Uniwersytet im. Adama Mickiewicza w Poznaniu
Wydział Matematyki i Informatyki

mgr Wojciech Włodarczyk

Modele ewaluacji poprawności danych lingwistycznych pozyskanych metodą crowdsourcing

Models for evaluating the correctness of linguistic data obtained using crowdsourcing

Rozprawa doktorska w dziedzinie:

NAUKI MATEMATYCZNE

dyscyplinie:

INFORMATYKA

Promotor:

prof. dr hab. Krzysztof Jassem

Poznań 2023

Streszczenie

Crowdsourcing pozwala na wykorzystanie zbiorowej inteligencji dużej grupy ludzi do rozwiązywania zadań z dziedzin takich jak sztuczna inteligencja, uczenie maszynowe i rozwój badań naukowych za pośrednictwem platform internetowych. Współcześnie badania nad rozwojem metody crowdsourcingu skupione są przede wszystkim w trzech obszarach: optymalizacji jakości pozyskanych danych, optymalizacji kosztu procesu oraz optymalizacji czasu trwania procesu.

Niniejsza rozprawa skupia się na zagadnieniach związanych z optymalizacją jakości procesu crowdsourcingu dla zadań dotyczących danych lingwistycznych. Praca opisuje autorski model Dynamicznej Informacji Zwrotnej (*DIZ*), którego zadaniem jest generowanie informacji zwrotnej w sposób automatyczny. Rozprawa weryfikuje skuteczność tego modelu dla danych empirycznych oraz danych symulacyjnych.

Analiza wyników przeprowadzonego eksperymentu wykazuje skuteczność modelu *DIZ* w poprawie jakości generowanej informacji zwrotnej, jednak jakość tego rozwiązana jest zależna od jakości oznaczeń tworzonych przez anotatorów.

Słowa kluczowe

crowdsourcing, obliczenia ludzkie, nauczanie maszynowe, przetwarzenie języka naturalnego

Summary

Crowdsourcing uses the collective intelligence of a large group of people to solve tasks in fields such as artificial intelligence, machine learning and scientific research development through online platforms. Nowadays, research on the development of the crowdsourcing method is focused primarily in three areas: optimization of the quality of acquired data, optimization of the cost of the process and optimization of the duration of the process.

This dissertation focuses on issues related to optimizing the quality of the crowdsourcing process for tasks related to linguistic data. The work describes the author's Dynamic Feedback Model, whose task is to generate feedback automatically. The dissertation verifies the effectiveness of this model for empirical and simulation data.

The analysis of the results of the experiment shows the effectiveness of the author's model in improving the quality of the generated feedback, but the quality of this solution depends on the quality of the markings created by annotators.

Keywords

crowdsourcing, human-based computation, machine teaching, natural language processing

Spis treści

Wprowadzenie

Mimo że termin „crowdsourcing” po raz pierwszy został użyty dopiero w 2006 roku [Howe, 2006b], to z metody tej zaczęto korzystać znacznie wcześniej. Metoda crowdsourcingu stanowi przykład zastosowania idei „ludzkich obliczeń” (ang. *human-based computation*), w której to człowiek zamiast maszyny wykorzystywany jest do wykonywania obliczeń. Już w XVIII wieku tworzone były projekty zorganizowanych obliczeń (ang. *organized computation*), w ramach których grupa ludzi angażowana była do równoległego rozwiązywania problemów obliczeniowych ówczesnego społeczeństwa [Grier, 2005, s. 1-8]. Jednym z efektów takiego projektu było stworzenie pierwszych dziesiętnych tablic trygonometrycznych. W późniejszych czasach zorganizowane projekty obliczeniowe były wykorzystywane przez wojsko Stanów Zjednoczonych i miały znaczący wpływ na przebieg drugiej wojny światowej [Grier, 2005]. Po pojawieniu się cyfrowych komputerów w połowie XX wieku idea ludzkich obliczeń zeszła na dalszy plan.

Dzięki powstaniu technologii internetowych ludzkie obliczenia ponownie zyskały na popularności. Jednym ze sposobów realizacji ludzkich obliczeń jest crowdsourcing – metoda, która pozwala organizacjom wykorzystać zbiorową inteligencję dużych grup ludzi za pośrednictwem platform internetowych. Crowdsourcing w dzisiejszych czasach wykorzystywany jest m.in. w sztucznej inteligencji, uczeniu maszynowym oraz badaniach naukowych o charakterze społecznym. Projekty crowdsourcingowe są związane z pracą nad zadaniami takimi jak: anotacja danych, oznaczanie obrazów, tworzenie treści, rozwiązywanie problemów, generowanie pomysłów czy podejmowanie decyzji.

Współcześnie badania nad rozwojem metody crowdsourcingu skupione są przede wszystkim w trzech obszarach: optymalizacji jakości pozyskanych danych, optymalizacji kosztu procesu oraz optymalizacji czasu trwania procesu. W ramach niniejszej rozprawy skupiłem się na zagadnieniach związanych z obszarem optymalizacji jakości dla zadań, które dotyczą pracy z danymi lingwistycznymi. Jako główny temat moich badań wybrałem wykorzystanie mechanizmu przekazywania informacji zwrotnej, która ma na celu edukowanie anotatorów podczas ich pracy. Wykonany przeze mnie przegląd literatury związanej z zastosowaniem mechanizmu informacji zwrotnej w metodzie crowdsourcingu wykazał niewielką liczbę badań, dotyczących zadań związanych z przetwarzaniem języka naturalnego.

Celem mojej rozprawy doktorskiej było potwierdzenie skuteczności stosowania informacji zwrotnej w procesie crowdsourcingu jako mechanizmu poprawy jakości danych dla zadań zawierających dane lingwistyczne. W ramach badań przeprowadziłem eksperyment, w którym porównałem jakość danych pozyskanych w procesie crowdsourcingu dla dwóch sytuacji: z wykorzystaniem i bez wykorzystania informacji zwrotnej. Wyniki przeprowadzonego przeze mnie eksperymentu wskazują, że stosowanie informacji zwrotnej w procesie crowdsourcingu ma pozytywny wpływ na jakość danych pozyskiwanych dla zadań związanych z przetwarzaniem języka naturalnego.

W większości eksperymentów opisywanych w literaturze źródłem informacji zwrotnej była ocena ekspertów lub predefiniowany zbiór referencyjny. W ramach niniejszej rozprawy zaproponowałem autorski model Dynamicznej Informacji Zwrotnej (*DIZ*), którego zadaniem było generowanie informacji zwrotnej w sposób automatyczny – na podstawie dostarczanych anotacji. W ramach niniejszej rozprawy opisany został eksperyment weryfikujący skuteczność tego modelu dla danych empirycznych oraz danych symulacyjnych.

Niniejsza rozprawa podzielona została na cztery rozdziały. Pierwszy rozdział zawiera opis metody crowdsourcingu. Przedstawiony został ciąg kolejnych kroków pozyskiwania danych za pomocą metody crowdsourcingu. Scharakteryzowani zostali uczestnicy tego procesu. Dokonano szczegółowego przeglądu literatury przedmiotu. Przedstawiono przykłady badań prowadzonych w zakresie trzech podstawowych obszarów badawczych: optymalizacji kosztu, optymalizacji czasu oraz optymalizacji jakości. Szczegółowo omówione zostało zagadnienie optymalizacji jakości, które bezpośrednio związane jest z tematem moich badań. Przedstawiono taksonomię kontroli jakości w procesie crowdsourcingu. Omówiono mechanizm informacji zwrotnej, stanowiący jedną z metod kontroli jakości w procesie crowdsourcingu i przedstawiono istniejące implementacje tego mechanizmu.

W Rozdziale 2 przedstawiono formalny opis mechanizmu informacji zwrotnej. Zdefiniowano proces nauczania maszynowego, leżący u podstaw mechanizmu informacji zwrotnej. Omówiono trzy podstawowe komponenty nauczania maszynowego: model reprezentacji ucznia, metody obliczania parametrów modelu ucznia oraz metody wyboru sygnałów nauczających. Przedstawiono przykłady konkretnych implementacji każdego z tych trzech komponentów. Omówione zostało umiejscowienie powyższych komponentów w taksonomii kontroli jakości procesu crowdsourcingu.

W rozdziale trzecim opisany został eksperyment autorski, którego celem było zweryfikowanie skuteczności mechanizmu informacji zwrotnej w procesie crowdsourcingu dla zadań związanych z przetwarzaniem języka naturalnego. Eksperyment przeprowadzony został na sześciu różnych zbiorach danych. Każdy ze zbiorów dotyczył innego typu zadania, ale wszystkie związane były z przetwarzaniem języka naturalnego. W eksperymencie wykorzystano autorski system informatyczny, stworzony specjalnie dla potrzeb prowadzonych badań. Rozdział 3 zakończono omówieniem wyników eksperymentu oraz przedstawieniem wniosków dotyczących wpływu obecności, oraz jakości informacji zwrotnej na jakość anotacji danych.

W czwartym rozdziale rozprawy zaprezentowany został autorski model Dynamicznej informacji Zwrotnej (*DIZ*), którego zadaniem było generowanie informacji zwrotnej w sposób automatyczny. W celu wygenerowania informacji zwrotnej model ten wykorzystywał komponenty używane w procesie nauczania maszynowego oraz aktualne oznaczenia anotatora. Rozdział zawiera opis architektury oraz sposobu działania tego modelu. Opisano eksperyment, którego celem było zweryfikowanie skuteczności modelu *DIZ* w procesie crowdsourcingu, przedstawiono wyniki tego eksperymentu oraz płynące z nich wnioski.

Do rozprawy dołączone zostały również załączniki, które stanowią jej istotny element. W załącznikach zamieszczona została: treść protokołów anotacyjnych oraz treść algorytmów zniekształcenia anotacji referencyjnych użytych w eksperymencie omawianym w rozdziale trzecim.

ROZDZIAŁ 1

Metody ewaluacji jakości w metodzie crowdsourcing

1.1. Metoda crowdsourcing

Termin *crowdsourcing*[1] został spopularyzowany w 2006 roku za sprawą artykułu Jeffa Howe'a „The Rise of Crowdsourcing"[Howe, 2006b] oraz postów, które Jeff Howe publikował na swoim blogu [Howe, 2006a]. Zgodnie z opublikowaną na blogu definicją, sam termin crowdsourcing bezpośrednio nawiązuje do terminu *outsourcing*. *Outsourcing* opisuje proces, w którym firma lub organizacja zleca wykonanie danej usługi (lub jej części) podmiotom zewnętrznym, mimo że sama usługa była do tej pory wykonywana wewnątrz tej firmy [Dolgui & Proth, 2013]. Według definicji Howe'a crowdsourcing może być rozumiany jako szczególny przypadek *outsourcingu*, w ramach którego wykonawcą usługi jest grupa ludzi niepowiązanych ze sobą żadną strukturą organizacyjną[2].

Proces pozyskiwania danych metodą crowdsourcingu może być prowadzony zarówno w sposób manualny, jak i przy użyciu systemów informatycznych. W niniejszej rozprawie, pojęcie to ogranicza się tylko do procesów, w których osoby wykonujące powierzoną pracę otrzymują dostęp do zadań za pośrednictwem internetowych platform crowdsourcingowych.

W ramach tego rozdziału niniejszej rozprawy wprowadzone zostały terminy oraz klasyfikacja pojęć, które stanowią postawę dla rozważań opisywanych w dalszej części niniejszej rozprawy. W pierwszej części rozdziału przedstawiłem ogólny zarys metody crowdsourcing: elementy oraz uczestników procesu pozyskiwania danych za pomocą tej metody. W dalszej części rozdziału omówiłem również aktualne obszary badań prowadzonych przez badaczy crowdsourcingu. W szczególności skupiłem się na temacie kontroli jakości danych, a także na tym, jak zapewnienie informacji zwrotnej używane jest jako narzędzie kontroli i zapewnienia wyższej jakości w procesie crowdsourcingu.

[1] Ponieważ w polskiej literaturze brak polskiego odpowiednika dla słowa „crowdsourcing" w ramach niniejszej rozprawy słowo to będzie używane w takiej formie wraz z polską odmianą przez przypadki.

[2] „*Crowdsourcing represents the act of a company or institution taking a function once performed by employees and outsourcing it to an undefined (and generally large) network of people in the form of an open call.*" Howe [2006a].

1.1.1. Uczestnicy procesu crowdsourcingu

Podstawowymi uczestnikami procesu crowdsourcingu są: zleceniodawca (zob. Definicja 1) oraz anotatorzy (zob. Definicja 2).

Definicja 1 (Zleceniodawca)
Zleceniodawca (ang. requester) to osoba/organizacja inicjująca zadanie crowdsourcingowe. Jego/jej zadaniem jest przygotowanie odpowiedniej definicji zadania, dekompozycja zadania na podproblemy, tzw. mikro-zadania, zebranie oraz zagregowanie wyników mikro-zadań, a także zapewnienie odpowiedniej zapłaty za wykonaną pracę.

Definicja 2 (Anotator)
Anotator (ang. annotator lub end-worker) należy do grupy osób odpowiedzialnych za wykonanie mikro-zadań przygotowanych przez zleceniodawcę. W procesie crowdsourcingu anotator jest osobą anonimową, niebędącą częścią żadnej wspólnej organizacji. Do jego/jej decyzji należy wybór, jakie zadanie będzie wykonywać, a także ile pracy chce wykonać.

1.1.2. Platforma crowdsourcingowa

Główną funkcją platformy crowdsourcingowej jest udostępnienie interfejsu, w którym anotatorzy rozwiązują mikro-zadania. Współczesne platformy dostarczają również narzędzia, które ułatwiają zarządzanie zadaniami, ewaluację wyników, moderowanie pracy anotatorów oraz opłacenie ich pracy.

Obecnie dostępnych jest wiele platform crowdsourcingowych takich jak: *Amazon Mechanical Turk*[3], *Appen*[4] lub *Prolific*[5]. Skorzystanie z takiej platformy daje zleceniodawcy dostęp do szerokiej grupy anotatorów pracujących już na platformie. Użycie gotowej platformy nie jest jednak darmowe[6], ale i tak może oznaczać dla zleceniodawcy niższe koszty niż w przypadku tworzenia własnej.

Zleceniodawca, który chce zlecić wykonanie zadania, używa do tego odpowiedniego formularza, za pomocą którego przekazuje zbiór danych, określa definicję treści zadania oraz kwotę za jego wykonanie. Anotatorzy za pośrednictwem listy dostępnych zadań wybierają to, nad którym w danym momencie chcą pracować.

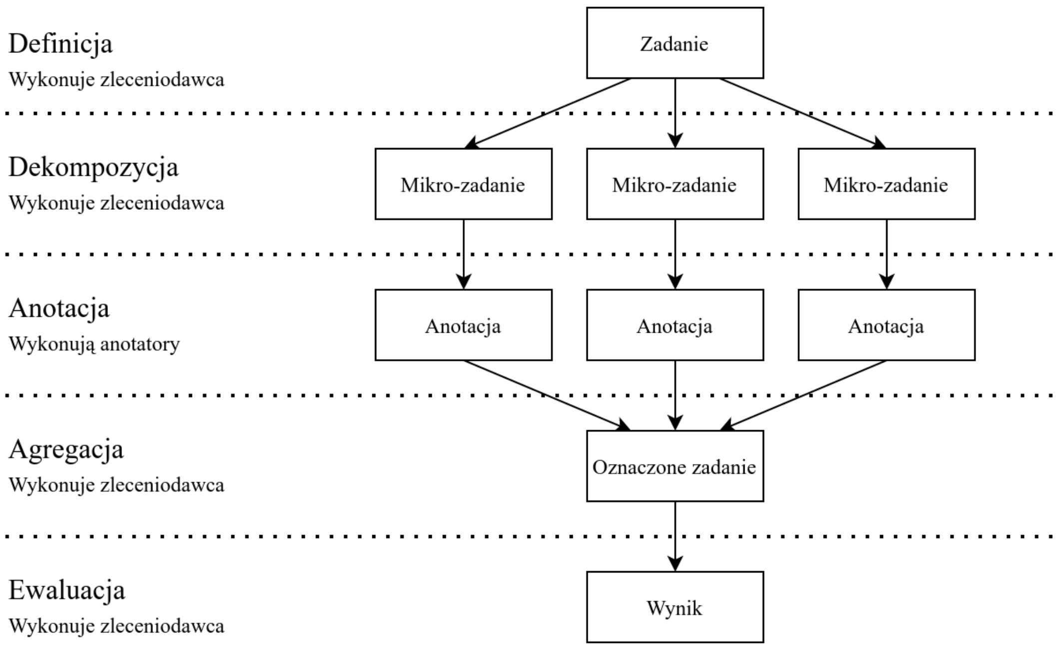
Alternatywnym rozwiązaniem jest stworzenie przez zleceniodawcę własnej platformy. W tej sytuacji zleceniodawca sam odpowiada za obsługę całego procesu, a także znalezienie anotatorów chętnych do pracy. Rozwiązanie to może być korzystne zwłaszcza w przypadku,

[3]https://www.mturk.com/; dostęp: 05.05.2021 r.
[4]https://appen.com/; dostęp: 05.05.2021 r.
[5]https://www.prolific.co/; dostęp: 05.05.2021 r.
[6]Przykładowo, platforma *Amazon Mechanical Turk* pobiera dodatkowe opłaty stanowiące 20% kwoty oferowanej za wykonanie każdego mikro-zadania

gdy zleceniodawca planuje dodać własne mechanizmy, które nie są dostępne na istniejących platformach. Przykładowo, niektórzy zleceniodawcy wprowadzają mechanizmy mające za zadanie zachęcić anotatorów do pracy w inny sposób niż poprzez oferowanie korzyści finansowych[7].

1.1.3. Proces crowdsourcingu

Bez względu na typ oznaczanych danych, proces ich oznaczania w metodzie crowdsourcingu jest zawsze bardzo zbliżony. Proces ten można podzielić na pięć kluczowych kroków: definicja, dekompozycja, anotacja, agregacja, ewaluacja (zob. Rysunek 1.1).



Rysunek 1.1: Diagram procesu oznaczania danych w metodzie crowdsourcingu (opracowanie własne)

Poniżej omówiłem formalny opis poszczególnych kroków procesu oznaczania danych widoczny na Rysunku 1.1.

1. Definicja

W pierwszym kroku procesu crowdsourcingu zleceniodawca definiuje treść zadania oraz określa zbiór danych, który ma zostać oznaczony (zob. Rysunek 1.1). W tym kroku może również dokonać wyborów związanych z logistyką przeprowadzenia samego procesu, np.

[7] Alternatywne mechanizmy motywacji anotatorów zostały opisane w dalszej części pracy, zob. Paragraf 1.3.4.

wyborem platformy crowdsourcingowej, formy oraz kwoty wynagrodzenia, czy sposobu pozyskiwania grupy anotatorów.

2. Dekompozycja

W związku z tym, że nie każde zadanie może być wykonane przez osobę bez specjalistycznego przeszkolenia, takie zadania powinny zostać uproszczone i przygotowane w odpowiedni sposób. Z tego powodu, w procesie crowdsourcingu zadanie jest dekomponowane na mniejsze mikro-zadania (zob. Rysunek 1.1). Zleceniodawca dokonuje dekompozycji złożonego zadania poprzez określenie n-elementowego zbioru treści mikro-zadań $\mathcal{D} = \{x_1, \cdots, x_n\}$, gdzie pojedyncze mikro-zadanie x_i zdefiniowane jest według poniższej definicji (zob. Definicja 3 oraz Przykład 1).

W przypadku gdy dekompozycja zadania nie jest wykonalna, to zadanie powinno zostać przeformułowane. Możliwa jest też sytuacja, gdy dekompozycja nie jest możliwa nawet po przeformułowaniu zadania, wtedy takie zadanie nie jest odpowiednie dla metody crowdsourcingu.

Definicja 3 (Mikro-zadanie)
Niech:
$\mathcal{D} = \{x_1, \cdots, x_n\}$ *– zbiór treści mikro-zadań.*
X *– przestrzeń możliwych wartości cech opisujących treść mikro-zadania.*
J *– zbiór możliwych wartości wyjściowych, które mogą być przypisane do treści mikro-zadania.*
Mikro-zadaniem nazywamy problem polegający na przypisaniu odpowiedniej wartości wyjściowej $j \in J$ *do wskazanej treści* $x_i \in X$.

Przykład 1 (Klasyfikacja irysów)
Rozważmy przykład klasyfikacji zbioru danych zawierającego informacje na temat irysów[8]*.*
Zbiór treści mikro-zadań $\mathcal{D} = \{x_1, \cdots, x_n\}$ *zawiera* $n = 150$ *elementów. Treść mikro-zadań* $x_i \in X$*, określona jest przez 4-wymiarową przestrzeń wektorową rzeczywistą* $X = \mathbb{R}^4$*. Każdy wymiar* X *opisuje inną cechę irysa: szerokość płatka, długość płatka, szerokość działki kielicha, długość działki kielicha.*
Zbiór możliwych wartości wyjściowych składa się z 3 wartości kategorycznych, które określają różne gatunki irysów: $J = \{setosa, versicolor, virginica\}$.
Mikro-zadanie dla tak zdefiniowanego problemu polega na przypisaniu odpowiedniego gatunku $j \in J$ *do wskazanej treści mikro-zadania* $x_i \in X$.

[8]Przykładowy zbiór danych, zob.: https://archive.ics.uci.edu/ml/datasets/Iris, dostęp: 14.02.2022 r.

Jednym z najprostszych sposobów dekompozycji zadania oznaczenia zbioru danych jest zdefiniowanie tylko jednego typu mikro-zadań dla całego zbioru danych. W takim podejściu przestrzeń wartości cech X oraz zbiór możliwych wartości J są wspólne dla wszystkich mikro-zadań, a same mikro-zadania różnią się jedynie treścią, która ma być oznaczona (zob. Przykład 2 oraz Przykład 3).

Przykład 2 (Oznaczanie jednostek nazwanych)
Rozważmy zadanie oznaczania jednostek nazwanych (ang. named entities) w zbiorze danych pozyskanych z platformy Twitter[9]*. Zadaniem anotatora jest wskazanie w tekście fraz (składających się z jednego lub więcej słów), które mogą być przypisane do jednej z trzech możliwych kategorii: lokalizacja, organizacja, osoba. W takim przypadku, zadanie oznaczenia pełnego zbioru może być podzielone tak, by jedno mikro-zadanie obejmowało oznaczenie jednego wpisu (ang. tweet; zob. Fromreide et al., 2014).*

Przykład 3 (Oznaczanie zbioru grafik)
Rozważmy zadanie oznaczenia obiektów w zbiorze zawierającym pliki graficzne (np. zdjęcia). Dla każdego zdjęcia znajdującego się w zbiorze danych utworzone zostanie jedno mikro-zadanie. Treścią takiego mikro-zadania może być zaznaczenie obszaru, w którym znajduje się wyszczególniony obiekt, np. postać (zob. Su et al., 2012).

Innym podejściem dekompozycji zadania jest zaprojektowanie wieloetapowego procesu oznaczania danych, w którym dla każdego wpisu ze zbioru danych wykonywana jest sekwencja kilku różnych mikro-zadań. W takim podejściu mikro-zadania nie są od siebie niezależne. Anotacje wykonane dla zbioru mikro-zadań z jednego etapu mają bezpośredni wpływ na pracę w kolejnym etapie (zob. Przykład 4).

Przykład 4 (Tłumaczenie tekstu)
Rozważmy zadanie tłumaczenia zbioru zdań z języka urdu na język angielski. Zadanie to może być zrealizowane w procesie crowdsourcingu, w którym tłumaczenia przeprowadzane są w kilku etapach. W pierwszym etapie mikro-zadania obejmują wykonanie tłumaczenia. W drugim etapie anotatorzy nanoszą propozycję korekcji tłumaczeń. W ramach ostatniego etapu tworzą oni ranking alternatywnych tłumaczeń dla tego samego tekstu (zob. Zaidan & Callison-Burch, 2011).

[9] `https://twitter.com/`

3. Anotacja

Trzecim korkiem procesu crowdsourcingu jest anotacja (zob. Rysunek 1.1). Anotatorzy rozwiązują mikro-zadania za pośrednictwem platformy crowdsourcingowej. Pojedyncze rozwiązanie danego mikro-zadania nazywane jest anotacją:

Definicja 4 (Anotacja)

Niech:

$A = \{a_1, \cdots, a_n\}$ *– zbiór* n *anotatorów.*

$\mathcal{D} = \{x_1, \cdots, x_n\}$ *– zbiór treści mikro-zadań.*

J *– zbiór możliwych wartości wyjściowych, które mogą być przypisane do treści mikro-zadania.*

Anotacją nazywamy pojedyncze przypisanie wartości ze zbioru J *do treści mikro-zadania* x_i *przez anotatora* k *i oznaczamy jako:* $\hat{y}_i^{(k)} \in J$.

Proces anotacji pełnego zbioru treści mikro-zadań $\mathcal{D}$ odbywa się według następującej procedury (zob. Procedura 1):

Procedura 1: Procedura anotacji mikro-zadań

Niech :

$\mathcal{D}$ – zbiór treści mikro-zadań,

A – zbiór wszystkich anotatorów

Kroki :

1 Dla każdego $x_i \in \mathcal{D}$:

- Wybierany jest podzbiór anotatorów:
  $A_i \subset A$ – anotatorzy, którzy rozwiążą mikro-zadanie x_i.
- Anotatorzy tworzą zbiór anotacji dla mikro-zadania x_i:
  $Y_i = \{\hat{y}_i^{(k)} | \, k \in A_i\}$.

Wyjście:

Y – Finalny zbiór anotacji dla wszystkich mikro-zadań

W najprostszym przypadku podzbiór anotatorów A_i rozwiązujących mikro-zadanie x_i zawiera tylko jednego anotatora. Aby zwiększyć prawdopodobieństwo otrzymania prawidłowej odpowiedź dla danego mikro-zadania, możliwe jest zastosowanie podejścia, w którym jedno mikro-zadanie rozwiązywane jest przez więcej niż jednego anotatora [Su et al., 2012].

Takie podejście stanowi jeden z podstawowych mechanizmów zapewnienia wysokiej jakości danych w procesie crowdsourcingu (zob. Paragraf 1.3.4).

Sposób rozwiązywania mikro-zadań określony jest przez zleceniodawcę w ramach procedury anotacji danych (zob. Definicja 5). Procedura ta udostępniana jest anotatorom, którzy rozpoczynają pracę nad danym mikro-zadaniem. Procedura anotacji oprócz opisu kroków oznaczania danych może zawierać również przykłady rozwiązań oraz odpowiedzi na często zadawane pytania.

Definicja 5 (Procedura anotacji danych)
Procedura anotacji danych to instrukcja, która w precyzyjny sposób definiuje kroki prawidłowej anotacji mikro-zadania.

4. Agregacja

W czwartym kroku procesu crowdsourcingu wykonywana jest agregacja anotacji stworzonych w poprzednim kroku (zob. Rysunek 1.1). Zleceniodawca pobiera dane z systemu i agreguje je – uzyskuje finalne rozwiązanie dla każdego mikro-zadania. Proces agregacji odbywa się według następującej procedury (zob. Procedura 2):

Procedura 2: Procedura agregacji anotacji

Niech :

$\mathcal{D}$ – zbiór treści mikro-zadań,

AG – wybrana funkcja agregująca,

Y – zbiór anotacji dla wszystkich mikro-zadań

Kroki :

1 Dla każdego $x_i \in \mathcal{D}$:

- Obliczana jest finalna odpowiedź $\hat{y}_i \in \hat{Y}$ dla treści mikro-zadania x_i:
  $\hat{y}_i = AG(Y_i)$, gdzie a Y_i to zbiór zebranych anotacji dla treści mikro-zadania x_i.

Wyjście:

$\hat{Y}$ – Finalny zbiór zagregowanych odpowiedzi dla wszystkich mikro-zadań

Szczegóły implementacji powyższej procedury mogą różnić się w zależności od typu agregowanych anotacji oraz od preferencji zleceniodawcy. Jednym z najprostszych sposób agregacji anotacji, które zawierają dane liczbowe, jest obliczenie ich średniej (zob. Przykład 5). Bardziej złożone algorytmy agregacji anotacji zostały omówione w drugim rozdziale niniejszej rozprawy (zob. Podrozdział 2.3).

Przykład 5 (Agregacja anotacji liczbowych)
Dla pewnego mikro-zadania x_i uzyskany został zbiór anotacji $Y_i = \{10, 12, 12, 11\}$. Jeżeli jako funkcję agregującą AG wybrana zostanie średnia arytmetyczna, to:

$$y_i = \frac{\sum_{j=0}^{n} Y_{ij}}{n} = \frac{45}{4} = 11{,}25$$

gdzie n to liczba elementów zbioru Y_i.

5. Ewaluacja

W ostatnim kroku procesu crowdsourcingu zleceniodawca dokonuje ewaluacji jakości pozyskanych danych oraz płatności za wykonaną pracę (zob. Rysunek 1.1). Ewaluacja może dotyczyć zarówno oceny jakości finalnego zbioru danych, jak również oceny samych anotatorów. Na tym kroku zleceniodawca może zdecydować się uzależnić kwotę wynagrodzenia wypłacanego anotatorom od jakości ich pracy. Jakość anotatorów może być oszacowana na przykład na podstawie losowej próbki oznaczeń wykonanych przez danego anotatora. Szczegóły dotyczące mechanizmów ewaluacji w procesie crowdsourcingu zostały opisane w dalszej części tego rozdziału (zob. Paragraf 1.3.3).

1.2. Sposoby optymalizacji procesu crowdsourcingu

W tej części niniejszej rozprawy przedstawiłem przegląd aktualnych badań związanych z optymalizacją procesu crowdsourcingu. Celem jest zarysowanie ogólnego stanu badań prowadzonych w celu rozwoju metody crowdsourcing, a także umiejscowienie prowadzonych przeze mnie eksperymentów w szerszym kontekście.

W ramach pracy Li et al. [2016] zaproponowana została klasyfikacja, która wyróżnia trzy podstawowe grupy eksperymentów nad optymalizacją procesu oznaczania danych w metodzie crowdsourcing:

– optymalizacja kosztu,

– optymalizacja czasu,

– optymalizacja jakości.

Mimo że każda ze wspomnianych grup skupia się bezpośrednio tylko na jednym problemie procesu oznaczania danych, to w praktyce proponowane rozwiązania obejmują więcej niż jedną grupę. Poniżej szczegółowo omówiłem każdą z wymienionych grup.

1.2.1. Koszt wykonania zadania

Najprostszy sposób opłacania pracy anotatorów w metodzie crowdsourcingu polega na ustaleniu stałej kwoty, którą anotator otrzyma za oznaczenie pojedynczego mikro-zadania. Koszt wykonania pojedynczego mikro-zadania jest zazwyczaj niewielki. W zależności od stopnia skomplikowania wykonywanej pracy zleceniodawcy decydują się na ustalenie kwoty wynagrodzenia w przedziale od kilku centów do kilku dolarów [Li et al., 2016][10]. Mimo że opłata za wykonanie pojedynczego mikro-zadania jest niewielka, to ze względu na rozmiar całego zbiorów danych, optymalizacja kosztów procesu crowdsourcingu jest jednym z kluczowych tematów badań w tej dziedzinie.

Ustalenie odpowiedniej opłaty może mieć również bezpośredni wpływ na skuteczność procesu oznaczania danych. Wyższa opłata może zachęcić do uczestnictwa większą liczbę anotatorów i dzięki temu umożliwić szybsze zakończenie procesu pozyskiwania danych. Z kolei zbyt niska opłata może zniechęcić anotatorów i wydłużyć proces oznaczania danych. Wnioski z eksperymentów przeprowadzonych przez Li et al. [2016] wskazują, że zwiększenie opłaty może pozytywnie wpływa na jakość pozyskiwanych danych. Efekt ten zachodzi jednak tylko do pewnego progu, powyżej którego zwiększenie opłaty nie przekłada się już na wzrost jakości danych [Li et al., 2016].

Autorzy istniejących eksperymentów dotyczących redukcji kosztów procesu crowdsourcingu opisują metody związane z optymalizacją metod doboru anotatorów, formy i treści mikro-zadań, a także optymalizacją działania samego systemu używanego do anotacji [Li et al., 2016, s.6-7].

Przykładem sposobu optymalizacji kosztów procesu anotacji jest zastosowanie algorytmów dedukcji rozwiązania dla części mikro-zadań. Podejście to stosowane jest w sytuacjach, gdy możliwe jest określenie podobieństwa pomiędzy elementami wejściowego zbioru danych. Pogrupowanie powiązanych mikro-zadań zmniejsza ilość redundantnej pracy potrzebnej do oznaczenia całego zbioru [S. Wang et al., 2015]. Takie podejście wiąże się jednak z ryzykiem pogorszenia jakości pozyskanych danych. W przypadku gdy anotator dokonał błędnego rozwiązania danego mikro-zadania, błędna anotacja zostanie rozpropagowana do powiązanych mikro-zadań.

Innym rozwiązaniem, które jest stosowanie w celu redukcji kosztów procesu crowdsourcingu, jest zmniejszenie liczby anotatorów równocześnie oznaczających zbiór danych. Wstępna selekcja anotatorów, której celem jest uzyskanie grupy o wyższych kwalifikacjach (np. wybór anotatorów o wysokiej skuteczności lub wysokiej szybkości pracy) może

[10]Przykładowo, mediana kosztu wykonania jednego mikro-zadania w systemie *Amazon Mechanical Turk* to 0,35 $ czyli około 1,30 zł; `https://worker.mturk.com/`, dostęp: 31.12.2020 r.

w pozytywny sposób wpłynąć na redukcję kosztów oznaczenia całego zbioru [Boutsis & Kalogeraki, 2014].

