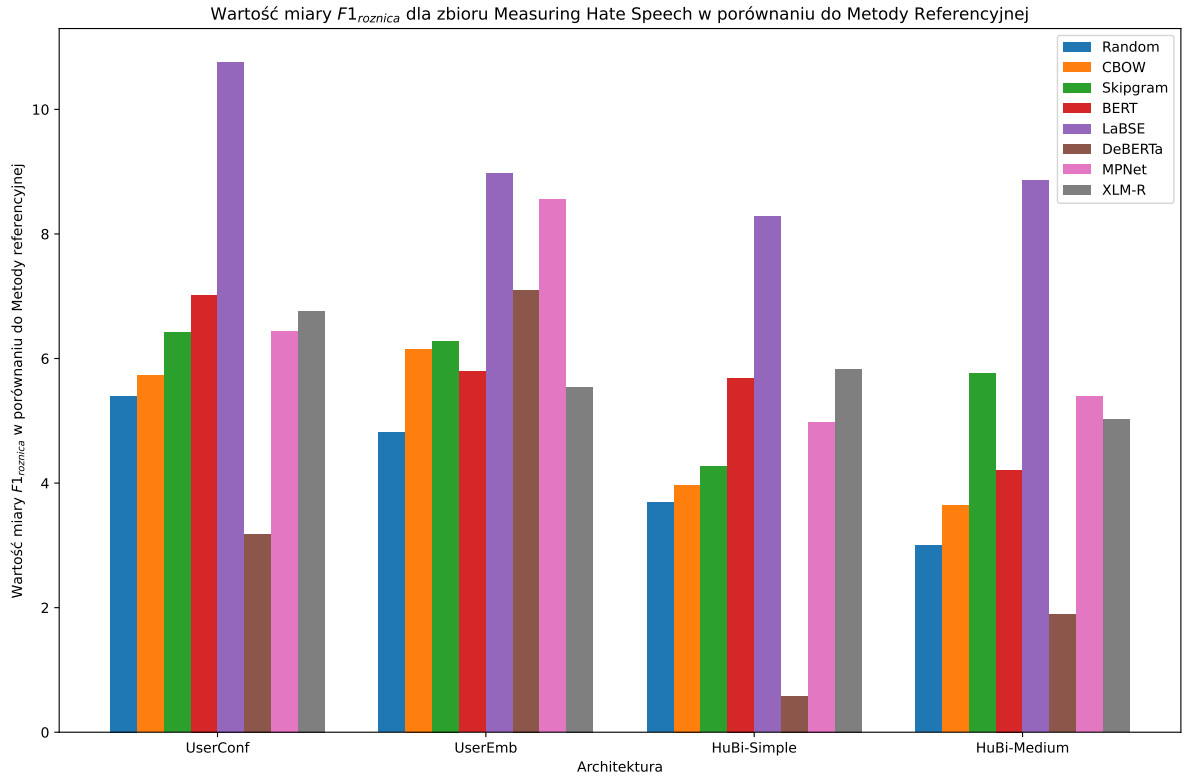


Rysunek 13: Wartość miary $F1_{roznica}$ względem Metody Referencyjnej dla zbioru WikiDetox: Toxicity. [Źródło: opracowanie własne]

W analizie przeprowadzonej dla zbioru Doccano 1 wykazano, że modele spersonalizowane, takie jak UserConf i UserEmb, znacząco poprawiają jakość predykcji obraźliwości tekstu w miarę zwiększania rozmiaru zbioru treningowego (Rys. 24). Już przy niskiej liczbie foldów (np. 1-3 foldy), modele te osiągały lepsze wyniki niż metoda referencyjna. Różnica ta była szczególnie widoczna przy pełnym zestawie foldów (1-8), gdzie modele personalizowane znacznie przewyższały metodę referencyjną w zakresie miary F1 macro. Na przykład, metoda UserConf osiągała wyraźnie wyższe wyniki F1 macro w porównaniu z metodą referencyjną, co pokazuje efektywność personalizacji względem zróżnicowanych preferencji anotatorów i specyficznych wzorców obraźliwości w zbiorze Doccano 1.

Podobnie, w przypadku zbioru Doccano 2, modele UserConf i UserEmb wykazały wyraźną przewagę nad metodą referencyjną (Rys. 25). Modele spersonalizowane systematycznie poprawiały swoje wyniki F1 macro wraz ze wzrostem liczby foldów w zbiorze uczącym. Dla pełnego zestawu foldów (1-8) różnica między metodami personalizowanymi a metodą referencyjną była najbardziej znacząca. Model UserConf wykazywał większą stabilność i przewagę w miarę wzrostu liczby danych, co potwierdzał wysoki wynik F1 macro, przewyższający znacząco wartość uzyskiwaną przez metodę referencyjną, zwłaszcza w przypadku pełnego zbioru danych.

Wszystkie te zbiory, mimo różnorodności treści i specyfiki danych, potwierdzają przewagę metod spersonalizowanych w zadaniu predykcji obraźliwości tekstów. Wzrost liczby foldów



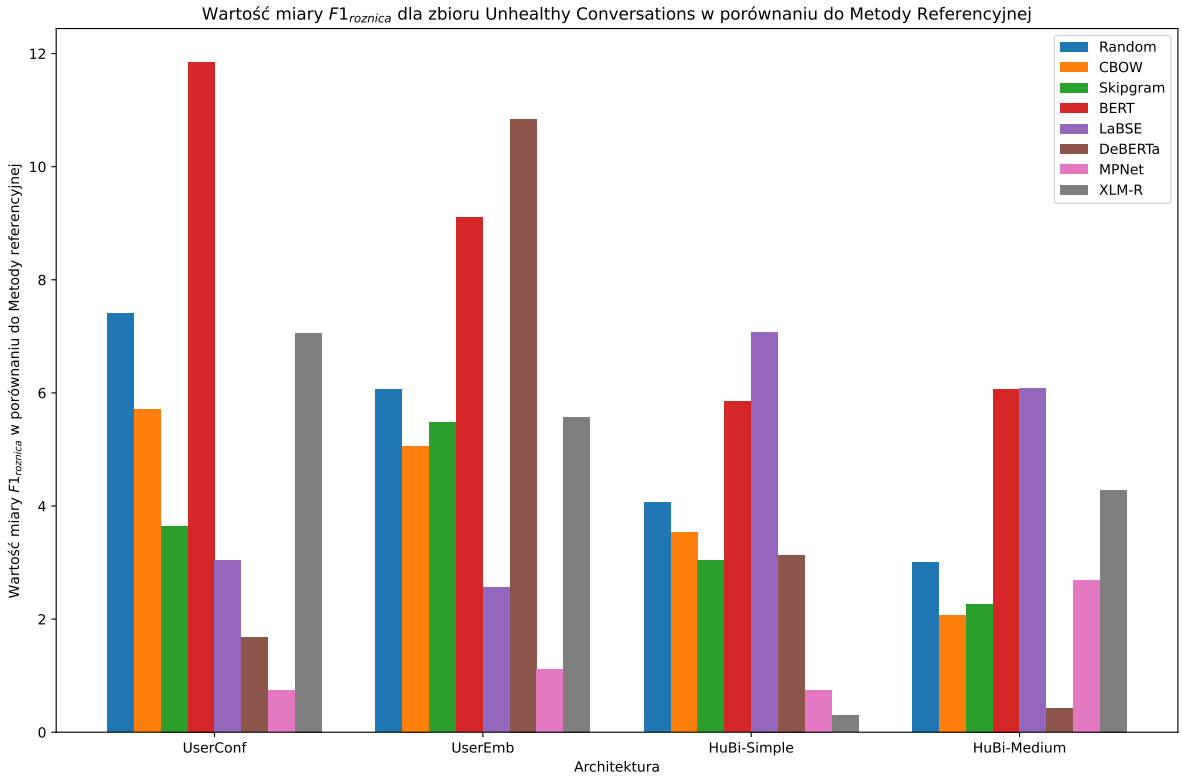
Rysunek 14: Wartość miary $F1_{roznica}$ względem Metody Referencyjnej dla zbioru Measuring Hate Speech. [Źródło: opracowanie własne]

w zbiorze uczącym systematycznie podnosił wartość F1 macro, co wskazuje na rosnącą efektywność modeli wraz ze zwiększaniem liczby danych.

Metoda referencyjna zaczynała od wyniku F1 macro 31.09 przy jednym foldzie, stopniowo poprawiając się do 43.27 przy ośmiu foldach. Były to jednak najniższe wyniki wśród wszystkich testowanych metod, co potwierdza ograniczoną skuteczność podejścia opartego na generalizacji bez uwzględnienia specyficznych cech użytkowników. Z drugiej strony, metoda UserConf osiągnęła lepsze wyniki od samego początku – wartość F1 macro dla jednego folda wyniosła 37.89, a przy pełnej liczbie foldów model ten uzyskał F1 macro na poziomie 53.12, co oznacza różnicę 9.85 punktów względem metody referencyjnej. Takie wyniki wskazują, że model personalizowany, który uwzględnia kontekst użytkownika, jest znacznie bardziej skuteczny.

Model UserEmb również wykazywał znaczące przewagi w stosunku do metody referencyjnej. Przy jednym foldzie wartość F1 macro wyniosła 37.01, a przy ośmiu foldach wzrosła do 51.89, co oznacza różnicę o 8.62 punktów w stosunku do metody referencyjnej. Choć model ten był nieco mniej skuteczny niż UserConf, to jednak jego stabilny wzrost skuteczności wskazuje, że zastosowanie uśrednionych reprezentacji semantycznych również przyczynia się do lepszych wyników.

Modele HuBi-Simple i HuBi-Medium uzyskały niższe wartości F1 macro niż modele UserConf i UserEmb, jednak i tak przewyższały metodę referencyjną. HuBi-Simple zaczynał od wyniku 35.19 przy jednym foldzie i osiągnął 49.89 przy ośmiu foldach, co daje różnicę o

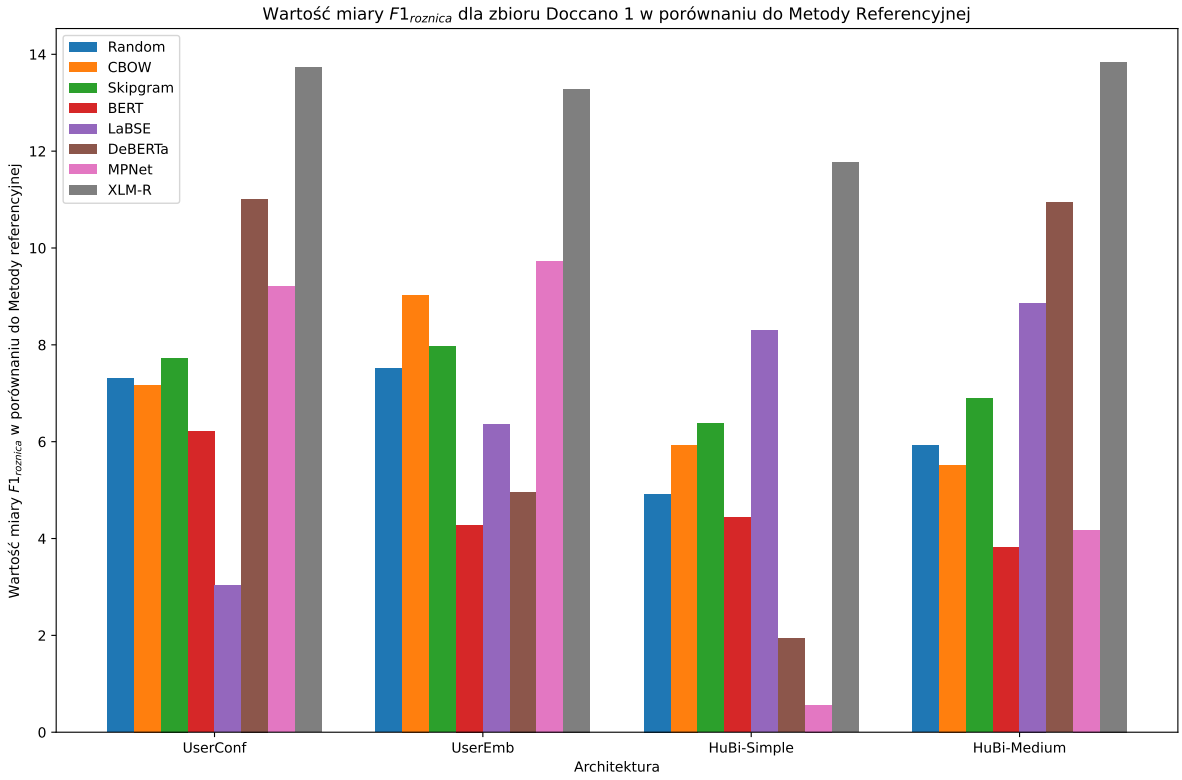


Rysunek 15: Wartość miary $F1_{roznica}$ względem Metody Referencyjnej dla zbioru Unhealthy Conversations. [Źródło: opracowanie własne]

6.62 punktów względem metody referencyjnej. HuBi-Medium, z wynikiem startowym 34.81 dla jednego folda, zakończył na poziomie 48.89 przy ośmiu foldach, co oznacza różnicę o 5.62 punktów w porównaniu do metody referencyjnej. Choć te modele były nieco mniej skuteczne, to również wykazały wyraźną poprawę w miarę wzrostu liczby foldów.

Analiza wyników wyraźnie pokazuje, że zwiększenie rozmiaru zbioru treningowego prowadzi do znaczącej poprawy jakości predykcji modeli spersonalizowanych w porównaniu z metodą referencyjną, która nie uwzględnia indywidualnych preferencji użytkowników. Modele takie jak UserConf i UserEmb wykazały stabilny wzrost skuteczności, a różnice w wartościach F1 macro były wyraźne już od jednego folda, osiągając maksymalną przewagę przy pełnym zbiorze danych. Co ważne, nawet modele HuBi-Simple i HuBi-Medium, które stosują mniej zaawansowane mechanizmy personalizacji, przewyższały metodę referencyjną w każdej iteracji. Wyniki te dowodzą, że modele spersonalizowane, poprzez lepsze dopasowanie do indywidualnych wzorców użytkowników, radzą sobie lepiej nawet na mniejszych zbiorach danych, a ich przewaga rośnie wraz z powiększaniem zbioru treningowego.

Metody spersonalizowanego rozpoznawania tekstów obraźliwych charakteryzują się istotnie lepszą jakością predykcji niż metody zgeneralizowane, niezależnie od rozmiaru zbioru uczącego, co potwierdza hipotezę H5..



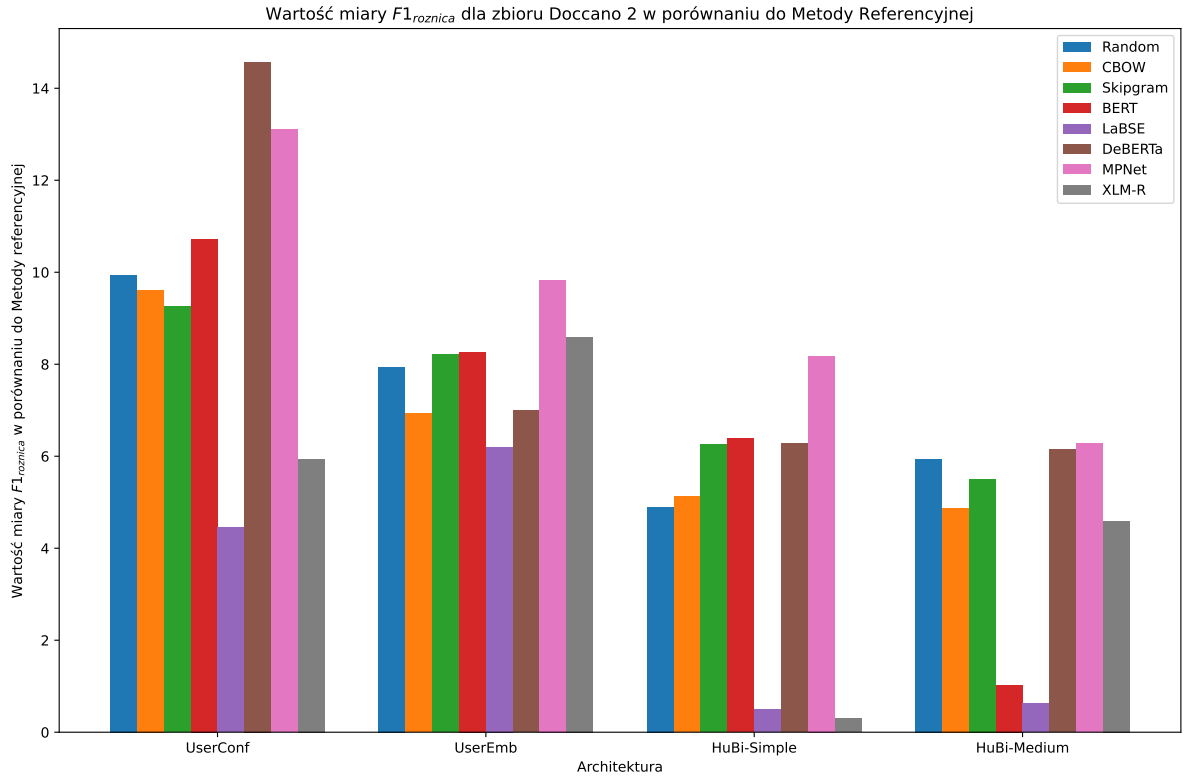
Rysunek 16: Wartość miary $F1_{roznica}$ względem Metody Referencyjnej dla zbioru Doccano 1. [Źródło: opracowanie własne]

9.5.3 Analiza wpływu wykorzystania metody aktywnego uczenia na jakość predykcji obraźliwości tekstu

W kontekście analizy efektywności różnych metod aktywnego uczenia w zadaniu predykcji obraźliwości tekstu, szczególną uwagę zwraca się na pięć wyróżnionych podejść: Random, Kontrowersyjność, VarRatio, Ratio Distance oraz Stranger Count. Każda z tych metod została poddana szczegółowej ocenie na podstawie wyników uzyskanych w kolejnych iteracjach na zbiorach danych. Analiza ta pozwala na wyodrębnienie zarówno mocnych, jak i słabych stron poszczególnych metod w kontekście detekcji agresji i toksyczności w tekstach.

Pierwszym aspektem analizy jest wysoka jakość wnioskowania metody VarRatio, która już od pierwszej iteracji wykazuje wyraźną przewagę nad pozostałymi metodami, osiągając wynik 3.88 względem metody referencyjnej w zbiorze WikiDetox: Aggression (Rys. 26). Ten rezultat potwierdza zdolność tej metody do skutecznego identyfikowania obraźliwych treści, co jest szczególnie istotne w kontekście dynamicznie zmieniających się wzorców komunikacji. Jej najwyższy wynik względem metody referencyjnej równy 8.81 w 13. iteracji, podkreśla, że VarRatio nie tylko osiąga istotnie lepsze wyniki na etapie wczesnym, ale także potrafi utrzymać wysokie wyniki wnioskowania w dłuższym okresie analizy.

Warto również zwrócić uwagę na metodę Kontrowersyjność, która, mimo że początkowo nieznacznie ustępuje VarRatio, z czasem zaczyna zyskiwać na efektywności (Rys. 27). Już w zbiorze WikiDetox: Toxicity, w 4. iteracji, osiąga wyniki porównywalne z czołową metodą, a jej



Rysunek 17: Wartość miary $F1_{roznica}$ względem Metody Referencyjnej dla zbioru Doccano 2. [Źródło: opracowanie własne]

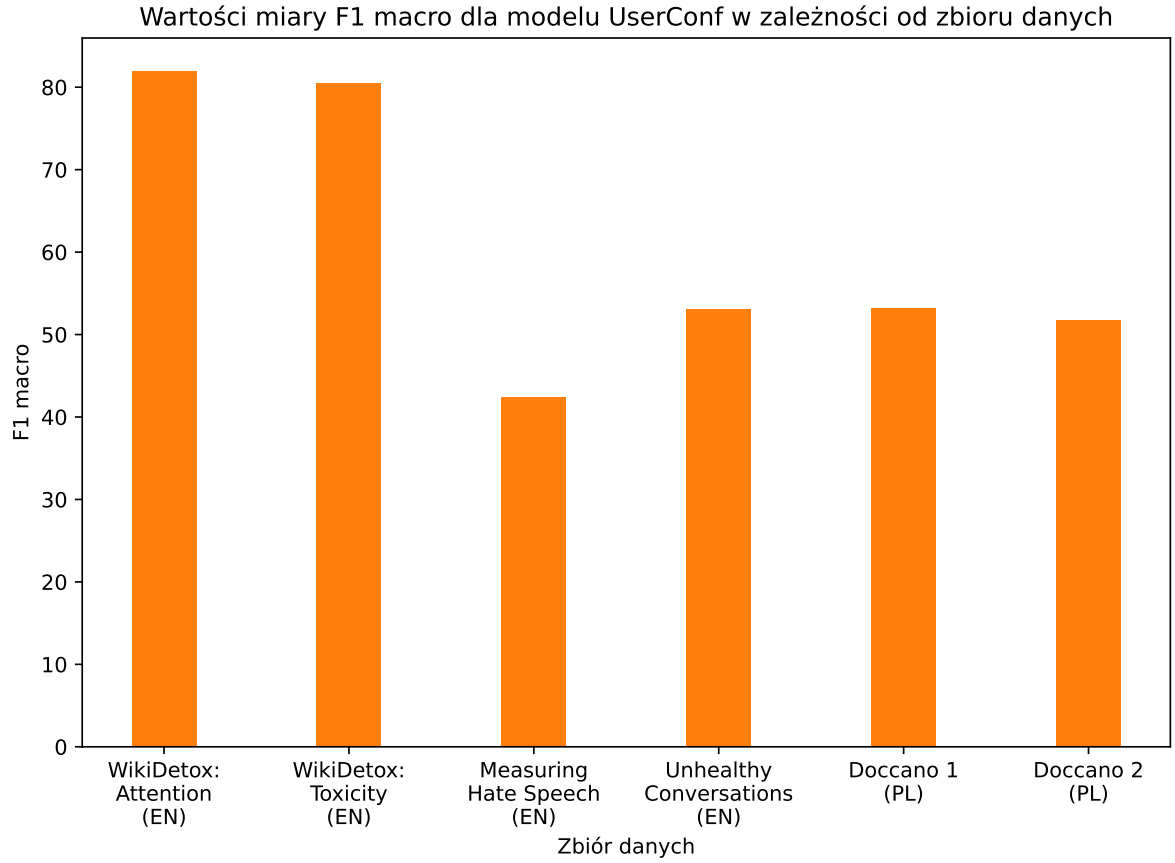
maksymalny wynik 6.85 w 12. iteracji względem metody referencyjnej świadczy o jej ogromnym potencjale. To zjawisko sugeruje, że Kontrowersyjność ma zdolność do efektywnego dostosowywania się do specyfiki analizowanych danych, co czyni ją wartościowym narzędziem w kontekście aktywnego uczenia.

Analizując wyniki uzyskane na zbiorze Measuring Hate Speech, również metoda Kontrowersyjność wykazała się wyższą jakością wnioskowania na tle innych metod, osiągając najwyższy wynik 3.98 w 14. iteracji względem metody referencyjnej, co dodatkowo potwierdza jej stabilność i efektywność (Rys. 28). Z kolei w zbiorze Unhealthy Conversations, metoda ta od pierwszej iteracji osiągnęła najwyższe wyniki, a jej maksymalny wynik 5.10 w 12. iteracji jeszcze bardziej podkreśla jej przewagę nad innymi podejściami (Rys. 29).

Dla zbioru Doccano 1 wyniki metody wykorzystującej kontrowersyjność tekstów okazały się najlepsze. Natomiast metoda VarRatio pozwoliła na osiągnięcie najlepszych wyników względem metody referencyjnej dla 3, 4, 5 i 6 anotacji per użytkownik (Rys. 30).

Dla zbioru Doccano 2 wyniki metody opartej o kontrowersyjność tekstu były istotnie lepsze od wszystkich pozostałych metod dla każdej rozważanej liczby anotacji dla poszczególnych użytkowników (Rys. 31).

Z kolei metody Ratio Distance oraz Stranger Count wykazują najniższe wyniki w zestawieniach. Pomimo że osiągnęły swoje maksymalne wartości, odpowiednio 1.87 i 0.88, są one wyraźnie w tyle za dwoma czołowymi metodami. Ich relatywna stabilność w późniejszych ite-



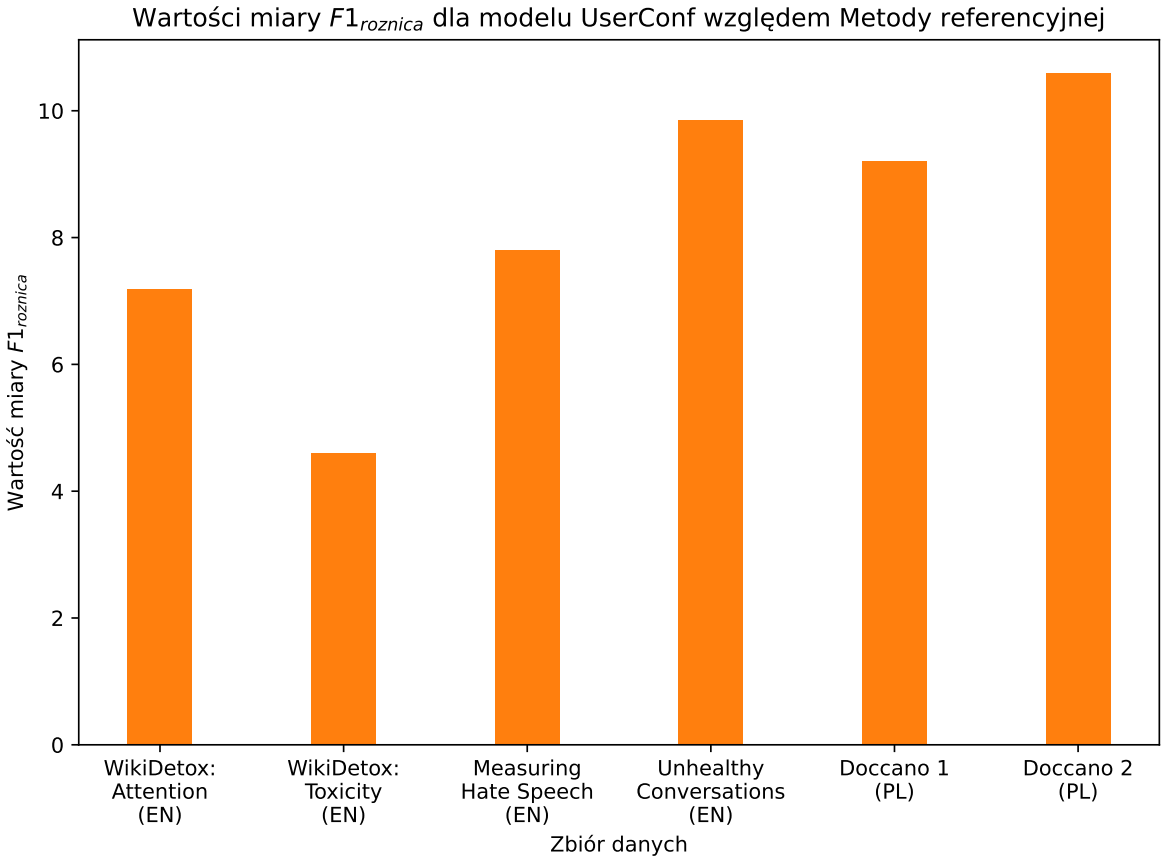
Rysunek 18: Wartość miary F1 macro dla modelu UserConf dla wszystkich wykorzystanych zbiorów danych obejmujących zarówno zbiory anglojęzyczne (EN), jak i polskojęzyczne (PL). [Źródło: opracowanie własne]

racjach sugeruje, że mogą one pełnić funkcję uzupełniającą, ale nie powinny być traktowane jako podstawowe strategie w detekcji obraźliwości tekstu.

Analiza wyników wskazuje na istotne różnice w efektywności metod aktywnego uczenia w zadaniu predykcji obraźliwości tekstu. Spersonalizowane metody, takie jak VarRatio i Kontrowersyjność, wykazują wyraźną przewagę nad innymi strategiami, co podkreśla ich potencjał w kontekście tworzenia bardziej efektywnych modeli predykcyjnych. Spersonalizowane metody aktywnego uczenia, w szczególności metody zaprezentowane w pracy, pozwalają na osiągnięcie zbliżonej jakości predykcji modelu przy istotnie mniejszym zbiorze danych w porównaniu do metody nieuwzględniającej wiedzy na temat indywidualnych anotacji użytkowników, co potwierdza hipotezę H6..

9.5.4 Analiza jakości predykcji generatywnego modelu ogólnego przeznaczenia w zadaniu predykcji obraźliwości tekstu

W badaniach nad generatywnymi modelami językowymi, takimi jak ChatGPT-3.5, kluczową kwestią było porównanie ich efektywności w zadaniu predykcji obraźliwości tekstu w stosunku do bardziej wyspecjalizowanych modeli, takich jak spersonalizowany model UserConf.

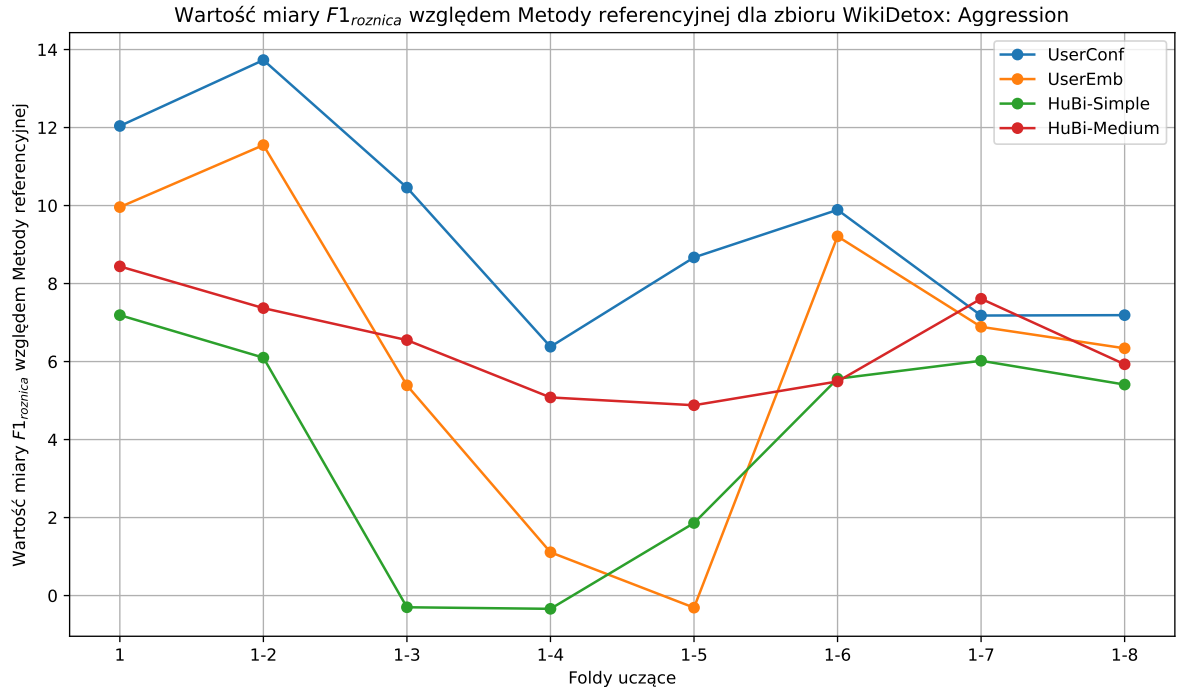


Rysunek 19: Wartość miary $F1_{roznica}$ dla modelu UserConf względem Metody referencyjnej dla wszystkich wykorzystanych zbiorów danych obejmujących zarówno zbiory anglojęzyczne (EN), jak i polskojęzyczne (PL). [Źródło: opracowanie własne]

Badania te miały na celu ocenę, w jakim stopniu generatywny model ogólnego przeznaczenia może konkurować z zaawansowanymi, specjalnie dostosowanymi metodami w zadaniach o wysokim poziomie subiektywności, takich jak rozpoznawanie agresji i toksyczności.