W przypadku gdy zleceniodawca używa własnego interfejsu anotacyjnego, możliwe jest zastosowanie mechanizmu, który w sposób dynamiczny oblicza liczbę anotacji niezbędnych do uzyskania finalnej odpowiedzi o odpowiedniej jakości dla danego mikro-zadania. W mechanizmie zaproponowanym przez Franklin et al. [2013], proces anotacji kontynuowany jest tylko wtedy gdy, rozszerzanie zbioru danych przekłada się na wyraźny wzrost jakości zbioru.

1.2.2. Czas zbierania danych

Eksperymenty związane z optymalizacją czasu trwania procesu crowdsourcingu obejmują zarówno rozwiązania mające na celu obniżenie czasu anotacji, jak również estymację czasu potrzebnego do stworzenia pełnego zbioru danych.

Czas wykonywania mikro-zadań może mieć wpływ zarówno na koszt pracy anotatorów, jak i jakość tworzonych danych. Poprzez ograniczenie czasu wykonania pojedynczego zadania skracany jest czas pracy anotatorów, a tym samym czas potrzebny na zakończenie całego procesu anotacji. Jednak jakość anotacji, które tworzone są w krótszym czasie, może być niska. Autorzy przeprowadzonych eksperymentów starają się opracować rozwiązania, które pozwalają na optymalizację czasu anotacji bez znacznego obniżenia jakości pozyskanych danych.

Jednym z podstawowych podejść estymacji czasu oznaczania danych w metodzie crowdsourcingu jest prowadzenie procesu anotacji w turach. Podczas każdej tury zlecane jest oznaczenie tylko wydzielonej części zbioru danych. Po tym, jak opublikowane w danej turze mikro-zadania zostaną zakończone, zleceniodawca rozpoczynana kolejną turę [Li et al., 2016]. W ten prosty sposób zleceniodawca może wykorzystać informację o czasie trwania minionej tury do estymacji czasu potrzebnego do wykonania kolejnych tur.

Bardziej rozbudowane metody estymacji czasu trwania procesu crowdsourcingu tworzone są na podstawie modeli statystycznych. Przykładem algorytmu stosującego to podejście jest rozwiązanie zaproponowane przez Faridani et al. [2011], w którym czas wykonania mikro-zadania oszacowywany jest za pomocą modelu proporcjonalnego hazardu (ang. *Cox proportional hazard*) znanego ze statystycznej analizy przeżycia (ang. *survival analysis*). Precyzyjna estymacja czasu trwania zadania pozwoliła autorom na osiągnięcie poprawę dystrybucji zlecanych mikro-zadań w ciągu dnia.

Istnieją również badania, które wykazują, że systemowe wydłużenie czasu oznaczania danych może pozytywnie wpłynąć na ich jakość. Przykładowo, Rzeszotarski et al. [2013] w swojej pracy opisują interfejs, który nadzoruje czas pracy anotatora. Po zakończeniu określonej liczby mikro-zadań system blokuje anotatorowi dostęp do interfejsu na krótki czas, tak aby wymusić w jego pracy krótkie przerwy. Takie podejście pozwoliło badaczom poprawić jakość pozyskanych danych.

1.2.3. Jakość danych

Podstawowym celem badań związanych z jakością danych w procesie crowdsourcingu jest zapewnienie jak najwyższej jakości danych. Jednym z głównych tematów badań dotyczących jakości danych jest opracowywanie algorytmów agregacji zebranych anotacji. Algorytmy te używają anotacji uzyskanych w procesie crowdsourcingu w celu uzyskania finalnych oznaczeń o wyższej jakości. Inne eksperymenty związane z optymalizacją jakości danych obejmują również tworzenie algorytmów, które służą ewaluacji i kontroli jakości pracy anotatorów, projektowanie optymalnego interfejsu anotacji lub nawet optymalizacji działania całej platformy crowdsourcingowej [Li et al., 2016, s.3-6].

1.2.4. Podsumowanie

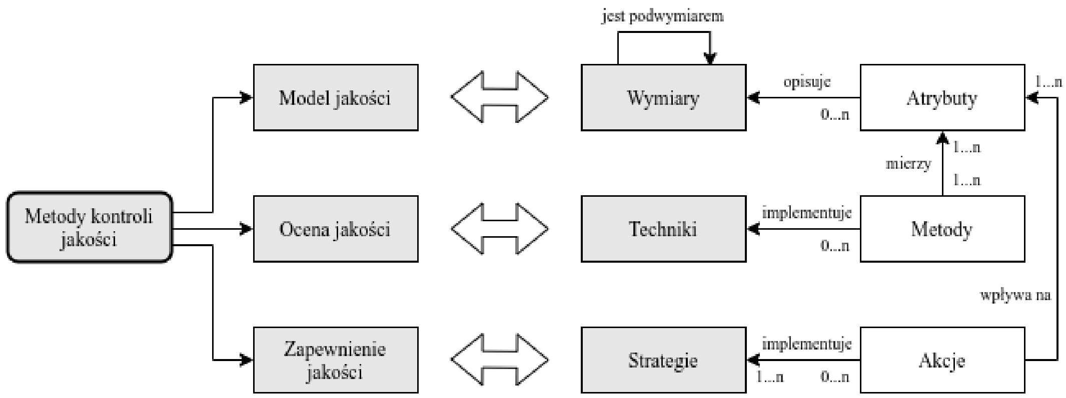
Podsumowując, badania dotyczące optymalizacji procesu crowdsourcingu dotyczą: optymalizacji kosztu wykonania zadania, czasu procesu anotacji oraz jakości pozyskanych danych. W mojej pracy skupiłem się na problemach związanych z jakością danych. W związku z tym w kolejnym podrozdziale (zob. Podrozdział 1.3) szczegółowo scharakteryzowałem metody ewaluacji danych oraz zapewnienia ich jakości.

1.3. Kontrola jakości danych

W tym podrozdziale omówiłem charakterystykę metod kontroli jakości za pomocą taksonomii zaproponowanej przez Daniel et al. [2018], która wyszczególnia trzy bazowe komponenty kontroli jakości w metodzie crowdsourcing: model jakości, ocena jakości oraz zapewnienie jakości. W niniejszej rozprawie przybliżyłem ogólny opis tej taksonomii oraz komponenty definiujące taksonomię, wartości komponentów, a także przykłady implementacji i rozwiązań używanych w istniejących platformach crowdsourcingowych.

1.3.1. Taksonomia kontroli jakości danych

W pracy Daniel et al. [2018] przedstawiona została rozbudowana taksonomia kontroli jakości danych w procesie crowdsourcingu. Jako główne składowe taksonomii wyróżnione są: model jakości, ocena jakości oraz zapewnienie jakości (zob. Rysunek 1.2[11]).



Rysunek 1.2: Komponenty metod kontroli jakości danych i ich wewnętrzna struktura (źródło: [Daniel et al., 2018])

Model jakości

Komponent *model jakości* zdefiniowany jest w oparciu o wymiary oraz atrybuty jakości:

– **Wymiary** – opisują elementy, z których składa się mikro-zadnie, takie jak: dane wejściowe, dane wyjściowe lub osoby zaangażowane w proces. Pomiar jakości wymiarów odbywa się za pomocą atrybutów.

– **Atrybuty** – obejmują cechy mikro-zadania; mogą przyjmować formę skonkretyzowaną lub abstrakcyjną:

  – **Atrybuty skonkretyzowane** to takie, które można zmierzyć bezpośrednio, np.: jakość anotacji, poziom wiedzy anotatorów.

  – **Atrybuty abstrakcyjne** to takie, których nie można zmierzyć bezpośrednio. Atrybuty te mierzone są w sposób pośredni poprzez powiązane z nimi atrybuty skonkretyzowane. Przykładem atrybutów abstrakcyjnych są operacje agregacji (np. operacja obliczania średniej odpowiedzi dla mikro-zadania). Jakość takiej operacji mierzona jest w sposób pośredni poprzez jakość danych (atrybuty skonkretyzowane) powstałych w wyniku agregacji.

[11] O ile nie zaznaczono inaczej, wszystkie tłumaczenia cytowanych rysunków i tabel zostały wykonane przez autora niniejszej rozprawy.

Ocena jakości

Komponent *ocena jakości* obejmuje metody, które stosowane są w celu oceny atrybutów wybranego modelu jakości. Ocena jakości odbywa się za pomocą technik i metod oceny, które rozumiane są w następujący sposób:

- **Techniki** – definiują podmiot, który jest odpowiedzialny za wykonanie oceny. Przykładowo, ocenę może wykonać pojedyncza osoba (np. zleceniodawca oceniający wybrany podzbiór anotacji) lub grupa osób (np. w przypadku gdy odpowiedź dla mikrozadania wybierana jest na podstawie głosowania grupy anotatorów).

- **Metody oceny** – opisują dokładny sposób mierzenia jakości atrybutów. Przykładowo, dokładność (ang. *accuracy*) danych może być obliczona poprzez porównanie oznaczonych danych ze zbiorem referencyjnym, a poziom ekspertyzy anotatora może zostać określony w oparciu o test kwalifikacyjny.

Zapewnienie jakości

Komponent ten określa zbiór akcji, które powinny zostać wykonane w celu zapewnienia oczekiwanej jakości danych. Każda akcja realizuje jedną z możliwych strategii.

- **Akcje** – opisują podstawowe operacje implementujące wybraną strategię wykonywaną w celu uzyskania oczekiwanej jakości pozyskiwanych danych (podjęte akcje wpływają na atrybuty określone przez model jakości). Przykładowo, przeprowadzenie szkolenia dla anotatorów może zostać zrealizowane poprzez wyświetlanie filmu szkoleniowego przed rozpoczęciem anotacji.

- **Strategie** – opisują wysokopoziomowe decyzje, które są podejmowane w celu poprawy jakości. Strategie określają akcje, które powinny zostać podjęta oraz moment, kiedy powinno to nastąpić. Przykładowo, w celu zapewniania wysokiej jakości danych zleceniodawca może dokonać selekcji tylko dobrych anotatorów (np. za pomocą testu kwalifikacyjnego) albo może przeprowadzić szkolenie anotatorów, którzy mają pracować nad zadaniem.

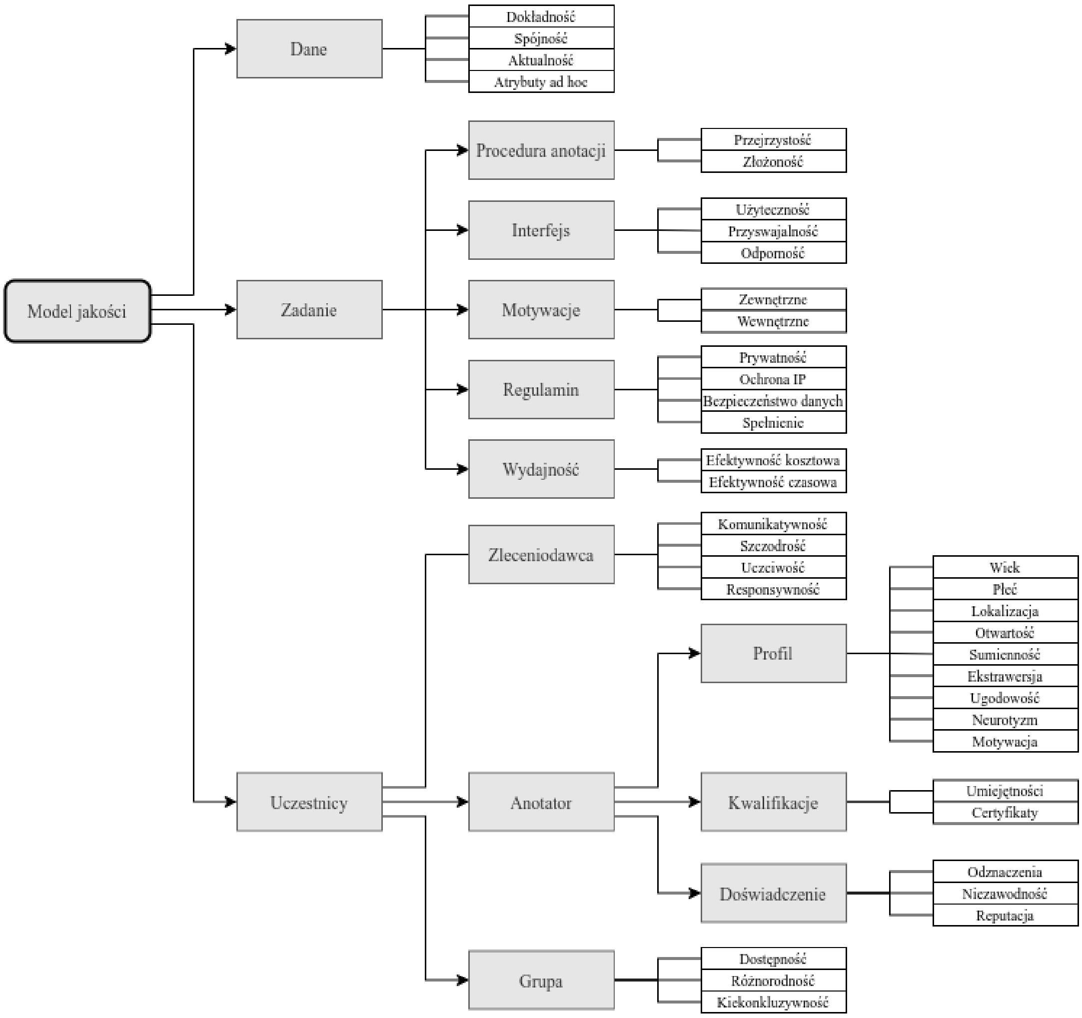
Wybór konkretnej akcji może zostać uzależniony od pomiarów atrybutów określonych przez model jakości.

Interakcja pomiędzy komponentami

Kontrola jakości w procesie crowdsourcingu oparta jest na interakcji pomiędzy powyżej opisanymi komponentami (zob. Rysunek 1.2). Model jakości określa atrybuty, które opisują przyjęte wymiary definiujące jakość danych. Atrybuty obliczane są za pomocą metod oceny jakości realizujących wybraną technikę. Po dokonaniu oceny jakości danych wykonywane są konkretne akcje, mające na celu poprawę jakości. Wybór konkretnych akcji dokonywany jest przez aktualnie realizowaną strategię na podstawie wartości wybranych atrybutów. Po zakończeniu danej akcji możliwa jest ponowna ocena danych w celu weryfikacji efektów zastosowanej akcji.

1.3.2. Model jakości

Pierwszym komponentem taksonomii kontroli jakości w procesie crowdsourcingu jest model jakości (zob. Rysunek 1.2). W tej części pracy omówiłem dokładniej zdefiniowane w taksonomii wymiary modelu jakości oraz odpowiadające im atrybuty. Podstawowe wymiary modelu jakości obejmują: *dane*, *zadanie* oraz *uczestników* (zob. Rysunek 1.3).



Rysunek 1.3: Diagram wymiarów modelu jakości (zaciemnione bloki) i odpowiadających im atrybutów (jasne bloki) (źródło: [Daniel et al., 2018])

Dane

Wymiar ten obejmuje *dane*, które pojawiają się w procesie anotacji zbioru danych. Są to dane wejściowe, które potrzebne są do rozpoczęcia zadania (np. zbiór zdań, które powinny zostać przetłumaczone, zbiór plików graficznych, które będą kategoryzowane) oraz dane dodatkowe (np. zbiór możliwych kategorii dla danych wyjściowych). Wymiar ten obejmuje również dane wyjściowe, które powstaną w wyniku anotacji. W taksonomii kontroli jakości wyszczególnione zostały cztery atrybuty określające wymiar danych:

- dokładność (ang. *accuracy*, metryka określające poprawność danych wyjściowych),
- spójność (ang. *consistency*, np. współczynnik zgodności wielu anotacji [Eickhoff et al., 2012]),
- aktualność (ang. *timeliness*, np. czas wykonywania mikro-zadania),

– atrybuty *ad-hoc* (pozostałe atrybuty, zazwyczaj charakterystyczne dla danego zadania).

Zadanie

Wymiar *zadanie* obejmuje elementy związane z implementacją procesu oznaczania wybranego zbioru danych. Wymiar ten podzielony został na pięć podwymiarów:

– procedura anotacji (np. atrybuty dotyczące przejrzystości opisu procedury oznaczania danych),

– interfejs użytkownika (wygląd interfejsu, jego łatwość użycia i użyteczność),

– motywacje (np. opis sposobu, w jaki anotatorzy zachęcani są do wykonywania zadania),

– regulamin (np. dokument opisujący między innymi ochronę prywatności anotatorów),

– wydajność (np. wskaźniki związane z kosztem i czasem tworzenia oznaczeń).

Konkretne decyzje podjęte podczas implementacji interfejsu anotacji mają bezpośredni wpływ na jakość pozyskanych danych [Marcus et al., 2012]. Atrakcyjna forma zadania może również zachęcić większą liczbę anotatorów do pracy nad zadaniem. Możliwe jest również użycie mechanizmów grywalizacji (które występują zamiast motywacji finansowej) jako formy motywacji antotorów [Eickhoff et al., 2012; Poesio et al., 2013].

Uczestnicy

Wymiar *uczestnicy* obejmuje wszystkie osoby zaangażowane w proces anotacji, czyli zarówno osoby odpowiedzialne za tworzenie zadań, jak i te, które je rozwiązują. Wymiar ten został podzielony na trzy podwymiary:

– zleceniodawca (zawiera w sobie atrybuty opisujące warianty współpracy ze zleceniodawcą, takie jak uczciwość czy szczodrość),

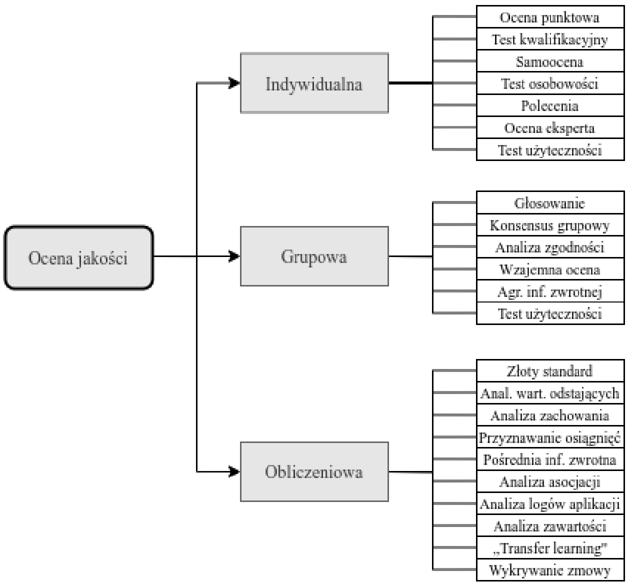
– anotator (opisuje jakość anotacji tworzonych przez anotatora, jego kwalifikacje oraz dane osobowe),

– grupy (określa sposoby analizy anotatorów w podgrupach np.: jedną grupę mogą stanowić wszyscy anotatorzy, który rozwiązywali to samo mikro-zadanie).

1.3.3. Ocena jakości

Kolejnym komponentem taksonomii kontroli jakości jest *ocena jakości* (zob. Rysunek 1.2). Komponent ten obejmuje techniki, które mogą zostać użyte w celu ustalenia jakości elementów występujących w procesie crowdsourcingu (np.: ocena poprawności danych wyjściowych, ocena zgodności pomiędzy anotatorami, ocena użyteczności interfejsu platformy crowdsourcingowej). W ramach każdej techniki wyszczególnione zostały konkretne metody, które ją realizują (zob. Rysunek 1.4).

Ocena jakości może odbywać się w sposób automatyczny, na podstawie algorytmu lub w sposób manualny przez samego zleceniodawcę, grupę ekspertów bądź innych anotatorów. Metody oceny jakości mogą zostać podzielone na techniki: *indywidualne*, *grupowe* oraz *obliczeniowe*. W przypadku technik indywidualnych i grupowych ocena dokonywana jest w sposób manualny, natomiast w przypadku technik obliczeniowych ocena jest automatyczna.



Rysunek 1.4: Diagram technik ocen jakości (zaciemnione bloki) oraz odpowiadających im metod ich realizacji (jasne bloki) (źródło: [Daniel et al., 2018])

Mimo że nie jest to objęte przez taksonomię Daniel et al. [2018], w ramach niniejszej rozprawy zaproponowałem dodatkowy podział metod realizujących techniki oceny jakości. Podział ten wyszczególnia element procesu crowdsourcingu, który jest oceniany w danej metodzie. Stworzony przeze mnie podział obejmuje trzy grupy metod oceny jakości:

– **metody oceny danych** – służą do oceny jakości stworzonych anotacji,

– **metody oceny anotatorów** – używane są do oceny samych anotatorów, ich kwalifikacji, zachowania na platformie oraz jakości ich pracy[12],

– **metody oceny platformy** – używane do oceny elementów platformy wykorzystywanego do anotacji.

Na podstawie przeprowadzony przeze mnie przegląd literatury dotyczącej istniejących implementacji metod oceny jakości w procesie crowdsourcingu uznałem, że wprowadzenie dodatkowego podziału pozwoli na ich precyzyjniejszą klasyfikację.

Poniżej przedstawiłem przykłady implementacji różnych metod oceny jakości dla każdej z trzech technik. Dodatkowo przykłady metod opisane w ramach każdej techniki zostały przeze mnie podzielone w zależności od grupy metod oceny jakości, do których należą.

Indywidualna ocena jakości

Technika *indywidualna ocena jakości* obejmuje metody, w których ocena jakości wykonywana jest przez pojedynczą osobę. Ocenę może wykonać zarówno zleceniodawca, wyznaczony ekspert lub inny anotator.

W przypadku metod służących do oceny danych często stosowanym rozwiązaniem jest użycie *oceny punktowej* (ang. *rating*) [Dalvi et al. 2013; Sakurai et al. 2013]. W ramach tej metody osoba dokonująca oceny przypisuje stworzonej anotacji wartość na podstawie ustalonej wcześniej skali [Dalvi et al., 2013]. Ocenę punktową można również przyznać opierając się na: skali binarnej (w której ocena jest pozytywna albo negatywna), skali porządkowej (w której dopuszczalne oceny pochodzą z dyskretnego zbioru dopuszczalnych punktów) lub skali ciągłej (w której możliwe jest użycie dowolnej wartości z ustalonego zakresu).

Ocena punktowa stosowana jest również w metodach oceny jakości pracy poszczególnych anotatorów (oszacowanie jakości pracy anotatora możliwe jest poprzez ocenę próbki stworzonych przez niego anotacji) [Sakurai et al., 2013]. Indywidualne metody oceny anotatora mogą również odbywać się za pośrednictwem testów kwalifikacyjnych lub testów osobowości. Testy kwalifikacyjne używane są w celu weryfikacji wiedzy domenowej anotatora. Zazwyczaj mają one formę pytań zamkniętych, co ułatwia automatyczną weryfikację danych [Alonso et al., 2008]. Możliwe jest również użycie testów osobowości lub innych testów, które nie są związane z wiedzą domenową danego zadania. Testy te używane są

[12]Zauważam, że użyty tu termin *jakość anotatora* jest pojęciem szerszym niż *jakość pracy anotatora*. *Jakość pracy anotatora* dotyczy jakości stworzonych przez niego anotacji, natomiast *jakość anotatora* oprócz jakości pracy obejmuje dodatkowo ogół cech, które opisują samego anotatora np. jego deklarowane kwalifikację.

do pozyskania ogólnych informacji o anotatorze, takich jak cechy jego charakteru (np. sumienność) lub deklarowane kwalifikacje (np. znajomość języków) [Kazai et al., 2012].

W celu oceny jakości samej platformy crowdsourcingowej możliwe jest przeprowadzenie testów użyteczności (ang. *usability test*). Testy tego typu stosowane są w celu weryfikacji czy stworzony interfejs używany jest przez użytkowników systemu w zamierzony sposób. Innym podejściem jest przeprowadzenie testów porównujących skuteczność różnych interfejsów stworzonych dla tego samego zadania [W. Willett et al., 2012].

Grupowa ocena jakości

Technika *grupowa ocena jakości* obejmują metody, w których decyzja o ocenie pojedynczej anotacji podejmowana jest przez grupę osób. Analogicznie jak w przypadku metod indywidualnych, ocenę mogą wykonywać zarówno zleceniodawcy (w przypadku gdy w rolę zleceniodawcy wciela się wiele osób), eksperci lub inni anotatorzy.

W przypadku metod grupowej oceny danych i metod oceny platformy możliwe jest zastosowanie tych samych metod, co w przypadku techniki indywidualnej. Przykładowo, ocena na skali punktowej może zostać użyta w sposób analogiczny, z tą różnicą, że finalna ocena stanowi agregację ocen wykonywanych przez grupę osób [Waggoner & Chen, 2014]. W przypadku metod oceny anotatora istnieją pewne wyjątki, przykładowo nie jest możliwe wykonanie testu osobowości grupowo.

Przykładem grupowej metody oceny danych jest użycie rozkładu odpowiedzi udzielonych przez anotatorów dla pojedynczego mikro-zadania do oszacowania jakości finalnej odpowiedzi. W sytuacji, gdy mikro-zadania zostały oznaczone przez więcej niż jednego anotatora[13], częstość wybrania najpopularniejszej odpowiedzi użyta jest do zdefiniowania prawdopodobieństwa poprawności oznaczeń [Kuncheva et al., 2003; Cao et al., 2012]. Istnieją również metody, które pozwalają na grupową ocenę zgodności dla wszystkich anotacji poprzez zastosowanie metryk zgodności anotacji stworzonych dla całego zbioru danych wyjściowych. Zastosowanie metryk zgodności (np. współczynnik *kappa Fleiss'a* [Fleiss et al., 1969]) dostarcza informacji o tym, jak często anotatorzy wybierali te same odpowiedzi dla tych samych mikro-zadań.

Zbiór, w którym mikro-zadania zostały oznaczone przez więcej niż jednego anotatora, może również zostać użyty do obliczenia grupowej oceny konkretnego anotatora. Jakość pracy anotatora obliczana jest poprzez porównanie jego anotacji z odpowiedziami innych anotatorów [Huang & Fu 2013; Eickhoff et al. 2012].

[13]Takie podejście stanowi jeden z podstawowych mechanizmów zapewnienia wyżej jakości w procesie crowdsourcingu (zob. Paragraf 1.3.4).

Obliczeniowa ocena jakości

Technika *obliczeniowa ocena jakości* obejmuje metody, za pomocą których dokonuje się oceny w sposób automatyczny w oparciu wybrany algorytm, bez potrzeby angażowania ludzi. Jedno z najbardziej podstawowych rozwiązań realizujących tę technikę polega na użyciu *złotego standardu*, czyli referencyjnego zbioru zawierającego wcześniej zweryfikowane odpowiedzi dla podzbioru mikro-zadań [Huang & Fu, 2013].

Obliczeniowa ocena jakości danych używana jest między innymi przy wyborze algorytmu agregacji anotacji i ewaluacji finalnego zbioru danych. Przykładowo, wybrany algorytm agregacji anotacji ewaluowany jest poprzez porównanie zagregowanych danych wyjściowych z dopowiedziami referencyjnymi, które pochodzą ze *złotego standardu* [Bu et al. 2019; Fornaciari et al. 2020].

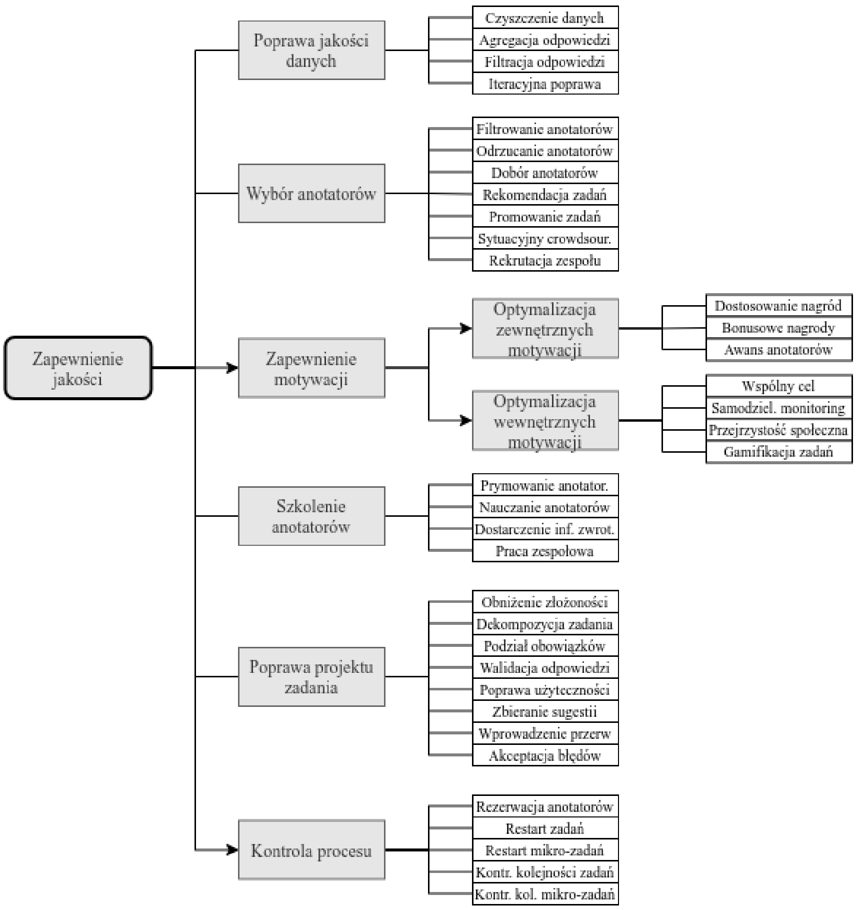
Złoty standard używany jest również w metodach oceny anotatorów. W tej metodzie anotatorzy w czasie pracy rozwiązują pewien ustalony zbiór dodatkowych mikro-zadań należących do *złotego standardu*. Jakość anotatora obliczana poprzez porównanie jego odpowiedzi z odpowiedziami referencyjnymi [Donmez et al., 2009]. Innym podejściem jest dokonanie automatycznej oceny jakości anotatorów poprzez analizę logów platformy crowdsourcingowej lub logów opisujących zachowanie anotatora w interfejsie anotacji. W ten sposób możliwe jest wykrycie szkodliwych anotatorów, którzy próbują oszukać zleceniodawcę (np. stosując narzędzia do automatycznego wprowadzania losowych odpowiedzi) [Rzeszotarski & Kittur, 2011].

Możliwe jest również zastosowanie techniki *obliczeniowej oceny jakości* w metodach oceny jakości platformy. Dudley et al. [2019] opisują algorytm automatycznej optymalizacji interfejsu anotacyjnego. Algorytm automatycznie ewaluuje aktualną skuteczność interfejsu na podstawie analizy zachowania anotatorów. Następnie algorytm wprowadza w interfejsie zmiany w celu zwiększenia jego skuteczności.

1.3.4. Zapewnienie jakości

Finalnym komponentem taksonomii kontroli jakości jest *zapewnienie jakości* (zob. Rysunek 1.2). Komponent ten opisuje wysokopoziomowe strategie, które mogą zostać wybrane w celu poprawy jakości pozyskiwanych danych. Strategie te realizowane są przez akcje, które mają postać konkretnych algorytmów bądź czynności wpływających bezpośrednio na dane, anotatorów lub platformę używaną do anotacji (zob. Rysunek 1.5).

W ramach taksonomii kontroli jakości wyszczególnione zostało sześć strategii zapewnienia jakości: poprawa jakości danych, wybór anotatorów, zapewnienie motywacji, szkolenie anotatorów, poprawa projektu zadania oraz kontrola procesu. Poniżej przedstawiłem przykłady akcji realizujących każdą z sześciu strategii.



Rysunek 1.5: Diagram strategii zapewnienie jakości (zaciemnione bloki) oraz realizujących je akcji (jasne bloki) (źródło: [Daniel et al., 2018])

W zależności od sytuacji, w której dane strategie są używane, mogą być one dodatkowo podzielone na:

– **strategie reaktywne** – dla których akcje aktywowane są przez zewnętrzne bodźce (np. po obliczeniu oceny jakości anotatorów możliwe jest wykluczenia anotatorów o niskiej jakości),

– **strategie proaktywne** – które nie są aktywowane przez bodźce zewnętrzne (np. odpowiednie zaprojektowanie interfejsu anotacji).

Poprawa jakości danych

Jedną z podstawowych metod stosowanych w celu podwyższenia jakości pozyskiwanych danych jest agregacja anotacji. W takim podejściu jedno mikro-zadanie rozwiązywane jest przez więcej niż jednego anotatora, a finalna odpowiedzi dla danego mikro-zadania wybierana jest na podstawie głosowania większościowego. Metoda ta bazuje na założeniu, że odpowiedź uzyskana jako efekt agregacji wielu anotacji pozwala uzyskać wzrost jakości. Założenie to było potwierdzone przez szereg niezależnych eksperymentów i stanowi jedno z podstawowych sposobów poprawy jakości danych stosowanych w procesie crowdsourcingu [Littlestone & Warmuth, 1994; Kuncheva et al., 2003; Cao et al., 2012; Ho et al., 2013].

W bazowym wariancie tego rozwiązania, odpowiedzi każdego anotatora mają taką samą wagę w głosowaniu większościowym [Kuncheva et al., 2003; Cao et al., 2012]. Istnieją również bardziej złożone warianty tej metody przypisujące anotatorom wagi na podstawie jakość ich pracy [Littlestone & Warmuth, 1994; Ho et al., 2013]. Opis implementacji algorytmów głosowania większościowego oraz innych algorytmów agregacji danych przedstawiłem w Paragrafie 2.3.2.

Wybór anotatorów

Kolejną strategią zapewnienia jakości danych jest *wybór anotatorów*. Akcje realizujące tę strategię mogą zostać użyte do wyeliminowania tzw. *spamerów* (czyli osób celowo wprowadzających dane złej jakości) lub anotatorów tworzących oznaczenia o najniższej jakości [Ipeirotis et al., 2010; Raykar & Yu, 2012].

Możliwe jest również zastosowanie akcji, które realizują omawianą strategię w odwrotny sposób – do wybrania najbardziej wykwalifikowanych anotatorów. W literaturze wyszczególnione zostały dwa alternatywne podejścia przypisywania zadań do anotatorów o najwyższych predyspozycjach [Daniel et al., 2018]:

– **podejście bazujące na anotatorach** (ang. *worker-based*) – zawierające metody, które dla danego typu zadania określają grupę anotatorów najlepiej do niego dopasowanych [Zhao et al., 2013; Zheng et al., 2015],

– **podejście bazujące na zadaniu** (ang. *task-based*) – zawierające metody, które dla zadanego anotatora wybierają zadania, do których jest najlepiej dopasowany [Boim et al., 2012].

Jednym z mechanizmów używanych do wyboru anotatorów jest system reputacji stanowiący przykład wyszczególnionego wyżej podejścia bazującego na anotatorach. Reputacja stosowana jest zarówno do oceny pracy anotatorów (definiowana jest wtedy np. jako procent zaakceptowanych mikro-zadań), jak i zleceniodawców (w takim przypadku reputacja zleceniodawcy może zależeć np. od liczby niesłusznie odrzuconych mikro-zadań). Zleceniodawca może ustalić minimalny próg reputacji, który anotator musi osiągnąć by móc rozpocząć pracę nad zadaniem. Oprócz systemu reputacji, platformy crowdsourcingowe udostępniają również możliwość filtrowania anotatorów pochodzących z wybranego kraju lub znających wybrany język (funkcjonalność ta dostępna jest np. na platformie *Amazon Mechanical Turk*. `https://www.mturk.com/`; dostęp: 05.06.2021 r.).

Zapewnienie motywacji

Jednym z kluczowych elementów motywujących anotatora do tworzenia wysokiej jakości oznaczeń jest oferowanie mu odpowiedniej nagrody za wykonaną pracę. Temat ten jest bezpośrednio powiązany z opisywanym wcześniej problemem optymalizacji kosztów w procesie crowdsourcingu (zob. Paragraf 1.2.1).

Oprócz nagród materialnych niektóre projekty crowdsourcingowe motywują anotatorów do pracy poprzez próbę wpłynięcia na ich wewnętrzną motywację. Przykładem wewnętrznej motywacji jest wspólny cel, który dzielą razem ze zleceniodawcami [Prestopnik & Crowston, 2012]. Jednym z projektów, który stosuje taką strategię jest *Galaxy Zoo*[14], w którym użytkownicy przypisują zdjęcia galaktyk do odpowiednich kategorii [K. W. Willett et al., 2013]. Możliwość wzięcia udziału w ciekawym projekcie jest często wystarczającą motywacją do oznaczania danych. Wielu anotatorów decyduje się na nieodpłatną pracę, gdy projekt jest wystarczająco ciekawy.

Innym przykładem wykorzystania elementów grywalizacji do zwiększenia zaangażowania anotatorów jest platforma *Sprawdzamy Jak Jest*[15], w którym anotatorzy nieodpłatnie weryfikują dokumenty przesłane przez instytucje publiczne. Oprócz możliwości udziału w ciekawym projekcie anotatorzy są dodatkowo motywowani przez mechanizmy grywalizacji zastosowane w serwisie. Zastosowanie mechanizmów takich jak: ranking użytkowników lub otrzymywanie punktów i odznak za wykonane mikro-zadania pozwala na uatrakcyjnienie procesu anotacji.

Szkolenie anotatorów

Dzięki zapewnieniu szkolenia możliwe jest zwiększenie kwalifikacji anotatorów pracujących nad danym zadaniem. W zależności od formy i złożoności zadania szkolenia mogą przyjmować różną formę. W przypadku prostych zadań możliwe jest wprowadzenie interaktywnych szkoleń, które anotator wykonuje przed rozpoczęciem zadania [Dontcheva et

[14] `https://www.zooniverse.org/projects/zookeeper/galaxy-zoo/`; dostęp: 05.06.2021 r.
[15] https://sprawdzamyjakjest.pl/; dostęp: 05.06.2021 r.

al., 2014]. W przypadku bardziej złożonych zadań pomocne okazuje się udostępnienie anotatorom przykładowych rozwiązań mikro-zadań wykonanych przez ekspertów [Doroudi et al., 2016].

Inną metodą szkolenia jest przeprowadzenie walidacji danych oznaczonych przez anotatora i przekazanie informacji zwrotnej na temat jego pracy. Informacja zwrotna otrzymana do wykonanych mikro-zadań pozwala anotatorom dokonać walidacji własnej wiedzy, co wpływa na podwyższenie ich kompetencji. W związku z tym, że weryfikacja dużej liczby mikro-zadań jest procesem czasochłonnym, jednym z rozwiązań jest zlecenie weryfikacji stworzonych anotacji innym anotatorom w formie kolejnego mikro-zadania. Proces weryfikacji anotacji pozwala zwiększyć kompetencje zarówno osoby otrzymującej informację zwrotną, jak i tej dokonującej weryfikacji wykonanych oznaczeń [Yu et al. 2014; H. Zhu et al. 2014]. Bardziej szczegółowy opis możliwości zastosowania informacji zwrotnej w celu poprawy jakości danych został opisany w Podrozdziale 1.4, poświęconemu w całości tej tematyce.

Wprowadzenie do procesu crowdsourcingu szkoleń może być kosztowne, ponieważ zleceniodawca nie tylko musi zapłacić za mikro-zadania oznaczone przez anotatora, ale również za czas, który anotator poświęcił na szkolenie. Skuteczności tej strategi jest zatem bezpośrednio związana ze znalezieniem balansu pomiędzy edukacją anotatorów a optymalizacją kosztów zleceniodawcy.

Poprawa projektu zadania

Strategia *poprawa projektu zadania* realizowana jest przez akcje, które służą optymalizacji interfejsu anotacji lub projektu zadania (np. formatu danych czy procedury anotacji).

Eksperymenty opisane w ramach pracy Rogstadius et al. [2011] wykazały, że wysoka złożoność zadania wpływa na pogorszenie się jakości pozyskiwanych danych. Jednym ze sposobów uproszczenia złożoności mikro-zadania jest jego dekompozycja na mniejsze zadania [Bernstein et al., 2010]. Dekompozycja zazwyczaj wykonywana jest przez zleceniodawcę, możliwe jest jednak również przeniesienie odpowiedzialności za dekompozycję zadań na anotatorów. W pracy Kittur et al. [2011] przedstawiono rozwiązanie, które pozwala na dynamiczną dekompozycję zadań przez samych anotatorów. W pierwszej kolejności zadaniem anotatorów jest wyznaczenie podziałów zadania na mikro-zadania. Nowe mikro-zadania są następnie rozwiązywane przez anotatorów. Po zakończeniu anotacji następuje etap agregacji. Agregacja może odbywać się automatycznie lub w sposób manualny, wykonywany przez anotatorów.

Optymalizacja interfejsu anotacji może być związana między innymi z poprawą użyteczności samego interfejsu, dodaniem mechanizmów automatycznej walidacji wprowadzanych przez anotatorów danych lub udoskonaleniem protokołu anotacyjnego zadania. Praca Kazai et al. [2011] wykazała pozytywny wpływ poprawy projektu interfejsu na jakość pozyskanych danych. Interfejs zawierający dokładniejszy opis zadania oraz automatyczną walidację danych w pozytywny sposób wpłynął na jakość pozyskiwanych danych.

Zmiana interfejsu może dotyczyć również dodania mechanizmów sterujących pracą anotatora. Przykładowo, w pracy Rzeszotarski et al. [2013] opisany został eksperyment, w którym wprowadzone zostały obowiązkowe, krótkie przerwy pomiędzy zadaniami.

Kontrola procesu

Ostatnią strategią zapewnienia jakości jest *kontrola procesu*. Strategia ta dotyczy kontroli samego procesu zbierania danych. Jednym z kluczowych zadań tej strategii jest utrzymanie zainteresowania u najlepszych anotatorów. Przykładowo, eksperymenty przeprowadzone przez Chilton et al. [2010] wykazały, że w przypadku dużych platform crowdsourcingowych (na których wielu różnych zleceniodawców ma dostęp do tej samej puli anotatorów) anotatorzy bardziej zainteresowani są nowymi zadaniami. Korzystnym rozwiązaniem jest zatem usuwanie i ponowne dodawanie zadań.

Dodatkowo kontrola procesu obejmuje również akcje związane bezpośrednio z wyborem odpowiedniej kolejności mikro-zadań lub ich grupowaniem. Eksperymenty przeprowadzone w ramach pracy Lasecki et al. [2015] jak i te opisane w Newell & Ruths [2016] wykazały, że grupowanie mikro-zadań o podobnej treści przekłada się na wzrost jakości pozyskiwanych danych. Dzięki zmniejszeniu różnorodności kolejnych mikro-zadań anotator rzadziej zmienia kontekst pracy, co w praktyce przekłada się na wzrost jakości oznaczanych danych [Lasecki et al., 2015].

1.4. Informacja zwrotna jako narzędzie kontroli jakości

W niniejszym podrozdziale zawarty został przegląd literatury opisującej przykłady wykorzystania informacji zwrotnej w procesie pozyskiwania danych w ramach procesu crowdsourcingu. Ponadto, w podrozdziale tym przedstawiłem taksonomię, która klasyfikuje istniejące sposoby wykorzystania informacji zwrotnej ze względu na ich cechy szczególne.

Zapewnienie informacji zwrotnej podczas oznaczania danych jest jednym ze sposobów realizacji strategii *szkolenie anotatorów* z komponentu *zapewnienie jakości* pozyskiwanych danych, która została opisana w Paragrafie 1.3.4. Ze względu na powszechne użycie informacji zwrotnej w mechanizmach grywalizacji, może ona zostać również zakwalifikowana do strategii *zapewnienie motywacji* (zob. Paragraf 1.3.4). Samo wyrażenie *informacja zwrotna* jest pojęciem ogólnym, które odnosi się do dowolnej informacji mającej formę opinii lub oceny. W niniejszej rozprawie rozumienie tego terminu ograniczyłem do poniższej definicji (zob. Definicja 6).

Definicja 6 (Informacja zwrotna)
Informacja zwrotna w procesie crowdsourcingu to informacja zawierająca ewaluację wykonanego mikro-zadania. Informacja ta jest przekazywana anotatorowi za pośrednictwem systemu crowdsourcingowego.

1.4.1. Klasyfikacja informacji zwrotnej

W ramach Dow et al. [2012] zaproponowana została klasyfikacja, która określa wymiary definiujące cechy informacji zwrotnej. W niniejszej rozprawie zaproponowałem rozszerzenie powyższej taksonomii o dwa wymiary, które w mojej ocenie pozwalają na pełniejsze opisanie informacji zwrotnej. Moja modyfikacja wprowadza wymiar *moment efektu* oraz wymiar *kanał komunikacji.* (zob. Rysunek 1.6). Dodatkowo zdecydowałem się pominąć wymiar *format*, ponieważ podczas analizy literatury nie znalazłem przykładów użycia informacji zwrotnej w innym formacie niż tekstowy[16].

Wymiar	**Wartości**			
Czas	synchroniczny	asynchroniczny		
Źródło treści	ocena ekspertów	ocena innych anotatorów	automatycznie wygenerowana ocena	zbiór referencyjny
Szczegółowość	punktowa ocena	szablon oceny	otwarta forma ocen	
Liczba ocen	1:1	1:wiele	wiele:1	wiele:wiele
Moment efektu	natychmiastowy	długoterminowy		
Kanał komunikacji	bezpośredni	pośredni		

Rysunek 1.6: Klasyfikacja informacji zwrotnej (opracowanie własne na podstawie [Dow et al., 2012])

[16]Sami autorzy opisywanej klasyfikacji nie załączają opisu wymiaru *format* w swojej publikacji. Wymiar ten pojawia się jedynie na graficznej ilustracji klasyfikacji (zob. [Dow et al., 2012, s. 3]).

Czas

Wymiar *czas* definiuje moment, w którym anotator otrzymuje informację zwrotną na temat wykonanego mikro-zadania. W ramach tego wymiaru wyszczególnione zostały dwie możliwe opcje:

– **synchroniczny** – anotator otrzymuje informację zwrotną zaraz po wykonaniu mikro-zadania,

– **asynchroniczny** – informacja zwrotna przesłana jest do anotatora dopiero po pewnym czasie.

W związku, z tym że synchroniczna informacja zwrotna zostaje przekazana w krótkim odstępie czasu od zakończenia mikro-zadania, to posiada ona większy walor edukacyjny. Jest tak dlatego, że osoba ucząca się (anotator) nie wyszła jeszcze z kontekstu oznaczonego mikro-zadania i dzięki temu łatwiej jest jej odnieść przekazaną informację zwrotną do treści mikro-zadania. Z tego powodu ta forma informacji zwrotnej ma większy pozytywny wpływ na jakość pozyskiwanych danych niż podejście asynchroniczne [Dow et al., 2012].

Jednak pomimo wyraźnych korzyści związane z wykorzystaniem synchronicznego podejścia, asynchroniczne przekazywanie informacji zwrotnej jest dużo częstszym rozwiązaniem[17]. Asynchroniczne podejście daje zleceniodawcy znacznie więcej czasu na sprawdzenie wykonanych mikro-zadań i przygotowanie informacji zwrotnej. W praktyce proces ten może trwać nawet do kilku dni.

Źródło treści

Kolejnym wymiarem opisującym informacje zwrotną jest źródło jej treści. Wymiar ten może być realizowany przez jedną z następujących opcji:

– **ocenę ekspertów** – wykonywaną przez określoną wcześniej grupę ekspertów,

– **ocenę innych anotatorów** – wykonywaną w formie zadania crowdsourcingowego przez grupę anotatorów,

– **automatyczne wygenerowanie ocen** – informacja zwrotna generowana jest w sposób automatyczny bez udziału człowieka (np. za pomocą wybranego algorytmu lub zbioru reguł),

[17] W systemach takich jak *Amazon Mechanical Turk* przekazywanie informacji zwrotnej możliwe jest jedynie w asynchroniczny sposób. Zleceniodawca przekazuje informację zwrotną w momencie potwierdzania opłaty za wykonaną pracę (np. po zakończeniu anotacji całego zbioru).

- **zbiór referencyjny** – ocena odbywa się poprzez porównanie stworzonej anotacji z referencyjnymi odpowiedziami.

Treść informacji zwrotnej dla danego mikro-zadania może być stworzona w sposób manualny (przez człowieka) lub w sposób automatyczny. Metody manualne, które wymagają zaangażowania pracy ekspertów lub innych anotatorów pozwalają na osiągnięcie wysokiej jakości informacji zwrotnej. Mogą one jednak wymagać dużego nakładu czasu i pracy. Dodatkowy problem stanowi zorganizowanie procesu walidacji tak, by możliwe było zapewnienie informacji zwrotnej w formie synchronicznej[18].

Alternatywę stanowią metody automatycznego generowania treści informacji zwrotnej. Rozwiązania te pozawalają na zastosowanie synchronicznej informacji zwrotnej, ale wiążą się z innymi problemami. W przypadku użycia zbioru referencyjnego przekazywana jest informacja zwrotna o wysokiej jakości, jednak jest ona dostępna tylko dla części mikro-zadań – tych, dla których istnieją referencyjne oznaczenia. W przypadku gdy treść informacji zwrotnej jest generowana przez algorytm, może być ona dostępna dla wszystkich mikro-zadań, ale wyzwanie stanowi opracowanie takiego algorytmu, który jest w stanie wygenerować informację zwrotną o odpowiedniej jakości.

Szczegółowość

Wymiar ten określa liczbę szczegółów, które zostały przekazane w informacji zwrotnej. W procesie crowdsourcingu informacja zwrotna zawiera ocenę stworzonej anotacji. Jej treść można przyporządkować do jednej z trzech opcji:

- **ocena punktowa** – informacja zwrotna ma formę punktową ocenę anotacji zdefiniowaną w wybranej skali (np. ocena binarna, ocena w skali od 1 do 5),
- **szablon oceny** – bardziej rozbudowana forma informacji zwrotnej, zbudowana według wcześniej przygotowanego szablonu oceny (np. zawierającego pytania zamknięte pomagające w ewaluacji mikro-zadania),
- **otwarta forma oceny** – informacja zwrotna udzielana za pomocą otwartej odpowiedzi (np. szczegółowy opis popełnionych błędów); tworzona specjalnie dla każdego mikro-zadania z osobna.

W większości systemów crowdsourcingowych informacja zwrotna ma niską szczegółowość. Najpopularniejszym rozwiązaniem jest przekazywanie tylko oceny binarnej (tj. punkto-

[18]Tak jak w przypadku systemu *Shaphard* zaprezentowanego przez Dow et al. [2012].

wej), dzięki której anotator dowiaduje się, czy jego zadanie zostało zaakceptowane. Popularnym rozwiązaniem jest również przekazywanie prawidłowej odpowiedzi dla danego mikro-zadania. Jednak takie podejście stosowane jest przede wszystkim w sytuacji, w której anotator przed rozpoczęciem pracy wykonuje szkolenie zawierające mikro-zadania wybrane ze zbioru referencyjnego.

Rzadziej spotykana jest otwarta forma oceny mikro-zadania. Z uwagi na fakt, że automatyczne wygenerowanie szczegółowej ewaluacji jakości wykonanej pracy jest trudne do wykonania, forma ta wymaga dodatkowej pracy ze strony zleceniodawcy (bądź np. ekspertów, których praca również musi zostać opłacona przez zleceniodawcę). Z tego powodu rozwiązanie to stosowane jest zazwyczaj w przypadku bardziej złożonych zadań (np. tworzenia podsumowań tekstów, zob. Yu et al. [2014]).

Możliwe jest również podejście pośrednie, w którym mikro-zadania ewaluowane są według wcześniej przygotowanego szablonu ocen. Takie rozwiązanie pozwala na przekazanie pełniejszej oceny stworzonej anotacji. Użycie szablonu ocen wymaga dodatkowego nakładu pracy, lecz mniejszego niż w przypadku formy otwartej [Dow et al., 2012].

Liczba ocen

Kolejny wymiar klasyfikacji informacji zwrotnej określa liczbę indywidualnych ocen, które zostały użyte do wygenerowania informacji zwrotnej. Oceny te można przypisać do następujących kategorii:

- **1:1** – informacja zwrotna dotyczy jednego mikro-zadania, które zostaje zweryfikowane jeden raz,
- **1:wiele** – informacja zwrotna dotyczy jednego mikro-zadania, które zostaje zweryfikowane wielokrotnie,
- **wiele:1** – informacja zwrotna jest stworzona w oparciu o zagregowane oceny pochodzące z jednokrotnej oceny wielu mikro-zadań jednego anotatora,
- **wiele:wiele** – informacja zwrotna jest stworzona w oparciu o zagregowane oceny pochodzące z wielokrotnej oceny wielu mikro-zadań jednego anotatora.

Powyższe kategorie można w naturalny sposób podzielić na dwie grupy. W pierwszej grupie (pierwsze dwie kategorie) ocena w ramach informacji zwrotnej przygotowywana jest dla każdego wykonanego mikro-zadanie z osobna. Ocena może być wykonana przez jedną lub więcej osób. Zwiększenie liczby osób wykonujących ocenę zwiększa koszty procesu, ale pozwala na osiągnięcie wyżej jakości. Druga grupa (ostatnie dwie kategorie) zawiera kategorię, w których informacja zwrotna tworzona jest na podstawie oceny wielu mikrozadań jednego anotatora. Tak jak w przypadku pierwszej grupy, tak i w tym przypadku ocena również może być wykonana przez jedną lub więcej osób.

Moment efektu

Wymiar ten określa, w jakim momencie zauważalny jest efekt wprowadzenia informacji zwrotnej. Siła efektu oraz jego wpływ mogą różnić się w zależności od specyfiki prowadzonego eksperymentu i formy samej informacji zwrotnej. W literaturze przedmiotu można znaleźć badania, które wskazują na wpływ informacji zwrotnej zarówno na zwiększoną jakością pozyskiwanych danych, jak i na zwiększoną motywacją anotatorów [Horton 2010; Dow et al. 2012; Gaikwad et al. 2019].

Większość badań, które potwierdzają pozytywny wpływ przekazywania informacji zwrotnej na jakość tworzonych anotacji, przede wszystkim zwraca uwagę na edukacyjny efekt informacji zwrotnej (zob. Yu et al. [2014], Singla et al. [2014], Mamykina et al. [2016]). Jednakże eksperymenty przeprowadzone w ramach w niniejszej rozprawy wykazały, że wprowadzenie informacji zwrotnej ma efekt natychmiastowy (zob. Podrozdział 3.5) – podwyższenie jakości danych widoczne było już od pierwszych anotacji stworzonych przez danego anotatora. W przypadku gdy anotator dopiero rozpoczął pracę nad danym zadaniem, nie jest możliwe, aby zwiększona jakość anotacji była spowodowana efektem edukacyjnym. Efekt natychmiastowy musiał być więc spowodowany innymi czynnikami[19]. Obie z powyższych opcji (efekt natychmiastowy i efekt długoterminowy) nie wykluczają się i są obserwowane równocześnie z różnym natężeniem. Z tego powodu zdecydowałem się na rozszerzenie klasyfikacji Dow et al. [2012] o wymiar *moment efektu*, który zawiera dwie kategorie (zob. Rysunek 1.6):

- **natychmiastowy** – efekt, który następuje od razu po wprowadzeniu informacji zwrotnej,
- **długoterminowy** – efekt, który można zaobserwować dopiero po upływie dłuższego czasu.

[19]Zbadanie przyczyny wystąpienia efektu natychmiastowego wykracza poza zakres niniejszej rozprawy. W mojej opinii sam fakt wprowadzenia informacji zwrotnej może być dla anotatora sygnałem, że stworzone przez niego anotacje będą weryfikowane. Tym samym może być on bardziej zmotywowany do skrupulatniejszej pracy.

Kanał komunikacji

Wymiar ten określa formę, w jakiej informacja zwrotna została przekazana do anotatora. Kanał komunikacji podzielony został na dwie kategorie:

– **bezpośredni kanał komunikacji** – informacja zwrotna przekazywana anotatorowi w bezpośredni sposób odnosi się wykonanej przez niego pracy,

– **pośredni kanał komunikacji** – informacja o jakości wykonanej pracy jest przekazana w sposób pośredni (np. poprzez zmianę reputacji anotatora).

W systemach takich jak *Amazon Mechanical Turk* informacja zwrotna przekazywana jest w sposób bezpośredni (w momencie, kiedy zleceniodawca akceptuje płatność za wykonane anotacje). Istnieją również systemy, w których informacja zwrotna przekazywana jej w sposób pośredni. Popularnym sposobem implementacji takiego rozwiązania są systemy używające reputacji (zob. np. Gaikwad et al. [2019]). W takich systemach spadek poziomu reputacji u danego anotatora jest dla niego sygnałem, że powinien on poprawić jakość swojej pracy.

1.4.2. Informacja zwrotna – omówienie literatury

W niniejszym paragrafie przedstawiony został aktualny stan badań dotyczących skuteczności mechanizmu informacji zwrotnej. Opis ten został również rozszerzony o przykłady implementacji tego mechanizmu w procesie crowdsourcingu.

Skuteczność informacji zwrotnej, a teoria „przepływu”

Pozytywny wpływ informacji zwrotnej w kontekście poprawy jakości pracy oraz zwiększenia motywacji został zbadany w znacznie szerszym kontekście, wychodzącym poza badania dotyczące metody crowdsourcingu. W pracy Csikszentmihalyi [1991] natychmiastowa informacja zwrotna wymieniona została jako jeden z podstawowych warunków wymaganych do osiągnięcia stanu nazywanego „przepływem” (ang. *flow*). Stan *przepływu* opisuje moment całkowitej koncentracji i zaangażowania w wykonywane zadanie. Liczne badania empiryczne potwierdziły, że założenia teorii *przepływu* sprawdzają się w dziedzinach takich jak: sztuka i muzyka, sport, media, praca i życie osobiste, a także podczas aktywności wykonywanych *online* (zob. C. Nguyen et al. [2015]). Skuteczność informacji zwrotnej została również wykazana przez eksperymenty, które nie zajmowały się bezpośrednio teorią „przepływu”. Przykładowo, Wiggins [1998] opisał pozytywny wpływ informacji zwrotnej zapewnianej uczniom przez ich nauczyciela. Podobne wnioski zostały również potwierdzone podczas analizy wpływu zachowania pracodawców na rozwój pracowników [Lave & Wenger, 1991].

Założenia teorii *przepływu* stosowane są również w dziedzinie crowdsourcingu. Przykładowo, C. Nguyen et al. [2015] opisali implementację systemu, którego celem było utrzymanie anotatorów w stanie *przepływu* dzięki zastosowaniu natychmiastowej informacji zwrotnej oraz dostosowaniu trudności mikro-zadań do umiejętności samego anotatora. Podobne elementy stosowane są w systemach wykorzystujących mechanizmy grywalizacji. Chociaż celem wykorzystania grywalizacji nie zawsze jest wywołanie uczucia przypływu, to oczekuje się, że dołączenie do różnych zadań elementów gier pozytywnie wpłynie na zaangażowanie osób, które je wykonują oraz efektywność ich pracy. Zastosowanie mechanizmów grywalizacji w celu zapewnienia wyższej motywacji anotatorów w procesie crowdsourcingu został opisany we wcześniejszej części niniejszego rozdziału.

Informacja zwrotna w systemach crowdsourcingowych

Mechanizm informacji zwrotnej może zostać wykorzystany praktycznie we wszystkich systemach crowdsourcingowych. Przeprowadzone eksperymenty wykazały pozytywny wpływ wprowadzenia informacji zwrotnej na jakości oznaczeń wykonywanych dla danych lingwistycznych [Horton 2010; Dow et al. 2012; Yu et al. 2014; T. T. D. T. Nguyen et al. 2017] i plików graficznych [Singla et al. 2014]. Podobnych wniosków dostarczyły eksperymenty przeprowadzone dla innych zagadnień, takich jak ocena pożywności posiłków [Mamykina et al., 2016] czy wymyślanie pomysłów [Chan et al., 2016]. W przypadku tych eksperymentów informacja zwrotna również poprawiła jakość stworzonych anotacji. Poniżej przedstawiłem dokładniejszy opis tych eksperymentów.

Eksperymenty związane z oznaczaniem danych lingwistycznych dotyczyły przede wszystkim tworzenia dłuższej (np. kilkuzdaniowej) treści (zob. Dow et al. [2012], Yu et al. [2014]). Eksperymenty opisane w pracy Yu et al. [2014] dotyczyły zagadnienia tworzenia podsumowań tekstów, a eksperymenty opisane przez Dow et al. [2012] – tworzenia recenzji produktów z wybranych kategorii. W obu przypadkach anotatorzy po wykonaniu zadania otrzymywali informację zwrotną na temat swojego zadania. Wykazano, że w obu przypadkach anotatorzy otrzymujący informację zwrotną osiągali wyższą jakość tworzonych danych, niż grupa kontrolna, która po wykonaniu zadania nie otrzymała żadnej dodatkowej informacji.

Informacja zwrotna w tych eksperymentach tworzona była przez człowieka. Alternatywne podejście zastosowano w eksperymencie opisanym w ramach T. T. D. T. Nguyen et al. [2017], w którym informacja zwrotna była generowana w sposób losowy. Zadaniem anotatorów biorących udział w eksperymencie było przygotowanie treści artykułu promującego wskazanego przez zleceniodawcę produktu[20]. Eksperyment wykazało, że wpływ informacji zwrotnej na jakość danych zależy nie tylko od samej treści przekazywanej informacji, ale również od dodatkowych czynników [T. T. D. T. Nguyen et al., 2017]. Nawet gdy treść informacji zwrotnej była całkowicie losowa i nie zależała od treści anotacji, uzupełnienie informacji zwrotnej o pozytywne wyrażenia miało pozytywny wpływ na jakość tworzonych danych. W ramach tego eksperymentu analizowana była również reakcja anotatorów na otrzymaną informację zwrotną. Zgodnie z wynikami eksperymentu, czynnikiem istotnie wpływającym na odbiór informacji zwrotnej było przekazanie informacji o jej autorze. Anotatorzy znacznie lepiej odbierali informację zwrotną, której autorem jest zleceniodawca, niż gdy autorami są inni anotatorzy. Najlepiej odbierana jest anonimowa informacja zwrotna.

Bardziej formalne podejście do informacji zwrotnej zostało opisane przez Singla et al. [2014] w eksperymencie dotyczącym oznaczania plików graficznych. Opisano przebieg zalgorytmizowanego procesu nauczania anotatora. Algorytm bazuje na modelu statystycznym, którego celem było reprezentowanie aktualnego stanu wiedzy anotatora. Model ten wykorzystywany był przez algorytm w celu wyboru optymalnego zbioru przykładów uczących dla anotatora. Przykłady te wybierane były ze wcześniej przygotowanego zbioru referencyjnego, który zawierał poprawne odpowiedzi dla mikro-zadań. Za każdym razem, gdy anotator oznaczył jedno mikro-zadanie ze zbioru uczącego, otrzymywał on informację zwrotną w postaci poprawnej anotacji dla danego mikro-zadania. Następnie algorytm używał udzielonych dotychczas odpowiedzi do zaktualizowania modelu wiedzy anotatora. Proces ten stworzony był tak, by osiągnąć jak najwyższą jakość anotatora przy minimalnej liczbie przykładów. Proces zaprezentowany przez Singla et al. [2014] jest przykładem zastosowania mechanizmu tzw. *nauczania maszynowego*. Nauczanie maszynowe stanowi istotny element formalnego opisu informacji zwrotnej w procesie crowdsourcingu i dlatego mechanizm ten został obszernie opisany w następnym rozdziale. (zob. Podrozdział 2.1).

Przywołane przeze mnie powyżej eksperymenty opisują przykłady systemów, w których informacja zwrotna była wyświetlana anotatorom w sposób bezpośredni, po wykonaniu zadania (czas synchroniczny, bezpośredni kanał komunikacji; zob. Paragraf 1.4.1). Alternatywnym rozwiązaniem jest przekazywanie informacji zwrotnej w sposób pośredni –

[20]W ramach tego zadania promowanym produktem były lalki stworzone z recyklingowanych materiałów.

za pomocą mechanizmu reputacji (czas asynchroniczny, pośredni kanał komunikacji; zob. Paragraf 1.4.1). System reputacji wymaga systematycznej pracy zleceniodawców. W eksperymencie opisanym w ramach Gaikwad et al. [2019] opracowany został system *Boomerang.* W tym systemie nie tylko zleceniodawca dokonywał oceny anotatorów, ale również anotatorzy oceniali zleceniodawcę. Anotator posiadający wyższą reputację uzyskiwał wcześniejszy dostęp do nowych zadań. Dzięki temu mógł on wykonać więcej mikro-zadań niż anotatorzy z niższą reputacją. Natomiast zleceniodawcy o wysokiej reputacji mogli liczyć na większe zainteresowanie ich zadaniami ze strony dobrych anotatorów.

Systemy reputacji są również powszechnie używane w tzw. platformach *gig marketplace*, w których możliwe jest tymczasowe zatrudnianie niezależnych pracowników w celu wykonania krótkoterminowych usług (Do najbardziej znanych platform tego typu należą m.in. serwisy takie jak *Uber*[21], *Glovo*[22], *Bolt*[23]). Przykładowo, platforma *Uber* oferuję usługę transportu *crowdsourcowanego* (ang. *crowdsourced transportation*), w którym zleceniodawca zgłasza potrzebę przemieszczenia się we wskazane miejsce. Usługę wykonuje wykonawca, który zgłosi się jako pierwszy. Po zakończeniu usługi zleceniodawca może wystawić ocenę, która ma wpływ na reputację wykonawcy. Ze względu na skalę takich systemów, manualna weryfikacja wystawionych ocen nie jest możliwa. Niesłuszna ocena może pogorszyć reputację wykonawcy, a nawet zakończyć się zablokowaniem konta. Problem ten był motywacją dla eksperymentu opisanego w Toxtli et al. [2020], w którym stworzony system weryfikował słuszność wystawianych ocen, a tym samym zmniejszał liczbę niesłusznych ocen.