Należy zauważyć, że metoda UserConf opiera się na mierze kontrowersyjności inspirowaną entropią jako alternatywą dla tradycyjnych współczynników takich jak Kappa Fleissa, Kappa Cohena czy Alfa Krippendorffa. Zdecydowano się na to podejście, ponieważ entropia pozwalała na lepsze uchwycenie zróżnicowania anotacji, co jest kluczowe przy ocenie tekstów, których obraźliwość może być subiektywna. Metoda kontrowersyjności premiowała teksty, które charakteryzowały się większą rozbieżnością ocen anotatorów, co umożliwiło skuteczniejsze modelowanie predykcji w zadaniach wymagających spersonalizowanego podejścia

Wyniki uzyskane na zbiorach danych WikiDetox: Aggression oraz Unhealthy Conversations wskazują na istotne różnice w jakości predykcji pomiędzy ChatGPT-3.5 a modelem UserConf, zwłaszcza w kontekście zastosowanych strategii, takich jak zero-shot prompting oraz in-context learning. Miara F1 macro dla modelu UserConf była zauważalnie wyższa niż dla ChatGPT-3.5, co sugeruje, że modele specjalistyczne lepiej radzą sobie z wyzwaniami związanymi z personalizacją w predykcji obraźliwości (Rys. 32).

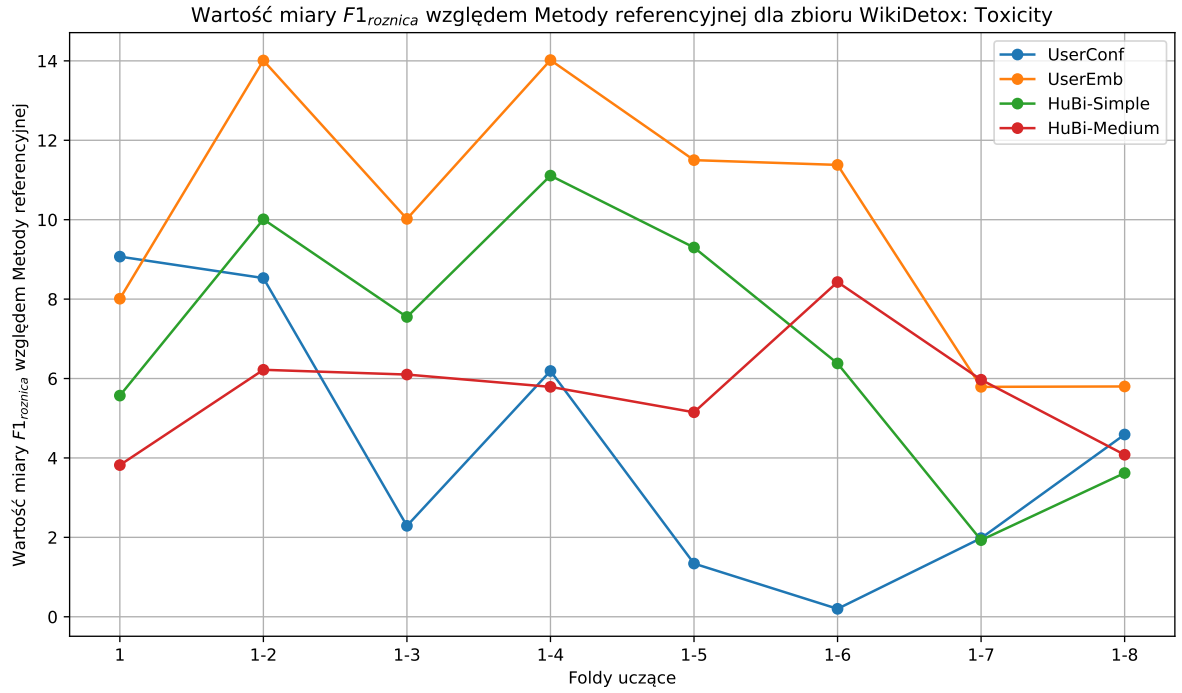


Rysunek 20: Wartość miary $F1_{roznica}$ względem Metody Referencyjnej dla zbioru WikiDetox: Aggression w zależności od liczby foldów znajdujących się w zbiorze uczącym. [Źródło: opracowanie własne]

W przypadku zbioru WikiDetox: Aggression, model UserConf osiągnął wartość miary F1 macro na poziomie 0.78, podczas gdy ChatGPT-3.5 w trybie zero-shot prompting uzyskał wynik jedynie 0.65. Podobna tendencja wystąpiła w kontekście in-context learning, gdzie ChatGPT-3.5 osiągnął wynik 0.68, co wciąż było istotnie niższe od wartości uzyskanej przez model spersonalizowany. Te wyniki wskazują, że choć generatywne modele językowe, takie jak ChatGPT-3.5, są wszechstronne i efektywne w wielu ogólnych zadaniach, to jednak ich zdolność do predykcji obraźliwości w specyficznych i subiektywnych zadaniach jest ograniczona w porównaniu do dedykowanych modeli.

Na zbiorze Unhealthy Conversations sytuacja była podobna. Model UserConf osiągnął miarę F1 macro na poziomie 0.81, co wyraźnie przewyższało wyniki ChatGPT-3.5 zarówno w trybie zero-shot prompting (0.62), jak i in-context learning (0.67). Różnice te można przypisać temu, że generatywne modele, takie jak ChatGPT-3.5, nie są w stanie w pełni uchwycić indywidualnych preferencji użytkowników oraz kontekstu, co stanowi kluczową cechę modelu UserConf, który został zaprojektowany z myślą o spersonalizowanej analizie.

Dodatkową miarą, która pozwoliła na głębszą analizę jakości predykcji, były wartości Loss i Gain dla modelu ChatGPT-3.5. W przypadku zbioru WikiDetox: Aggression, Loss dla ChatGPT-3.5 w trybie zero-shot prompting wyniósł 0.15, natomiast w trybie in-context learning było to 0.10 (Rys. 33). Analogicznie, dla zbioru Unhealthy Conversations, wartości Loss wynosiły odpowiednio 0.19 (zero-shot prompting) i 0.14 (in-context learning). Te stosunkowo wysokie wartości Loss wskazują na wyraźne ograniczenia modelu w uchwyceniu subiektywnych cech predykcji obraźliwości w porównaniu do dedykowanych metod.

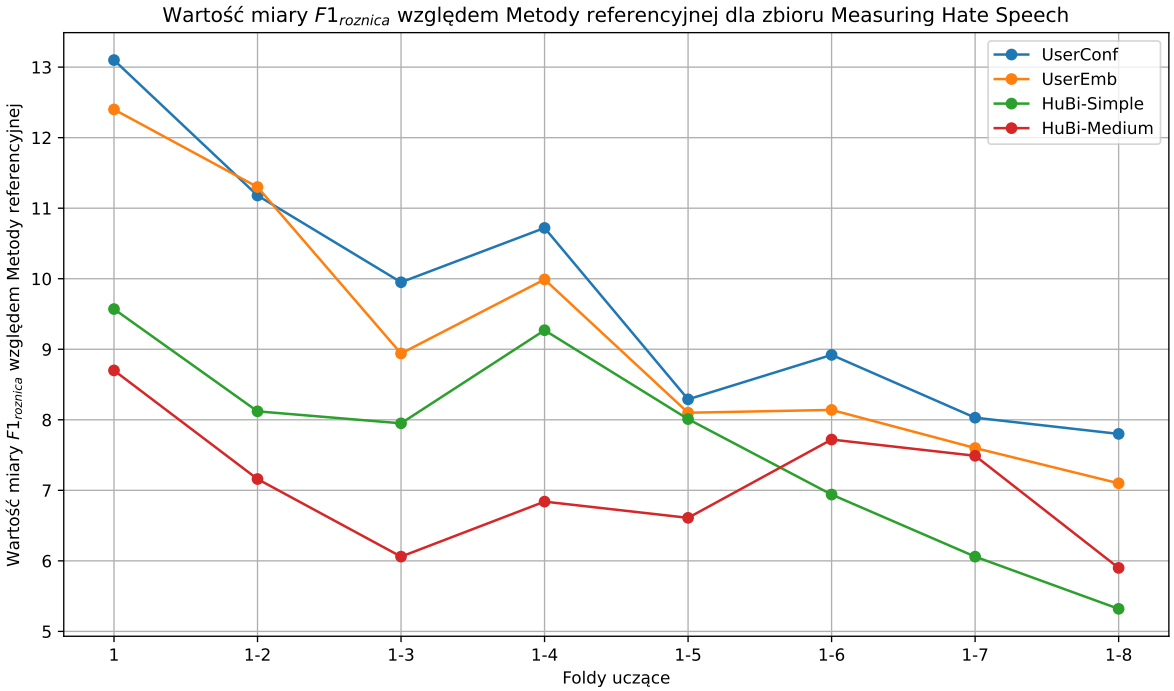


Rysunek 21: Wartość miary $F1_{roznica}$ względem Metody Referencyjnej dla zbioru WikiDetox: Toxicity w zależności od liczby foldów znajdujących się w zbiorze uczącym. [Źródło: opracowanie własne]

Jednocześnie wartości Gain, które reprezentują wzrost efektywności modelu po dodaniu większej liczby kontekstów do procesu uczenia, były ograniczone. Dla WikiDetox: Aggression, ChatGPT-3.5 osiągnął wartości Gain na poziomie +0.03 w trybie zero-shot prompting i +0.06 dla in-context learning, co pokazuje, że złożenie większej liczby przykładów nie przekładało się na drastyczną poprawę wyników. Dla Unhealthy Conversations wartości Gain wynosiły odpowiednio +0.05 i +0.07, co również było niższe niż oczekiwano w kontekście zaawansowanej personalizacji.

Wyniki te prowadzą do istotnych wniosków na temat przyszłości rozwiązań z zakresu przetwarzania języka naturalnego. Chociaż generatywne modele ogólnego przeznaczenia, takie jak ChatGPT-3.5, wykazują imponującą wszechstronność i zdolności adaptacyjne, to w zadaniach wymagających głębokiego zrozumienia kontekstu użytkownika oraz specyficznych wzorców zachowań i obraźliwości, nadal dominują spersonalizowane podejścia. Modele takie jak UserConf, zaprojektowane specjalnie do detekcji agresji i toksyczności, wykazują znacznie lepsze wyniki i mogą stanowić podstawę dla przyszłych rozwiązań w dziedzinie moderacji treści i zarządzania interakcjami w internecie.

Generatywne modele, mimo swoich zalet, wciąż ustępują spersonalizowanym technologiom w kontekście predykcji obraźliwości tekstu. O ile zastosowania metod takich jak *zero-shot prompting* i *in-context learning* w połączeniu z generatywnym dużym modelem językowym stanowią nowatorskie podejścia, to ich skuteczność w zadaniach wymagających głębszej analizy oraz personalizacji pozostaje ograniczona. Wyniki te wyraźnie wskazują na potrzebę



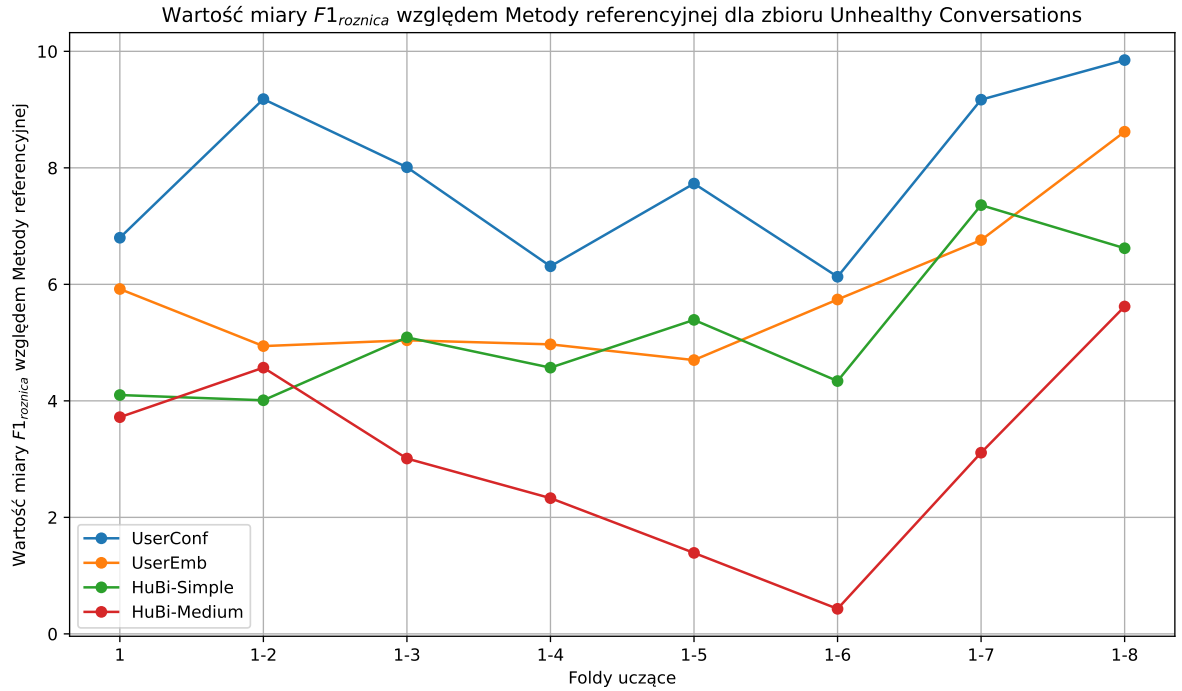
Rysunek 22: Wartość miary $F1_{roznica}$ względem Metody Referencyjnej dla zbioru Measuring Hate Speech w zależności od liczby foldów znajdujących się w zbiorze uczącym. [Źródło: opracowanie własne]

dalszych prac nad integracją personalizacji w modelach generatywnych, co mogłoby znacząco poprawić ich efektywność w bardziej wymagających zadaniach.

W zadaniu spersonalizowanego oraz uogólnionego rozpoznawania treści obraźliwych generatywne modele ogólnego przeznaczenia (np. ChatGPT-3.5) mogą mieć mniejszą skuteczność niż modele dedykowane nauczone na danym zbiorze danych, co potwierdza hipotezę H7..

9.6 DYSKUSJA

Rozdział Dyskusja podsumowuje najważniejsze obserwacje oraz omawia istotne wnioski płynące z zastosowanych w pracy metod. Zawarto w nim omówienie wpływu zastosowanych technik na jakość predykcji obraźliwości tekstu, szczególnie w kontekście generowania reprezentacji wektorowej oraz wyboru modeli neuronowych. Podkreślono również znaczenie rozmiaru zbioru treningowego oraz metod aktywnego uczenia, które pozwoliły na efektywne usprawnienie procesu predykcji. Wreszcie, szczególną uwagę poświęcono generatywnym modelom, takim jak ChatGPT-3.5, oceniając ich skuteczność w porównaniu z modelami dedykowanymi, w tym modelami spersonalizowanymi, co pozwoliło na sformułowanie wniosków dotyczących przyszłego rozwoju technologii i optymalizacji ich wykorzystania w moderacji treści.



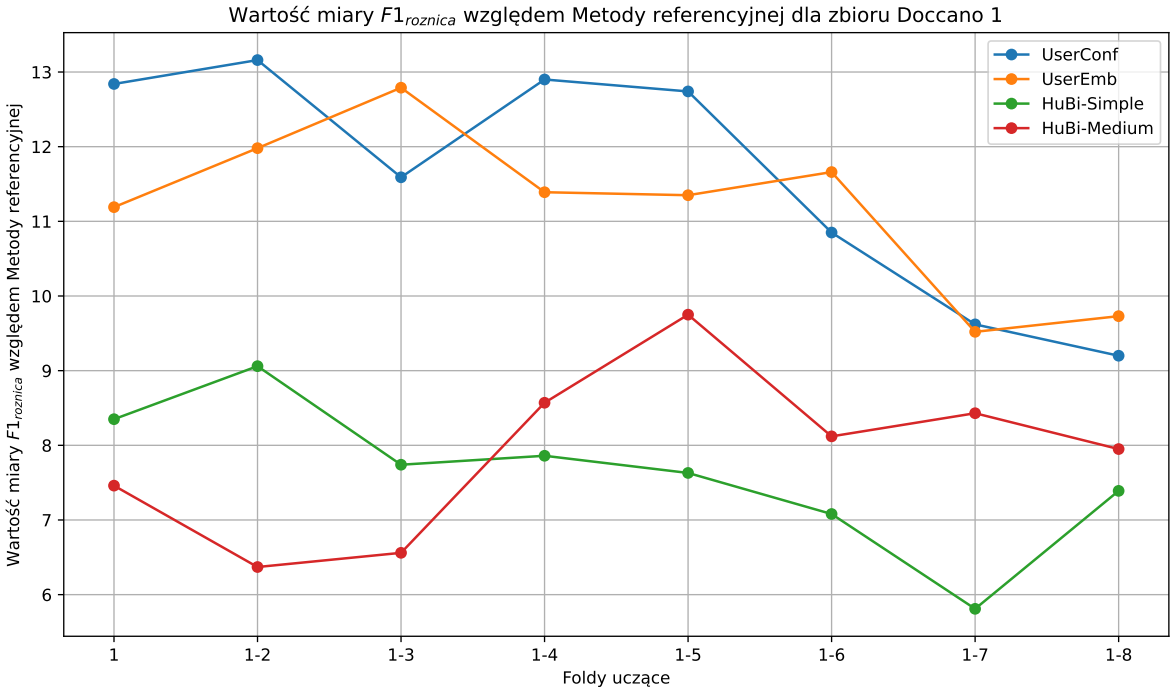
Rysunek 23: Wartość miary $F1_{roznica}$ względem Metody Referencyjnej dla zbioru Unhealthy Conversations w zależności od liczby foldów znajdujących się w zbiorze uczącym. [Źródło: opracowanie własne]

9.6.1 Wpływ metody generowania reprezentacji wektorowej tekstu oraz modelu głębokiej sieci neuronowej służącej do predykcji obraźliwości tekstu

Na podstawie wyników i analizy badań nad metodami można stwierdzić, że kluczową rolę w tym scenariuszu odegrały różnice między reprezentacjami kontekstowymi i bezkontekstowymi. Modele oparte na architekturach typu Transformer, takich jak BERT, DeBERTa, czy MPNet, wykazały istotną przewagę nad klasycznymi, bezkontekstowymi metodami generowania reprezentacji, takimi jak CBOW i Skipgram. Reprezentacje kontekstowe, umożliwiając pełniejsze zrozumienie złożonych relacji między słowami w zdaniu, okazały się bardziej skuteczne w zadaniach wymagających głębszego rozumienia treści.