Eksperymenty opisane w ramach Jiang & Huang [2016] wykazały, że otwarty dostęp do automatycznej informacji zwrotnej może mieć również negatywny wpływ na jakość wykonywanej pracy. W ramach pracy Jiang & Huang [2016] analizowali konkursy crowdsourcingowe, w których uczestnicy próbowali w jak najlepszy sposób rozwiązać zadany problem (np. poprzez opracowanie odpowiedniego algorytmu). Konkursy tego typu regularnie są organizowane przez platformy takie jak *Kaggle*[24] lub *HackerRank*[25]. eksperymenty Jiang & Huang [2016] wykazały, że uczestnicy konkursu, którzy mieli nieustanny dostęp do informacji zwrotnej, dostosowywali swoje rozwiązanie tak, aby zdobyło jak najwięcej punktów według metryki oceny przekazywanej w informacji zwrotnej. Natomiast rozwiązania tworzone w konkursach, w których dostęp do informacji zwrotnej był ograniczony (i

[21] `https://www.uber.com/`; dostęp: 26.01.2023 r.
[22] `https://glovoapp.com/`; dostęp: 26.01.2023 r.
[23] `https://bolt.eu/`; dostęp: 26.01.2023 r.
[24] https://www.kaggle.com/competitions; dostęp: 05.06.2021 r.
[25] https://www.hackerrank.com/contests; dostęp: 05.06.2021 r.

odbywa się tylko w określonych momentach, np. pod koniec konkursu), okazały się lepiej generalizować wiedzę i dzięki temu działały lepiej na niedostępnych wcześniej zbiorach danych.

Podsumowanie

Przywołane w niniejszym podrozdziale eksperymenty usystematyzowałem i przedstawiłem w Tabeli 1.1. Każde z nich scharakteryzowałem za pomocą wymiarów z klasyfikacji informacji zwrotnej opisanych w Paragrafie 1.4.1.

Większość z opisanych prac badała możliwość użycia synchronicznej informacji zwrotnej (zob. Tabela 1.1, kolumna *Czas inf. zwrotnej*). Źródłem informacji zwrotnej były przede wszystkim inni uczestnicy procesu (inni anotatorzy lub eksperci) lub zbiór referencyjny (zob. Tabela 1.1, kolumna *Źródło inf. zwrotnej*). Wszystkie opisywane eksperymenty stosowały informację zwrotną o niskiej szczegółowości (użycie punktowej oceny lub szablonu oceny).

Wśród analizowanych przeze mnie eksperymentów brak jednak takich, które porównywałby skuteczność informacji zwrotnej dla różnych typów mikro-zadań, a w szczególności mikro-zadań dotyczących anotacji danych lingwistycznych. Przeprowadzono niewiele eksperymentów, które sprawdzałyby wpływ jakości przekazywanej informacji zwrotnej na jakość pozyskiwanych danych. W ramach niniejszej rozprawy opracowałem eksperyment, w ramach którego zagłębiłem oba powyższe tematy (zob. Rozdział 3).

Warto również wspomnieć, że dotychczas przeprowadzono niewiele eksperymentów, które zajmowały się automatycznym generowaniem informacji zwrotnej. Podejście to może być wykorzystywane w sytuacji, gdy niemożliwe jest użycie innych źródeł informacji zwrotnej. Jakość informacji zwrotnej pochodzącej od innych anotatorów wzrasta wraz ze zwiększeniem liczby anotatorów, którzy anotowali dane mikro-zadanie. Użycie automatycznie wygenerowanej informacji zwrotnej może być korzystne zwłaszcza w sytuacji, gdy dane mikro-zadanie nie zostało jeszcze oznaczone przez wystarczającą liczbę anotatorów i nie jest możliwe użycie grupowych metod oceny jakości (zob. Paragraf 1.3.3). W niniejszej rozprawie przedstawiłem model dynamicznej informacji zwrotnej, którego celem było podwyższenie jakości automatycznie generowanej informacji zwrotnej (zob. Rozdział 4).

Tabela 1.1: *Przegląd istniejących implementacji informacji zwrotnej w systemach crowdsourcingowych (opracowanie własne)*

Eksperyment	**Problem badawczy**	**Czas inf. zwrotnej**	**Źródło inf. zwrotnej**	**Szczegółowość inf. zwrotnej**	**Liczba ocen**	**Moment efektu inf. zwrotnej**	**Kanał komunikacji inf. zwrotnej**	**Badanie efektów inf. zwrotnej**
Bauer & Popović [2017]	Klasyfikacja struktur białka	synchroniczny	ocena innych anotatorów	punktowa ocena	wiele:1	brak danych	bezpośredni	Tak
Chan et al. [2016]	Sugerowanie pomysłów	synchroniczny	ocena ekspertów	forma otwarta	1:1	natychmiastowy	bezpośredni	Tak
Dow et al. [2012]	Napisanie recenzji produktów	synchroniczny	ocena innych anotatorów (lub ocena własna)	szablon oceny	wiele:1	długoterminowy	bezpośredni	Tak
Dziedzic [2016]	Rozpoznawanie jednostek nazwanych	synchroniczny	zbiór referencyjny i ocena innych anotatorów	punktowa ocena	1:1	brak danych	bezpośredni	Nie
Eickhoff et al. [2012]	Klasyfikacja plików graficznych i tekstu	synchroniczny	ocena innych anotatorów	punktowa ocena	wiele:1	brak danych	bezpośredni	Nie
Jiang & Huang [2016]	Projekt graficzny logotypu	zależny od warunku eksperymentu	ocena ekspertów	punktowa ocena	1:1	zależny od warunku eksperymentu	bezpośredni	Tak
Mamykina et al. [2016]	Oznaczanie wartości odżywczych w posiłkach	synchroniczny	ocena ekspertów	zależny od warunku eksperymentu	1:1	brak danych	bezpośredni	Tak
		synchroniczny	ocena innych anotatorów	zależny od warunku eksperymentu	wiele:1	brak danych	bezpośredni	Tak

Kontynuacja na następnej stronie

Tabela 1.1: *Przegląd istniejących implementacji informacji zwrotnej w systemach crowdsourcingowych (opracowanie własne)* (Kontynuacja)

T. T. D. T. Nguyen et al. [2017]	Napisanie artykułu o produkcie	synchroniczny	automatycznie wygenerowana ocena	szablon oceny	1:1	natychmiastowy	bezpośredni	Tak
Singla et al. [2014]	Klasyfikacja plików graficznych	synchroniczny	zbiór referencyjny	punktowa ocena	1:1	brak danych	bezpośredni	Tak
Gaikwad et al. [2019]	Projekt systemu	asynchroniczny	ocena ekspertów	punktowa ocena	1:wiele	długoterminowy	pośredni	Tak
Toxtli et al. [2020]	Napisanie oceny za wykonaną usługę	synchroniczny	automatycznie wygenerowana ocena	szablon oceny	1:1	natychmiastowy	bezpośredni	Tak
Yu et al. [2014]	Podsumowywanie tekstów	asynchroniczny	ocena ekspertów	otwarta forma ocen	wiele:1	długoterminowy	bezpośredni	Tak
		asynchroniczny	ocena innych anotatorów	punktowa ocena	wiele:1	długoterminowy	bezpośredni	Tak

ROZDZIAŁ 2

Modelowanie procesu informacji zwrotnej

W niniejszym rozdziale zamieściłem opis mechanizmów, które używane są do implementacji procesu przekazywania informacji zwrotnej w metodzie crowdsourcingu. W pierwszej części rozdziału omówiłem mechanizm „nauczania maszynowego”, który nadaje ustrukturyzowaną formę temu procesowi i stanowi formalną podstawę dla omawianych zagadnień. Nauczanie maszynowe zdefiniowane jest w formie iteracyjnego procesu, w którym uczeń zdobywa wiedzę dzięki informacji zwrotnej przekazywanej mu do rozwiązanych przez niego zadań.

W dalszej części niniejszego rozdziału skupiłem się na tym, jak mechanizm nauczania maszynowego może zostać włączony do procesu crowdsourcingu jako mechanizm kontroli jakości. Opisałem, jak elementy procesu nauczania maszynowego realizują trzy główne komponenty taksonomii kontroli jakości w procesie crowdsourcingu (zob. Paragraf 1.3.1):

– **modele reprezentacji ucznia**, który realizuje komponent *model jakości*,

– **metody obliczania parametrów modelu ucznia**, który realizuje komponent *ocena jakości*,

– **metody wyboru sygnałów nauczających**, który realizuje komponent *zapewnienie jakości.*

Dla każdego z komponentów przedstawiłem przykłady algorytmów, które mogą być użyte w implementacji.

2.1. Nauczanie maszynowe

Badania dotyczące formalizacji procesu nauczania maszynowego zostały zapoczątkowane zarówno przez teoretyków z dziedzin psychologii i pedagogiki (zob. Patil et al. [2014]), jak i tych związanych z matematyką (zob. Shinohara & Miyano 1991, Goldman & Kearns 1995). Jednym z głównych badaczy i popularyzatorów teorii nauczania maszynowego jest profesor Xiaojin Zhu z Uniwersytetu Wisconsin-Madison. Jego prace stanowią według mojej wiedzy najpełniejszy, dostępny opis tego zagadnienia[1]. W niniejszym rozdziale opisana została definicja oraz charakterystyka nauczania maszynowego na podstawie pracy: *An Overview of Machine Teaching* [X. Zhu et al., 2018].

Z uwagi na to, że nauczanie maszynowe można rozpatrywać jako szczególny przypadek uczenia maszynowego, opis nauczania maszynowego przedstawiłem w odniesieniu do klasycznych metod stosowanych w uczeniu maszynowym.

Uczenie maszynowe

Rozważmy problem znalezienia pewnej nieznanej funkcji docelowej $f : X \to Y$ (ang. *target function*)[2], która przyporządkowuje elementy ze zbioru danych wejściowych X do wartości ze zbioru danych wyjściowych Y. Zakłada się, że dostępny jest zbiór przykładów uczących:

$$\mathcal{D} = \{(x_1, y_1), \cdots, (x_n, y_n)\}, \text{ gdzie } y_n = f(x_n). \tag{2.1}$$

Zadaniem uczenia maszynowego jest znalezienie hipotezy g będącą przybliżeniem funkcji f obliczonym na podstawie dostępnego zbioru przykładów uczących. Proces uczenia maszynowego może zostać opisany za pomocą kroków przedstawionych w ramach Procedury 3. Kroki procesu zostały również zaprezentowane w formie graficznej na Rysunku 2.1):

[1]Zestawienie badań poświęconych nauczaniu maszynowym można znaleźć na stronie internetowej profesora Xiaojin Zhu: `https://pages.cs.wisc.edu/~jerryzhu/machineteaching/`; dostęp: 27.01.2023 r.

[2]W uczeniu maszynowym istnieje również pojęcie funkcji celu (ang. *objective function*). Mimo podobnej nazwy termin ten odnosi się on do innego rodzaju funkcji. Funkcja celu to funkcja, która optymalizowana jest w procesie uczenia [Abu-Mostafa et al., 2012, s.28].

Procedura 3: Proces uczenia maszynowego

Niech:

X – zbiór wartości wejściowych (np. $X = \mathbb{R}$),

Y – zbiór wartości wyjściowych Y (np. w przypadku problemu klasyfikacji binarnej $Y = 0, 1$),

A – wybrany algorytm uczący,

H – zbiór wszystkich możliwych hipotez H,

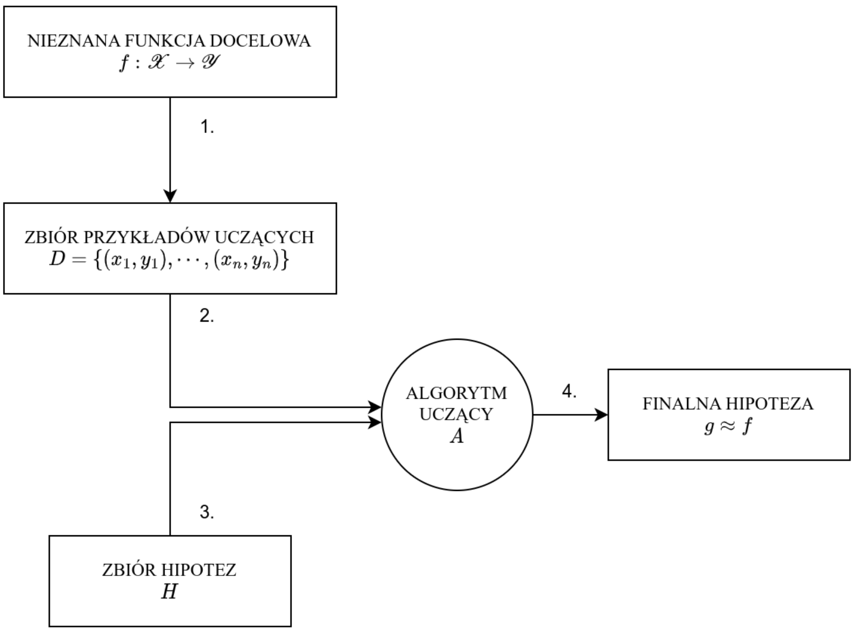
Kroki:

1 Określony zostaje zbiór przykładów uczących $\mathcal{D} = (x_1, y_1), \cdots, (x_n, y_n)$, gdzie $y_n = f(x_i)$ dla $i = 1., \cdots, n$,

2 Zbiór przykładów uczących zostaje przekazany do algorytmu uczącego A,

3 Algorytm uczący A używa zbioru przykładów uczących $\mathcal{D}$ w celu wybrania hipotezy $g : X \to Y$ będącej przybliżeniem funkcji docelowej f,

4 Po zakończeniu procesu uczenia wybrana zostaje finalna hipoteza g_D będąca najlepszym (według algorytmu $\mathcal{A}$) przybliżeniem dla funkcji f stworzonym na podstawie zbioru $\mathcal{D}$.



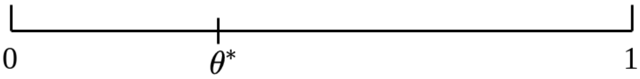
Rysunek 2.1: Proces uczenia maszynowego (źródło: [Abu-Mostafa et al., 2012])

Przykład 6

Rozważmy problem, w którym zadaniem jest zbudowanie jednowymiarowego klasyfikatora, który przyporządkowuje wartościom ze zbioru o rozkładzie jednorodnym $X \sim U(0, 1)$

wartość binarną $y \in \{0,1\}$ (zob. Rysunek 2.2). Wartość zmiennej y dla każdego $x \in X$ wyrażona jest za pomocą funkcji f zależnej od progu θ^*, która określona jest następującym wzorem:

$$y = f(x) = \begin{cases} 1, x \geqslant \theta^* \\ 0, x < \theta^* \end{cases}$$



Rysunek 2.2: Wizualizacja problemu jednowymiarowego klasyfikatora binarnego (źródło: [X. Zhu et al., 2018])

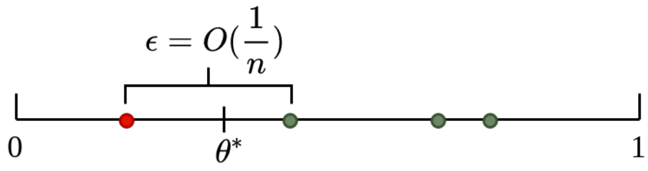
Załóżmy, że chcemy rozwiązać problem opisany w *Przykładzie 6* za pomocą uczenia maszynowego. W takim przypadku celem algorytmu uczącego będzie określenie hipotezy g zależnej od parametru $\hat{\theta}$, która jest przybliżeniem funkcji f. Wartości parametru $\hat{\theta}$ zostaje obliczona na podstawie dostępnego zbioru przykładów uczących $\mathcal{D}$. Sam proces uczenia maszynowego można uogólnić do minimalizowania błędu generalizacji[3]:

$$\min_{\hat{\theta}} ||\hat{\theta} - \theta^*|| \tag{2.2}$$

Przebieg procesu uczenia maszynowego może różnić się w zależności od wybranej strategii dostępu do przykładów uczących. Dwie podstawowe strategie to: uczenie pasywne (ang. *passive learning*) oraz uczenie aktywne (ang. *active learning*). W podejściu pasywnym zbiór przykładów uczących $\mathcal{D}$ składa się ze stałej liczby n elementów. Załóżmy, że dysponujemy modelem będącym idealnym uczniem – czyli takim, który nie popełnia błędów podczas uczenia. Biorąc pod uwagę to, że elementy zbioru $\mathcal{D}$ mają rozkład jednorodny, spodziewany błąd generalizacji takiego modelu można określić jako $||\hat{\theta} - \theta^*|| = O(n^{-1})$.

Łatwo zauważyć, że dla n elementów pochodzących z rozkładu jednorodnego średnia odległość pomiędzy elementami wynosi $\frac{1}{n}$. Taka też jest wielkość granicy decyzji, która określona jest w oparciu parę elementów: „negatywnego” (dla którego $f(x) = 0$) oraz „pozytywnego” ($f(x) = 1$) zbioru $\mathcal{D}$ będących najbliżej progu θ^*. Tym samym, by osiągnąć błąd generalizacji ϵ, zbiór $\mathcal{D}$ musi zawierać przynajmniej $n \geqslant O(\epsilon^{-1})$ (zob. Rysunek 2.3). Aby osiągnąć błąd równy $\epsilon = 0{,}001$ potrzebne jest przynajmniej $n = 1000$ elementów.

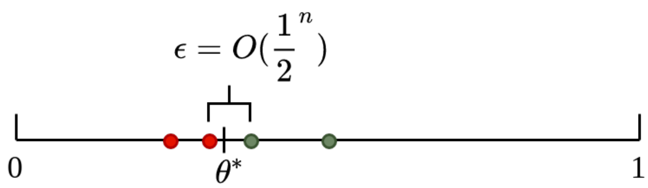
[3]Błąd generalizacji (nazywany też błędem uogólniania) jest to wartość oczekiwana błędu dla nowych danych [Geéron, 2017, s. 150].



Rysunek 2.3: Jednowymiarowy klasyfikator binarny stworzony w podejściu uczenia pasywnego (źródło: [X. Zhu et al., 2018])

Alternatywnym podejściem stosowanym w algorytmach uczenia maszynowego jest uczenie aktywne. W tym podejściu zbiór $\mathcal{D}$ nie jest określony na stałe, a sam model ma możliwość interaktywnego odpytywania tzw. „wyroczni"[4] (ang. *oracle*) o dodatkowe elementy zbioru uczącego. Model, używając odpowiedniego algorytmu, wybiera element $x \in X$, a *wyrocznia* przekazuje dla tego elementu wartość $f(x)$.

Dla rozpatrywanego wcześniej problemu klasyfikacji binarnej w jednowymiarowej przestrzeni za algorytm wyboru nowych przykładów uczących zbioru $\mathcal{D}$ można przyjąć algorytm przeszukiwania binarnego, którego celem jest znalezienie wartości progu θ^*. Z każdym kolejnym elementem algorytm zawęża pozostałą przestrzeń przeszukiwań o połowę. Błąd generalizacji takiego modelu zdefiniowany jest więc jako $||\hat{\theta} - \theta^*|| = O(2^{-n})$. Tym samym, by osiągnąć błąd generalizacji ϵ zbiór $\mathcal{D}$ musi zawierać przynajmniej $n \geqslant O(log(\epsilon^{-1}))$ (zob. Rysunek 2.4) elementów. Aby osiągnąć błąd równy $\epsilon = 0{,}001$ potrzebne jest przynajmniej $n = 10$ elementów.



Rysunek 2.4: Jednowymiarowy klasyfikator binarny stworzony w podejściu uczenia aktywnego (źródło: [X. Zhu et al., 2018])

Nauczanie maszynowe

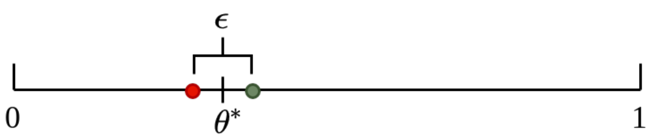
W przypadku nauczania maszynowego sytuacja ta jest odwrotna w stosunku do uczenia maszynowego. W procesie tym bierze udział dwóch uczestników:

[4]Wyrocznią w procesie uczenia aktywnego nazywane jest źródło danych, które dostarcza dane uczące. Rolę wyroczni może przyjąć np. grupa ekspertów oznaczających dane.

1. **nauczyciel** – ma informacje o prawdziwych wartościach parametrów definiujących funkcje f (np. wartość progu θ^*), jego zadaniem jest wybór najbardziej efektywnego zbioru $\mathcal{D}$ dla wybranego ucznia,

2. **uczeń** – na podstawie dostępnego zbioru $\mathcal{D}$ dokonuje on wyboru funkcji g będącej przybliżeniem funkcji f.

W sytuacji, gdy rozważamy idealnego ucznia, który nie popełnia żadnego błędu podczas uczenia, nauczyciel musi wybrać tylko dwa elementy dla zbioru $\mathcal{D}$: jeden pozytywny i jeden negatywny. Elementy te powinny definiować przedział wielkości ϵ, który zawiera θ^* na samym środku. Błąd generalizacji modelu stworzonego w oparciu o te dane równy jest ϵ. Warto zauważyć, że dla tak zdefiniowanego problemu, zbiór przykładów uczących będzie zawsze zawierał dokładnie dwa elementy, bez względu na to, jaka wielkość ϵ zostanie wybrana (zob. Rysunek 2.5).

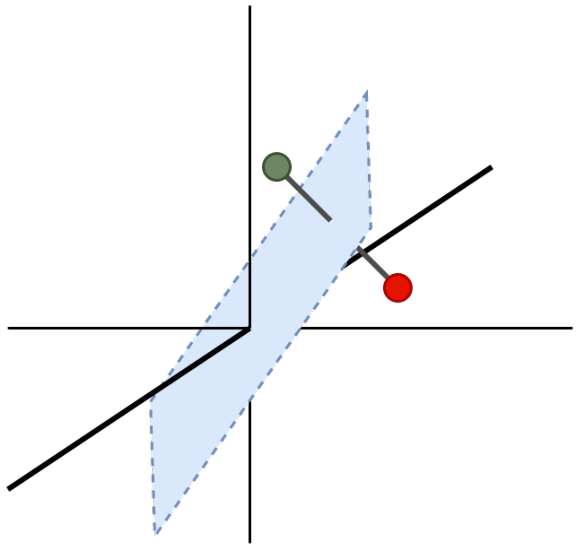


Rysunek 2.5: Jednowymiarowy klasyfikator binarny stworzony w podejściu nauczania maszynowego (źródło: [X. Zhu et al., 2018])

W przypadku problemu jednowymiarowego, określenie zbioru $\mathcal{D}$ okazało się to bardzo proste. Rozważmy teraz bardziej złożony przykład (zob. Przykład 7) znalezienia odpowiedniego zbioru przykładów uczących dla problemu o wyższym wymiarze.

Przykład 7

Rozważmy problem, w którym zadaniem jest zbudowanie d-wymiarowego klasyfikatora binarnego. Klasyfikator ten przyporządkowuje wartościom ze zbioru $X \sim \mathbb{R}^d$ wartość binarną $y \in \{0, 1\}$ (zob. Rysunek 2.6).



Rysunek 2.6: d-wymiarowy klasyfikator binarny stworzony za pomocą algorytmu maszyny wektorów nośnych (źródło: [X. Zhu et al., 2018])

W przypadku, gdy problem opisany w *Przykładzie 7* chcemy rozwiązać za pomocą podejścia nauczania maszynowego, a algorytm ucznia używa algorytmu maszyny wektorów nośnych[5] to celem nauki jest określenie granicy dla klasyfikacji binarnej w d-wymiarowej przestrzeni. W tej sytuacji, aby precyzyjnie określić położenie płaszczyzny, nauczycielowi również wystarczą dokładnie dwa elementy zbioru uczącego. Nauczyciel musi wybrać jeden z nieskończenie wielu zbiorów uczących zdefiniowanych tak, by zawierał on dwa punkty oddalone od siebie o odległość równą ϵ. Odcinek tworzony przez te punkty jest prostopadły do płaszczyzny, która przecina go dokładnie w połowie (zob. Rysunek 2.6).

Wielkość minimalnego zbioru uczącego $\mathcal{D}$ określona jest przez tzw. „wymiar nauczania" (zob. Definicję 7):

Definicja 7

Wymiar nauczania – TD (od ang. teaching dimension) – rozmiar najmniejszego wymaganego zbioru uczącego potrzebnego do nauczenia ucznia z wybraną dokładnością ϵ [X. Zhu et al., 2018, s. 3].

Tak więc $d_{TD}(H) = 2$, gdzie H to zbiór możliwych hipotez dla zadanego problemu. Wymiar nauczania określony jest jako analogia do wymiaru VC (*Vapnik-Chervonenkis*) zdefiniowanego jako wielość największego zbioru, który dzielony jest przez przestrzeń hipotez H (zob. Vapnik [1998]). W tej sytuacji wymiar VC równy jest $d_{VD}(H) = d + 1$.

[5] Algorytm maszyn wektorów nośnych (ang. *support vector machines*) to algorytm, którego celem jest określenie najszerszego możliwego marginesu dla hiperpłaszczyzny, która oddziela zbiory punktów. Dokładny opis działania algorytmu można znaleźć np. w Kecman [2005].

2.1.1. Proces nauczania maszynowego

Poniżej przedstawiłem charakterystykę procesu nauczania maszynowego. Opisany przeze mnie proces bazuje na opracowaniu przygotowanym przez X. Zhu et al. [2018], jednak proces ten został przeze mnie dostosowany tak, aby lepiej pasował do metody crowdsourcingu. Główną zmianą zaproponowaną przeze mnie jest wprowadzenie dodatkowego komponentu – modelu ucznia.

Podobnie jak w przypadku uczenia maszynowego, celem nauczania maszynowego jest stworzenie jak najdokładniejszego przybliżenia funkcji docelowej f. W uczeniu maszynowym funkcja docelowa f jest nieznana, a jej przybliżenie obliczane jest przez algorytm uczący A w oparciu o zbiór przykładów uczących $\mathcal{D} = \{(x_1, y_1), \cdots, (x_n, y_n)\}$, gdzie $y_n = f(x_n)$. W nauczaniu maszynowym funkcja docelowa f jest z góry znana przez algorytm nauczyciela (algorytm nauczający) A_n. Algorytm nauczyciela steruje iteracyjnym procesem uczenia algorytmu A_u. W każdej i-tej iteracji procesu nauczania maszynowego zadaniem algorytmu nauczyciela jest stworzenie „sygnału nauczającego" (zob. Definicja 8) będącego podzbiorem przykładów uczących $\mathcal{D}_i \subset \mathcal{D}$, który pozwoli algorytmowi ucznia na jak najdokładniejsze przybliżenie funkcji docelowej f.

Definicja 8 (Sygnał nauczający)

Sygnał nauczający to pojedynczy podzbiór przykładów uczących $\mathcal{D}_i \subset \mathcal{D}$ przekazywany przez nauczyciela do ucznia w i-tej iteracji procesu nauczania maszynowego. Elementy sygnału uczącego mają postać par (x_i, y_i), w których x to dane wejściowe dla nauczanego problemu, a y to dane wyjściowe z prawidłowym rozwiązaniem.

Sygnał nauczający przekazywany jest w dwóch krokach. W pierwszym kroku przekazywany jest tylko zbiór danych wejściowych $X_i = \{x_1, \cdots, x_n\}$, dla których algorytm ucznia generuje zbiór oznaczeń $\hat{Y}_i = \{\hat{y}_1, \cdots, \hat{y}_n\}$ na podstawie swojego aktualnego stanu wiedzy g_{i-1}. Następnie algorytm nauczający przekazuję zbiór prawidłowych danych wyjściowych $Y_i = \{y_1, \cdots, y_n\}$. Algorytm ucznia wykorzystuje przekazane sygnały uczące w celu aktualizacji swojego stanu wiedzy, czyli wybranie nowej hipotezy g_i ze zbioru wszystkich hipotez H.

Warto zauważyć, że ponieważ nauczanie maszynowe może być również użyte w celu opisania procesu nauczania człowieka, w praktyce bardzo często nie jest możliwe określenie dokładnej formy, w jakiej tworzona jest prawdziwa hipoteza ucznia g[6]. Z tego powodu zdecydowałem się na wprowadzenie komponentu model ucznia M_u (zob. Definicja 9).

[6] W przypadku gdy rolę ucznia pełni człowiek (jak ma to miejsce np. gdy nauczanie maszynowe używane jest w procesie crowdsourcingu) dokładne odwzorowanie hipotezy g wymagałoby pełnego zrozumienia procesu uczenia się człowieka.

Model ucznia M_u stanowi jedynie przybliżenie prawdziwej hipotezy ucznia g, które zostało zbudowane w oparciu o zbiór oznaczeń ucznia $\hat{Y}$. Przykłady implementacji modelu ucznia w procesie nauczania maszynowego zostaną szerzej opisane w dalszej części tego rozdziału (zob. Podrozdział 2.2).

Definicja 9 (Model ucznia)
Model matematyczny, którego zadaniem jest jak najdokładniejsze odwzorowanie obecnego stanu wiedzy ucznia na podstawie danych pozyskanych w procesie nauczania maszynowego.

Proces nauczania maszynowego opisany został za pomocą kroków przedstawionych w ramach Protokołu 4. Kroki procesu zostały również zaprezentowane w formie graficznej na Rysunku 2.7):

Procedura 4: Proces nauczania maszynowego

Niech:

$f : X \to Y$ – nieznana funkcja docelowa,

X – zbiór wartości wejściowych (np. $X = \mathbb{R}$),

Y – zbiór wartości wyjściowych Y (np. dla klasyfikacji binarnej $Y = \{0, 1\}$),

A_n – wybrany algorytm nauczający (algorytm nauczyciela),

A_u – wybrany algorytm uczący (algorytm ucznia),

H – zbiór wszystkich możliwych hipotez,

Kroki:

1 Określony zostaje zbiór wszystkich przykładów uczących

$\mathcal{D} = \{(x_1, y_1), \cdots, (x_n, y_n)\}$, gdzie $y_n = f(x_i)$ dla $i = 1., \cdots, n$.

2 Zbiór przykładów uczący zostaje przekazany do algorytmu nauczającego A_n.

3 Algorytm nauczający A_n rozpoczyna i-tą iterację nauczania. Algorytm A_n wykorzystuje obecny model ucznia M_u^{i-1} w celu wybrania sygnału uczącego $\mathcal{D}_i = \{X_i, Y_i\}$, gdzie $X_i = \{x_1, \cdots, x_n\}$ oraz $Y_i = \{y_1, \cdots, y_n\}$.

4 Algorytm A_u otrzymuje sygnał uczący $\mathcal{D}_i$:

– A_n przekazuje zbiór danych wejściowych X_i,

– A_u generuje zbiór oznaczeń $\hat{Y}_i = \{\hat{y}_1, \cdots, \hat{y}_n\}$,

– A_n przekazuje zbiór prawidłowych danych wyjściowych Y_i.

5 Algorytm A_u używa zaktualizowanego zbioru przykładów uczących $\mathcal{D}_u = D_u \cup D_i$ w celu wybrania nowej hipotezy ze zbioru hipotez H.

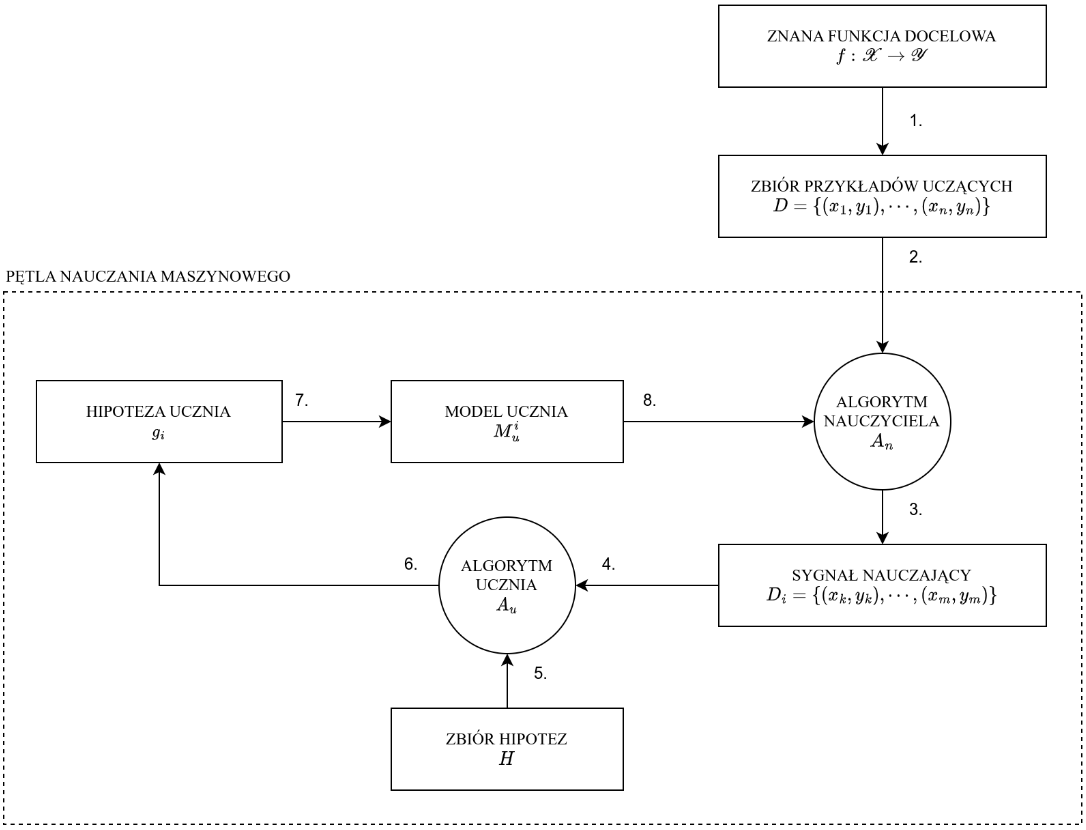
6 Algorytm A_u wybiera nową hipotezę ucznia g_i.