Modele transformacyjne, dzięki mechanizmowi uwagi, pozwalają na uchwycenie kontekstu zarówno na poziomie lokalnym, jak i globalnym, co prowadzi do wyższej skuteczności w identyfikacji treści obraźliwych. Przykładowo, model MPNet w połączeniu z metodą UserConf uzyskał najwyższe wyniki w zbiorze WikiDetox: Aggression, wskazując na wyjątkową zdolność do wykrywania agresji w komentarzach. Podobnie, DeBERTa osiągnęła znakomite rezultaty, co podkreśla efektywność zaawansowanych architektur typu Transformer.

Z drugiej strony, bezkontekstowe metody generowania reprezentacji tekstu, choć bardziej uproszczone, znacznie gorzej radziły sobie z uchwyceniem subtelnych niuansów językowych, co przełożyło się na niższe wyniki. Modele takie jak CBOW i Skipgram, opierające się jedynie na sąsiadujących słowach, okazały się mniej trafne w rozpoznawaniu obraźliwych treści, szczególnie w złożonych zbiorach danych. Analiza wykazała, że bogatsze reprezentacje kon-



Rysunek 24: Wartość miary $F1_{roznica}$ względem Metody Referencyjnej dla zbioru Doccano 1 w zależności od liczby foldów znajdujących się w zbiorze uczącym. [Źródło: opracowanie własne]

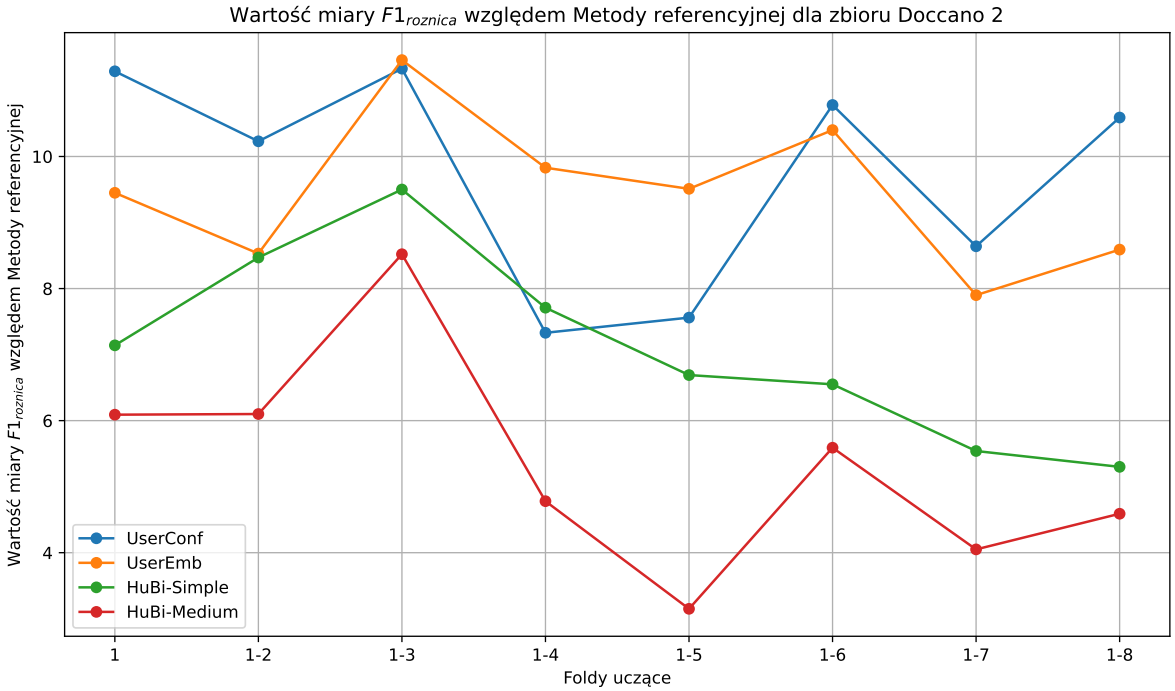
tekstowe są niezbędne do osiągnięcia wysokiej skuteczności w zadaniach moderacji treści online.

Zastosowanie zaawansowanych metod generowania reprezentacji wektorowej tekstu znacząco wpływa na poprawę jakości predykcji w kontekście identyfikacji obraźliwości, a modele transformacyjne są nieodzownym narzędziem w złożonych analizach treści.

9.6.2 Wpływ rozmiaru zbioru treningowego w zadaniu predykcji obraźliwości tekstu

W ramach analizy wpływu rozmiaru zbioru treningowego na predykcję obraźliwości tekstu wykazano, że zwiększenie liczby dostępnych danych treningowych ma istotny wpływ na poprawę jakości predykcji, szczególnie w przypadku spersonalizowanych modeli. Modele takie jak UserConf i UserEmb wykazały systematyczny wzrost skuteczności wraz ze wzrostem liczby foldów, co potwierdza znaczenie większych zbiorów danych w tego typu zadaniach. W szczególności w zbiorach takich jak WikiDetox: Aggression czy Unhealthy Conversations różnice pomiędzy metodami spersonalizowanymi a referencyjnymi były wyraźne już od pierwszych etapów treningu.

Modele spersonalizowane, nawet przy ograniczonej liczbie foldów, przewyższały metody referencyjne, co pokazuje, że indywidualne podejście do predykcji obraźliwości ma przewagę nad klasycznymi metodami generalizacyjnymi. Co więcej, zastosowanie bardziej zaawansowanych modeli, takich jak UserConf, pozwalało na uzyskanie jeszcze lepszych wyników przy



Rysunek 25: Wartość miary $F1_{roznica}$ względem Metody Referencyjnej dla zbioru Doccano 2 w zależności od liczby foldów znajdujących się w zbiorze uczącym. [Źródło: opracowanie własne]

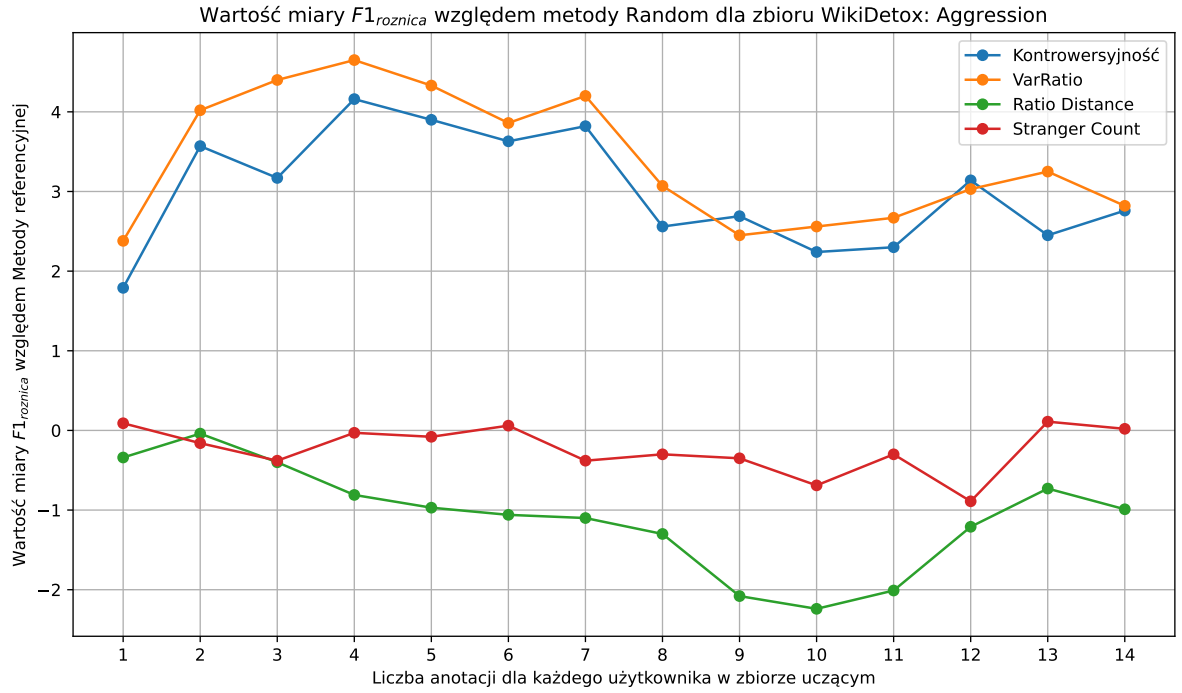
pełnej liczbie foldów, co pokazuje, że modele te dobrze radzą sobie z bardziej złożonymi strukturami danych.

Rozmiar zbioru treningowego miał również zróżnicowany wpływ w zależności od charakterystyki danych. W złożonych zbiorach, takich jak Measuring Hate Speech czy Doccano 1, zwiększenie ilości danych treningowych prowadziło do znaczących wzrostów miary F1 macro, co wskazuje, że większe zbiory są kluczowe dla osiągnięcia stabilnych i precyzyjnych wyników w trudniejszych zadaniach związanych z identyfikacją treści obraźliwych.

9.6.3 Wpływ wykorzystania metody aktywnego uczenia na jakość predykcji obraźliwości tekstu

Analiza wpływu wykorzystania metod aktywnego uczenia na jakość predykcji obraźliwości tekstu wykazała istotną przewagę strategii personalizowanych nad losowym doborem próbek. Kluczowe wnioski wynikają z wyższej skuteczności metod takich jak VarRatio i Kontrowersyjność, które od pierwszych iteracji przyniosły znaczące korzyści w porównaniu do prostych metod losowych. VarRatio, uzyskując stabilnie wysokie wyniki, zwłaszcza w późniejszych iteracjach, okazała się jedną z najbardziej efektywnych metod, potwierdzając, że odpowiedni dobór przykładów do anotacji ma kluczowe znaczenie dla podniesienia jakości predykcji.

Co istotne, wraz ze zwiększaniem liczby anotacji przypisanych każdemu użytkownikowi ze zbioru *train*, stabilność przyrostu miary F1 była zauważalna. Modele spersonalizowane, takie jak UserConf, wykazały wyraźny wzrost jakości predykcji przy stosunkowo niewielkiej



Rysunek 26: Wartość miary $F1_{roznica}$ względem Metody Referencyjnej dla zbioru WikiDetox: Aggression w zależności od wybranej metody aktywnego uczenia oraz liczby anotacji dla każdego użytkownika znajdujących się w zbiorze uczącym. [Źródło: opracowanie własne]

liczbie anotacji, co sugeruje, że metoda aktywnego uczenia pozwala na znaczną optymalizację procesu treningowego, minimalizując potrzebę dużych zbiorów danych.

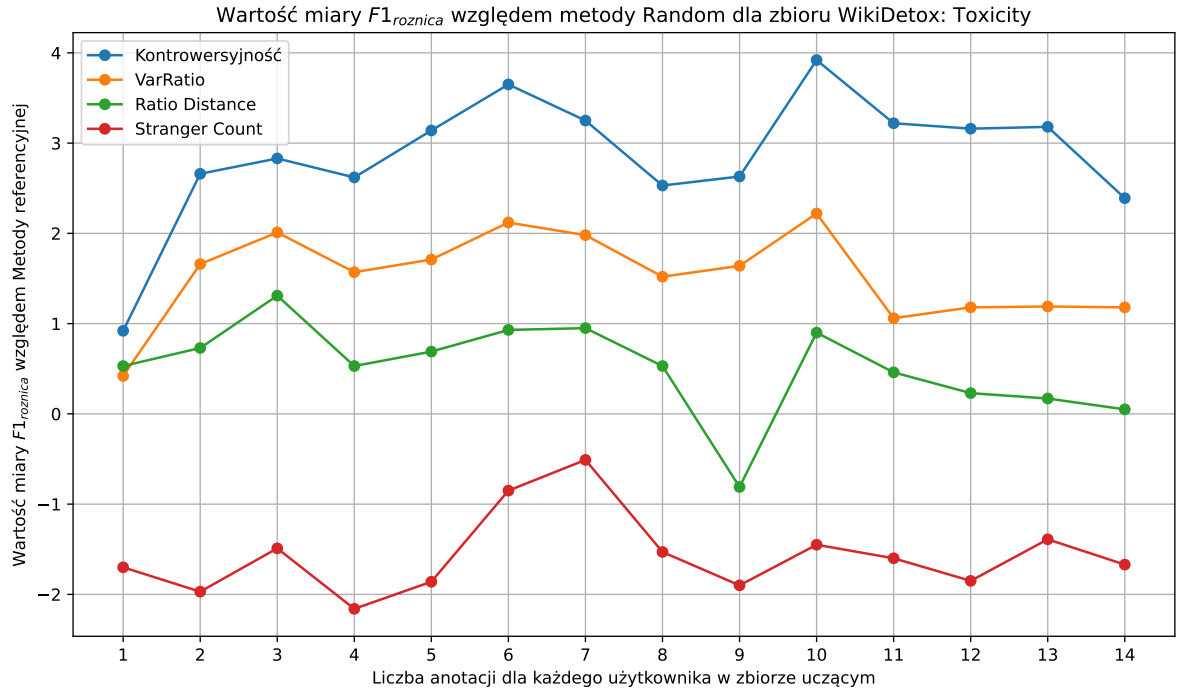
Jednakże nie wszystkie metody aktywnego uczenia sprawdziły się równie dobrze. Metody takie jak Ratio Distance i Stranger Count, mimo pewnych pozytywnych efektów, nie były w stanie osiągnąć porównywalnych wyników, co może sugerować, że ich zastosowanie ma sens jedynie w specyficznych scenariuszach lub w połączeniu z bardziej zaawansowanymi podejściami.

Aktywne uczenie, szczególnie w kontekście predykcji obraźliwych treści, nie tylko znacząco poprawia jakość predykcji, ale również optymalizuje proces anotacji, co prowadzi do efektywniejszego wykorzystania danych treningowych.

9.6.4 Wpływ jakości predykcji generatywnego modelu ogólnego przeznaczenia w zadaniu predykcji obraźliwości tekstu

Wyniki badań jednoznacznie wskazują, że mimo zaawansowania technologii LLM, modele takie jak ChatGPT-3.5 ustępują pod względem skuteczności w specyficznych zadaniach, takich jak predykcja obraźliwości, szczególnie w kontekście personalizacji.

Generatywny model ChatGPT-3.5, choć wszechstronny i efektywny w wielu ogólnych zadaniach, wykazuje niższe wyniki w predykcji obraźliwości tekstu w porównaniu do modeli specjalistycznych. Przykładowo, w zadaniach takich jak WikiDetox: Aggression, ChatGPT-3.5



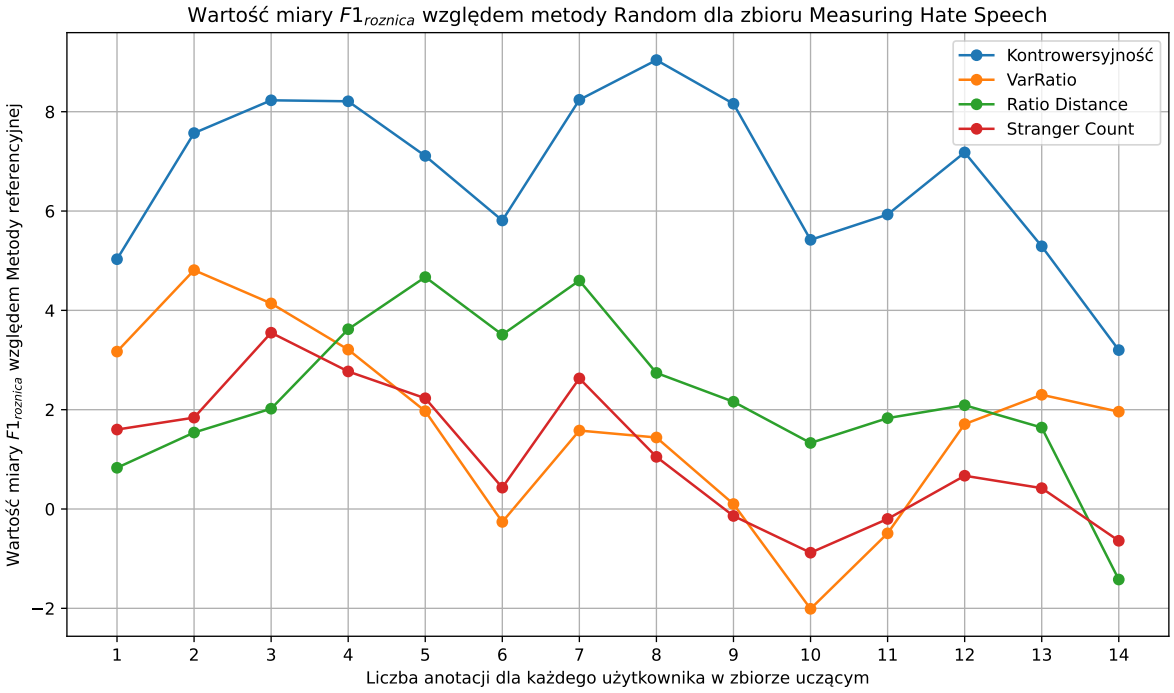
Rysunek 27: Wartość miary $F1_{roznica}$ względem Metody Referencyjnej dla zbioru WikiDetox: Toxicity w zależności od wybranej metody aktywnego uczenia oraz liczby anotacji dla każdego użytkownika znajdujących się w zbiorze uczącym. [Źródło: opracowanie własne]

osiągnął F1 macro na poziomie 69.10 w trybie zero-shot prompting, podczas gdy spersonalizowany model UserConf uzyskał wynik 81.87. W trybie in-context learning, ChatGPT-3.5 poprawił swoją skuteczność (F1 macro 72.57), jednak nadal nie dorównał wynikowi UserConf. Podobne tendencje zaobserwowano na zbiorze Unhealthy Conversations.

Kluczową zaletą modelu UserConf jest zdolność do personalizacji predykcji poprzez uwzględnienie indywidualnych preferencji użytkowników, co jest szczególnie ważne w zadaniach o dużej subiektywności, takich jak rozpoznawanie obraźliwości tekstu. ChatGPT-3.5, choć potrafi adaptować się do różnych kontekstów, wymaga znacznego dostosowania w postaci dostarczania większej liczby tokenów wejściowych, co wiąże się z istotnym wzrostem kosztów obliczeniowych. Konieczność wzbogacenia danych wejściowych o przykłady anotacji użytkownika w celu poprawy jakości predykcji zwiększa liczbę przetwarzanych tokenów, co prowadzi do wzrostu kosztów generowania odpowiedzi.

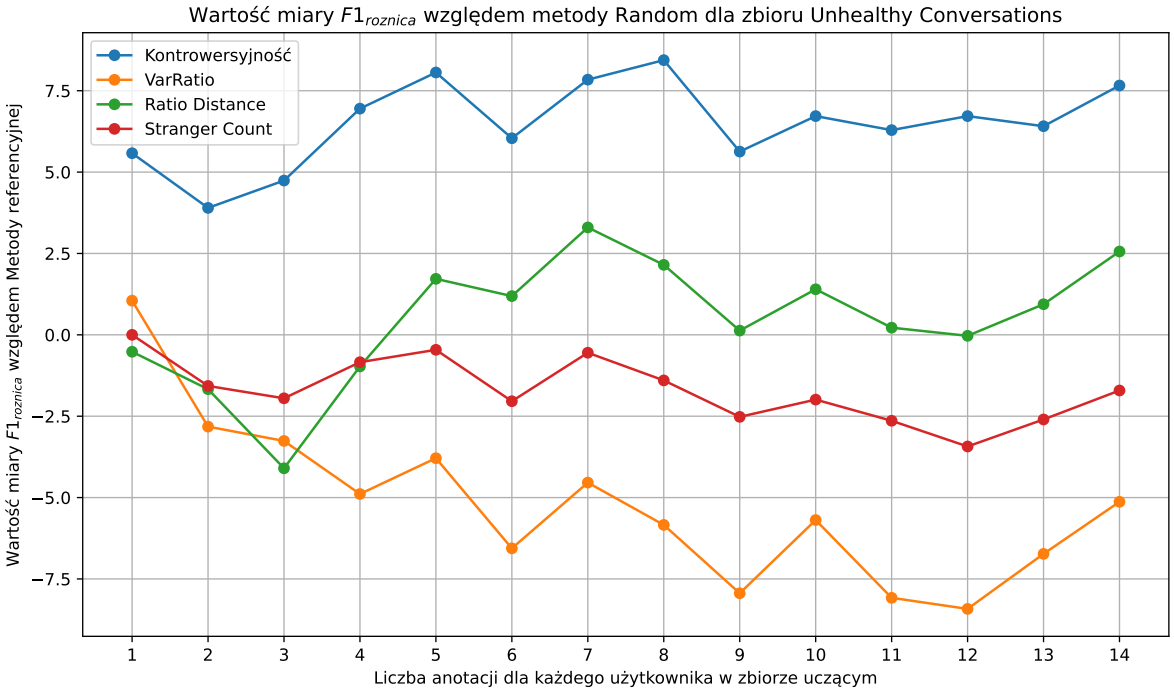
Ponadto, wyniki sugerują, że pomimo lepszej skuteczności in-context learning w ChatGPT-3.5 w stosunku do zero-shot prompting, metoda ta nadal nie dorównuje modelom wyspecjalizowanym. Oznacza to, że w zadaniach wymagających głębszej personalizacji i zaawansowanej analizy semantycznej, takie jak predykcja obraźliwości tekstu, bardziej wyspecjalizowane podejścia mogą okazać się bardziej efektywne, zarówno pod względem jakości predykcji, jak i kosztów obliczeniowych.

Wnioski te podkreślają potrzebę dalszego rozwoju modeli generatywnych w kierunku lepszego dostosowania do specyficznych, spersonalizowanych zadań. Dalsza integracja mechani-

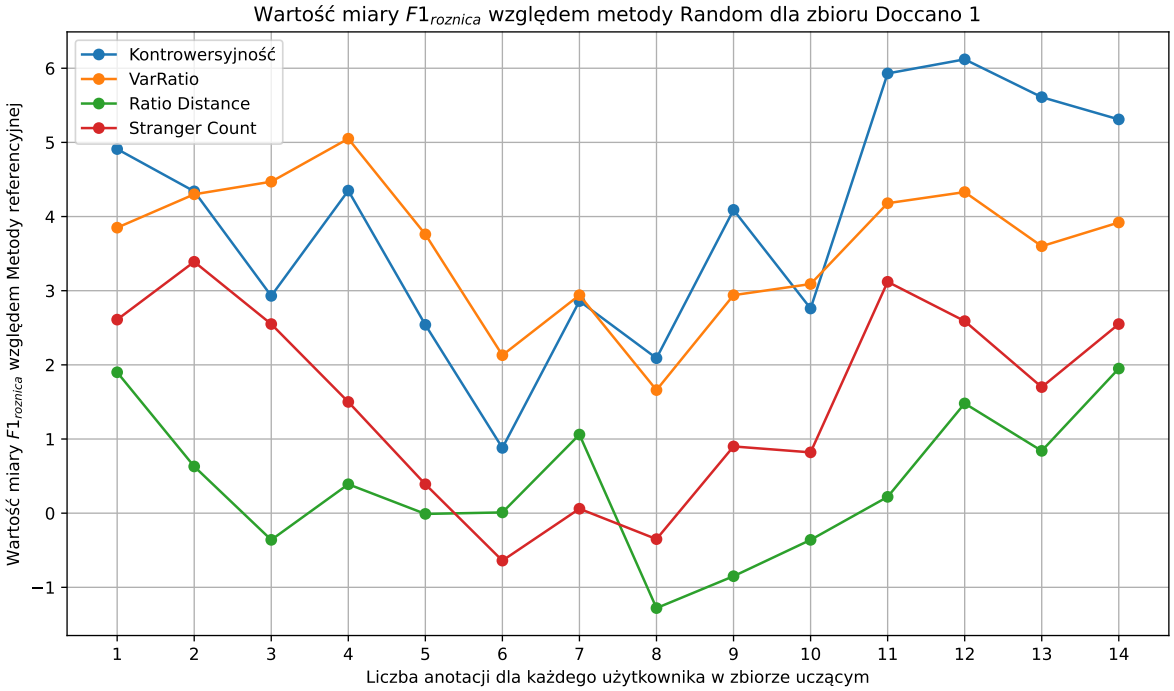


Rysunek 28: Wartość miary $F1_{roznica}$ względem Metody Referencyjnej dla zbioru Measuring Hate Speech w zależności od wybranej metody aktywnego uczenia oraz liczby anotacji dla każdego użytkownika znajdujących się w zbiorze uczącym. [Źródło: opracowanie własne]

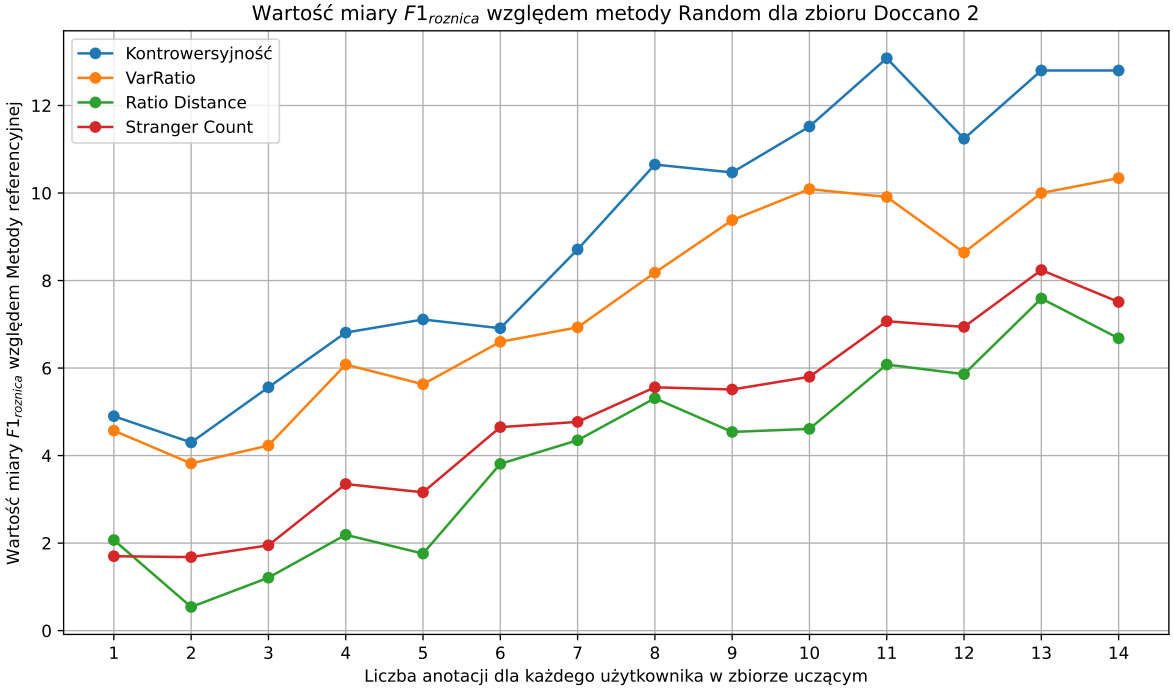
zmów personalizacji mogłaby znacząco poprawić efektywność modeli takich jak ChatGPT-3.5, szczególnie w aplikacjach wymagających indywidualnego podejścia do oceny tekstu.



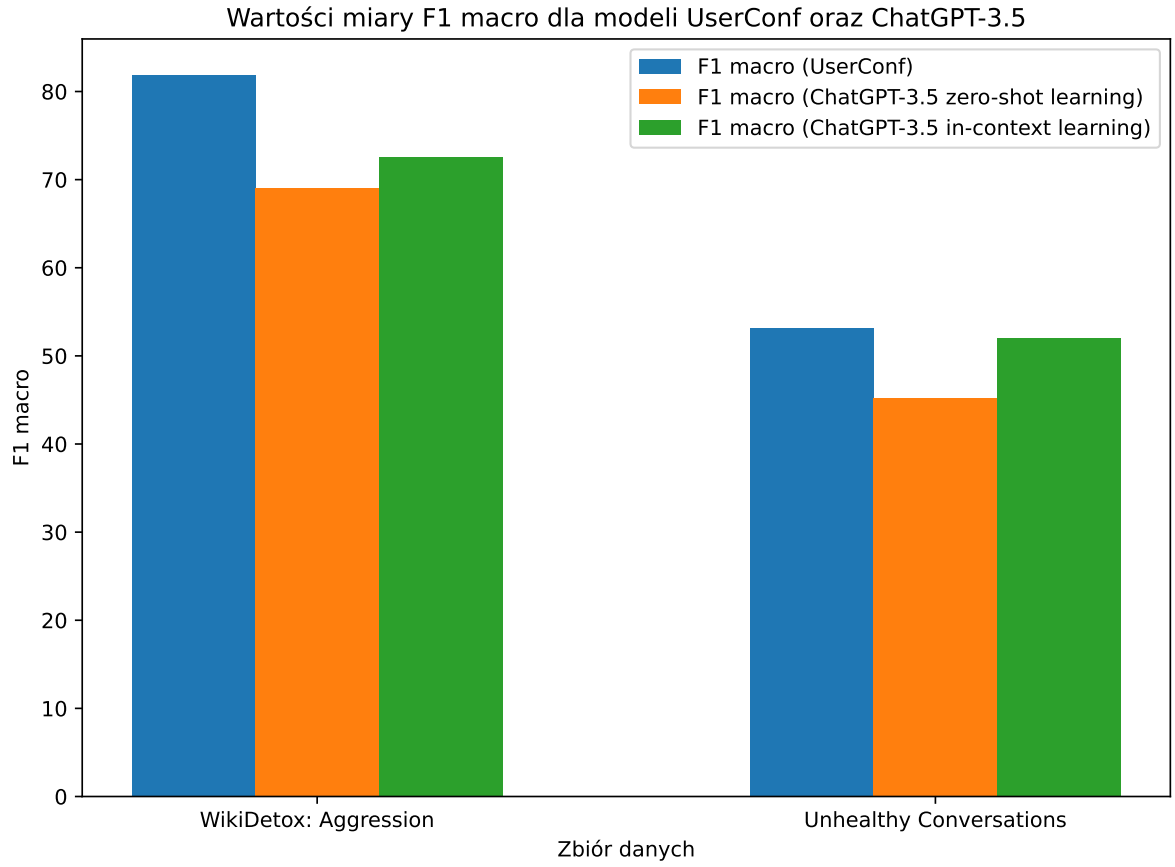
Rysunek 29: Wartość miary $F1_{roznica}$ względem Metody Referencyjnej dla zbioru Unhealthy Conversations w zależności od wybranej metody aktywnego uczenia oraz liczby anotacji dla każdego użytkownika znajdujących się w zbiorze uczącym. [Źródło: opracowanie własne]