7 W oparciu o zbiór oznaczeń ucznia $\hat{Y}$ aktualizowany jest model ucznia M_u^i stanowiący przybliżenie hipotezy ucznia g_i.

8 Model ucznia M_u^i przekazywany jest do algorytmu nauczyciela A_n.

Korki od 3 do 8 powtarzane są aż do spełnienia wybranego warunku stopu[a].

[a] Możliwe warunki stopu zależą od strategii wybranej przez algorytm nauczyciela. Warunek stopu może np. określać maksymalną wielkość zbioru przykładów uczących bądź maksymalny akceptowalny błąd modelu wiedzy ucznia.



Rysunek 2.7: Proces nauczenia maszynowego (opracowanie własne)

2.1.2. „Problem nauczyciela" i „problem ucznia"

Kluczowymi elementami powyższego procesu są: algorytm nauczyciela A_n oraz algorytm ucznia A_u. Zadaniem tych algorytmów jest kolejno: wybór optymalnego zbioru przykładów uczących $\mathcal{D}_u$ (w przypadku algorytmu nauczyciela) oraz użycie zbioru $\mathcal{D}_u$ do stworzenia hipotezy ucznia g będącej jak najlepszym przybliżeniem funkcji docelowej f (w przypadku algorytmu ucznia).

Rozważmy szczególny przypadek nauczania maszynowego, w którym nauczyciel ma pełną wiedzę o wiedzy ucznia, a model ucznia idealnie odwzorowuję hipotezę ucznia: $M_u = g$. Zgodnie z X. Zhu et al. [2018], w takiej sytuacji uogólniona postać nauczania maszynowego może być sformułowana jako optymalizacja poniższych dwóch formuł:

$$\begin{aligned} \min_{D,g} \; & RyzykoNauczania(g) - \eta \; KosztNauczania(D_u), \\ & \text{gdzie } \hat{g} = UczenieMaszynowe(D_u) \end{aligned} \tag{2.3}$$

„Problem nauczyciela"

Górne wyrażenie definiuje „problem nauczyciela" (zob. Wzór 2.3), który składa się z dwóch komponentów: $RyzykoNauczania$ i $KosztNauczania$. $RyzykoNauczania$ jest miarą skuteczności nauczania i określa stopień „niezadowolenia" nauczyciela. Optymalizacja tego komponentu związana jest z doprowadzeniem hipotezy ucznia g jak najbliżej do funkcji docelowej f.

Załóżmy, że funkcja f to funkcja liniowa zdefiniowana przez parametry θ^*, a hipoteza g to funkcja liniowa definiowana przez parametry $\hat{\theta}$. To $RyzykoNauczania$ może być definiowane jako:

$$RyzykoNauczania(g) = ||g - f|| = ||\hat{\theta} - \theta^*||^2 \tag{2.4}$$

Drugim komponentem definiującym „problem nauczyciela" jest $KosztNauczania$ (zob. Wzór 2.3). Określa on ogólnie rozumiany „koszt" związany z użyciem zbioru uczącego $\mathcal{D}_u$. W najprostszym przypadku, $KosztNauczania$ związany jest bezpośrednio z wielkością zbioru $\mathcal{D}_u$:

$$KosztNauczania(D_u) = ||D_u||_0 \tag{2.5}$$

Powyżej opisane formuły stanowią jedynie podstawową formę *problemu nauczyciela*. Zgodnie z X. Zhu et al. [2018], mogą one zostać rozszerzone według potrzeb. Przykładowo, możliwe jest rozszerzenie wzoru $KosztNauczania$ tak, aby dodatkowo uwzględniał złożoność elementów zbioru $\mathcal{D}_u$ (liczoną np. jako długość tekstu). Jak również przeformułowanie komponentu $RyzykoNauczania$ żeby był zdefiniowany bez użycia parametrów funkcji docelowej θ^*. Czyli chociażby jako błąd generalizacji hipotezy ucznia g obliczony w oparciu o zbiór przykładów uczących $\mathcal{D}_u$ [X. Zhu et al., 2018, s. 5].

„Problem ucznia"

Dolne wyrażenie Wzoru 2.3 definiuje „problem ucznia", na który składa się komponent $UczenieMaszynowe$. Problem ten polega na optymalizacji wybranego algorytmu uczenia maszynowego, który używany jest do wyboru hipotezy ucznia g przybliżającej funkcję docelową f. Nauczyciel zna zarówno obecny stan wiedzy ucznia, jak i algorytm uczący, który został użyty.

Załóżmy, że hipoteza ucznia g to funkcja liniowa definiowana przez parametry $\hat{\theta}$. Komponent $UczenieMaszynowe$ może być zdefiniowany jako:

$$\hat{\theta} = \underset{\theta \in \Theta}{argmin} \sum_{(x,y) \in D_u} \ell(x, y, \theta) + \lambda ||\theta||^2$$

Alternatywne postaci nauczania maszynowego

W zależności od tego, który komponent jest kluczowy w danej implementacji procesu nauczania maszynowego, możliwe jest wprowadzenie alternatywnej formy opisanych wyżej formuł (zob. Wzór 2.3). Przykładowo, załóżmy sytuację, w której kluczowym elementem jest minimalizacja kosztów związanych z nauczaniem przy jednoczesnym utrzymaniu akceptowalnego stanu wiedzy ucznia. Niech minimalny próg, który określa akceptowalny poziom wiedzy ucznia zdefiniowany jest przez ϵ. W takim przypadku nauczanie maszynowe może zostać przeformułowane do następującej formy:

$$\min_{D, \hat{\theta}} KoszNauczania(D_u), \tag{2.6}$$

gdzie

$$\begin{aligned} RyzykoNauczania(g) &< \epsilon \\ \hat{\theta} &= UczenieMaszynowe(D_u). \end{aligned} \tag{2.7}$$

Innym przykładem może być sytuacja, w którym celem nauczania maszynowego jest optymalizacja komponentu $RyzykoNauczania$ przy utrzymaniu kosztów nauczania poniżej pewnego określonego progu B. W takim przypadku nauczanie maszynowe może zostać przeformułowane do następującej formy:

$$\min_{D, \hat{\theta}} RyzykoNauczania(g), \tag{2.8}$$

gdzie

$$\begin{aligned} KoszNauczania(D_u) &\leqslant B \\ g &= UczenieMaszynowe(D_u). \end{aligned} \tag{2.9}$$

2.1.3. Charakterystyka nauczania maszynowego

W niniejszym paragrafie przybliżyłem charakterystykę procesu nauczania maszynowego. W pierwszej kolejności omówiłem główne zastosowania metod nauczania maszynowego, a następnie opisałem osiem głównych wymiarów, które mogą być użyte do dokładniejszego opisu konkretnych implementacji procesu nauczania maszynowego w praktyce.

X. Zhu et al. [2018] w swoim artykule wyszczególnił dwa główne zastosowania dla metod nauczania maszynowego:

– **Edukacja** – nauczyciel-edukator zna cel edukacyjny, który może być określony jako funkcja docelowa. Zadaniem nauczyciela jest przekazanie tej wiedzy uczniom. Jeżeli nauczyciel jest w stanie sformułować kognitywny model procesu uczenia się uczniów, to może użyć nauczania maszynowego w celu wybrania odpowiedniego zbioru uczącego. Przykładowo, nauczyciel próbuje nauczyć studentów medycyny jak rozpoznawać chorobę płuc. Dzięki zastosowaniu nauczania maszynowego wybiera optymalny zbiór uczący zawierający zdjęcia RTG, dzięki którym uczniowie szybciej nauczą się sami rozpoznawać nieprawidłowości na zdjęciu.

– **Atak** – nauczyciel-napastnik celowo zakłóca działanie modelu poprzez wprowadzenie szkodliwych przykładów uczących. Przykładowo, algorytmy działające w systemach wykrywających *e-maile* zawierające spam automatycznie dostosowują swoje działanie do treści *e-maili*, które otrzymuje użytkownik. Nauczyciel-napastnik przygotowuje odpowiednie wiadomości, które stopniowo zakłócają działanie algorytmu, aby przepuszczał on niektóre niepożądane wiadomości.

W pracy X. Zhu et al. [2018] wyszczególnione zostało również osiem wymiarów, które mogą być użyte do scharakteryzowania procesu nauczania maszynowego:

– Uczestnicy procesu,

– Sygnał nauczający,

– Granulacja sygnału nauczającego,

– Znajomość hipotezy ucznia,

– Świadomość bycia nauczanym,

– Liczba uczniów,

– Intencja nauczania,

– Cel nauczania.

Poniżej zamieściłem dokładniejsze omówienie każdego z tych wymiarów.

Uczestnicy procesu

Pierwszy wymiar nauczania maszynowego *uczestnicy procesu* służy do opisania uczestników procesu. Ponieważ obie role uczestników nauczania maszynowego (nauczyciel i uczeń) mogą być pełnione zarówno przez człowieka, jak i przez maszynę, to w ramach tego wymiaru wyszczególnione zostały cztery możliwe wartości:

- Maszyna naucza maszynę,
- Maszyna naucza człowieka,
- Człowiek naucza maszynę,
- Człowiek naucza człowieka.

Sygnał nauczający

Wymiar *sygnał nauczający* definiuje rodzaj sygnału przekazywanego w procesie nauczania maszynowego. Jako podstawowe rodzaje przekazywanego sygnału nauczającego można wyróżnić:

- **Sygnał predefiniowany** – nauczyciel wybiera elementy ze zdefiniowanego wcześniej zbioru uczącego,
- **Sygnał syntetyczny** – nauczyciel generuje syntetyczne elementy. Rozwiązanie to sprawdza się zwłaszcza w przypadku, gdy przestrzeń, w której zdefiniowane są te elementy, jest ciągła,
- **Sygnał hybrydowy** – nauczyciel używa rozwiązania, które łączy w sobie oba powyższe. Przykładem może być użycie automatycznego modyfikowania predefiniowanego zbioru obrazów poprzez wprowadzenie syntetycznych zniekształceń lub transformacji.

Granulacja sygnałów nauczających

Wymiar *granulacja sygnałów nauczających* definiuje liczbę sygnałów nauczających przesyłanych jednorazowo do ucznia. W ramach tego wymiaru wyszczególnione zostały dwie możliwe wartości, które różnią się przede wszystkim momentem przekazania informacji zwrotnej do ucznia:

- **Sygnały sekwencyjne** – w tym podejściu nauczyciel przesyła do ucznia zawsze tylko jeden sygnał. Uczeń otrzymuje informację zwrotną na temat swojej pracy od razu po rozwiązaniu danego zadania. Takie podejście pozwala nauczycielowi wpływać na proces uczenia nie tylko przez wybór zbioru uczącego, ale również kolejności przesyłanych sygnałów.

– **Sygnały grupowane** – w przypadku tego podejścia nauczyciel przesyła więcej niż jeden sygnał jednocześnie. Uczeń rozwiązuje wszystkie problemy z przesłanego zbioru w dowolnej kolejności. Po oznaczeniu danych przesyła je do nauczyciela, a nauczyciel przesyła informację zwrotną do wszystkich zadań równocześnie.

Znajomość hipotezy ucznia

Wymiar *znajomość hipotezy ucznia* określa stopień wiedzy nauczyciela na temat hipotezy ucznia, a tym samym formy, jaką przyjmuje model ucznia. W ramach tego wymiaru zdefiniowane są dwa skrajne przypadki znajomości modelu ucznia:

– **Pełna wiedza** – nauczyciel zna dokładny model, który określa obecny stan wiedzy ucznia oraz parametry definiujące przebieg proces uczenia. Oznacza to, że model ucznia używany przez nauczyciela jest tożsamy z hipotezą ucznia ($M_u = g$).

– **Brak wiedzy** – nauczyciel nie wie, w jaki sposób zamodelowana jest wiedza ucznia. Jedyne informacje na temat aktualnego stanu wiedzy zdobywa na podstawie odpowiedzi ucznia udzielanych po przesłaniu sygnału uczącego. W takiej sytuacji nauczyciel używa modelu ucznia (M_u) będący jedynie przybliżeniem rzeczywistego stanu wiedzy ucznia (czyli hipotezy ucznia g).

Świadomość bycia nauczanym

Wymiar *świadomość bycia nauczanym* służy do opisania tego, czy uczeń wie, że jest częścią procesu nauczania oraz czy zna cel tego procesu. W ramach wymiaru wyszczególnione zostały dwie podstawowe wartości:

– **Uczeń świadomy** – uczeń wie, że jest częścią procesu nauczania. Przykładem są uczestnicy szkoleń, którzy rozumieją, że przekazywane sygnały uczące mają im pomóc w nauce.

– **Uczeń nieświadomy** – uczeń nie wie, że jest częścią procesu nauczania. Przykładowo, w przypadku ataku na model uczenia maszynowego, model nie jest świadomy, że nowe sygnały uczące mogą zaszkodzić jakości jego działania.

Liczba uczniów

Wymiar *liczba uczniów* określa liczbę uczniów, których równocześnie naucza nauczyciel. W ramach tego wymiaru wyróżnione zostały dwa przypadki: nauczyciel naucza jednego ucznia lub nauczyciel naucza wielu uczniów:

– **Jeden uczeń** – nauczyciel naucza tylko jednego ucznia,

– **Wiele uczniów** – nauczyciel równocześnie naucza wielu uczniów.

Nauczyciel wpływa na stan wiedzy ucznia przede wszystkim przez ustalanie zbioru przykładów uczących $\mathcal{D}_u$. Dlatego przypadek, w którym nauczyciel przygotowuje osobny zbiór przykładów uczących dla każdego ucznia, jest analogiczny do sytuacji, w której nauczany jest tylko jeden uczeń.

W przypadku, gdy nauczyciel naucza wielu uczniów, możliwa jest sytuacja, w której poziom wiedzy różni się pomiędzy uczniami. Jeżeli nauczyciel chce użyć tylko jednego zbioru przykładów uczących dla wszystkich uczniów, to potrzebne jest przeformułowanie ogólnej postaci nauczania maszynowego (zob. Paragraf 2.1.2) tak, aby brała pod uwagę wielu uczniów. Poniżej przedstawiłem dwie przykładowe modyfikacje zaproponowane w ramach X. Zhu et al. [2018]:

Pierwsza z możliwych modyfikacji zakłada minimalizację $RyzykaNauczania$ dla najgorszego ucznia:

$$\underset{D,\{\hat{g}_\lambda\}}{min} \int_\lambda f(\lambda)\ RyzykoNauczania(\hat{g}_\lambda)\ d\lambda - \eta\ KoszNauczania(D_u), \tag{2.10}$$

gdzie

$$\hat{g}_\lambda = UczenieMaszynowe_\lambda(D_u), \tag{2.11}$$

a parametr λ określa identyfikator dla każdego z uczniów. Innym podejściem może być zastosowanie metody Bayesowskiej optymalizacji ryzyka w celu optymalizacji średniego $RyzykoNauczania$ dla wszystkich uczniów. Takie podejście wymaga określenia rozkładu *a priori* $f(\lambda)$ (np. rozkładu jednorodnego).

$$\underset{D,\{\hat{g}_\lambda\}}{min}\ \underset{\lambda}{max}\ RyzykoNauczania(\hat{\theta}_\lambda) - \eta\ KoszNauczania(D), \tag{2.12}$$

gdzie

$$\hat{g}_\lambda = UczenieMaszynowe_\lambda(D_u). \tag{2.13}$$

Intencja nauczyciela

Wymiar *intencje nauczyciela* określa intencję nauczyciela. Czyli tego, czy jego działania mają na celu wpłynięcie (pozytywnie lub negatywnie) na działanie modelu. Wymiar ten zawiera dwie skrajne wartości [X. Zhu et al., 2018, s. 11]:

– **Intencja pozytywna** – proces nauczania ma pozytywny wpływ na działanie modelu. Przykładami są procesy, których celem jest: edukacja, optymalizacja modelu uczenia maszynowego, badanie modeli kognitywnych.

– **Intencja negatywna** – proces nauczania ma negatywny wpływ na działanie modelu. Przykładem takiego procesu jest atak na model uczenia maszynowego w celu ominięcia zabezpieczeń.

Cel nauczania

W ramach wymiaru *cel nauczania* definiowany jest cel prowadzenia procesu nauczania maszynowego. Jako podstawowe cele nauczania można wyszczególnić [X. Zhu et al., 2018, s. 11]:

– **Teoretyczny cel nauczania** – głównym celem procesu nauczania jest przeprowadzenie eksperymentów.

– **Praktyczny cel nauczania** – głównym celem procesu nauczania jest poprawienie wiedzy uczniów.

W praktyce większość procesów nauczania maszynowego łączy w sobie obie te wartości z różnym natężeniem.

2.1.4. Nauczanie maszynowe w crowdsourcingu

Nauczanie maszynowe jest stosunkowo niszową dziedziną, a możliwość zastosowania tego podejścia w kontekście crowdsourcingu nie została jeszcze w pełni przebadania. Istniejąca literatura opisuje przede wszystkim wyniki eksperymentów, w których nauczanie maszynowe zastosowane jest w celu zwiększenia ekspertyzy anotatorów pracujących nad danym zadaniem (zob. Zhou et al. [2018], P. Wang et al. [2021b], Yang et al. [2017]). Większość publikowanych wyników dotyczy rozwiązań związanych z anotacją plików graficznych. Przykładowo, Zhou et al. [2018] opisuje eksperyment, w którym nauczanie maszynowe zostało użyte w projekcie przypisywania odpowiednich kategorii do zdjęć zwierząt.

Artykuł ten opisuje algorytm, który rozszerza model ucznia o dodatkowy komponent symulujący stopniowy zanik wiedzy.

W artykule P. Wang et al. [2021b] zaprezentowany został algorytm nauczania, w ramach którego selekcja zbioru nauczającego opiera się na optymalizacji gradientu funkcji opisującej $RyzykoNauczania$ (opisanej w Paragrafie 2.1.2). Skuteczność tej metody została potwierdzona w eksperymencie, w którym zadaniem anotatorów była klasyfikacja zdjęć motyli.

W ramach artykułu Yang et al. [2017] autorzy opisali podejście nauczania maszynowego, w którym model ucznia oparty jest na algorytmie stosującym wnioskowanie Bayesowskie. Zaproponowany przez autorów model zakłada istnienie wiedzy *a priori* ucznia stanowiącej jego początkowe „przekonania". Takie podejście umożliwia wnioskowanie w oparciu o model ucznia (np. do wyboru optymalnego sygnału nauczającego) nawet przy niewielkiej liczbie zaobserwowanych odpowiedzi ucznia.

Inne podejście do nauczania maszynowego zostało zaprezentowane w pracy Hong et al. [2020], w którym to anotatorzy wcielali się w rolę nauczyciela. W ramach pojedynczego mikro-zadanie anotator (nauczyciel) stara się nauczyć algorytm klasyfikujący obrazy (uczeń) automatycznego rozróżniania trzech obiektów na podstawie ich zdjęć. W opisanym eksperymencie anotator sam buduje zbiór przykładów uczących poprzez samodzielnie wykonanie zdjęć wybranym obiektom.

Pomimo że zastosowanie nauczania maszynowego w procesie crowdsourcingu nie jest szeroko reprezentowane w literaturze, to połączenie tych dwóch metod odbywa się w bardzo naturalny sposób. Zgodnie z założeniami taksonomii kontroli jakości w procesie crowdsourcingu (zob. Paragraf 1.3.1), wysoka jakość danych zapewniona jest poprzez zdefiniowanie trzech składowych: *modelu jakości*, *oceny jakości* oraz *zapewnienia jakości*.

Te trzy składowe realizowane są przez elementy procesu nauczania maszynowego. Przykładowo, model ucznia używany w nauczaniu maszynowym bezpośrednio realizuję wymiar *jakość anotatora* komponentu *model jakości* z taksonomii kontroli jakości. W Tabeli 2.1 przedstawiłem zestawienie komponentów taksonomii wraz z realizującymi je elementami procesu nauczania maszynowego.

Równie naturalnie odbywa się dostosowanie procesu nauczania maszynowego (opisanego w pierwszej części niniejszego rozdziału zob. Podrozdział 2.1.1) do projektu crowdsourcingowego. W tym przypadku, w rolę ucznia wciela się anotator, który w czasie procesu nauczania maszynowego nauczany jest przez platformę crowdsourcingową. Sam proces na-

Tabela 2.1: *Zestawienie komponentów taksonomii kontroli jakości z realizującymi je elementami procesu nauczania maszynowego (opracowanie własne)*

Crowdsourcing		**Nauczanie maszynowe**
Komponent	**Podkomponent**	**Element procesu**
model jakości	jakość anotatora	model ucznia
ocena jakości	metoda grupowa i obliczeniowa	obliczanie parametrów modelu ucznia
zapewnienie jakości	szkolenie anotatorów	sygnały nauczające

uczania maszynowego odbywa się w formie interaktywnych iteracji. Kroki przykładowej, i-tej iteracji procesu przestawiłem poniżej:

1. Na podstawie wybranego algorytmu nauczającego A_n oraz obecnego modelu anotatora (model ucznia) M_u^{i-1} system (platforma crowdsourcingowa, nauczyciel) wybiera sygnał nauczający $s_i = (x_i, y_i)$.
2. System przekazuje dane wejściowe x_i dla aktualnego sygnału s_i.
3. Anotator (uczeń) dokonuje predykcji $\hat{y}_i$ w oparciu o obecną, nieznaną przez system, hipotezę g_{i-1} (rzeczywisty stan wiedzy anotatora).
4. System aktualizuje aktualny model wiedzy anotatora M_u^i w oparciu o odpowiedź udzieloną przez anotatora $\hat{y}_i$.
5. System przekazuje informację zwrotną y_i.
6. Anotator aktualizuje swój obecny stan wiedzy g_i w oparciu o otrzymaną informację zwrotną y_i.

Opisane powyżej kroki definiują jedynie szablon obrazujący zastosowanie procesu nauczania maszynowego w metodzie crowdsourcingu. W celu pełnego zaadoptowania procesu nauczania maszynowego do danego projektu wymagany jest wybór konkretnych algorytmów używanych w trakcie procesu. W kolejnej części niniejszego rozdziału przedstawiłem przykłady rozwiązań, które mogą być użyte w celu implementacji trzech kluczowych elementów: *model ucznia*, *metoda obliczenia parametrów modelu ucznia* oraz *metoda wyboru sygnałów nauczających*.

2.2. Modele reprezentacji ucznia

Jednym z podstawowych elementów procesu nauczania maszynowego jest model ucznia. W procesie crowdsourcingu element ten realizuje wymiar *jakość anotatora* komponentu *model jakości* z taksonomii kontroli jakości (zob. Paragraf 1.3.2). Zadaniem modelu ucznia jest odwzorowanie obecnego stanu wiedzy ucznia (zob. Definicja 9). Metody tworzenia modeli ucznia obejmują szeroki przekrój rozwiązań, które mogą się różnić złożonością i dokładnością w odwzorowaniu rzeczywistej wiedzy ucznia. Wybór konkretnej metody zależy od czynników takich jak format oznaczanych danych czy dostępne funkcjonalności używanej platformy crowdsourcingowej. W ramach niniejszej rozprawy zdecydowałem się podzielić modele ucznia na dwie kategorie:

– **modele wiedzy ucznia** – kategoria ta obejmuje złożone modele ucznia, które pozwalają na odwzorowywanie wiedzy ucznia. W skład tej kategorii wchodzą algorytmy, które pozwalają na predykcje oznaczeń anotatora dla elementów zbioru wejściowego, które nie zostały przez niego oznaczone. Modele te implementowane są między innymi przez zastosowanie algorytmów uczenia maszynowego.

– **modele jakości ucznia** – kategoria ta obejmuje uproszczone modele ucznia, które opisują jakość wykonanych dotychczas anotacji. Metody te nie pozwalają na predykcje odpowiedzi dla nowych elementów zbioru wejściowego. Modele jakości ucznia implementowane są między innymi przez użycie predefiniowanych metryk lub zbioru parametrów.

W dalszej części tego rozdziału przedstawione zostaną przykłady oraz dokładniejszy opis obu kategorii.

2.2.1. Modele wiedzy ucznia

Wykorzystanie *modeli wiedzy ucznia* pozwala na przeprowadzenie pełnej analizy oceny jakości jego odpowiedzi. Określenie obecnego stanu wiedzy ucznia może zostać wykonane w oparciu o historyczne odpowiedzi ucznia, jak również na podstawie estymowanych (przy użyciu modelu) odpowiedzi stworzonych dla całego dostępnego zbioru danych. Podstawowym sposobem na stworzenie modelu wiedzy ucznia jest zastosowanie algorytmów uczenia maszynowego.

Algorytmy uczenia maszynowego

Tak jak przedstawiłem w pierwszej części tego rozdziału (zob. Paragraf 2.1.2), w klasycznej konfiguracji nauczania maszynowego problem ucznia zazwyczaj rozwiązywany jest poprzez zastosowanie metod uczenia maszynowego. W takim podejściu, w każdej i-tej iteracji nauczania, zaktualizowany stan wiedzy ucznia M_u^i, który stanowi przybliżenie rzeczywistej hipotezy ucznia g^i. Hipoteza ta wybierana jest przez algorytm uczenia A_u na podstawie dostępnego zbioru przykładów uczących (zob. Paragraf 2.1.1).

Mimo wspomnianych wcześniej zalet, zastosowanie modeli wiedzy ucznia nie sprawdza się we wszystkich sytuacjach. W przypadku zadań, w których dane wejściowe zdefiniowane są przez wektory cech o wysokim wymiarze (np. w zadaniach związanych z przetwarzaniem tekstu lub obrazu), złożony model uczenia maszynowego nie jest w stanie w poprawny sposób odwzorować obecnego stanu wiedzy ucznia. Jest tak, ponieważ model wiedzy ucznia budowany jest zazwyczaj w oparciu o niewielki zbiór przykładów uczących, które zostały oznaczone przez danego ucznia (np. na podstawie tylko 20 elementów). Zastosowanie uczenia maszynowego w celu reprezentacji obecnego stanu wiedzy ucznia sprawdza się najlepiej w sytuacji, gdy dane wejściowe opisywane są przez wektory cech o niskim wymiarze (np. zadań związanych tylko z przetwarzaniem wartości liczbowych).

Dodatkowo wybór konkretnego modelu, który użyty zostanie do reprezentacji wiedzy ucznia jest w dużej mierze uzależniony od wiedzy nauczyciela na temat hipotezy ucznia g (zob. Paragraf 2.1.3). W sytuacji, gdy nauczyciel ma pełną wiedzę na temat hipotezy ucznia, możliwe jest zamodelowanie jego obecnego stanu wiedzy w ten sam sposób (wtedy model ucznia równy jest hipotezie ucznia $M_u = g$). Jeżeli nauczyciel nie posiada informacji na temat algorytmu ucznia, to model wiedzy ucznia oparty jest o dowolny, wybrany przez nauczyciela algorytm (zob. Paragrafie 2.1.3).

W związku z tym, że zagadnienie uczenia maszynowego jest bardzo szeroką i ciągle rozwijaną dziedziną, szczegółowe omówienie współcześnie stosowanych modeli daleko wykracza poza zakres niniejszej rozprawy. W celu demonstracji zasady użycia modelu wiedzy ucznia w kolejnym paragrafie przedstawiłem przykład zastosowania modelu regresji logistycznej jako modelu wiedzy ucznia dla problemu klasyfikacji binarnej.

Regresja logistyczna

Rozważmy problem stworzenia klasyfikatora binarnego, który przyporządkowuje wielowymiarowym wektorom liczb rzeczywistych $X \in \mathbb{R}^d$ wartość binarną $y \in \{0, 1\}$[7]. Do rozwiązania powyższego problemu użyjemy modelu regresji logistycznej. Niech f będzie liniową funkcją taką, że: $f(x) = w^\top x + b$, gdzie x to wektor wejściowy $x \in \mathbb{R}^d$, w to wek-

[7]Problem ten stanowi rozszerzenie przykładu opisanego w pierwszej części tego rozdziału; zob. Przykład 6, Paragraf 2.1.

tor parametrów funkcji $w \in \mathbb{R}^d$, a $b \in \mathbb{R}$ to dodatkowy parametr definiujący przesunięcie klasyfikatora (ang. *bias*). Ostateczny klasyfikator przyjmuję formę:

$$\hat{y} = \begin{cases} 1, w^\top x + b \geqslant \gamma \\ 0, \text{w p. p.} \end{cases} \tag{2.14}$$

gdzie γ to wartość progowa określająca punkt zmiany przypisywanej kategorii binarnej[8]. Prawdopodobieństwo pozytywnej klasy $\hat{y} = 1$ obliczane jest poprzez zaaplikowanie funkcji sigmoidalnej σ na f:

$$P(y = 1|x,w,b) = \sigma(w^\top x + b) \tag{2.15}$$

gdzie funkcja σ zdefiniowana jest jako:

$$\sigma(z) = \frac{1}{1 + e^{-z}}. \tag{2.16}$$

Rozważmy Przykład 8 polegający na klasyfikacji pacjentów, u których występuje choroba serca[9]. Załóżmy, że do oznaczenia danych z tego przykładu użyty został proces nauczania maszynowego, a do reprezentacji modelu wiedzy ucznia wybrany został algorytm regresji linowej. Proces tworzenia modelu ucznia odbywa się zgodnie ze schematem opisanym w pierwszej części tego rozdziału (zob. Podrozdział 2.1.1). Szczegółowy opis zastosowania algorytmu regresji liniowej do zamodelowania obecnej wiedzy ucznia został opisany w Przykładzie 9.

Przykład 8

Zbiór wektorów wejściowych $x_i \in X$ określony jest przez 2-wymiarową przestrzeń wektorową rzeczywistą $X = \mathbb{R}^2$. Każdy wymiar X opisuje inną cechę pacjenta: ciśnienie, poziom cholesterolu. Zbiór możliwych wartości wyjściowych składa się z 2 wartości kategorycznych określających występowanie choroby u pacjenta: $J = \{chory, zdrowy\}$. Na potrzeby przykładu załóżmy, że zbiór przykładów uczących $\mathcal{D} = \{(x_0, y_0), ..., (x_n, y_n)\}$ składa się z 10 elementów. Zawartość zbioru $\mathcal{D}$ została przedstawiona w Tabeli 2.2.

Przykład 9

Załóżmy, że proces nauczania jest obecnie w iteracji: $i = 5$. Uczeń do tej pory oznaczył pierwsze 5 elementów zbioru, które zostały wybrane przez algorytm nauczyciela. Uczeń dokonał selekcji odpowiedzi w oparciu o swój aktualny (nieznany nauczycielowi) stan

[8] Zazwyczaj $\gamma = 0{,}5$.

[9] Opisany przykład stanowi uproszczoną wersję problemu *Cleveland Heart Disease*, `https://archive.ics.uci.edu/ml/datasets/heart+disease`; dostęp: 10.04.2022 r.

Tabela 2.2: *Przykładowy zbiór przykładów uczących dla problemu klasyfikacji binarnej*

i	**ciśnienie**	**cholesterol**	**chory**
1	145	233	0
2	160	286	1
3	120	229	1
4	130	250	0
5	130	204	0
6	120	236	0
7	140	268	1
8	120	354	0
9	130	254	1
10	140	203	1

wiedzy – hipotezę g_i. Po oznaczeniu każdego elementu uczeń otrzymał informację zwrotną z poprawną odpowiedzią. Obecny zbiór oznaczeń ucznia zawiera następujące elementy:

$$Y_u = \{a_1, ..., a_5\} = \{0, 0, 1, 0, 1\}$$

gdzie a_i to anotacja wykonana przez ucznia dla mikro-zadania x_i.

Po zakończeniu iteracji zebrane oznaczenia Y_u używane są do stworzenia modelu wiedzy ucznia według wybranego algorytmu uczenia. Załóżmy, że parametry M_u zostały obliczone w oparciu o algorytm metody gradientu prostego[10]. Model wiedzy ucznia w piątej iteracji procesu nauczania maszynowego M_u^5 przyjmuje formę:

$$w = [-9{,}3490, 5{,}0376]$$

$$b = -0{,}1897$$

$$M_u^5 = \sigma(-9{,}3490, 5{,}0376]^\top x - 0{,}1897)$$

Dzięki zastosowaniu powyższego modelu możliwa jest estymacja odpowiedzi ucznia dla mikro-zadań, które nie zostały jeszcze przez niego oznaczone. Przykładowo, dla pozostałych pięciu elementów zbioru $\mathcal{D}$ otrzymujemy estymowany zbiór anotacji Y_u:

$$\hat{Y}_u = [1, 1, 1, 1, 0]$$

[10]Gradient prosty to jeden z popularnych algorytmów optymalizacyjnych używanych do obliczania parametrów modeli uczenia maszynowego. Szczegóły działania algorytmu zostały opisane między innymi w pracy [Goodfellow et al., 2016, s. 82-93].

Podczas gdy zbiór Y_u może zostać użyty do oszacowania dotychczasowej jakości pracy ucznia, zbiór $\hat{Y}_u$ pozwalana na oszacowanie jakości dla nieoznaczonych danych. W tym przypadku dotychczasowa i szacowana dokładność to odpowiednio 60% *i* 40%.

2.2.2. Modele jakości ucznia

Alternatywnym podejściem reprezentacji wiedzy ucznia jest zastąpienie pełnego modelu wiedzy ucznia (opartego np. o algorytm uczenia maszynowego) prostszym rozwiązaniem – *modelem jakości ucznia*. W ramach tego modelu, obecny stan wiedzy ucznia zdefiniowany jest za pomocą zbioru parametrów lub metryk opisujących jakość jego dotychczasowych odpowiedzi. Modele jakości z powodzeniem stosowane są dla złożonych zadań, ponieważ ich parametry mogą zostać obliczone dla mniejszego zbioru danych. W dalszej części tego rozdziału opisałem przykłady modeli, które mogą zostać wykorzystane do opisania jakości ucznia.

Tablica pomyłek

W przypadku zadań związanych z klasyfikacją (np. wybór części mowy dla podanych słów, analiza wydźwięku opinii o produktach) możliwe jest zastosowanie tablicy pomyłek do zamodelowania jakości pracy danego ucznia[11].

Niech zbiór $C = \{c_1, c_2, \cdots, c_n\}$ określa wszystkie możliwe n kategorii, wtedy macierz pomyłek zdefiniowana jest wówczas jako macierz $n \times n$:

$$Q = \begin{bmatrix} Q_{1,1} & Q_{1,1} & \cdots & Q_{1,n} \\ Q_{2,1} & Q_{2,2} & \cdots & Q_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ Q_{n,1} & Q_{n,1} & \cdots & Q_{n,n} \end{bmatrix} \tag{2.17}$$

gdzie każdy j-ty wiersz ($1 \leqslant j \leqslant n$), określony jako $Q_j = [Q_{j,1}, Q_{j,1}, \cdots, Q_{j,n}]$ zawiera rozkład prawdopodobieństwa kategorii, którą wybierze uczeń, gdy prawidłową kategorią jest c_j. Każdy element macierzy $Q_{j,k}$ określa prawdopodobieństwo wyboru kategorii c_k ($1 \leqslant k \leqslant n$) w przypadku, gdy poprawną odpowiedzią jest kategoria c_j.

W Przykładzie 10 przedstawiłem sposób użycia macierzy pomyłek do reprezentacji modelu jakości ucznia dla problemu klasyfikacji.

[11] W przypadku zadania polegającego na wyborze wartości liczbowych, zastosowanie tej metody jest możliwe, ale wymaga wykonania transformacji danych wyjściowych. Przykładem takiej transformacji jest sztuczne utworzenie zakresów wartości, którym przypisane zostały kategorie.

Tabela 2.3: *Tablica pomyłek dla problemu klasyfikacji binarnej*

		Klasy przewidywane	
		$\hat{\mathbf{y}} = \mathbf{1}$	$\hat{\mathbf{y}} = \mathbf{0}$
Klasy rzeczywiste	$\mathbf{y} = \mathbf{1}$	tp (prawdziwie pozytywna)	fn (fałszywie ujemna)
	$\mathbf{y} = \mathbf{0}$	fp (fałszywie pozytywna)	tn (prawdziwie ujemna)

Przykład 10

Rozważmy przykład zadania analizy wydźwięku wypowiedzi, w którym zbiór możliwych kategorii składa się z $n = 3$ elementów $C = \{Pozytywny, Neutralny, Negatywny\}$. Macierz pomyłek dla przykładowego ucznia może mieć następującą formę:

$$Q = \begin{bmatrix} 0{,}6 & 0{,}1 & 0{,}3 \\ 0{,}1 & 0{,}8 & 0{,}1 \\ 0{,}2 & 0{,}5 & 0{,}4 \end{bmatrix}$$

Jeżeli założymy, że indeks danej kolumny i danego wiersza odpowiada kolejnym kategoriom ze zbioru C, to przykładowo wartość $Q_{2,1} = 0{,}1$ oznacza, że w 10% dotychczasowych odpowiedzi uczeń wybierał kategorie Pozytywny gdy poprawną kategorią była Neutralny.

Szczególnie istotna jest tablica pomyłek tworzona dla problemu klasyfikacji binarnej, której elementy używane są do zdefiniowana wielu popularnych metryk (przykładowe metryki zostaną opisane w dalszej części niniejszego rozdziału, zob. Paragraf 2.2.2). Do każdej komórki tablicy przypisana została odpowiednia nazwa w zależności od poprawności klasyfikacji (zob. Tabela 2.3).

Prawdopodobieństwo prawidłowej odpowiedzi

Najprostszym sposobem, w jaki można przedstawić jakość ucznia jest pojedyncza zmienna $q \in [0, 1]$, której wartość równa jest prawdopodobieństwu, że dany uczeń wykona prawidłową anotację. Wartość zmiennej q obliczona jest jako:

$$q = P(\hat{y} = y) = \frac{tp}{tp + fp} \tag{2.18}$$

W Przykładzie 11 przedstawiłem sposób obliczania prawdopodobieństwa prawidłowej odpowiedzi ucznia.

Przykład 11

Niech tablica pomyłek ucznia będzie równa:

$$Q = \begin{bmatrix} 0{,}7 & 0{,}4 \\ 0{,}3 & 0{,}6 \end{bmatrix}$$

Wartość $q = \frac{0{,}7}{0{,}7+0{,}3} = 0{,}7$ *oznacza, że w przypadku* 70% *dotychczas oznaczonych mikro-zadań uczeń dokonał prawidłowego oznaczenia. Tym samym zgodnie z tym modelem istnieje* 70% *prawdopodobieństwo, że wybrany dokona prawidłowego oznaczenia nowego mikro-zadania.*

Istnieją również podejścia, które rozszerzają opisany powyżej model:

– rozszerzenie zakresu możliwych wartości zmiennej q z $[0, 1]$ do szerszego zbioru: np. $(-\infty, \infty)$. W takim podejściu wartość ta nie jest interpretowana jako prawdopodobieństwo, a arbitralnie dobraną punktacją, w której wyższa wartość oznacza wyższą jakość tworzonych danych.

– reprezentacja zmiennej q w formie przedziału ufności (bądź rozkładu prawdopodobieństwa) możliwych wartości[12].

Metryki oceny jakości klasyfikacji

Kolejnym sposobem reprezentacji modelu jakości ucznia jest użycie predefiniowanych metryk jakości. Metryki te pozwalają na przedstawienie wybranego aspektu jakości pracy ucznia w formie łatwiej w interpretacji liczby. Wybór metryk bezpośrednio zależy od typu zadania – inne metryki używane są w przypadku zadań klasyfikacji, a inne w zadaniach związanych z estymacją wartości liczbowych [Sokolova & Lapalme, 2009].

W przypadku zadań klasyfikacji używane są metryki takie jak: dokładność, precyzja, czułość, swoistość i $F1$. Wartości metryk obliczane są dla wybranej kategorii na podstawie tablicy pomyłek (zob. Paragraf 2.2.2) dla klasyfikacji binarnej[13]:

$$\text{dokładność (ang. } \textit{accuracy}\text{)} = \frac{tp + tn}{tp + tn + fp + fn} \tag{2.19}$$

$$\text{precyzja (ang. } \textit{precision}\text{)} = \frac{tp}{tn + fp} \tag{2.20}$$

$$\text{czułość (ang. } \textit{sensitivity, recall}\text{)} = \frac{tp}{tp + fn} \tag{2.21}$$

[12] Takie podejście może być stosowane na przykład, gdy uczeń pracował w kilku oddzielnych sesjach. Rozkład wartości opisuje zróżnicowanie jakości pracy podczas kolejnych sesji.

[13] W przypadku zadania klasyfikacji wykonywanej dla więcej niż dwóch kategorii, problem upraszczany jest do klasyfikacji binarnej poprzez rozpatrywanie tylko dwóch kategorii: „poprawna” i „niepoprawna”.

$$\text{swoistość (ang. \textit{specificity})} = \frac{tn}{fp + tn} \tag{2.22}$$

$$\text{F1} = \frac{2}{\text{czułość}^{-1} + \text{precyzja}^{-1}} \tag{2.23}$$

Metryki przestawione powyżej nie są przeznaczone wyłącznie dla eksperymentów związanych z analizą jakości uczniów w procesie nauczania maszynowego. Są one powszechnie stosowane w badaniach związanych ze statystyką, uczeniem maszynowym lub analizą i przetwarzaniem danych.

Błąd przesunięcia i wariancji

W przypadku zadań związanych z estymacją wartości liczbowej (np. oszacowanie ceny mieszkania, oszacowanie wymiarów produktu przedstawionego na zdjęciu), model jakości ucznia może być przedstawiony w formie rozkładu popełnianych błędów.

Załóżmy, że błąd ucznia zamodelowany jest za pomocą rozkładu normalnego $\mathcal{N}(\epsilon, \sigma^2)$, który definiowany jest za pomocą dwóch parametrów:

– **błąd przesunięcia** (ang. *bias*) $\epsilon \in (-\inf, +\inf)$ opisuje średni błąd ucznia. Jeżeli uczeń ma tendencje do przeszacowywania wartości to błąd ten będzie większy od zera $\epsilon > 0$, a w przypadku gdy uczeń nie doszacowuje $\epsilon < 0$.

– **błąd wariancji** $\sigma^2 \in [0, +\inf)$ oznacza wariancję średniego błędu ucznia skupioną wokół ϵ.

Metryki oceny jakości estymacji wartości liczbowych

W przypadku zadań związanych z estymacją wartości liczbowych (np. przewidywanie ceny mieszkań lub oszacowywanie wagi przedmiotów spożywczych) stosowane są inne metryki niż w przypadku zadań związanych z klasyfikacją. Metryki te nie są obliczane na podstawie tablicy pomyłek, a poprzez porównanie prawdziwej wartości z wartością wybraną przez ucznia:

Niech X będzie zbiorem wartości wejściowych wielkości n, Y zbiorem wartości wyjściowych (np. zbiór $\mathbb{R}$), $\hat{y}_i \in Y$ to wartość wybrana przez ucznia dla elementu x_i, a $y_i \in Y$

to prawdziwa wartość dla tego elementu. Przykładowe metryki oceny jakości pracy ucznia mogą być zdefiniowane jako:

$$\text{błąd średniokwadratowy (ang. } \textit{mean square error}, MSE) = \frac{1}{n}\sum_{i=1}^{n}(y_i^* - \hat{y}_i)^2 \tag{2.24}$$

$$\text{pierwiastek błędu średniokwadratowego (ang. } \textit{root mean square error}, RMSE) = \sqrt{MSE} \tag{2.25}$$

$$\text{średni błąd względny (ang. } \textit{mean error}, ME) = \frac{1}{n}\sum_{i=1}^{n}(y_i^* - \hat{y}_i) \tag{2.26}$$

$$\text{średni błąd bezwzględny (ang. } \textit{mean absolute error}, MAE) = \frac{1}{n}\sum_{i=1}^{n}|y_i^* - \hat{y}_i| \tag{2.27}$$

Pozostałe metryki

W przypadku projektów crowdsourcingowych stosowane są również metryki, które wykorzystują informacje dotyczące sposobu interakcji anotatora (ucznia) z platformą crowdsourcingową. Metryki te pozwalają w sposób pośredni zamodelować obecny stan wiedzy anotatora poprzez analizę parametrów takich jak: średni czas spędzony na anotacji jednego mikro-zadania, liczba kliknięć, szybkość przewijania strony z interfejsem anotacyjnym [Rzeszotarski & Kittur, 2011].

2.3. Metody obliczania parametrów modelu ucznia

Kolejnym elementem określanym w ramach implementacji procesu nauczania maszynowego jest wybór metody obliczania parametrów modelu ucznia. W każdej iteracji procesu nauczania maszynowego dokonywana jest ocena dotychczasowej wiedzy ucznia, a wynik tej oceny zapisywany jest w formie parametrów wybranego modelu ucznia. Wybór konkretnej metody oceny ucznia zależy od tego, jaki model ucznia został wybrany do odwzorowania jego wiedzy. W tym podrozdziale niniejszej pracy przedstawione zostaną metody obliczania parametrów modelu ucznia, charakterystyczne dla metody crowdsourcingu.

W przypadku, gdy proces nauczania maszynowego implementowany jest jako część projektu crowdsourcingowego, ocena wiedzy ucznia (anotatora) powinna odbywać się automatycznie po przesłaniu udzielonej przez niego odpowiedzi. Wygenerowana ocena używana jest w dwóch elementach procesu:

– aktualizacja parametrów modelu ucznia,

– zapewnienie informacji zwrotnej.

Oznacza to, że nauczanie maszynowe wprowadza ograniczenia dotyczące formy informacji zwrotnej, która musi zostać przekazana w formie synchronicznej (Anotator otrzymuje informację zwrotną zaraz po wykonaniu mikro-zadania; zob. Paragraf 1.4.1). Z tego powodu w projekcie crowdsourcingowym, obliczania parametrów modelu ucznia realizowane jest przez techniki *obliczeniowej i grupowej oceny jakości* komponentu *ocena jakości* z taksonomii kontroli jakości (zob. Paragraf 1.3.3).

Techniki *obliczeniowej oceny jakości* pozwalają na automatyczną ocenę dotychczasowej pracy ucznia na podstawie wybranego algorytmu i dostępnego zbioru danych referencyjnych. W sytuacji, gdy referencyjny zbiór danych nie jest dostępny, anotacje referencyjne mogą zostać wygenerowane w sposób automatyczny. Dzięki zastosowaniu takiego rozwiązania możliwe jest dokonanie automatycznej oceny anotatora, poprzez porównanie jego anotacji z wygenerowaną anotacją referencyjną. Jednym z powszechnie stosowanych sposobów automatycznego generowania anotacji referencyjnych w procesie crowdsourcingu jest zastosowanie algorytmów agregacji (zob. Paragraf 1.3.4). Algorytmy te używane są zarówno w ramach strategii *zapewnienia jakości* (zob. Paragraf 1.3.4), jak i w ramach techniki *grupowej oceny jakości* (zob. Paragraf 1.3.3).

W podstawowej wersji algorytmy agregacji odpowiadają jedynie za wygenerowanie finalnej odpowiedzi dla danego mikro-zadania w oparciu o zbiór zawierający anotacje wykonane dla tego mikro-zadania przez różnych anotatorów. W tej wersji algorytmy agregacji anotacji mogą zostać użyte do automatycznej oceny anotatora. Ze względu na prostą budowę tych algorytmów, użycie agregacji jako źródło informacji zwrotnej w procesie nauczania maszynowego możliwe jest tylko w sytuacji, gdy anotator rozwiązuje mikro-zadań, dla których istnieją oznaczenia innych anotatorów,

Bardziej złożone algorytmy agregacji anotacji rozszerzone są o model ucznia (anotatora). Algorytm te równocześnie wybierają finalną odpowiedź dla mikro-zadań ze zbioru dostępnych anotacji oraz dokonują ewaluacji pracy anotatorów. Efektem działania algorytmu jest zbiór finalnych odpowiedzi wygenerowanych dla wszystkich mikro-zadań oraz zbiór parametrów modelu ucznia wygenerowany dla wszystkich anotatorów. W przypadku, gdy algorytmy te implementowane są w formie modeli uczenia maszynowego, możliwe jest zastosowanie stworzonych modeli w celu estymacji rozwiązań dla mikro-zadań, dla których nie zostały jeszcze stworzone żadne anotacje.

Ponieważ algorytmy agregacji anotacji stanowią istotny element metody crowdsourcing, w dalszej części niniejszego podrozdziału zamieściłem dokładniejszy opis sposób ich działania oraz konkretne przykłady implementacji prostych i złożonych algorytmów.

2.3.1. Komitety klasyfikatorów

Algorytmy agregacji anotacji bezpośrednio bazują na podejściu nazywanym „komitetem klasyfikatorów" (ang. *ensemble*) [G. Wang et al., 2011]. Celem tych algorytmów jest połączenie predykcji wielu „klasyfikatorów bazowych" (nazywanych również „słabymi klasyfikatorami") w celu uzyskania końcowego klasyfikatora o wyższej jakości [Polikar, 2006]. Klasyfikatory wchodzące w skład komitetu mogą (ale nie muszą) używać tego samego algorytmu. Przykładowo, możliwe jest stworzenie komitetu złożonego z klasyfikatorów używających algorytmów sieci neuronowej, maszyn wektorów nośnych czy drzew decyzyjnych [G. Wang et al., 2011]. Nawet w przypadku gdy klasyfikatory trenowane są na tym samym zbiorze danych, połączenie ich predykcji pozwala na redukcję ryzyka związanego ze słabą generalizacją pojedynczego klasyfikatora [Polikar, 2006].

Literatura wyróżnia wiele alternatywnych strategii łączenia klasyfikatorów w metodach *ensamble*. Jako trzy najczęściej pojawiające się metody można wyróżnić: *bagging*, *boosting*, *stacking* [G. Wang et al., 2011].

– **bagging** – w tym podejściu każdy z klasyfikatorów uczony jest niezależnie, a końcowa predykcja tworzona jest poprzez połączenie predykcji klasyfikatorów przy użyciu wybranego algorytmu,

– **boosting** – w tym podejściu klasyfikatory uczone są w sposób sekwencyjny, tak że każdy kolejny klasyfikator wykorzystuje predykcje poprzedniego klasyfikatora,

– **stacking** – w tym podejściu każdy klasyfikator jest uczony niezależnie, a końcowa predykcja tworzona jest przy użyciu odpowiednio wytrenowanego klasyfikatora „wysokopoziomowego", którego zadaniem jest łączenie odpowiedzi klasyfikatorów bazowych.

W przypadku procesu crowdsourcingu rolę klasyfikatorów bazowych przyjmują pojedynczy anotatorzy. Połączenie ich predykcji dokonywane jest poprzez zastosowanie algorytmu agregacji ich anotacji. Ze względu na specyfikę tego procesu, odpowiedzi antotorów zbierane są w sposób niezależny. Z tego powodu najpopularniejszymi strategiami łączenia predykcji anotatorów są podejścia: *bagging* i *stacking*. Przykłady tych metod omówiłem w dalszej części niniejszego podrozdziału.

2.3.2. Algorytmy głosowania większościowego

Podstawowym sposobem agregacji anotacji jest użycie algorytmów bazujących na głosowaniu większościowym (ang. *majority vote*). W algorytmach głosowania większościowego finalna odpowiedź dla danego mikro-zadania wybierana jest w głosowaniu, w którym pojedyncze odpowiedzi anotatorów pełnią funkcję głosów. Decyzja o finalnej odpowiedzi podejmowana jest niezależnie dla każdego mikro-zadania. Algorytmy głosowania większościowego stanowią przykład metody *bagging* [G. Wang et al., 2011, s. 224]. Poniżej przedstawione zostały dwie podstawowe wersje tych algorytmów: bazowy *algorytm głosowania większościowego* oraz *algorytm ważonego głosowania większościowego*.

Algorytm głosowania większościowego

W najbardziej podstawowej wersji algorytmu głosowania większościowego, jako finalna odpowiedź wybrana zostaje tak, która została wybrana przez największą liczbę anotatorów. Poniżej przestawiony został opis algorytm głosowania większościowego (zob. Procedura 5) oraz przykład działania algorytmu (zob. Przykład 12).

Procedura 5: Algorytm głosowania większościowego

Niech :

X – przestrzeń cech definiujących problem,

Y – zbiór możliwych kategorii,

$\mathcal{D} = \{x_1, \cdots, x_n\}$ – zbiór treści mikro-zadań, gdzie $x_i \in X$,

$\{h_1, \cdots, h_k\}$ – zbiór J anotatorów, gdzie $h_i : X \rightarrow Y$,

Kroki :

1 Każdy anotator h_j rozwiązuje podzbiór mikro-zadań:

– $\{(x_i, h_j(x_i))|\ x_i \in D_{(h_j)}\}$, gdzie: $\mathcal{D}_{(h_j)} \subset \mathcal{D}$,

– $h_j(x_i)$ – kategoria wybrana przez anotatora h_j dla mikro-zadania i.

2 Dla każdego mikro-zadania x_i finalna anotacja $H(x_i)$ wybierana jest według następującej formuły:

$$H(x_i) = \underset{y \in Y}{argmax} \sum_{t=0}^{J} 1(y = h_t(x_i))$$

$$\text{gdzie } 1(\alpha) = \begin{cases} 1, & \text{jeśli } \alpha \text{ jest prawdziwe} \\ 0, & \text{w p. p.} \end{cases}$$

Wyjście:

$H(x)$ określa finalne anotacje dla każdego mikro-zadania

Przykład 12

Rozważmy sytuację, w której zbiór $\mathcal{D}$ składa się z $n = 4$ elementów. Zadaniem anotatorów jest wybór jednej z dwóch kategorii $Y = \{a, b\}$. Zbiór został oznaczony przez czterech anotatorów. W celu wyboru finalnej odpowiedzi $H(x)$ zastosowany został algorytm głosowania większościowego. W Tabeli 2.4 przedstawiłem wyniki działania algorytmu głosowania większościowego. W kolumnie $H(x)$ zawarta została finalna odpowiedź (bądź odpowiedzi) wybrane przez algorytm.

Analizując wyniki powyższego przykładu, można zauważyć wadę tego algorytmu. W przypadku, gdy dwie odpowiedzi otrzymają taką samą liczbę głosów, algorytm głosowania większościowego nie jest w stanie wybrać finalnej odpowiedzi. Z tego powodu powszechnie stosowane są bardziej rozbudowane modyfikacje tego algorytmu (np. algorytm ważonego głosowania większościowego omówiony poniżej).

Tabela 2.4: *Wybór końcowej anotacji za pomocą Algorytmu głosowania większościowego*

$\mathbf{x_i}$	$\mathbf{h_1(x)}$	$\mathbf{h_2(x)}$	$\mathbf{h_3(x)}$	$\mathbf{h_4(x)}$	$\mathbf{H(x)}$
x_1	a	a	b	a	a
x_2	b	b	b	a	b
x_3	a	b	a	a	a
x_4	a	b	b	a	a lub b

Algorytm ważonego głosowania większościowego

Najprostszym sposobem rozbudowy bazowej wersji algorytmu głosowania większościowego jest zastosowanie algorytmu ważonego głosowania większościowego [Littlestone & Warmuth, 1994]. Algorytm ten wprowadza dodatkowe wagi, które zostają przypisane do anotacji. Wagi obliczone są na podstawie wybranego modelu jakości anotatora (np. metryki oceny jakości: *dokładność*). W zależności od użytego modelu, wartości wag mogą być wspólne dla wszystkich odpowiedzi anotatora lub mogą zostać obliczone osobno dla każdej anotacji.

Rozważmy zbiór wag: $\gamma = \{\gamma_1, \cdots, \gamma_k\}$, w którym element γ_t zawiera dotychczasową dokładność anotatora h_t obliczoną w oparciu o wybrany model jakości. Algorytm ważonego głosowania większościowego działa w sposób analogiczny do wersji bazowej algorytmu (opisanej powyżej, zob. Paragraf 2.3.2). Jedyną różnicą pomiędzy tymi algorytmami jest formuła, która używana jest do wyboru anotacji. Dla każdego mikro-zadania x_i finalna anotacja $H(x_i)$ wybierana jest według następującej formuły:

$$H(x_i) = \underset{y \in Y}{argmax} \sum_{t=0}^{J} \gamma_t 1(y = h_t(x_i)) \tag{2.28}$$

$$\text{gdzie } 1(\alpha) = \begin{cases} 1, & \text{jeśli } \alpha \text{ jest prawdziwe} \\ 0, & \text{w p. p.} \end{cases} \tag{2.29}$$

2.3.3. Algorytmy nienadzorowane EM

Kolejną grupą algorytmów agregacji anotacji są algorytmy maksymalizacji wartości oczekiwanej (ang. *Expectation-Maximization*, EM) [Raykar et al., 2010]. W ramach tych algorytmów dokonywana jest estymacja nieznanych parametrów T na podstawie zaobserwowanego zbioru danych $\mathcal{D}$. Wartość T zależy od parametrów ukrytych π (ang. *latent variables*), które również nie są znane. Algorytmy EM działają w iteracjach, których celem jest maksymalizacja prawdopodobieństwa *a posteriori* dla parametrów T poprzez marginalizacje parametrów π. Każda iteracja składa się z dwóch kroków:

1. **Krok E (oczekiwanie)** – obliczenie wartości parametrów π w oparciu o aktualne wartości parametrów T.
2. **Krok M (maksymalizacja)** – obliczenie wartości parametrów T w oparciu o parametry π obliczone w kroku E.

Przykładowo, w przypadku zadań crowdsourcingowych, parametry T mogą określać finalne odpowiedzi dla mikro-zadań, a parametry π wartości modelu jakości anotatorów. Podejście to zostało po raz pierwszy zaproponowane w ramach algorytmu *Dawid-Skene* [Dawid et al., 1979]. Poniżej został zaprezentowany dokładniejszy opis działania tego algorytmu.

Algorytm *Dawid-Skene*

Załóżmy, że w ramach procesu crowdsourcingu rozwiązywane jest I mikro-zadań, którym przypisujemy kolejne identyfikatory ze zbioru $i \in \{1, \cdots, I\}$. Każde mikro-zadanie i jest oznaczone jest przez co najmniej jednego z K anotatorów, $k \in \{1, \cdots, K\}$. Anotator wybiera dla mikro-zadania jedną z J możliwych kategorii, $l \in \{1, \cdots, J\}$. Anotator nie musi rozwiązywać wszystkich mikro-zadań oraz może oznaczyć jedno mikro-zadanie więcej niż jeden raz. Dla każdego anotatora definiowany jest model jakości w formie tablicy pomyłek (zob. Paragraf 2.2.2) $n^{(k)}$ o wymiarach $I \times J$, gdzie każdy element $n_{il}^{(k)}$ oznacza liczbę wyboru kategorii l dla mikro-zadania i przez anotatora k. T to macierz $I \times J$, w której pojedynczy element T_{ij} zawiera prawdopodobieństwo, że kategoria j jest prawidłową odpowiedzią dla mikro-zadania i.

Celem algorytmu *Dawid-Skene* jest obliczenie współczynnika błędu anotatora, który oznaczony jest jako $\pi_{jl}^{(k)}$. Współczynnik ten definiowany jest jako prawdopodobieństwo tego, że anotator k dokona wyboru kategorii l dla mikro-zadania, dla którego poprawną odpowiedzią jest kategoria j. Poniżej przedstawiłem kroki algorytmu (zob. Procedura 6) oraz przykład jego działania (zob. Przykład 13).

Procedura 6: Algorytm *Dawid-Skene*

Niech :

$\mathcal{D} = \{(x_i, \{y_i^1, \cdots, y_i^R\})\}_{n=1}^{n}$ – zbiór przykładów uczących,

K – zbiór anotatorów

J – zbiór kategorii

n – tablica pomyłek (model jakości)

Kroki :

1 Inicjalizacja wartości początkowych:

– T na podstawie zbioru $\mathcal{D}$ za pomocą algorytmu głosowania większościowego,

2 Krok E

– Obliczenie estymacji dla p

$$p_j = \frac{\sum_i Tij}{I}$$

– Obliczenie estymacji dla π:

$$\pi_{jl}^{(k)} = \frac{\sum_i T_{ij} n_{il}^{(k)}}{\sum_l \sum_i T_{ij} n_{il}^{(k)}}$$

3 Krok M

– Aktualizacja macierzy T:

$$P(T_{ij} = 1|\text{D}) = \frac{\prod_{k=1}^{K} \prod_{l=1}^{J} (\pi_{jl}^{(k)})^{n_{il}^{(k)}} p_j}{\sum_{q=1}^{J} \prod_{k=1}^{K} \prod_{l=1}^{J} (\pi_{ql}^{(k)})^{n_{il}^{(k)}} p_q}$$

4 Kroki E i M powtarzane są do osiągnięcia zbieżności (lub przez ustaloną, stałą liczbę kroków).

Wyjście:

T zawiera finalny rozkład prawdopodobieństwa odpowiedzi dla każdego mikro-zadania,

π zawiera finalne wartości współczynników błędu dla anotatorów.

Przykład 13

Rozważmy sytuację analogiczną do Przykładu 12, w której rozwiązywany jest zbiór $\mathcal{D}$ składający się z 4 elementów. W celu wyboru finalnej odpowiedzi $H(x)$ zastosowany został algorytm Dawid-Skene.

W Tabeli 2.5 przedstawione zostały odpowiedzi anotatorów oraz finalne anotacja ob-

Tabela 2.5: *Wybór końcowej anotacji za pomocą algorytmu Dawid-Skene (opracowanie własne)*

$\mathbf{x_i}$	$\mathbf{h_1(x)}$	$\mathbf{h_2(x)}$	$\mathbf{h_3(x)}$	$\mathbf{h_4(x)}$	$\mathbf{H(x)}$
x_1	a	a	b	a	a
x_2	b	b	b	a	b
x_3	a	b	a	a	a
x_4	a	b	b	a	a

Tabela 2.6: *Przebieg zmiany prawdopodobieństwa dla algorytmu Dawid-Skene (opracowanie własne)*

iteracja	$\mathbf{P(x_1 = a)}$	$\mathbf{P(x_2 = a)}$	$\mathbf{P(x_2 = a)}$	$\mathbf{P(x_2 = a)}$
1	0,83	0,10	0,83	0,57
2	0,88	0,05	0,88	0,60
3	0,92	0,04	0,92	0,65
5	0,98	0,00	0,98	0,73
10	1,00	0,00	1,00	0,93
27	1,00	0,00	1,00	1,00

liczone za pomocą tego algorytmu. Każdy wiersz tabeli zawiera anotacje wykonane przez anotatorów oraz finalną odpowiedź dla pojedynczego mikro-zadania.

Tabela 2.6 zawiera wartości prawdopodobieństwa wyboru kategorii a dla każdego z mikro-zadań, które zostały obliczone w kolejnych iteracjach działania algorytmu dla macierzy T. Każdy wiersz tabeli przedstawia prawdopodobieństwo wyboru odpowiedzi a dla każdego z czterech mikro-zadań w danej iteracji.

Algorytm regresji logistycznej dla anotacji

Kolejnym przykładem algorytmu typu EM używanym do agregacji anotacji jest opisany w ramach Raykar et al. [2010] algorytm regresji logistycznej działającej dla wielu anotatorów. Opisany algorytm działa zarówno dla klasyfikacji binarnej, jak i dla klasyfikacji wielu kategorii. Poniżej omówiłem sposób działania tego algorytmu dla klasyfikacji binarnej.

Zgodnie z algorytmem regresji logistycznej dla klasyfikacji binarnej (zob. Paragraf 2.2.1), prawdopodobieństwo wyboru kategorii $y = 1$ zdefiniowane jest w oparciu o dwa parametry $T = \{w, b\}$ i określone jest wzorem:

$$P(y = 1|x, w, b) = \sigma(w^\top x + b). \tag{2.30}$$

W ramach omawianego algorytmu zbiór parametrów regresji logistycznej T został rozszerzony o dodatkowe komponenty definiujące model jakości anotatora:

$$T = \{w, b, \alpha, \beta\}, \tag{2.31}$$

gdzie $\alpha = \{\alpha^1, \cdots, \alpha^R\}$ i $\beta = \{\beta^1, \cdots, \beta^R\}$ to odpowiednio czułość i swoistość (zob. Paragraf 2.2.2), które definiują model jakości dla każdego z R anotatorów.

W przypadku danych zebranych w procesie crowdsourcingu każdemu elementowi ze zbioru wejściowego x_i przypisany zostaje zbiór wartości wyjściowych $\{y_i^1, \cdots, y_i^R\}$ powstały z anotacji wykonanych przez R anotatorów. Proces uczenia modelu polega na wyborze parametrów T, które maksymalizują poniższe równanie:

$$P[D|T] = \prod_{i=1}^{N} P[y_i^1, \cdots, y_i^R | x_i, T] \tag{2.32}$$

W przypadku opisywanego algorytmu, równanie to przyjmuje postać:

$$P[D|T] = \prod_{i=1}^{N} [a_i p_i + b_i(1 - p_i)], \tag{2.33}$$

gdzie

$$\begin{aligned} p_i &= \sigma(w^\top x_i + b), \\ a &= \prod_{j=1}^{R} [\alpha^j]^{y_i^j} [1 - \alpha^j]^{y_i^j}, \\ b &= \prod_{j=1}^{R} [\beta^j]^{y_i^j} [1 - \beta^j]^{y_i^j}. \end{aligned} \tag{2.34}$$

Procedura 7 opisuje algorytm wyboru parametrów θ za pomocą metody EM:

Procedura 7: Algorytm Regresji logistycznej dla anotacji

Niech :

$\mathcal{D} = \{(x_i, \{y_i^1, \cdots, y_i^R\})\}_{i=1}^n$ – zbiór przykładów uczących,

$x_i \in X$ – gdzie X to zbiór wartości wejściowych,

$y_i \in Y$ – gdzie Y to zbiór wartości wyjściowych,

n – liczba mikro-zadań,

R – liczba anotatorów.

Kroki :

1 Inicjalizacja wartości początkowych:

– μ na podstawie zbioru $\mathcal{D}$ za pomocą algorytmu głosowania większościowego,

– α^j i β^j w oparciu o μ_i.