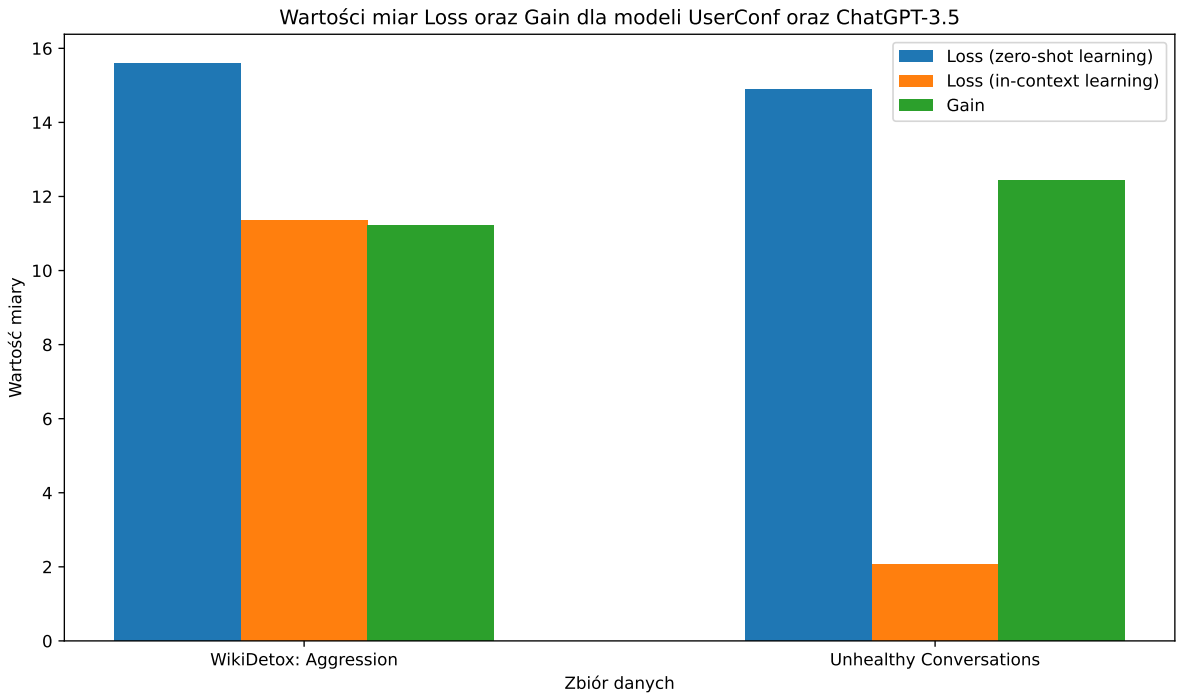
Rysunek 30: Wartość miary $F1_{roznica}$ względem Metody Referencyjnej dla zbioru Doccano 1 w zależności od wybranej metody aktywnego uczenia oraz liczby anotacji dla każdego użytkownika znajdujących się w zbiorze uczącym. [Źródło: opracowanie własne]



Rysunek 31: Wartość miary $F1_{roznica}$ względem Metody Referencyjnej dla zbioru Doccano 2 w zależności od wybranej metody aktywnego uczenia oraz liczby anotacji dla każdego użytkownika znajdujących się w zbiorze uczącym. [Źródło: opracowanie własne]



Rysunek 32: Wartość miary F1 macro dla modelu UserConf oraz ChatGPT-3.5 z użyciem metod *zero-shot prompting* oraz *in-context learning* dla zbiorów WikiDetox Aggression oraz Unhealthy Conversations. [Źródło: opracowanie własne]



Rysunek 33: Wartości miar Loss oraz Gain dla modelu ChatGPT-3.5 z użyciem metod *zero-shot prompting* oraz *in-context learning* dla zbiorów WikiDetox Aggression oraz Unhealthy Conversations. [Źródło: opracowanie własne]

10 PODSUMOWANIE

W niniejszej pracy rozważono zagadnienie spersonalizowanego rozpoznawania tekstów obraźliwych przy użyciu zaawansowanych metod uczenia maszynowego oraz głębokich sieci neuronowych. Celem badań było określenie, w jaki sposób indywidualne preferencje użytkowników mogą wpłynąć na dokładność rozpoznawania obraźliwości treści w tekstach. Badania te obejmowały zastosowanie zarówno klasycznych technik, jak i nowoczesnych modeli generatywnych, takich jak modele typu transformer.

W trakcie przeprowadzonych eksperymentów wykazano, że uwzględnienie reprezentacji użytkownika, jego preferencji oraz kontekstu, w którym odbiera on daną wypowiedź, istotnie poprawia skuteczność predykcji. Przedstawione metody pozwoliły na precyzyjniejsze rozpoznawanie tekstów obraźliwych w porównaniu do tradycyjnych podejść uogólnionych, które nie brały pod uwagę indywidualnych różnic pomiędzy użytkownikami. Zastosowanie metod aktywnego uczenia oraz modeli generatywnych również znacząco wpłynęło na poprawę efektywności klasyfikacji, szczególnie w kontekście mniejszych i bardziej zróżnicowanych zbiorów danych.

Wyniki badań wykazały, że modele spersonalizowane, uwzględniające indywidualne preferencje użytkowników oraz kontekst, w jakim przetwarzają oni informacje, mogą stanowić skuteczne narzędzie w zadaniach moderacji treści online, szczególnie w walce z mową nienawiści, agresją i innymi formami obraźliwych treści.

W ramach pracy przeprowadzono liczne eksperymenty, w których testowano wpływ różnych metod generowania reprezentacji wektorowych tekstu, liczby foldów uczących oraz wykorzystania aktywnego uczenia. W eksperymentach użyto takich metod generowania reprezentacji tekstu, jak Random, CBOW, Skipgram, BERT, DeBERTa, MPNet, oraz XLM-RoBERTa. Wnioski z tych badań wskazują, że zaawansowane metody, takie jak BERT oraz DeBERTa, istotnie przewyższają klasyczne techniki, takie jak CBOW czy Skipgram, zarówno pod względem dokładności predykcji, jak i zdolności do uchwycenia kontekstu tekstu.

Badania dotyczące liczby foldów uczących wykazały, że wraz ze wzrostem liczby foldów znacząco poprawia się jakość modeli predykcyjnych. Metody takie jak UserConf i UserEmb osiągnęły najlepsze rezultaty przy ośmiu foldach, przy czym metoda UserEmb okazała się najskuteczniejsza, osiągając wartość miary F1 macro na poziomie 81.68 na zbiorze WikiDetox: Toxicity. Z kolei metoda referencyjna, pomimo pewnej poprawy przy większej liczbie foldów, pozostawała wyraźnie mniej efektywna od pozostałych metod.

Aktywne uczenie było również istotnym elementem badań. Wykorzystano różne strategie aktywnego uczenia, takie jak metoda VarRatio oraz Ratio Distance, które pozwalały na dynamiczne dostosowanie modelu do nowych danych. Najważniejszym wnioskiem płynącym z tych eksperymentów jest potwierdzenie hipotezy na temat istotnej poprawy jakości predykcji modeli za pomocą spersonalizowanych metod aktywnego uczenia, zwłaszcza w przypadku

ograniczonych zbiorów danych, co czyni je efektywnym narzędziem w kontekście moderacji treści online.

Przeprowadzone eksperymenty udowodniły również, że zastosowanie generatywnych modeli językowych, takich jak ChatGPT-3.5, może być użyteczne w kontekście spersonalizowanego rozpoznawania treści obraźliwych, jednak ich efektywność bywa niższa w porównaniu z wyspecjalizowanymi modelami trenowanymi na specyficznych zbiorach danych.

Podsumowując, praca ta stanowi istotny wkład w rozwój metod spersonalizowanego przetwarzania języka naturalnego, a w szczególności w rozpoznawanie obraźliwości treści tekstowych. Wykorzystanie reprezentacji użytkownika oraz personalizacji pozwala na lepsze dostosowanie systemów moderacji treści do potrzeb użytkowników, co może przyczynić się do bardziej efektywnego eliminowania niepożądanych treści w przestrzeni internetowej.

11 KIERUNKI DALSZYCH BADAŃ

Niniejsza rozprawa otwiera szerokie możliwości kontynuacji badań w zakresie spersonalizowanego rozpoznawania tekstów obraźliwych. Przede wszystkim, istotnym kierunkiem rozwoju jest dalsza optymalizacja metod generowania reprezentacji wektorowych tekstów, z uwzględnieniem rosnącej liczby danych i ich zróżnicowania. Modele takie jak BERT czy DeBERTa dostarczają obiecujących wyników, jednak wciąż pozostaje przestrzeń do badania alternatywnych architektur, które mogą lepiej uchwycić subtelne różnice w treściach tekstowych oraz specyficzne preferencje użytkowników.

W przyszłych badaniach warto uwzględnić zastosowanie podejścia wielomodalnego, które pozwoli na łączenie różnych rodzajów danych, takich jak tekst, obraz, dźwięk czy wideo, w procesie rozpoznawania obraźliwych treści. Integracja danych wielomodalnych umożliwi stworzenie bardziej złożonych i holistycznych modeli rozpoznawania obraźliwości, które uwzględnią nie tylko kontekst tekstowy, ale także dodatkowe wskazówki kontekstualne wynikające z innych mediów. Wielomodalne podejście mogłoby poprawić jakość predykcji w przypadkach, gdzie obraźliwość jest trudna do oceny wyłącznie na podstawie tekstu, na przykład w postach w mediach społecznościowych, które zawierają zarówno tekst, jak i obrazy lub nagrania wideo.

Również zastosowanie technik aktywnego uczenia wymaga dalszej eksploracji. Przeprowadzone eksperymenty dowiodły, że takie metody mogą istotnie poprawiać skuteczność klasyfikacji w kontekście ograniczonych zbiorów danych, jednak badania nad ich efektywnością w sytuacjach, gdy dostępne są bardzo duże bazy danych, pozostają niewystarczająco rozpoznane. Analiza różnych strategii aktywnego uczenia, takich jak VarRatio czy Ratio Distance, może prowadzić do opracowania jeszcze bardziej precyzyjnych metod selekcji próbek, co z kolei pozwoli na dalszą poprawę wyników w kontekście personalizacji.

Kolejnym ważnym aspektem, który wymaga dalszych badań, jest zastosowanie generatywnych modeli językowych, takich jak ChatGPT-3.5, w zadaniach spersonalizowanego rozpoznawania obraźliwości tekstów. Chociaż pierwsze wyniki eksperymentów sugerują, że te modele mogą być efektywne, ich skuteczność w porównaniu z bardziej dedykowanymi podejściami wciąż pozostaje kwestią otwartą. W szczególności warto zbadać możliwości adaptacji tych modeli do specyficznych kontekstów i indywidualnych preferencji użytkowników.

W przyszłych badaniach warto również rozważyć zastosowanie techniki *Retrieval Augmented Generation* (RAG) (Lewis i in., 2020), która łączy w sobie generatywne modele językowe z mechanizmami wyszukiwania informacji w zewnętrznych bazach danych. RAG może być szczególnie przydatna w kontekście spersonalizowanego rozpoznawania tekstów obraźliwych, ponieważ pozwala modelowi nie tylko generować odpowiedzi na podstawie wbudowanej wiedzy, ale także odnosić się do aktualnych i dynamicznych źródeł informacji.

Technika RAG mogłaby zostać wykorzystana do wzbogacenia modeli o dodatkowe informacje kontekstowe, które mogłyby okazać się użyteczne w scenariuszach, w których konieczne jest odniesienie się do specyficznych przypadków użycia lub przykładów – model mógłby, na przykład wyszukiwać podobne sytuacje z przeszłości i na ich podstawie generować odpowiedź, dostosowaną do indywidualnych potrzeb użytkownika.

Zastosowanie RAG w kontekście multimodalności oraz zaawansowanej inżynierii podpowiedzi mogłoby jeszcze bardziej zwiększyć efektywność rozpoznawania obraźliwych treści. Dzięki możliwości wyszukiwania zewnętrznych danych multimodalnych (np. obrazów, wideo), RAG mogłoby umożliwić bardziej zaawansowaną analizę treści, integrując różne formy przekazu i pozwalając na lepsze zrozumienie obraźliwości w szerszym kontekście.

Włączenie techniki RAG do systemów moderacji treści pozwoliłoby na bardziej dynamiczne, aktualne i spersonalizowane rozpoznawanie obraźliwości, co mogłoby znaleźć szerokie zastosowanie w platformach społecznościowych i innych środowiskach online.

Interesującym kierunkiem dalszych badań może być również analiza wpływu różnorodnych kontekstów użytkownika na jakość predykcji. Wprowadzenie bardziej złożonych profili użytkowników, które uwzględniają szersze spektrum informacji, takich jak emocje czy intencje, może prowadzić do jeszcze bardziej precyzyjnych wyników. Badania nad personalizacją w czasie rzeczywistym, z możliwością dynamicznego dostosowywania modeli do zmieniających się preferencji użytkowników, stanowią obiecujący obszar dalszych eksploracji.

BIBLIOGRAFIA

Akhtar, Sohail, Valerio Basile i Viviana Patti (2020). „Modeling Annotator Perspective and Polarized Opinions to Improve Hate Speech Detection". W: *Proceedings of the Eighth AAAI Conference on Human Computation and Crowdsourcing*, s. 151–154. URL: https://www.aaai.org/ojs/index.php/HCOMP/article/view/7473/7260.

Al Kuwatly, Hala, Maximilian Wich i Georg Groh (list. 2020). „Identifying and Measuring Annotator Bias Based on Annotators' Demographic Characteristics". W: *Proceedings of the Fourth Workshop on Online Abuse and Harms*. Online: Association for Computational Linguistics, s. 184–190. DOI: 10.18653/v1/2020.alw-1.21. URL: https://www.aclweb.org/anthology/2020.alw-1.21.

Alrehili, A. (2019). „Automatic Hate Speech Detection on Social Media: A Brief Survey". W: *2019 IEEE/ACS 16th International Conference on Computer Systems and Applications (AICCSA)*, s. 1–6. DOI: 10.1109/AICCSA47632.2019.9035228.

Aroyo, Lora i Chris Welty (2013). *Harnessing disagreement in crowdsourcing a relation extraction gold standard*. Spraw. tech. Technical Report.

Augustyniak, Łukasz, Tomasz Kajdanowicz i Przemysław Kazienko (2019). „Aspect Detection using Word and Char Embeddings with (Bi) LSTM and CRF". W: *2019 IEEE Second International Conference on Artificial Intelligence and Knowledge Engineering (AIKE)*. IEEE, s. 43–50.

Bartlett, Maurice Stevenson (1937). „Properties of sufficiency and statistical tests". W: *Proceedings of the Royal Society of London. Series A-Mathematical and Physical Sciences* 160.901, s. 268–282.

Ben-David, Anat i Ariadna Matamoros Fernández (2016). „Hate speech and covert discrimination on social media: Monitoring the Facebook pages of extreme-right political parties in Spain". W: *International Journal of Communication* 10, s. 27.

Bernard, Jürgen, Matthias Zeppelzauer, Markus Lehmann, Martin Müller i Michael Sedlmair (2018). „Towards User-Centered Active Learning Algorithms". W: *Computer Graphics Forum*. T. 37. Wiley Online Library, s. 121–132.

Bielaniewicz, Julita, Kamil Kanclerz, Piotr Miłkowski, Marcin Gruza, Konrad Karanowski, Przemysław Kazienko i Jan Kocoń (2022). „Deep-SHEEP: Sense of Humor Extraction from Embeddings in the Personalized Context". W: *2022 IEEE International Conference on Data Mining Workshops (ICDMW)*, s. 967–974. DOI: 10.1109/ICDMW58026.2022.00125.

Bielaniewicz, Julita i Przemysław Kazienko (2023). „From Generalized Laughter to Personalized Chuckles: Unleashing the Power of Data Fusion in Subjective Hu-

mor Detection". W: *2023 IEEE International Conference on Data Mining Workshops (ICDMW)*. IEEE, s. 716–725.

Binns, Reuben, Michael Veale, Max Van Kleek i Nigel Shadbolt (2017). „Like Trainer, Like Bot? Inheritance of Bias in Algorithmic Content Moderation". W: *Social Informatics*, 405–415. ISSN: 1611-3349. DOI: 10.1007/978-3-319-67256-4_32. URL: http://dx.doi.org/10.1007/978-3-319-67256-4_32.

Blodgett, Su Lin i Brendan T. O'Connor (2017). „Racial Disparity in Natural Language Processing: A Case Study of Social Media African-American English". W: *ArXiv* abs/1707.00061.

Bloehdorn, Stephan i Andreas Hotho (2004). „Text classification by boosting weak learners based on terms and concepts". W: *Fourth IEEE International Conference on Data Mining (ICDM'04)*. IEEE, s. 331–334.

Bojanowski, Piotr, Edouard Grave, Armand Joulin i Tomas Mikolov (2017a). „Enriching word vectors with subword information". W: *Transactions of the Association for Computational Linguistics* 5, s. 135–146.

— (2017b). „Enriching Word Vectors with Subword Information". W: *Transactions of the Association for Computational Linguistics* 5. Red. Lillian Lee, Mark Johnson i Kristina Toutanova, s. 135–146. DOI: 10.1162/tacl_a_00051. URL: https://aclanthology.org/Q17-1010.