2 Krok E

– Aktualizacja predykcji finalnych anotacji:

$$\mu_i = \frac{a_i p_i}{a_i p_i + b_i(1 - p_i)}$$

3 Krok M

– Aktualizacja parametrów modelu jakości anotatorów:

$$\alpha^j = \frac{\sum_{i=1}^N \mu_i y_i^j}{\sum_{i=1}^N \mu_i}, \quad \beta^j = \frac{\sum_{i=1}^N (1 - \mu_i)(1 - y_i^j)}{\sum_{i=1}^N (1 - \mu_i)}$$

– Aktualizacja parametrów modelu regresji logistycznej (np. za pomocą metody gradientu prostego).

4 Kroki E i M powtarzane są do osiągnięcia zbieżności.

Wyjście:

5 μ – zawiera finalne anotacje wybrane dla wszystkich mikro-zadań,

6 α i β – określają jakość anotatorów.

Algorytm *AggNet*

Kolejnym algorytmem należącym do grupy *EM* jest algorytm *AggNet* opisany w ramach pracy Albarqouni et al. [2016]. *AggNet* stanowi przykład współcześnie rozwijanych metod nazywanych *Deep Crowd Learning*. W praktyce algorytm rozbudowuje wyżej opisany algorytm regresji logistycznej dla anotacji, zamieniając model regresji logistycznej na głęboką sieć neuronową.

W ramach projektu opisywanego w Albarqouni et al. [2016] stworzony został model klasyfikacji binarnej zdjęć raka piersi. Algorytm *AggNet* stanowi modyfikację architektury sieci neuronowej *Multi-Scale CNN Model*, która stosowana jest do przetwarzania obrazów[14]. Algorytm *AggNet* rozszerza *Multi-Scale CNN Model* o dodatkową, zaproponowaną w ramach pracy Albarqouni et al. [2016] warstwę *AggNet*[15], której zadaniem jest agregowanie zebranych anotacji w celu wybrania finalnych anotacji referencyjnych. W zaproponowanym algorytmie, parametry modelu obliczane są za pomocą algorytmu EM, który działa analogicznie do algorytmów opisanych wcześniej. Poniżej przedstawiłem opis procedury algorytmu *AggNet* (zob. Procedura 8).

[14]Publikacja zawiera dokładniejszy opis użytej architektury (zob. [Albarqouni et al., 2016]).

[15]Nazwa warstwy jest tożsama z nazwą algorytmu.

Procedura 8: Algorytm AggNet

Niech :

$\mathcal{D} = \{(x_i, \{y_i^1, \cdots, y_i^R\})\}_{i=1}^n$ – zbiór przykładów uczących,

$x_i \in X$ – gdzie X to zbiór wartości wejściowych,

$y_i \in Y$ – gdzie Y to zbiór wartości wyjściowych,

n – liczba mikro-zadań,

R – liczba anotatorów,

Kroki :

1 Inicjalizacja wartości początkowych:

– μ na podstawie zbioru $\mathcal{D}$ za pomocą algorytmu głosowania większościowego,

– α^j i β^j w oparciu o μ_i.

2 Krok E

– W oparciu o zbiór danych $\mathcal{D}$ oraz aktualne wartości parametrów ϕ aktualizowane predykcje finalnych anotacji μ.

$$\mu_i = \frac{a_i p_i}{a_i p_i + b_i(1 - p_i)}$$

3 Krok M

– W oparciu o zbiór $\mathcal{D}$ i aktualne wartości μ aktualizowane są parametry modelu jakości anotatorów: α i β oraz parametry sieci neuronowej.

4 Korki E i M powtarzane są aż do osiągnięcia zbieżności zgodnej z wyznaczonym progiem.

Wyjście:

5 μ – zawiera finalne anotacje wybrane dla wszystkich mikro-zadań,

6 α i β – określają jakość anotatorów.

2.4. Metody wyboru sygnałów nauczających

Ostatnim elementem kluczowym dla procesu nauczania maszynowego jest wybór algorytmu nauczyciela (zob. Paragraf 2.1.2). Algorytm nauczyciela odpowiada za wybór sygnałów nauczających, które będą przekazane do ucznia w czasie nauczania. Sygnał nauczający ma formę zadania, które ma rozwiązać uczeń (zob. Definicja 8). Po udzieleniu odpowiedzi otrzymuje on informację zwrotną i dowiaduje się czy udzielił poprawnej odpowiedzi (zob. Paragraf 2.1.1). Sygnały nauczające mogą przyjmować różną formą (predefiniowane, syntetyczne) i granulacją (przekazywane pojedynczo lub w grupach)[16]. W procesie crowdsourcingu przekazanie informacji zwrotnej w formie sygnałów uczących realizuje strategię *szkolenie anotatorów* komponentu *zapewnienie jakości* z taksonomii kontroli jakości (zob. Paragraf 1.3.4).

W ramach niniejszej pracy skupiłem się przede wszystkim na omówieniu zagadnienia modelu ucznia (zob. Podrozdział 2.2) i algorytmów obliczania jego parametrów (zob. Podrozdział 2.3) ponieważ są one bezpośrednio powiązane z tematem moich badań (zob. Rozdział 3 i Rozdział 4). Szczegółowe omówienie metod wyboru sygnałów nauczających wykracza poza ramy niniejszej rozprawy, dlatego temat ten omówiony będzie w ograniczonym stopniu, wystarczającym do pełniejszego zrozumienia procesu nauczania maszynowego. W następnym podrozdziale przedstawiłem opis przykładowych algorytmów wyboru sygnału nauczającego.

2.4.1. Algorytmy nauczyciela

Algorytmy nauczyciela działają w iteracjach, dokonując wyboru sygnału nauczającego na podstawie aktualnych odpowiedzi ucznia. Przykładowo, w podejściu zaproponowanym przez Rafferty et al. [2011], proces nauczania został opisany w formie częściowo obserwowalnego procesu decyzyjnego *Markowa* (ang. *partially observable Markov decision process, POMDP*) [Rafferty et al., 2011]. W tak zdefiniowanym procesie obecny stan wiedzy ucznia stanowi stan ukryty, który obserwowany jest w sposób pośredni podczas analizy udzielanych przez ucznia odpowiedzi. Nauczyciel wpływa na stan wiedzy ucznia poprzez wykonywanie akcji: przekazywanie sygnałów nauczających.

Algorytm nauczyciela *MaxGrad* zaproponowany przez P. Wang et al. [2021a] w każdej iteracji nauczania wybiera sygnał nauczający zawierający „najtrudniejsze" elementy zbioru przykładów uczących, czyli takie, które maksymalizują gradient $RyzykoNauczania$ (zob. Paragraf 2.1.2). Wybrane elementy przekładają się tym samym na optymalny wzrost obecnego stanu wiedzy ucznia (wartości $RyzykoNauczania$ obliczonej dla pełnego zbioru przykładów uczących). W celach demonstracyjnych poniżej szczegółowo opisałem działanie algorytmu *MaxGrad*.

[16]Charakterystyka sygnałów nauczających została omówiona na początku niniejszego rozdziału, zob. Paragraf 2.1.3.

Algorytm *MaxGrad*

W każdej iteracji t algorytm *MaxGrad* tworzy nowy sygnał nauczający $\mathcal{L}_t$ wybierając zbiór τ elementów, które w największym stopniu zminimalizują wybraną funkcję celu (np. błąd średniokwadratowy dla problemów regresji). Wartość funkcji celu obliczana jest dla zbioru $\mathcal{D}^{t-1}$ zawierającego elementy zbioru przykładów uczących, które nie są częścią zbioru nauczającego $\mathcal{L}$. Wartości referencyjne ze zbioru $\mathcal{D}^{t-1}$ porównywane są z estymacjami obliczonymi przez obecny model wiedzy ucznia. Ponieważ algorytm *MaxGrad* używa estymowanych odpowiedzi ucznia obliczonych dla zbioru $\mathcal{D}^{t-1}$, dlatego też jednym z wymagań tego algorytmu jest to, aby uczeń reprezentowany był za pomocą modelu wiedzy ucznia (zob. Paragraf 2.2). Opis działania algorytmu został zamieszczony poniżej (zob. Procedura 9).

Procedura 9: Algorytm MaxGrad (źródło P. Wang et al. [2021a])

Niech :

$\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}$ – zbiór przykładów uczących,

n – liczba mikro-zadań,

T – liczba iteracji nauczania,

$\mathcal{L}$ – zbiór nauczający,

τ – liczba przykładów nauczających w jednej iteracji,

f – model wiedzy ucznia,

Kroki :

1 $\mathcal{D}_u \leftarrow \emptyset$

2 $D^0 \leftarrow D$

3 Dla każdej, t-tej iteracji wykonaj następujące kroki:

– oblicz ξ_i dla każdego elementu D^{t-1}

– posortuj przykłady uczące malejąco względem parametru ξ_i

– stwórz zbiór $\mathcal{N}^t$ zawierający top τ z posortowanych przykładów uczących

– aktualizuj zbiór nauczający: $\mathcal{L}^t \leftarrow \mathcal{L}^{t-1} \cup \mathcal{N}^t$

– zaktualizuj model wiedzy ucznia: $f^{t+1} = f^*(\mathcal{L}^t)$

– $D^t \leftarrow D^{t-1} \setminus \mathcal{N}^t$

Wyjście:

4 $\mathcal{L}^t$ – sygnał nauczający w iteracji t

Dobór ξ oraz f^* może być dostosowany w zależności od nauczanego problemu. Przykładowo dla problemu klasyfikacji binarnej funkcje ξ oraz f^* mogą mieć następującą postać:

$$\begin{aligned} \xi_i &= (\phi(y_i f^t(x_i)))^2 \\ f^*(\mathcal{L}^t) &= \underset{f}{argmin} \sum_{(x_i,y_i)\in\mathcal{L}^t} \phi(y_i f(x_i)) \end{aligned} \tag{2.35}$$

2.5. Podsumowanie modelowania procesu informacji zwrotnej

W ramach niniejszego rozdziału przedstawiony został opis zastosowania procesu nauczania maszynowego w celu formalnego zamodelowania mechanizmu informacji zwrotnej w metodzie crowdsourcingu. Zastosowanie tego procesu w praktyce wymaga określenie konkretnych elementów, które zostaną użyte w trakcie implementacji. W niniejszym rozdziale opisane zostały trzy podstawowe elementy: modele ucznia, algorytmy używane do obliczenia parametrów modelu ucznia (oraz oceny jego pracy), a także algorytmy wyboru sygnału nauczającego.

W następnym rozdziale omówiłem autorski eksperyment, który miał na celu wykazanie zasadności użycia mechanizmu informacji zwrotnej w procesie crowdsourcingu. W szczególności celem eksperymentu było zweryfikowanie pozytywnego wpływu informacji zwrotnej na jakość danych pozyskiwanych dla problemów związanych z przetwarzaniem języka naturalnego.

ROZDZIAŁ 3

Wpływ informacji zwrotnej na jakość danych lingwistycznych

W niniejszym rozdziale pracy przedstawiłem opis autorskiego eksperymentu, którego celem było potwierdzenie skuteczności zastosowania mechanizmu informacji zwrotnej do podwyższenia jakości danych pozyskiwanych w ramach procesu crowdsourcingu.

W pierwszej części rozdziału omówiłem testowane hipotezy oraz pytania badawcze, a także opis przebiegu eksperymentu. W dalszej części szczegółowo opisałem zbiory danych, które testowane były w eksperymencie. W ostatniej części znalazł opis analizy wyników eksperymentów oraz podsumowanie.

3.1. Omówienie eksperymentu

3.1.1. Problem badawczy

Celem przeprowadzonego przeze mnie eksperymentu była weryfikacja wpływu informacji zwrotnej na jakość danych pozyskiwanych w procesie crowdsourcingu. W ramach niniejszej rozprawy problem ten został zbadany dla mikro-zadań, które związane były z przetwarzaniem tekstu[1]. Jak zostało to opisane wcześniej w ramach przeglądu istniejącej literatury przedmiotu, (zob. Paragraf 1.4.2) zagadnienie informacji zwrotnej w crowdsourcingu nie zostało jeszcze w pełni przebadane. W szczególności przeprowadzono niewiele eksperymentów związanych z oznaczaniem danych lingwistycznych. Jednak istniejące badania wykazały pozytywny wpływ informacji zwrotnej na jakość danych pozyskiwanych w zadaniach, które związane były z oznaczaniem danych innego typu. Z tego powodu przypuszczać można, że podobny efekt może zostać odtworzony dla zadań związanych z przetwarzaniem języka naturalnego. W ramach eksperymentu określone zostały dwie główne hipotezy badawcze:

H_1: Zapewnienie synchronicznej informacji zwrotnej w pozytywny sposób wpływa na jakość pozyskiwanych danych lingwistycznych w procesie crowdsourcingu.

[1] Część mikro-zadań zawierała również dane innego typu (np. graficzne), ale wszystkie analizowane zadania zawierają dane tekstowe w języku naturalnym.

H_2: Jakość przekazywanej informacji zwrotnej w procesie crowdsourcingu ma pozytywny wpływ na jakość pozyskiwanych danych.

W ramach omawianego eksperymentu powyższe hipotezy zostały zweryfikowane dla sześciu różnych zbiorów danych. Każdy ze zbiorów danych wykorzystanych w eksperymencie związany był z innym typem mikro-zadania. Dla każdego zbioru danych rozpatrzone zostaną hipotezy pomocnicze nazywane odpowiednio H_1^i oraz H_2^i. W ramach hipotez pomocniczych weryfikowana była zasadność hipotez głównych dla poszczególnych zbioru danych o numerze porządkowym i.

Hipotezy te są bezpośrednio związane z zasadnością użycia informacji zwrotnej w procesie crowdsourcingu. Prawdziwość hipotezy H_1 może stanowić argument na rzecz powszechniejszego użycia synchronicznej informacji zwrotnej w procesie crowdsourcingu. Natomiast prawdziwość hipotezy H_2 jest szczególnie ważna w przypadku informacji zwrotnej wygenerowanej automatycznie, ponieważ przeważnie jej jakość jest niższa niż w przypadku informacji zwrotnej stworzonej na podstawie oceny ekspertów.

Dodatkowo w ramach przeprowadzonych eksperymentów zweryfikowane zostały pytania badawcze, które rozszerzają zakres opisywanych analiz:

P_1: Czy informacja zwrotna w procesie crowdsourcingu ma efekt długoterminowy (edukacyjny)?

P_2: Czy informacja zwrotna ma wpływ na szybkość tworzenia anotacji w procesie crowdsourcingu?

P_3: Czy informacja zwrotna w procesie crowdsourcingu pozytywnie wpływa na zaangażowanie anotatorów?

Pytania pomocnicze zostały oznaczone odpowiednio jako P_1^i oraz P_2^i, gdzie i to numer porządkowy danego zbioru danych.

3.1.2. Forma przekazywanej informacji zwrotnej

Zgodnie z klasyfikacją opisaną we wcześniejszej części rozprawy (zob. Paragraf 1.4.1) forma informacji zwrotna może być opisana przez sześć wymiarów. Poniżej zostały opisane konkretne wartości, które opisują informację zwrotną stosowaną w ramach eksperymentu:

– czas: **asynchroniczny** – informacja zwrotna przekazywana była od razu po wysłaniu anotacji,

– źródło treści: **zbiór referencyjny** – w celu zminimalizowania wpływu czynników zewnętrznych źródłem informacji zwrotnej w eksperymencie były zbiory referencyjne o wysokiej jakości,

– szczegółowość: **punktowa** – informacja zwrotna generowana była automatycznie, miała formę zamkniętą i zawierała dwa elementy: informację binarną o poprawności anotacji (poprawna, niepoprawna) oraz odpowiedź ze zbioru referencyjnego,

– moment efektu: **natychmiastowy oraz długoterminowy** - oba efekty zostały zaobserwowane w eksperymencie (szczegóły omówione zostały w dalszej części rozdziału, zob. Paragraf 3.5),

– kanał: **bezpośredni** – informacja zwrotna przekazywana była bezpośrednio do anotatora.

Wartości wymiarów zostały również przedstawione w formie graficznej na diagramie pełnej klasyfikacji informacji zwrotnej (zob. Rysunek 3.1; opis wymiarów zob. Paragraf 1.4.1).

Wymiar	**Wartości**			
Czas	synchroniczny	asynchroniczny		
Źródło treści	ocena ekspertów	ocena innych anotatorów	automatycznie wygenerowana ocena	zbiór referencyjny
Szczegółowość	punktowa ocena	szablon oceny	otwarta forma ocen	
Liczba ocen	1:1	1:wiele	wiele:1	wiele:wiele
Moment efektu	natychmiastowy	długoterminowy		
Kanał komunikacji	bezpośredni	pośredni		

Rysunek 3.1: Klasyfikacja informacji zwrotnej użytej w eksperymencie (odpowiednie wartości są oznaczone na niebiesko).

3.1.3. Zbiory danych

Aby przeprowadzić bardziej zróżnicowaną analizę dotyczącą założonych hipotez, eksperyment został przeprowadzony dla sześciu różnych zbiorów danych. Każdy zbiór związany był z innym typem mikro-zadania. Użyte zbiory danych pokrywają różne typy mikro-zadań (np. klasyfikacja tekstu, oznaczanie wybranych fragmentów tekstu, estymacja wartości liczbowych), a ich treść pochodzi z różnych obszarów życia codziennego (np. *e-commerce*, usługi bankowe, usługi hotelowe). Dodatkowo praca z niektórymi zbiorami wymagała przyswojenia wiedzy domenowej charakterystycznej dla danego zadania (cztery zbiory), a do pracy z niektórymi wystarczała ogólna wiedza zdroworozsądkowa (dwa zbiory). Wszystkie zbiory danych, które zostały użyte w eksperymencie zostały przeze mnie udostępnione w publicznie dostępnym repozytorium projektu *Funcrowd*[2].

W Tabeli 3.1 przedstawione zostały szczegółowe informacje dotyczące każdego zbioru danych. Dla każdego zbioru podany został problem anotacyjny (typ mikro-zadania), opis danych, a także źródło pozyskania zbioru. Wszystkie powyższe zbiory danych zawierały tekst w języku angielskim. Język ten został wybrany ze względu na ułatwiony dostęp do dużej grupy potencjalnych anotatorów, którzy posługują się tym językiem.

[2] `https://github.com/heolin/funcrowd-dataset`; dostęp: 13.07.2022 r.

Tabela 3.1: *Zbiory danych użyte w eksperymencie*

Nazwa zbioru danych	Opis zbioru danych	Problem anotacyjny	Opis problemu	Źródło
skargi usług bankowych	Zbiór skarg zgłoszonych do produktów bankowych.	klasyfikacja	Anotatorzy dokonywali klasyfikacji tekstu skarg zgłoszonych do produktów. bankowych. Klasyfikacja polegała na dopasowaniu skargi do jednego z pięciu typów produktów bankowych (np. kredyt, karta kredytowa).	Zbiór publicznie dostępny za pośrednictwem oficjalnej strony amerykańskiej organizacji *Consumer Financial Protection Bureau*[a].
atrybuty produktów *eBay*[b]	Zbiór zawierający zdania wyekstrahowane z tytułów i opisów wybranych produktach dostępnych na platformie *e-commerce*, *eBay*.	oznaczanie fraz	Anotatorzy dokonywali oznaczania fraz będących atrybutami produktów (np. kolor lub marka produktu). Anotacja polegała na wybraniu jednego lub większej liczby słów i wyborze jednej z pięciu dostępnych kategorii.	Autorski zbiór stworzony na podstawie danych udostępnionych mi przez firmę *Webinterpret*[c].
waga produktów *eBay*	Zbiór zawierający informację o wybranych produktach dostępnych na platformie *e-commerce*, *eBay*.	szacowanie wagi	Anotatorzy dokonywali oszacowania wagi w gramach produktów z platformy *eBay*. Estymacja dokonywana była na podstawie informacji o tytule produktu, jego kategorii oraz zdjęciu.	Autorski zbiór stworzony na podstawie danych udostępnionych mi przez firmę *Webinterpret*.
wydźwięk opinii o hotelach	Zbiór zawierający opinie klientów hoteli. Zbiór zawierał tekst opinii oraz ocenę wystawioną przez klienta.	analiza wydźwięku	Anotatorzy oceniali wydźwięk opinii napisanych przed klientów hoteli. Ocena dokonywana była na pięciostopniowej skali.	Zbiór dostępny na stronie firmy *Datafinity* dostępny na ich oficjalnej stronie internetowej[d].
jednostki nazwane	Zbiór *GMB* (*Groningen Meaning Bank*). Zbiór stworzony specjalnie na potrzebę trenowania modeli oznaczających jednostki nazwane.	oznaczanie jednostek nazwanych	Anotatorzy dokonywali oznaczania jednostek nazwanych w podanym tekście. Anotacja polegała na wybraniu jednego lub większej liczby słów i wyborze jednej z pięciu dostępnych kategorii.	Zbiór publicznie dostępny za pośrednictwem oficjalnej strony internetowej projektu[e].
wyrazy bliskoznaczne	Autorski zbiór zawierający pary słów będące synonimami wraz z przykładowymi zdaniami, które pokazują użycie tych słów. Przykładowe zdania zostały wybrane ze zbioru *Gutenberg*.	klasyfikacja	Anotatorzy dokonywali klasyfikacji binarnej. Dla każdej pary słów dokonywali wyboru, czy podana para słów ma podobne znaczenie, czy nie.	Autorski zbiór danych.

[a] `https://www.consumerfinance.gov/data-research/consumer-complaints/`; dostęp: 06.06.2022 r.
[b] `https://www.ebay.com/`; dostęp: 29.03.2023 r.
[c] `https://www.webinterpret.com/`; dostęp: 29.03.2023 r.
[d] `https://datafiniti.co/products/business-data/`; dostęp: 06.06.2022 r.
[e] `https://gmb.let.rug.nl/data.php`; dostęp: 06.06.2022 r.

3.1.4. Rekrutacja uczestników eksperymentu

Uczestnicy eksperymentu (anotatorzy) rekrutowani byli za pomocą platformy crowdsourcingowej *Amazon Mechanical Turk* (dalej: *MTurk*) będącą obecnie największą platformą, na której wykonywane są zadania w tym modelu.

Mikro-zadania dostępne na platformie *MTurk* określane są terminem *HIT* od ang. *Human Intelligence Task*. *HIT* wykonywane są odpłatnie przez anotatorów, którzy zostali zatwierdzeni w procesie rejestracji[3]. Każdy anotator sam decyduje o tym, które zadanie chce wykonać. Wybór nie jest narzucany odgórnie przez system. Tworząc zadanie *HIT* zleceniodawca określa parametry definiujące strukturę zadania: nazwa zadania, interfejs anotacyjny, protokół anotacyjny, opłatę za wykonanie pojedynczego mikro-zadania. Zleceniodawca określa także dodatkowe metadane, takie jak tagi i opis, które ułatwiają anotatorom wyszukanie interesujących ich zadań. Dodatkowo zleceniodawca może określić filtry, które ograniczają dostęp do danego zadania tylko dla wybranych anotatorów (np. kraj pochodzenia, znane języki, osiągnięte progi dostępnych miar jakości). Zleceniodawca nie ma jednak wpływu na to, którzy anotatorzy zdecydują się wykonać jego zadanie, ani jaka będzie ich liczba.

3.2. Środowisko eksperymentu

3.2.1. Ograniczenia istniejących narzędzi

Głównym celem omawianego eksperymentu było zweryfikowanie skuteczności mechanizmu informacji zwrotnej. Jednak żadna z dostępnych platform crowdsourcingowych nie dostarcza mechanizmu, który pozwalałby na niezbędną modyfikację procesu anotacji, dostosowanie interfejsu, a także zebranie szczegółowych logów[4]. Z tego powodu, do eksperymentu użyte zostały dwa systemy. Do rekrutacji oraz obsługi anotatorów użyta została platforma crowdsourcingowa *MTurk*, a sam interfejs anotacyjny znajdował się w autorskim systemie *Funcrowd*, który został stworzony na potrzeby niniejszego eksperymentu. W systemie *Funcrowd* wdrożone zostały dwie funkcjonalności niedostępne w systemie *MTurk*, które były niezbędne do przeprowadzenia eksperymentu:

– **mechanizm przekazywania dynamicznej informacji zwrotnej**

Mechanizm ten pozwalał na przekazywanie informacji zwrotnej po każdym wypełnio-

[3] Rejestracja w systemie *Amazon Mechanical Turk* dostępna jest jedynie dla obywateli wybranych krajów. W dniu pisania niniejszej rozprawy nie jest ona dostępna dla obywateli Polski. Stan na: 06.06.2022 r.

[4] Stan na 28.02.2023 r.

nym mikro-zadaniu. Zaimplementowany mechanizm obsługiwał informację zwrotną o różnej szczegółowości (zob. Paragraf 1.4.1).

– **mechanizm grupowania mikro-zadań**

  Mechanizm ten został zaprojektowany tak, aby możliwe było wykonywanie mikro-zadań w partiach (np. jedna partia składa się z 50 mikro-zadań). Dzięki tej konfiguracji jeden anotator mógł wykonywać wiele mikro-zadań. Rozwiązanie to zostało zaimplementowane, aby ułatwić analizę długoterminowego efektu informacji zwrotnej.

Dokładniejszy opis działania systemu *Funcrowd* przedstawiony został w dalszej części tego rozdziału.

3.2.2. Platformy użyte w eksperymencie

Eksperyment przeprowadzony został za pomocą dwóch platform crowdsourcingowych. Anotatorzy rekrutowani byli na platformie *MTurk*. Uczestnicy, którzy decydowali się na wzięcie udział w eksperymencie, przenoszeni byli do zewnętrznej, autorskiej platformy – *Funcrowd*.

Platforma *Funcrowd* udostępniała uniwersalny interfejs, który umożliwiał przeprowadzanie kampanii w modelu crowdsourcingowym. Funkcjonalności systemu zostały również rozszerzone o dodatkowe elementy, które nie zostały użyte w eksperymencie (np. mechanizmy generowania raportów, mechanizm osiągnięć, rankingu anotatorów). System zbudowany był z dwóch głównych modułów:

– **funcrowd**[5]

  Jest to podstawowy moduł stanowiący serwer projektu. W tej części systemu zaimplementowano algorytmy obsługujące proces crowdsourcingu oraz algorytmy obsługi informacji zwrotnej. Moduł również odpowiadał za kontakt z bazą danych oraz komunikację za pomocą *REST API*.

– **funcrowd-frontend**[6]

  Ta część systemu obsługiwała interfejs użytkownika dla platformy dostępny w formie strony internetowej. Oprócz interfejsu anotacji moduł ten zawierał dodatkowe wido-

[5] Moduł ten został napisany w języku programowania *Python* za pomocą biblioteki *Django*. Kod źródłowy dostępny jest pod linkiem: `https://github.com/heolin/funcrowd`; dostęp: 22.06.2022 r.

[6] Moduł ten został napisany w języku programowania *JavaScript* za pomocą technologii *React*. Kod źródłowy dostępny jest pod linkiem: `https://github.com/heolin/funcrowd-frontend`; dostęp: 22.06.2022 r.

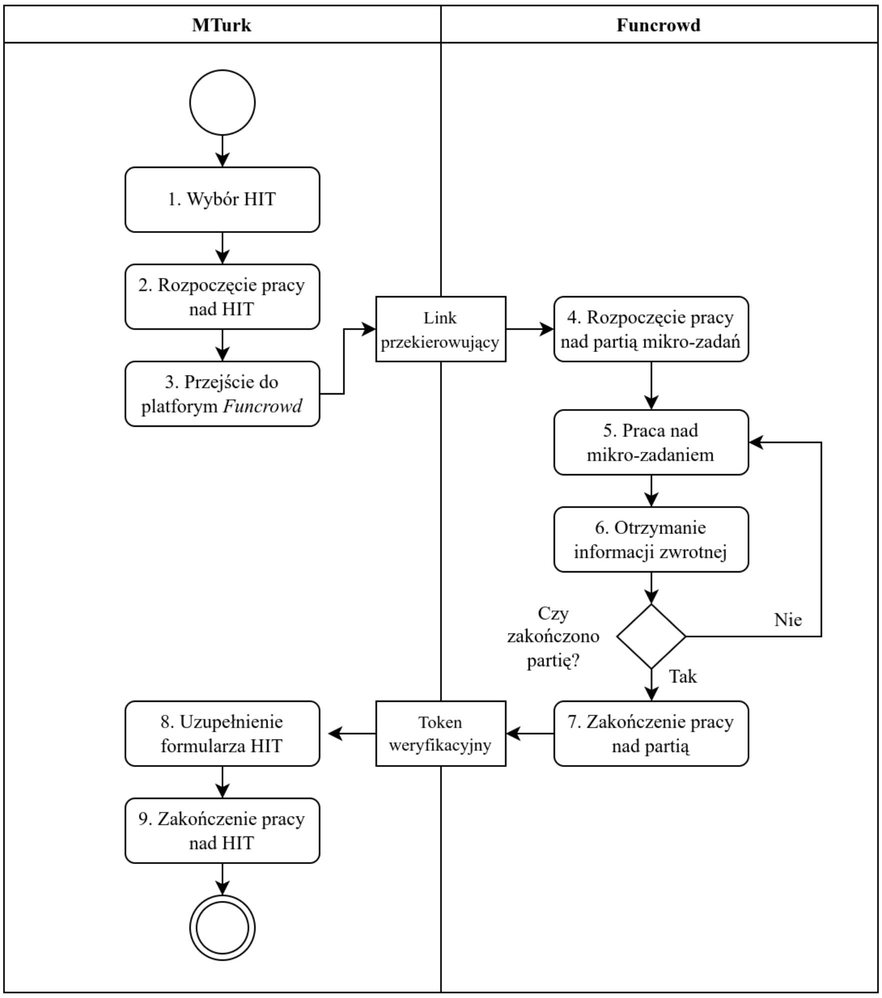
ki, np. interfejs wyboru kampanii, ranking anotatorów, panel ustawień użytkownika systemu.

Kod źródłowy obu modułów został w całości udostępniony publicznie na licencji otwartej.

Platforma *Funcrowd* została również użyta w innych eksperymentach i projektach, które nie stanowiły części niniejszej rozprawy, m.in. w projekcie *Sprawdzamy Jak Jest*[7], w którym anotatorzy nieodpłatnie weryfikują dokumenty przesłane przez polskie instytucje publiczne.

3.2.3. Przebieg eksperymentu

Udział w eksperymencie zdefiniowany był przez wieloetapowy proces, w którym użytkownik wykonujący określone zadanie, przemieszczał się pomiędzy dwoma platformami. Proces przebiegał w ten sam sposób dla każdego z 6 badanych zbiorów danych. Poniżej opisane zostały kluczowe kroki definiujące jego przebieg (zob. Rysunek 3.2):



Rysunek 3.2: Diagram przebiegu procesu eksperymentu

[7]https://sprawdzamyjakjest.pl/; dostęp: 05.06.2021 r.

1. **Wybór zadania HIT**

   Anotator wybierał jedno zadanie *HIT* z listy zadań dostępnych na platformie *MTurk*.

2. **Rozpoczęcie pracy nad zadaniem HIT**

   Anotator zapoznawał się z instrukcją określającą sposób pracy nad danym zadaniem. Opis ten zawierał techniczne informacje, które przedstawiają proces wykonywania zadania, ogólny zarys tematu zadania oraz sposób odebrania wynagrodzenia.

3. **Przejście do platformy *Funcrowd***

   W treści każdego zadania *HIT* znajdował się unikalny link, który przekierowywał anotatora do platformy *Funcrowd*. Po przekierowaniu anotator zostawał automatycznie zalogowany od systemu *Funcrowd* przy pomocy metadanych przekazanych w linku. Zalogowany użytkownik otrzymywał dostęp do interfejsu anotacji mikro-zadań.

4. **Rozpoczęcie pracy nad partią mikro-zadań**

   Jeżeli użytkownik po raz pierwszy wykonywał dany typ mikro-zadania, to system automatycznie wyświetlał okno z pełnym protokołem anotacyjnym. Po zapoznaniu się z protokołem anotator rozpoczynał anotacje partii mikro-zadań. Pojedyncza partia mikro-zadań zawierał dokładnie 50 elementów.

5. **Rozwiązywanie mikro-zadań**

   Anotator używał interfejsu anotacyjnego do oznaczenia przekazanego mikro-zadania.

6. **Otrzymanie informacji zwrotnej**

   Po przesłaniu wykonanej anotacji anotator otrzymywał informację zwrotną. Treść wyświetlonej informacji zwrotnej zależała od wariantu eksperymentu, do którego przypisany został anotator. Możliwe były dwie sytuację:

   – anotator otrzymywał informację zwrotną do stworzonej anotacji,

   – anotator otrzymywał tylko informację o zapisaniu anotacji.

7. **Zakończenie pracy nad partią mikro-zadań**

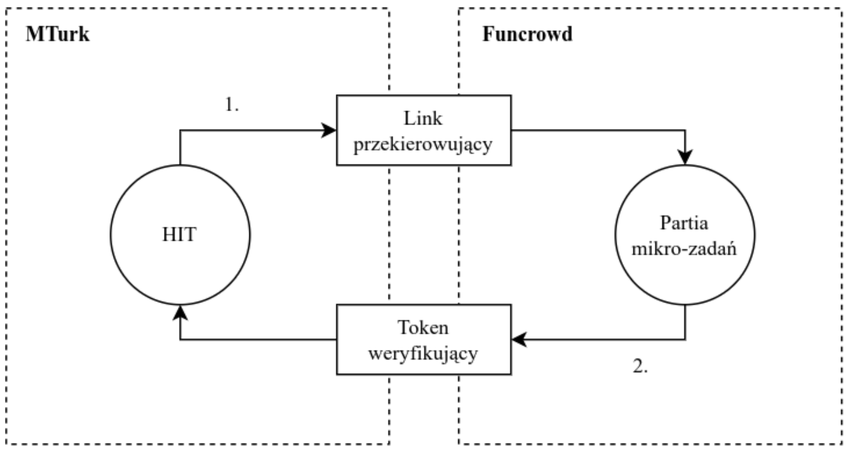
   Po zakończeniu pracy nad wszystkimi mikro-zadaniami z danej partii system *Funcrowd* przekazywał unikalny token weryfikacyjny, który potwierdzał wykonanie zadań.