Bolukbasi, Tolga, Kai-Wei Chang, James Zou, Venkatesh Saligrama i Adam Kalai (2016). „Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings". W: *Proceedings of the 30th International Conference on Neural Information Processing Systems*. NIPS'16. Barcelona, Spain: Curran Associates Inc., 4356–4364. ISBN: 9781510838819.

Breckheimer, Peter J (2001). „A haven for hate: the foreign and domestic implications of protecting Internet hate speech under the first amendment". W: *S. Cal. L. Rev.* 75, s. 1493.

Bridle, John (1989). „Training Stochastic Model Recognition Algorithms as Networks can Lead to Maximum Mutual Information Estimation of Parameters". W: *Advances in Neural Information Processing Systems*. Red. D. Touretzky. T. 2. Morgan-Kaufmann. URL: https://proceedings.neurips.cc/paper_files/paper/1989/file/0336dcbab05b9d5ad24f4333c7658a0e-Paper.pdf.

Brown, Alexander (2018). „What is so special about online (as compared to offline) hate speech?" W: *Ethnicities* 18.3, s. 297–326.

Brown, Tom i in. (2020). „Language Models are Few-Shot Learners". W: *Advances in Neural Information Processing Systems*. Red. H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan i H. Lin. T. 33. Curran Associates, Inc., s. 1877–1901.

Chen, Ying, Yilu Zhou, Sencun Zhu i Heng Xu (wrz. 2012). „Detecting Offensive Language in Social Media to Protect Adolescent Online Safety". W: *2012 Interna-*

tional Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing, s. 71–80. ISBN: 978-1-4673-5638-1. DOI: 10.1109/SocialCom-PASSAT.2012.55.

Chou, H. i C. Lee (2019). „Every Rating Matters: Joint Learning of Subjective Labels and Individual Annotators for Speech Emotion Classification". W: *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, s. 5886–5890. DOI: 10.1109/ICASSP.2019.8682170.

Clarivate (1997). *Web of Science*. https://clarivate.com/products/scientific-and-academic-research/research-discovery-and-workflow-solutions/webofscience-platform/. Dostęp: 27.09.2024.

Cohen, Jacob (1960). „A Coefficient of Agreement for Nominal Scales". W: *Educational and Psychological Measurement* 20, s. 37 –46. URL: https://api.semanticscholar.org/CorpusID:15926286.

Conneau, Alexis, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer i Veselin Stoyanov (lip. 2020). „Unsupervised Cross-lingual Representation Learning at Scale". W: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Red. Dan Jurafsky, Joyce Chai, Natalie Schluter i Joel Tetreault. Online: Association for Computational Linguistics, s. 8440–8451. DOI: 10.18653/v1/2020.acl-main.747. URL: https://aclanthology.org/2020.acl-main.747.

Davidson, Thomas, Debasmita Bhattacharya i Ingmar Weber (sierp. 2019). „Racial Bias in Hate Speech and Abusive Language Detection Datasets". W: *Proceedings of the Third Workshop on Abusive Language Online*. Florence, Italy: Association for Computational Linguistics, s. 25–35. DOI: 10.18653/v1/W19-3504. URL: https://www.aclweb.org/anthology/W19-3504.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee i Kristina Toutanova (2019a). „BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". W: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, s. 4171–4186.

— (czer. 2019b). „BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". W: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Red. Jill Burstein, Christy Doran i Thamar Solorio. Minneapolis, Minnesota: Association for Computational Linguistics, s. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: https://aclanthology.org/N19-1423.

— (czer. 2019c). „BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". W: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,*

Volume 1 (Long and Short Papers). Red. Jill Burstein, Christy Doran i Thamar Solorio. Minneapolis, Minnesota: Association for Computational Linguistics, s. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: https://aclanthology.org/N19-1423.

Dixon, Lucas, John Li, Jeffrey Sorensen, Nithum Thain i Lucy Vasserman (2018). „Measuring and Mitigating Unintended Bias in Text Classification". W: *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. AIES '18. New Orleans, LA, USA: Association for Computing Machinery, 67–73. ISBN: 9781450360128. DOI: 10.1145/3278721.3278729. URL: https://doi.org/10.1145/3278721.3278729.

Dorsey, Jack, Christopher Stone, Noah Glass i Evan Williams (2006). *Twitter*. https://x.com/. Dostęp: 20.09.2024.

Dugas, Charles, Yoshua Bengio, François Bélisle, Claude Nadeau i René Garcia (2000). „Incorporating Second-Order Functional Knowledge for Better Option Pricing". W: *Advances in Neural Information Processing Systems*. Red. T. Leen, T. Dietterich i V. Tresp. T. 13. MIT Press. URL: https://proceedings.neurips.cc/paper_files/paper/2000/file/44968aece94f667e4095002d140b5896-Paper.pdf.

Dunn, Olive Jean (1961). „Multiple comparisons among means". W: *Journal of the American statistical association* 56.293, s. 52–64.

Elsevier (2004). *Scopus*. https://www.sciencedirect.com/science/article/pii/S1566253523000167. Dostęp: 27.09.2024.

Fawcett, Tom (2006). „An introduction to ROC analysis". W: *Pattern recognition letters* 27.8, s. 861–874.

Feng, Fangxiaoyu, Yinfei Yang, Daniel Cer, Naveen Arivazhagan i Wei Wang (maj 2022). „Language-agnostic BERT Sentence Embedding". W: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Red. Smaranda Muresan, Preslav Nakov i Aline Villavicencio. Dublin, Ireland: Association for Computational Linguistics, s. 878–891. DOI: 10.18653/v1/2022.acl-long.62. URL: https://aclanthology.org/2022.acl-long.62.

Ferdinan, Teddy i Jan Kocoń (2023). „Personalized Models Resistant to Malicious Attacks for Human-centered Trusted AI". W: *The AAAI-23 Workshop on Artificial Intelligence Safety (SafeAI 2023)*. CEUR Workshop Proceedings.

Fortuna, P. i S. Nunes (2018). „A Survey on Automatic Detection of Hate Speech in Text". W: *ACM Computing Surveys (CSUR)* 51, s. 1 –30.

Fukushima, Kunihiko (1969). „Visual Feature Extraction by a Multilayered Network of Analog Threshold Elements". W: *IEEE Transactions on Systems Science and Cybernetics* 5.4, s. 322–333. DOI: 10.1109/TSSC.1969.300225.

Harinarayan, Venkatesh (2001). *Amazon Mechanical Turk*. https://www.mturk.com/. Dostęp: 20.09.2024.

Harpale, Abhay S i Yiming Yang (2008). „Personalized active learning for collaborative filtering". W: *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, s. 91–98.

Harris, Zellig S (1954). „Distributional structure." W: *Word*.

He, Pengcheng, Xiaodong Liu, Jianfeng Gao i Weizhu Chen (2020). „Deberta: Decoding-enhanced bert with disentangled attention". W: *arXiv preprint arXiv:2006.03654*.

Heyman, Steven J (2008). „Hate speech, public discourse, and the first amendment". W: *Oxford University Press, Forthcoming*.

Huffman, Steve, Aaron Swartz i Alexis Ohanian (2005). *Reddit*. https://www.reddit.com/. Dostęp: 20.09.2024.

Jacobs, James B (2002). „Hate Crime: Criminal Law and Identity Politics: Author's summary". W: *Theoretical Criminology* 6.4, s. 481–484.

Janz, Arkadiusz, Jan Kocoń, Maciej Piasecki i Zaśko-Zielińska Monika (list. 2017). „plWordNet as a Basis for Large Emotive Lexicons of Polish". W: *LTC'17 8th Language and Technology Conference*. Poznań, Poland: Fundacja Uniwersytetu im. Adama Mickiewicza w Poznaniu. ISBN: 978-83-64864-94-0.

Kanclerz, Kamil, Julita Bielaniewicz, Marcin Gruza, Jan Kocoń, Stanisław Woźniak i Przemysław Kazienko (2023a). „Towards Model-Based Data Acquisition for Subjective Multi-Task NLP Problems". W: *2023 IEEE International Conference on Data Mining Workshops (ICDMW)*. IEEE, s. 726–735.

Kanclerz, Kamil, Alicja Figas, Marcin Gruza, Tomasz Kajdanowicz, Jan Kocon, Daria Puchalska i Przemyslaw Kazienko (sierp. 2021). „Controversy and Conformity: from Generalized to Personalized Aggressiveness Detection". W: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Red. Chengqing Zong, Fei Xia, Wenjie Li i Roberto Navigli. Online: Association for Computational Linguistics, s. 5915–5926. DOI: 10.18653/v1/2021.acl-long.460. URL: https://aclanthology.org/2021.acl-long.460.

Kanclerz, Kamil, Marcin Gruza, Konrad Karanowski, Julita Bielaniewicz, Piotr Milkowski, Jan Kocon i Przemyslaw Kazienko (czer. 2022). „What If Ground Truth Is Subjective? Personalized Deep Neural Hate Speech Detection". W: *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022*. Red. Gavin Abercrombie, Valerio Basile, Sara Tonelli, Verena Rieser i Alexandra Uma. Marseille, France: European Language Resources Association, s. 37–45. URL: https://aclanthology.org/2022.nlperspectives-1.6.

Kanclerz, Kamil, Konrad Karanowski, Julita Bielaniewicz, Marcin Gruza, Piotr Miłkowski, Jan Kocon i Przemyslaw Kazienko (grud. 2023b). „PALS: Personalized Active Learning for Subjective Tasks in NLP". W: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Red. Houda Bouamor, Juan

Pino i Kalika Bali. Singapore: Association for Computational Linguistics, s. 13326–13341. DOI: 10.18653/v1/2023.emnlp-main.823. URL: https://aclanthology.org/2023.emnlp-main.823.

Kanclerz, Kamil, Piotr Miłkowski i Jan Kocoń (2020). „Cross-lingual deep neural transfer learning in sentiment analysis". W: *Procedia Computer Science* 176, s. 128–137.

Kanclerz, Kamil i Maciej Piasecki (maj 2022). „Deep Neural Representations for Multiword Expressions Detection". W: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*. Red. Samuel Louvan, Andrea Madotto i Brielen Madureira. Dublin, Ireland: Association for Computational Linguistics, s. 444–453. DOI: 10.18653/v1/2022.acl-srw.36. URL: https://aclanthology.org/2022.acl-srw.36.

Kaplan, Jared, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu i Dario Amodei (2020). „Scaling laws for neural language models". W: *arXiv preprint arXiv:2001.08361*.

Kara, Yunus Emre, Gaye Genc, Oya Aran i Lale Akarun (lip. 2015). „Modeling Annotator Behaviors for Crowd Labeling". W: *Neurocomput.* 160.C, 141–156. ISSN: 0925-2312. DOI: 10.1016/j.neucom.2014.10.082. URL: https://doi.org/10.1016/j.neucom.2014.10.082.

Karim, Jawed, Steve Chen i Chad Hurley (2005). *YouTube*. https://www.youtube.com/. Dostęp: 20.09.2024.

Kazienko, Przemysław, Julita Bielaniewicz, Marcin Gruza, Kamil Kanclerz, Konrad Karanowski, Piotr Miłkowski i Jan Kocoń (2023). „Human-centered neural reasoning for subjective content processing: Hate speech, emotions, and humor". W: *Information Fusion* 94, s. 43–65.

Kennedy, Chris J, Geoff Bacon, Alexander Sahn i Claudia von Vacano (2020). „Constructing interval variables via faceted rasch measurement and multitask deep learning: a hate speech application". W: *arXiv preprint arXiv:2009.10277*.

Kocoń, Jan, Joanna Baran i Kamil Kanclerz (2023). „Multi-Modal Personalized Hate Speech Analysis using Differential Dataset Cartography". W: *Second Workshop on Multimodal Fact Checking and Hate Speech Detection at AAAI 2023*.

Kocoń, Jan, Joanna Baran, Kamil Kanclerz, Michał Kajstura i Przemysław Kazienko (2023a). „Differential dataset cartography: Explainable artificial intelligence in comparative personalized sentiment analysis". W: *International Conference on Computational Science*. Springer, s. 148–162.

Kocoń, Jan, Igor Cichecki, Oliwier Kaszyca, Mateusz Kochanek, Dominika Szydło, Joanna Baran, Julita Bielaniewicz, Marcin Gruza, Arkadiusz Janz, Kamil Kanclerz i in. (2023b). „ChatGPT: Jack of all trades, master of none". W: *Information Fusion* 99, s. 101861.

Kocoń, Jan, Marcin Gruza, Julita Bielaniewicz, Damian Grimling, Kamil Kanclerz, Piotr Miłkowski i Przemysław Kazienko (2021). „Learning personal human biases and representations for subjective tasks in natural language processing". W: *2021 IEEE International Conference on Data Mining (ICDM)*. IEEE, s. 1168–1173.

Kocoń, Jan, Arkadiusz Janz i Maciej Piasecki (2018). „Classifier-based Polarity Propagation in a Wordnet". W: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Kocoń, Jan i Marek Maziarz (2021). „Mapping WordNet onto human brain connectome in emotion processing and semantic similarity recognition". W: *Information Processing & Management* 58.3, s. 102530.

Kocoń, Jan, Piotr Miłkowski i Kamil Kanclerz (2021). „Multiemo: Multilingual, multilevel, multidomain sentiment analysis corpus of consumer reviews". W: *International Conference on Computational Science*. Springer, s. 297–312.

Kocoń, Jan i in. (2019a). „Propagation of emotions, arousal and polarity in WordNet using Heterogeneous Structured Synset Embeddings". W: *Proceedings of the 10th International Global Wordnet Conference (GWC'19)*. Wrocław, Poland.

Kocoń, Jan i in. (2019b). „Recognition of emotions, valence and arousal in large-scale multi-domain text reviews". W: *Human Language Technologies as a Challenge for Computer Science and Linguistics*. Red. Zygmunt Vetulani i Patrick Paroubek. Poznań, Poland: Wydawnictwo Nauka i Innowacje, s. 274–280. ISBN: 978-83-65988-31-7.

Kolhatkar, Varada, Hanhan Wu, Luca Cavasso, Emilie Francis, Kavan Shukla i Maite Taboada (2020). „The SFU opinion and comments corpus: A corpus for the analysis of online news comments". W: *Corpus Pragmatics* 4.2, s. 155–190.

Krippendorff, Klaus (1970). „Estimating the reliability, systematic error and random error of interval data". W: *Educational and psychological measurement* 30.1, s. 61–70.

Kumar, Ritesh, Atul Kr Ojha, Bornini Lahiri, Marcos Zampieri, Shervin Malmasi, Vanessa Murdock i Daniel Kadar, red. (2020). *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*. European Language Resources Association (ELRA). ISBN: 979-10-95546-56-6.

Kumar, Ritesh, Atul Kr. Ojha, Marcos Zampieri i Shervin Malmasi, red. (2018). *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*. Association for Computational Linguistics. URL: https://www.aclweb.org/anthology/W18-4400.

Lan, Zhenzhong, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma i Radu Soricut (2019). „Albert: A lite bert for self-supervised learning of language representations". W: *arXiv preprint arXiv:1909.11942*.

Levmore, Saul i Martha Craven Nussbaum (2010). *The offensive Internet: Speech, privacy, and reputation*. Harvard University Press.

Lewis, Patrick, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel i in. (2020). „Retrieval-augmented generation for knowledge-intensive nlp tasks". W: *Advances in Neural Information Processing Systems* 33, s. 9459–9474.

Liu, Yinhan (2019). „Roberta: A robustly optimized bert pretraining approach". W: *arXiv preprint arXiv:1907.11692*.

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer i Veselin Stoyanov (2019). „Roberta: A robustly optimized bert pretraining approach". W: *arXiv preprint arXiv:1907.11692*.

Ljubešić, Nikola, Darja Fišer i Tomaž Erjavec (2019). „The FRENK Datasets of Socially Unacceptable Discourse in Slovene and English". W: *International Conference on Text, Speech, and Dialogue*. Springer, s. 103–114.

Lowell, David, Zachary C Lipton i Byron C Wallace (2018). „How transferable are the datasets collected by active learners". W: *arXiv preprint arXiv:1807.04801* 3.

Lu, Jinghui, Maeve Henchion i Brian Mac Namee (2019). „Investigating the effectiveness of representations based on word-embeddings in active learning for labelling text datasets". W: *arXiv preprint arXiv:1910.03505*.

Mandl, Thomas, Sandip Modha, Prasenjit Majumder, Daksh Patel, Mohana Dave, Chintak Mandlia i Aditya Patel (2019). „Overview of the HASOC Track at FIRE 2019: Hate Speech and Offensive Content Identification in Indo-European Languages". W: *Proceedings of the 11th Forum for Information Retrieval Evaluation*. FIRE '19. New York, NY, USA: Association for Computing Machinery, 14–17. ISBN: 9781450377508. DOI: 10.1145/3368567.3368584. URL: https://doi.org/10.1145/3368567.3368584.

Mann, Henry B i Donald R Whitney (1947). „On a test of whether one of two random variables is stochastically larger than the other". W: *The annals of mathematical statistics*, s. 50–60.

Mieleszczenko-Kowszewicz, Wiktoria, Kamil Kanclerz, Julita Bielaniewicz, Marcin Oleksy i Marcin Gruza (2023). „Capturing Human Perspectives in NLP: Questionnaires, Annotations, and Biases". W: *Second Workshop on Perspectivist Approaches to NLP (NLPerspectives) at ECAI 2023*.

Milkowski, Piotr, Marcin Gruza, Kamil Kanclerz, Przemyslaw Kazienko, Damian Grimling i Jan Kocon (sierp. 2021). „Personal Bias in Prediction of Emotions Elicited by Textual Opinions". W: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop*. Red. Jad Kabbara, Haitao Lin, Amandalynne Paullada i Jannis Vamvas. Online: Association for Computational Linguistics, s. 248–259. DOI: 10.18653/v1/2021.acl-srw.26. URL: https://aclanthology.org/2021.acl-srw.26.

Misiaszek, Andrzej, Przemysław Kazienko, Marcin Kulisiewicz, Łukasz Augustyniak, Włodzimierz Tuligłowicz, Adrian Popiel i Tomasz Kajdanowicz (2014). „Belief Propagation Method for Word Sentiment in WordNet 3.0". W: *Asian Conference on Intelligent Information and Database Systems*. Springer, s. 263–272.

Modha, Sandip, Prasenjit Majumder i Thomas Mandl (sierp. 2018). „Filtering Aggression from the Multilingual Social Media Feed". W: *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, s. 199–207. URL: https://www.aclweb.org/anthology/W18-4423.

Modha, Sandip, Prasenjit Majumder, Thomas Mandl i Chintak Mandalia (2020). „Detecting and visualizing hate speech in social media: A cyber Watchdog for surveillance". W: *Expert Systems with Applications* 161, s. 113725. ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2020.113725. URL: http://www.sciencedirect.com/science/article/pii/S0957417420305492.

Neviarouskaya, Alena, Helmut Prendinger i Mitsuru Ishizuka (2009). „Compositionality principle in recognition of fine-grained emotions from text". W: *Third International AAAI Conference on Weblogs and Social Media*.

OpenAI (2023). *GPT-3.5: Generative Pre-trained Transformer*. https://platform.openai.com/docs/models/gpt-3-5-turbo. Dostęp: 20.09.2024.

Piasecki, Maciej, Bernd Broda i Stanislaw Szpakowicz (2009). *A wordnet from the ground up*. Oficyna Wydawnicza Politechniki Wrocławskiej Wrocław.

Piasecki, Maciej i Kamil Kanclerz (2022a). „Is Context All You Need? Non-contextual vs Contextual Multiword Expressions Detection". W: *International Conference on Computational Science*. Springer, s. 248–261.

— (2022b). „Non-Contextual vs Contextual Word Embeddings in Multiword Expressions Detection". W: *International Conference on Computational Collective Intelligence*. Springer, s. 193–206.

Poletto, Fabio, Valerio Basile, M. Sanguinetti, Cristina Bosco i V. Patti (2020). „Resources and benchmark corpora for hate speech detection: a systematic review". W: *LREC 2020*.

Prabhu, Ameya, Charles Dognin i Maneesh Singh (2019). „Sampling bias in deep active classification: An empirical study". W: *arXiv preprint arXiv:1909.09389*.

Prabhu, Sumanth, Moosa Mohamed i Hemant Misra (2021). „Multi-class Text Classification using BERT-based Active Learning". W: *arXiv preprint arXiv:2104.14289*.

Price, Ilan, Jordan Gifford-Moore, Jory Flemming, Saul Musker, Maayan Roichman, Guillaume Sylvain, Nithum Thain, Lucas Dixon i Jeffrey Sorensen (list. 2020). „Six Attributes of Unhealthy Conversations". W: *Proceedings of the Fourth Workshop on Online Abuse and Harms*. Red. Seyi Akiwowo, Bertie Vidgen, Vinodkumar Prabhakaran i Zeerak Waseem. Online: Association for Computational Linguistics,

s. 114–124. DOI: 10.18653/v1/2020.alw-1.15. URL: https://aclanthology.org/2020.alw-1.15.

Ptaszyński, Michal, Agata Pieciukiewicz i Paweł Dybala (2019). „Results of the PolEval 2019 Shared Task 6: First Dataset and Open Shared Task for Automatic Cyberbullying Detection in Polish Twitter". W: *Proceedings of the PolEval 2019 Workshop*. Institute of Computer Science, Polish Academy of Sciences, s. 89–110. ISBN: 978-83-63159-28-3. URL: https://ruj.uj.edu.pl/xmlui/handle/item/152265.

Raykar, Vikas C. i Shipeng Yu (lut. 2012). „Eliminating Spammers and Ranking Annotators for Crowdsourced Labeling Tasks". W: *J. Mach. Learn. Res.* 13.1, 491–518. ISSN: 1532-4435.

Razavi, Amir H., Diana Inkpen, Sasha Uritsky i Stan Matwin (2010). „Offensive Language Detection Using Multi-level Classification". W: *Advances in Artificial Intelligence*. Red. Atefeh Farzindar i Vlado Kešelj. Berlin, Heidelberg: Springer Berlin Heidelberg, s. 16–27. ISBN: 978-3-642-13059-5.

Ren, Pengzhen, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Brij B Gupta, Xiaojiang Chen i Xin Wang (2021). „A survey of deep active learning". W: *ACM Computing Surveys (CSUR)* 54.9, s. 1–40.

Rijsbergen, Van (1979). „Information retrieval; ; Butterworth, 1978". W: *J. librariansh.* 11, s. 237.

Risch, Julian i Ralf Krestel (sierp. 2018). „Aggression Identification Using Deep Learning and Data Augmentation". W: *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, s. 150–158. URL: https://www.aclweb.org/anthology/W18-4418.

Rosenfeld, Michel (2002). „Hate speech in constitutional jurisprudence: a comparative analysis". W: *Cardozo L. Rev.* 24, s. 1523.

Sadiq, Saima, Arif Mehmood, Saleem Ullah, Maqsood Ahmad, Gyu Sang Choi i Byung-Won On (2021). „Aggression detection through deep neural model on Twitter". W: *Future Generation Computer Systems* 114, s. 120 –129. ISSN: 0167-739X. DOI: https://doi.org/10.1016/j.future.2020.07.050. URL: http://www.sciencedirect.com/science/article/pii/S0167739X19330717.

Safi Samghabadi, Niloofar, Parth Patwa, Srinivas PYKL, Prerana Mukherjee, Amitava Das i Thamar Solorio (maj 2020). „Aggression and Misogyny Detection using BERT: A Multi-Task Approach". W: *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*. Marseille, France: European Language Resources Association (ELRA), s. 126–131. ISBN: 979-10-95546-56-6. URL: https://www.aclweb.org/anthology/2020.trac-1.20.

Sahlgren, Magnus, Tim Isbister i Fredrik Olsson (paź. 2018). „Learning Representations for Detecting Abusive Language". W: *Proceedings of the 2nd Workshop on Abu-*

sive Language Online (ALW2). Brussels, Belgium: Association for Computational Linguistics, s. 115–123. DOI: 10.18653/v1/W18-5115. URL: https://www.aclweb.org/anthology/W18-5115.

Salminen, J., F. Veronesi, H. Almerekhi, S. Jung i B. J. Jansen (2018). „Online Hate Interpretation Varies by Country, But More by Individual: A Statistical Analysis Using Crowdsourced Ratings". W: *2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, s. 88–94. DOI: 10.1109/SNAMS.2018.8554954.

Sap, Maarten, Dallas Card, Saadia Gabriel, Yejin Choi i Noah A. Smith (lip. 2019). „The Risk of Racial Bias in Hate Speech Detection". W: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, s. 1668–1678. DOI: 10.18653/v1/P19-1163. URL: https://www.aclweb.org/anthology/P19-1163.

Schmidt, Anna i Michael Wiegand (kw. 2017). „A Survey on Hate Speech Detection using Natural Language Processing". W: *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*. Valencia, Spain: Association for Computational Linguistics, s. 1–10. DOI: 10.18653/v1/W17-1101. URL: https://www.aclweb.org/anthology/W17-1101.

Scott, Sam i Stan Matwin (1998). „Text classification using WordNet hypernyms". W: *Usage of WordNet in Natural Language Processing Systems*.

Seifert, Christin i Michael Granitzer (2010). „User-based active learning". W: *2010 IEEE International Conference on Data Mining Workshops*. IEEE, s. 418–425.

Shapiro, Samuel Sanford i Martin B Wilk (1965). „An analysis of variance test for normality (complete samples)". W: *Biometrika* 52.3-4, s. 591–611.

Snow, Rion, Brendan O'Connor, Daniel Jurafsky i Andrew Ng (paź. 2008). „Cheap and Fast – But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks". W: *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*. Honolulu, Hawaii: Association for Computational Linguistics, s. 254–263. URL: https://www.aclweb.org/anthology/D08-1027.

Soberón, Guillermo, Lora Aroyo, Chris Welty, Oana Inel, Hui Lin i Manfred Overmeen (2013). „Measuring Crowd Truth: Disagreement Metrics Combined with Worker Behavior Filters". W: *Proceedings of the 1st International Conference on Crowdsourcing the Semantic Web - Volume 1030*. CrowdSem'13. CEUR-WS.org, 45–58.

Song, Kaitao, Xu Tan, Tao Qin, Jianfeng Lu i Tie-Yan Liu (2020). „Mpnet: Masked and permuted pre-training for language understanding". W: *Advances in neural information processing systems* 33, s. 16857–16867.

Srivastava, Aarohi, Abhinav Rastogi, Abhishek Rao, Abu Awal Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adri Garriga-

Alonso i in. (2023). „Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models". W: *Transactions on machine learning research*.

Student (1908). „The probable error of a mean". W: *Biometrika*, s. 1–25.

Tan, Wei, Lan Du i Wray Buntine (2021). „Diversity Enhanced Active Learning with Strictly Proper Scoring Rules". W: *Advances in Neural Information Processing Systems* 34.

Tatman, Rachael (kw. 2017). „Gender and Dialect Bias in YouTube's Automatic Captions". W: *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*. Valencia, Spain: Association for Computational Linguistics, s. 53–59. DOI: 10.18653/v1/W17-1606. URL: https://www.aclweb.org/anthology/W17-1606.

Van Rossum, Guido i Fred L. Drake (2009). *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace. ISBN: 1441412697.

Wang, Keze, Dongyu Zhang, Ya Li, Ruimao Zhang i Liang Lin (2016). „Cost-effective active learning for deep image classification". W: *IEEE Transactions on Circuits and Systems for Video Technology* 27.12, s. 2591–2600.

Waseem, Zeerak (list. 2016). „Are You a Racist or Am I Seeing Things? Annotator Influence on Hate Speech Detection on Twitter". W: *Proceedings of the First Workshop on NLP and Computational Social Science*. Austin, Texas: Association for Computational Linguistics, s. 138–142. DOI: 10.18653/v1/W16-5618. URL: https://www.aclweb.org/anthology/W16-5618.

Waseem, Zeerak i Dirk Hovy (czer. 2016). „Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter". W: *Proceedings of the NAACL Student Research Workshop*. San Diego, California: Association for Computational Linguistics, s. 88–93. DOI: 10.18653/v1/N16-2013. URL: https://www.aclweb.org/anthology/N16-2013.

Wei, Jason, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai i Quoc V Le (2022). „Finetuned Language Models are Zero-Shot Learners". W: *International Conference on Learning Representations*.

Wich, Maximilian, Hala Al Kuwatly i Georg Groh (list. 2020). „Investigating Annotator Bias with a Graph-Based Approach". W: *Proceedings of the Fourth Workshop on Online Abuse and Harms*. Online: Association for Computational Linguistics, s. 191–199. DOI: 10.18653/v1/2020.alw-1.22. URL: https://www.aclweb.org/anthology/2020.alw-1.22.

Wich, Maximilian, Jan Bauer i Georg Groh (list. 2020). „Impact of Politically Biased Data on Hate Speech Classification". W: *Proceedings of the Fourth Workshop on Online Abuse and Harms*. Online: Association for Computational Linguistics, s. 54–64. DOI: 10.18653/v1/2020.alw-1.7. URL: https://www.aclweb.org/anthology/2020.alw-1.7.

Wiegand, Michael, Josef Ruppenhofer i Thomas Kleinbauer (czer. 2019). „Detection of Abusive Language: the Problem of Biased Datasets". W: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, s. 602–608. DOI: 10.18653/v1/N19-1060. URL: https://www.aclweb.org/anthology/N19-1060.

Wiegand, Michael, Melanie Siegel i Josef Ruppenhofer (2018). „Overview of the GermEval 2018 Shared Task on the Identification of Offensive Language". W: *Proceedings of GermEval 2018, 14th Conference on Natural Language Processing (KONVENS 2018)*, s. 1–10.

Wikipedia (2001). *Wikipedia the Free Encyclopedia*. https://en.wikipedia.org/wiki/Main_Page. Dostęp: 20.09.2024.

Wilcoxon, Frank (1945). „Individual Comparisons by Ranking Methods". W: *Biometrics Bulletin* 1.6, s. 80–83. ISSN: 00994987. URL: http://www.jstor.org/stable/3001968 (term. wiz. 24. 09. 2024).

Wojatzki, Michael, Tobias Horsmann, Darina Gold i Torsten Zesch (2018). „Do Women Perceive Hate Differently: Examining the Relationship Between Hate Speech, Gender, and Agreement Judgments". W: *KONVENS*.

Wolf, T (2019). „Huggingface's transformers: State-of-the-art natural language processing". W: *arXiv preprint arXiv:1910.03771*.

Wulczyn, Ellery, Nithum Thain i Lucas Dixon (2017). *Wikipedia Talk Labels: Aggression*. DOI: 10.6084/m9.figshare.4267550.v5. URL: https://figshare.com/articles/dataset/Wikipedia_Talk_Labels_Aggression/4267550/5.

Xia, Mengzhou, Anjalie Field i Yulia Tsvetkov (lip. 2020). „Demoting Racial Bias in Hate Speech Detection". W: *Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media*. Online: Association for Computational Linguistics, s. 7–14. DOI: 10.18653/v1/2020.socialnlp-1.2. URL: https://www.aclweb.org/anthology/2020.socialnlp-1.2.

Yang, Zhilin, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov i Quoc V Le (2019). „Xlnet: Generalized autoregressive pretraining for language understanding". W: *Advances in neural information processing systems*, s. 5753–5763.

Yenala, Harish, Ashish Jhanwar, Manoj Chinnakotla i Jay Goyal (grud. 2017). „Deep learning for detecting inappropriate content in text". W: *International Journal of Data Science and Analytics*. DOI: 10.1007/s41060-017-0088-4.

Zampieri, Marcos, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra i Ritesh Kumar (czer. 2019a). „Predicting the Type and Target of Offensive Posts in Social Media". W: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational

Linguistics, s. 1415–1420. DOI: 10.18653/v1/N19-1144. URL: https://www.aclweb.org/anthology/N19-1144.

Zampieri, Marcos, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra i Ritesh Kumar (czer. 2019b). „SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval)". W: *Proceedings of the 13th International Workshop on Semantic Evaluation*. Minneapolis, Minnesota, USA: Association for Computational Linguistics, s. 75–86. DOI: 10.18653/v1/S19-2010. URL: https://www.aclweb.org/anthology/S19-2010.

Zhou, Jin i Shiliang Sun (2014). „Improved margin sampling for active learning". W: *Chinese Conference on Pattern Recognition*. Springer, s. 120–129.

Ziems, Caleb, Ymir Vigfusson i Fred Morstatter (2020). „Aggressive, Repetitive, Intentional, Visible, and Imbalanced: Refining Representations for Cyberbullying Classification". W: *Proceedings of the International AAAI Conference on Web and Social Media* 14.1, s. 808–819. URL: https://ojs.aaai.org/index.php/ICWSM/article/view/7345.