8. **Zakończenie pracy nad zadaniem HIT**

   Aby otrzymać wynagrodzenie za wykonaną pracę, anotator przeklejał otrzymany w systemie *Funcrowd* token do formularza w systemie *MTurk*. Po podaniu tokenu anotator potwierdzał wykonanie danego *HIT*.

3.2.4. Integracja platform

Eksperyment został przeprowadzony przy użyciu dwóch niezależnych platform. Na Rysunku 3.3 zaprezentowany został ogólny schemat komunikacji pomiędzy systemami *Funcrowd* oraz *MTurk*.



Rysunek 3.3: Diagram integracji pomiędzy systemami używanymi w ramach eksperymentu

1. **Komunikacja z platformy *MTurk* do platformy *Funcrowd***

   Do każdego zadania *HIT* wykonywanego w ramach eksperymentu przypisany został link przekierowujący do systemu *Funcrowd*. W momencie wyświetlenia strony z zadaniem *HIT*, link zostaje automatycznie uzupełniony o dwa parametry:

   – **workerId** – unikalny identyfikator przypisany do każdego użytkownika platformy *MTurk*[8],

   – **packageId** – unikalny identyfikator przypisany do każdej partii mikro-zadań w systemie *Funcrowd*.

2. **Komunikacja z platformy Funcrowd do platformy *MTurk***

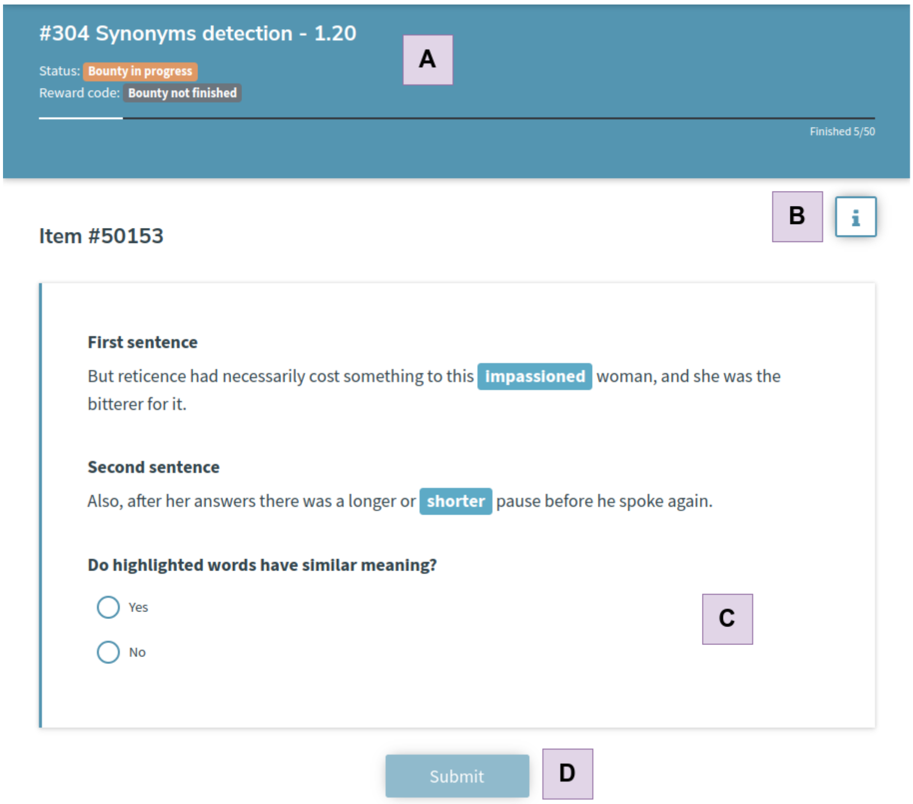
   System *Funcrowd* po otrzymaniu wyżej opisanych parametrów automatycznie tworzy konto dla danego użytkownika (jeżeli nie zostało ono wcześniej stworzone), a następnie loguje go do systemu. W celu weryfikacji wykonania danego zadania użyty

[8]Identyfikator *workerId* pozostawał stały dla użytkowników systemu *MTurk*, dzięki czemu możliwe była porównywanie anotacji tych samych anotatorów w różnych zadaniach.

jest unikalny token weryfikacyjny w formie *UUID* (ang. *Universally unique identifier*). Token ten przypisany jest do zadania *HIT* oraz użytkownika, wyświetlając się dopiero po zakończeniu pracy nad partią mikro-zadań, dzięki czemu możliwa była weryfikacja wykonanej pracy przed przekazaniem wynagrodzenia[9].

3.2.5. Interfejsu anotacji

Interfejs anotacji składał się z dwóch głównych elementów: panelu postępów pracy oraz panelu zadania. W interfejsie używane są również dwa pomocnicze okna: okno informacji zwrotnej oraz okno protokołu anotacyjnego. Na Rysunku 3.4 zaprezentowany został przykład interfejsu użytego w badaniu. Kluczowe elementy interfejsu oznaczone zostały literami na rysunku, a poniżej zamieszczony został opis każdego z nich:



Rysunek 3.4: Przykład interfejsu anotacyjnego dla jednego z badanych zbiorów danych

A. Panel postępów pracy

W górnej części znajduje się panel postępów pracy, który zawierał informację o aktualnej liczbie wykonanych mikro-zadań oraz liczbę wszystkich mikro-zadań w danej partii

[9]Podczas trwania eksperymentu wielokrotnie zdarzały się próby oszukania systemu poprzez podanie niepoprawnego tokenu lub użycie tego samego tokenu wielokrotnie.

(zob. Rysunek. 3.4, oznaczenie **A**). Po zakończeniu anotacji wszystkich mikro-zadań na panelu był wyświetlony token weryfikacyjny (zob. Rysunek. 3.5).

Rysunek 3.5: Panel postępów pracy z odblokowanym tokenem weryfikacyjnym dla ukończonej partii mikro-zadań.

B. Panel zadania

W dolnej części interfejsu znajdował się panel zadania. Panel ten zawierał komponenty wyświetlające dane wejściowe oraz formularz umożliwiający wprowadzenie oznaczeń dla mikro-zadań (zob. Rysunek. 3.4, oznaczenie **C**). Po wypełnieniu formularza anotator przesyłał swoją odpowiedź poprzez naciśnięcie odpowiedniego przycisku (zob. Rysunek. 3.4, oznaczenie **D**). Komponenty wyświetlane na panelu zadania różniły się w zależności od treści zadania.

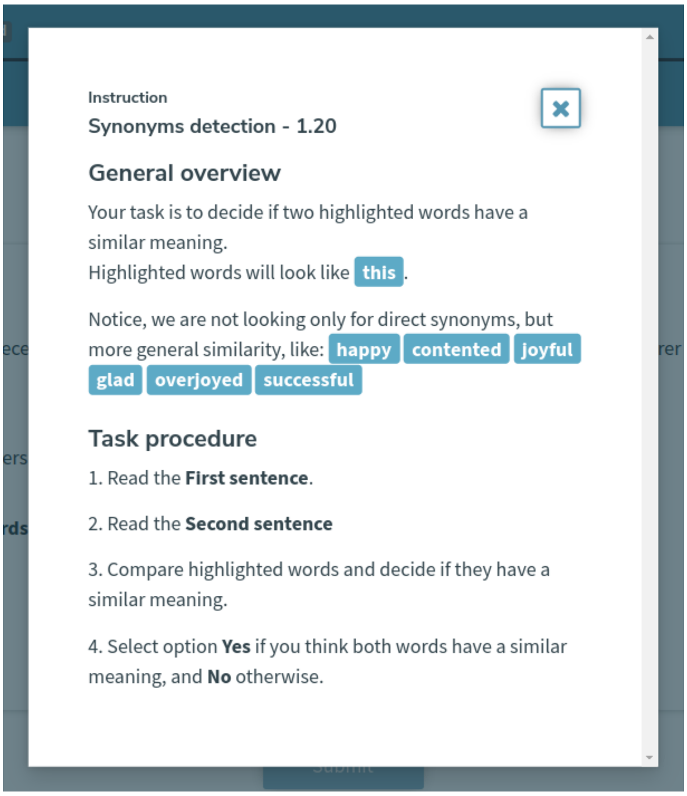
C. Okno informacji zwrotnej

Okno informacji zwrotnej pojawiało się po przesłaniu odpowiedzi przez anotatora. W zależności od wariantu eksperymentu okno mogło zawierać informację zwrotną do udzielonej anotacji lub informację o poprawnym zapisaniu udzielonej odpowiedzi. Komponenty wyświetlane w tym oknie różniły się w zależności od treści zadania.

D. Okno protokołu anotacyjnego

Okno zawierające protokół anotacyjny (zob. Rysunek 3.6) wyświetlało się w sytuacji, gdy anotator po raz pierwszy rozpoczynał pracę nad danym typem mikro-zadania. Dodatkowo anotator mógł wyświetlić okno protokołu w dowolnym momencie pracy nad mikro-zadaniem za pomocą odpowiedniego przycisku (zob. Rysunek. 3.4, oznaczenie **B**).

Ponieważ eksperyment przeprowadzany został dla odbiorców posługujących się językiem angielskim, wszystkie protokoły napisane zostały w tym języku. Treść wszystkich protokołów anotacyjnych została zamieszczona w Dodatku A.



Rysunek 3.6: Przykład zawartości okna protokołu anotacyjnego dla jednego z badanych zbiorów danych

3.3. Warianty eksperymentu

Eksperyment został przeprowadzony w czterech różnych wariantach: trzy warianty, w ramach których przekazywana była informacja zwrotna o różnej jakości oraz wariant bez informacji zwrotnej, który pełnił funkcję grupy kontrolnej.

Proces pozyskiwania danych dla różnych wariantów eksperymentu rozłożony był w czasie. Z tego powodu nie było możliwe użycie, tylko jednego, wspólnego wariantu kontrolnego. Aby upewnić się, że zmiany wprowadzone w ramach danego wariantu były jedynym czynnikiem różnicującym badane grupy (a nie np. czas wykonywania zadania), dla każdego z badanych wariantów stworzony został podzbiór danych z wariantu kontrolnego. Każdy podzbiór zawierał anotację zebrane dla grupy kontrolnej w tym samym czasie co dane dla danego wariantu eksperymentu. W celu zwiększenia czytelności wszystkie opisane powyżej warianty zostały podzielone na dwie grupy:

1. **Przekazywanie informacji zwrotnej**

   Grupa pierwsza zawierała warianty, w których anotatorzy otrzymywali informację zwrotną dotyczącą jakości ich oznaczenia. W ramach tej grupy wyróżnione zostały trzy warianty, które określały jakość przekazywanej informacji zwrotnej:

(a) Wysoka jakość (W_1)

Informacja zwrotna była stworzona w oparciu o poprawne anotacje referencyjne danego mikro-zadania.

(b) Umiarkowana jakość (W_2)

Informacja zwrotna była stworzona w oparciu o anotacje referencyjne, których część (około jednej czwartej) została zastąpiona błędnymi anotacjami.

(c) Niska jakość (W_3)

Informacja zwrotna była stworzona w oparciu o anotacje referencyjne, których znaczna część (około połowy) została zastąpiona błędnymi anotacjami.

Dokładny sposób określenia jakości anotacji dla każdego wariantu różni się dla każdego zbioru danych.

2. **Grupa kontrolna**

W ramach tej drugiej grupy wariantów anotatorom nie była przekazywana informacja zwrotna. Kategoria ta obejmowała warianty kontrolny składający się z trzech podwariantów – jeden dla każdego wariantu z grupy pierwszej:

(a) Brak informacji zwrotnej (W_1^c)

Wariant kontrolny dla wariantu W_1.

(b) Brak informacji zwrotnej (W_2^c)

Wariant kontrolny dla wariantu W_2.

(c) Brak informacji zwrotnej (W_3^c)

Wariant kontrolny dla wariantu W_3.

Każdy anotator został przydzielony do jednego z czterech wariantów w sposób losowy podczas pierwszego logowania na platformę *Funcrowd.* Wariant, do którego przypisany był dany anotator, pozostaje niezmienny dla wszystkich mikro-zadań anotatora.

3.4. Zbiory danych

Każdy z badanych zbiorów danych zawierał mikro-zadania, które związane były z interpretacją treści danych lingwistycznych. W ramach niniejszego podrozdziału, dla każdego zbioru danych przedstawione zostały: opis treści danych oraz ich źródła, opis interfejsu anotacyjnego, a także sposób generowania informacji zwrotnej.

3.4.1. Zbiór: skargi usług bankowych

Zbiór danych

Zbiór danych użyty w eksperymencie to podzbiór publicznie dostępnego zbioru *Consumer Complaint Database*. Zbiór ten stanowi bazę danych zawierającą skargi składane przez amerykańskich klientów usług z zakresu bankowości. Zawarte są w nim szczegółowe informacje dotyczące złożonej skargi, które określają m.in. nazwę firmy świadczącą usługę, na którą została złożona skarga, kategorię określającą typ produktu bankowego (np. karta kredytowa, konto bankowe), temat oraz pełną treść skargi. Zbiór ten jest dostępny za pośrednictwem oficjalnej strony amerykańskiej organizacji *Consumer Financial Protection Bureau* (zob. Tabela 3.1).

Opis mikro-zadania

W ramach mikro-zadania stworzonego dla tego zbioru danych anotatorzy dokonywali przypisania jednej z pięciu dostępnych kategorii do treści skargi. Treść skargi była wyświetlana w formie kilkuzdaniowego tekstu. Każda z dostępnych kategorii określa inny produkt bankowy, czyli szeroko rozumianą usługę oferowaną przez dany bank. Dostępne kategorie obejmowały:

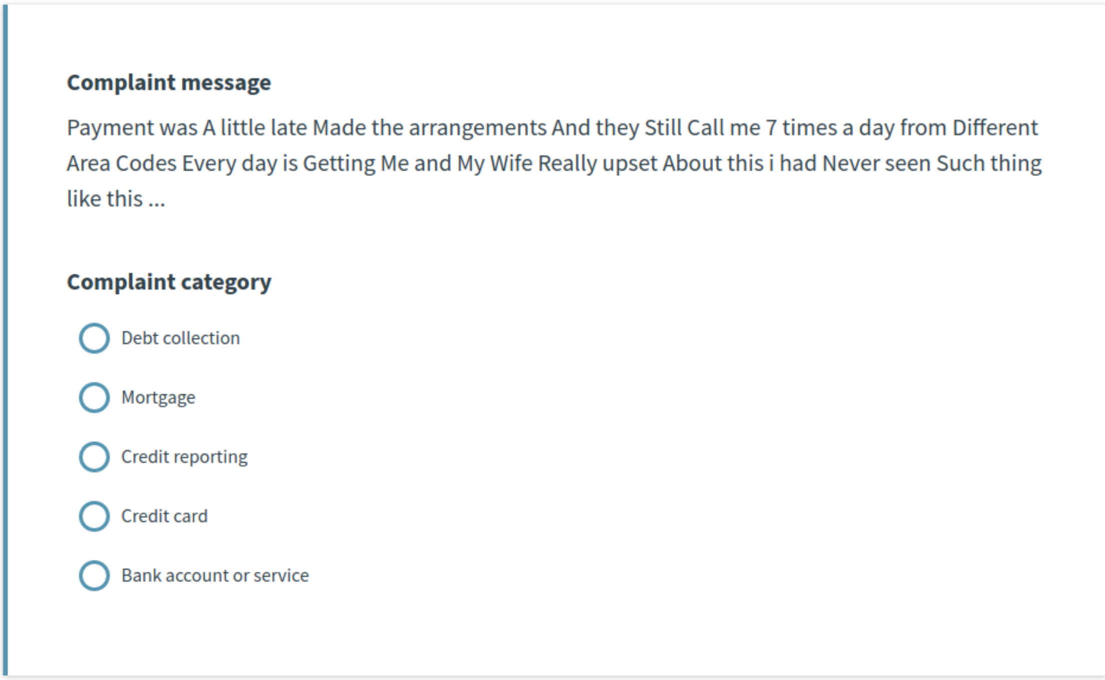
- *Debt collection* (tłum. pobór długów) – np. otrzymywanie telefonów od komornika,
- *Mortgage* (tłum. kredyt hipoteczny) – np. dwukrotne pobranie tej samej raty kredytu,
- *Credit reporting* (tłum. raporty kredytowe) – np. błędna nazwa pracodawcy,
- *Credit card* (tłum. karta kredytowa) – np. nieprawidłowy stan zadłużenia karty,
- *Bank account or service* (tłum. konto bankowe lub usługi) – np. sprawdzenie salda konta.

Treść skargi podobnie jak kategorie, które wybierane były w ramach anotacji, są w języku angielskim.

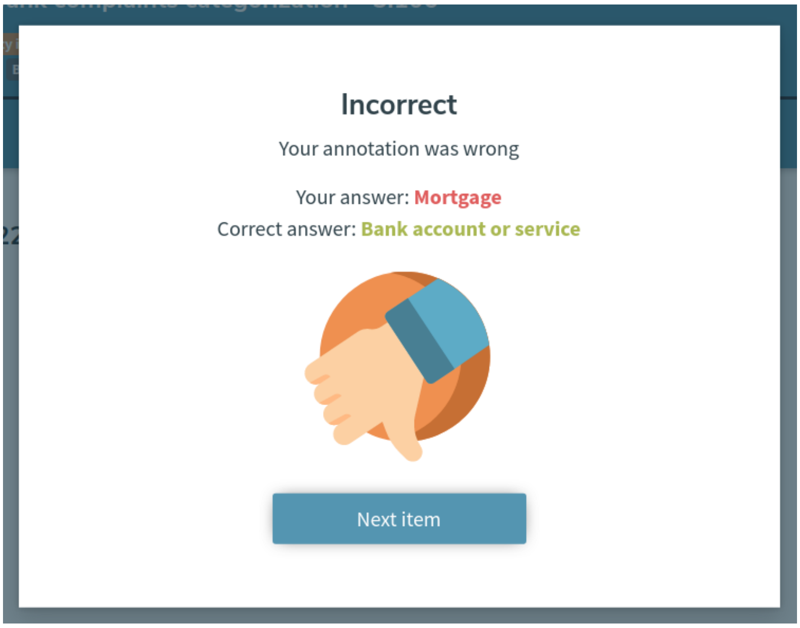
Mimo że zadanie to dotyczyło tematu usług bankowych, które anotatorzy mogą znać z życia codziennego, to prawidłowe oznaczanie kategorii skarg wkraczało poza wiedzę zdroworozsądkową. Podczas pracy nad kolejnymi mikro-zadaniami z tego zbioru, anotatorzy nabywali wiedzę domenową dotyczącej prawidłowej identyfikacji każdej z kategorii na podstawie treści załączonej skargi.

Przed rozpoczęciem pracy nad mikro-zadaniem system wyświetla protokół anotacyjny. Protokół anotacyjny zawierał techniczną instrukcję opisującą sposób pracy nad zadaniem, a także tabelę zawierającą opis każdej z kategorii produktów bankowych. Dokładna treść protokołu anotacyjnego została zamieszczona w załączniku (zob. Dodatek A.1).

Interfejs anotacyjny mikro-zadania obejmował dwa pola: *complaint message* (tłum. treść skargi), *complaint category* (tłum. kategoria skargi) (zob. Rysunek 3.7). Anotator zapoznawał się z treścią mikro-zadania, a następnie wybierał kategorię w polu *complaint category*. Po przesłaniu anotacji system wyświetlał okno zawierające informację zwrotną (zob. Rysunek 3.8).



Rysunek 3.7: Interfejs anotacyjny dla zadania „skargi usług bankowych”



Rysunek 3.8: Okno zawierające informację zwrotną dla zadania „skargi usług bankowych”

Przygotowanie zbioru danych

Pełen zbiór danych *Consumer Complaint Database* był zbyt złożony (zawierał zbyt wiele pól i możliwych kategorii), by użyć go w mikro-zadaniu w niezmienionej formie. Z tego powodu zbiór został poddany wstępnej obróbce, w celu dostosowania go do formatu, który mógł być oznaczony przy pomocy metody crowdsourcingu. Wytyczne dotyczące optymalnej konfiguracji zadań crowdsourcingowych nie są precyzyjnie określone przez literaturę, dlatego kroki opisane poniżej wybrane zostały na podstawie własnych doświadczeń eksperckich.

1. Ze zbioru *Consumer Complaint Database* wybrane zostały dwie kolumny (z 18 dostępnych kolumn): *Consumer complaint narrative* (tłum. treść skargi w formie tekstowej) oraz *Product* (tłum. nazwa produktu bankowego).

2. Ze zbioru wybrane zostały skargi należące do podzbioru pięciu (z 18 dostępnych) kategorii: *Debt collection*, *Mortgage*, *Credit reporting*, *Credit card*, *Bank account or service*[10].

3. Ze zbioru usunięte zostały skargi, w których długość treści skargi nie była odpowiednia do użycia w mikro-zadaniu:

   – Minimalna akceptowana długość treści to 30 znaków,

[10]Ostateczny zbiór kategorii został wybrany przeze mnie w sposób arbitralny. Zbiór ten zostały wybrany tak, aby uniknąć zbyt podobnych do siebie kategorii, a tym samym ułatwić zadanie dla anotatorów.

  - Maksymalna długość treści to 250 znaków.

4. W celu zapewnienia minimalnego poziomu trudności mikro-zadania usunięte zostały wszystkie wiersze, których treść wprost sugerowała właściwą kategorię[11]. Usunięte zostały skarg, w których opisy bezpośrednio zawierały słowa pojawiające się w nazwach kategorii. Łącznie wybrane zostało 7 słów, które:

   - *mortgage* (tłum. kredyt hipoteczny),
   - *card* (tłum. karta),
   - *collect* (tłum. pobór),
   - *account* (tłum. konto),
   - *report* (tłum. raport),
   - *debt* (tłum. dług),
   - *service* (tłum. serwis).

5. Dla każdej z 5 kategorii wylosowane zostało 400 mikro-zadań. W ten sposób finalny zbiór zawierał 2000 mikro-zadań.

6. Wszystkie mikro-zadania w zbiorze końcowym zostały posortowane w sposób losowy.

Informacja zwrotna

Treść informacji zwrotnej zależna była od wariantu eksperymentu, do którego przypisany był anotator. W każdym wariancie przekazywana była informacja zwrotna o innej jakości:

- Wysoka jakość (W_1)
  Informacja zwrotna przekazywana anotatorom zawierała anotacje referencyjne.

- Umiarkowana jakość (W_2)
  Informacja zwrotna zawierała losowo zniekształcone anotacje referencyjne. Anotacje zostały zniekształcone tak, by przy porównywaniu z wartościami referencyjnymi otrzymywały dokładność (ang. *accuracy*) na poziomie $ACC = 0{,}75$.

- Niska jakość (W_3)
  Informacja zwrotna zawierała losowo zniekształcone anotacje referencyjne. Anotacje zostały zniekształcone tak, by przy porównywaniu z wartościami referencyjnymi otrzymywały dokładność (ang. *accuracy*) na poziomie $ACC = 0{,}55$.

[11] W przypadku zbyt prostych mikro-zadań efekt eksperymentu był trudniejszy do zaobserwowania.

Zniekształcenie anotacji

W celu zmniejszenia jakości przekazywanej informacji zwrotnej referencyjne anotacje zostały zniekształcone tak, by zawierały błędy. Algorytm zniekształcenia referencyjnych miał formę iteracyjnej procedury. W każdej iteracji w sposób losowy wybierane było jedno, niezmodyfikowane wcześniej mikro-zadanie. Dla wybranego mikro-zadania obecna anotacja referencyjna zastępowana była losową, niewłaściwą kategorią usług bankowych. Proces ten powtarzany był aż do osiągnięcia określonej wartości wybranej metryki jakości (dokładność). W przypadku obu wariantów zawierających zniekształcone dane (W_2, W_3) został zastosowany ten sam algorytm różniący się tylko docelową wartością metryki jakości.

3.4.2. Zbiór: atrybuty produktów *eBay*

Zbiór danych

Zbiór użyty w tym eksperymencie zawierał dane produktów dostępnych na platformie *e-commerce eBay*. Jest to autorski zbiór stworzony na potrzebę tego eksperymentu. Surowe dane, które zostały użyte do przygotowania zbioru zostały udostępnione przez firmę *Webinterpret* (zob. Tabela 3.1).

Opis mikro-zadania

Celem mikro-zadania stworzonego dla omawianego zbioru było oznaczanie atrybutów produktów w przedstawionym tekście. Jako atrybuty rozumiane są fragmenty tekstu, które oznaczają cechy i znaki szczególne danego produktu (np. jego kolor, marka). W ramach tego mikro-zadania anotatorzy oznaczali w tekście atrybuty należące do jednej z sześciu kategorii:

– *Brand* (tłum. marka produktu),

– *Material* (tłum. materiał) – np. bawełna,

– *Size* (tłum. rozmiar produktu) – np. duży rozmiar,

– *Pattern* (tłum. wzór) – np. paski,

– *Color* (tłum. kolor produktu) – np. czerwony,

– *Deparment* (tłum. dział produktu) – np. męski, dziecięcy.

W związku z tym, że zadanie dotyczyło identyfikacji atrybutów przedmiotów z platformy *e-commerce*, część anotatorzy mogła mieć już styczność z niektórymi atrybutami.

Podczas pracy nad kolejnymi mikro-zadaniami z tego zbioru, anotatorzy nabywali wiedzę domenową dotyczącą prawidłowej identyfikacji każdej kategorii atrybutów, a ich wiedza domenowa była dodatkowo wspierana przez wiedzę zdroworozsądkową zdobytą wcześniej (np. podczas dokonywania zakupów na takiej platformie).

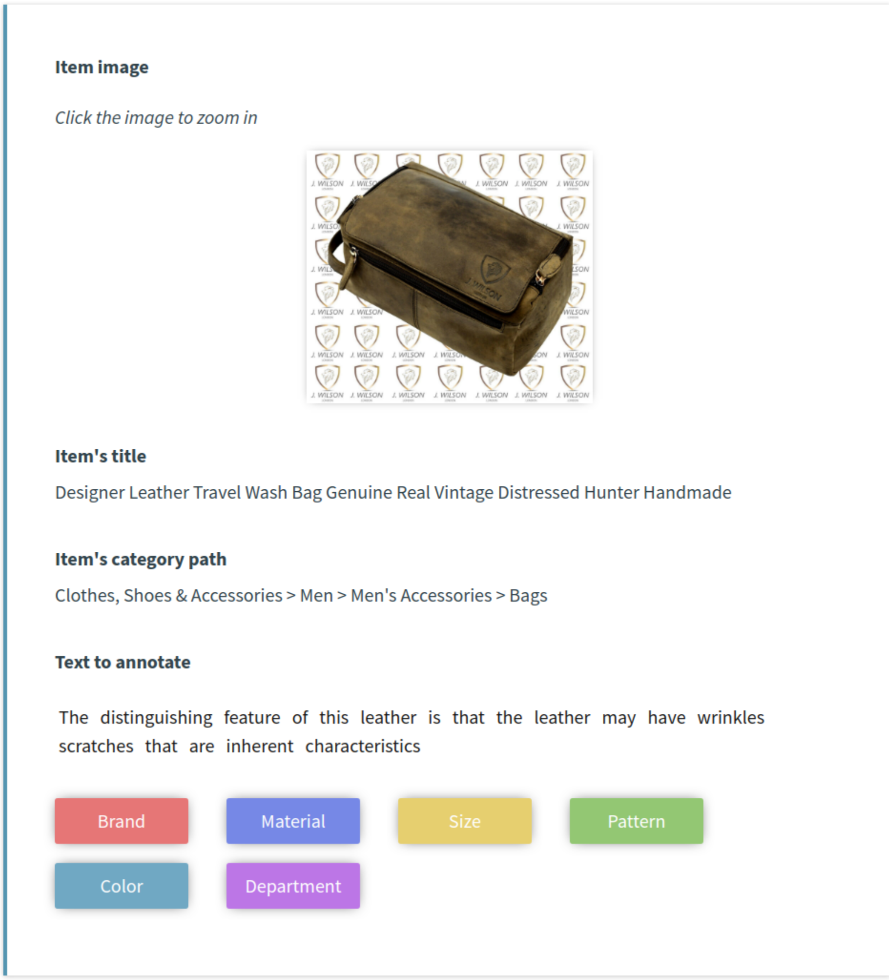
Przed rozpoczęciem pracy nad mikro-zadaniem system wyświetlał protokół anotacyjny. Protokół anotacyjny zawierał techniczną instrukcję opisującą sposób pracy nad zadaniem. Dodatkowo w protokole znalazła się tabela zawierającą przykładowe wartości dla każdej z kategorii atrybutów. Dokładna treść protokołu anotacyjnego została zamieszczona w załączniku (zob. Dodatek A.2).

Interfejs anotacyjny mikro-zadania zawierał cztery pola: *Item's image* (tłum. zdjęcie produktu), *Item's title* (tłum. tytuł produktu), *Item's category path*[12] (tłum. ścieżka kategorii produktu) oraz *Text to annotate* (tłum. anotowany tekst). Tekst anotowany w ramach mikro-zadania został wyekstrahowany z tytułu lub opisu danego produktu. Do zbioru wybrane zostały mikro-zadania, które według zbioru referencyjnego zawierały przynajmniej jeden atrybut.

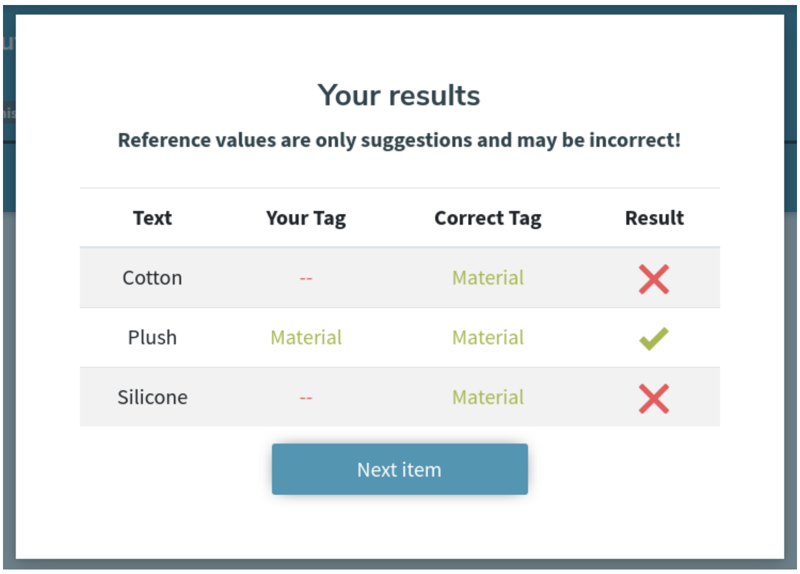
Anotator oznaczał atrybuty poprzez wybranie jednego lub więcej tokenów[13] oraz przypisanie ich do wybranej kategorii opisującej atrybut (zob. Rysunek 3.9). Po przesłaniu anotacji system wyświetlał okno zawierające informację zwrotną (zob. Rysunek 3.10).

[12]Taksonomia kategorii produktów sprzedawanych na platformie *eBay* ma formę drzewa. Głębokość kategorii w drzewie określa stopień ogólności danej kategorii. Każdy produkt przypisany jest do jednej kategorii będącej liściem w drzewie kategorii. Ścieżka kategorii zawierała kategorię od liścia do korzenia drzewa.

[13]Jako tokeny rozumiane są dowolne ciągi znaków oddzielone spacją.



Rysunek 3.9: Interfejs anotacyjny dla zadania „atrybuty produktów *eBay*”



Rysunek 3.10: Okno zawierające informację zwrotną dla zadania „atrybuty produktów *eBay*”

Przygotowanie zbioru danych

Z pełnego zbioru danych został wylosowany podzbiór mikro-zadań. Losowanie przeprowadzone zostało tak, aby dla każdego z pięciu kategorii atrybutów występowało $n = 400$ mikro-zadań. Łącznie wybrane zostało 2000 elementów.

Informacja zwrotna

Treść informacji zwrotnej zależna była od wariantu eksperymentu, do którego przypisany był anotator. W każdym wariancie przekazywana była informacja zwrotna o innej jakości:

- Wysoka jakość (W_1)
  Informacja zwrotna przekazywana anotatorom zawierała anotacje referencyjne.
- Umiarkowana jakość (W_2)
  Informacja zwrotna zawierała losowo zniekształcone anotacje referencyjne. Anotacje zostały zniekształcone tak, by przy porównywaniu z wartościami referencyjnymi otrzymywały jakość mierzoną metryką F_1 na poziomie $F_1 = 0{,}75$.
- Niska jakość (W_3)
  Informacja zwrotna zawierała losowo zniekształcone anotacje referencyjne. Anotacje zostały zniekształcone tak, by przy porównywaniu z wartościami referencyjnymi otrzymywały jakość mierzoną metryką F_1 na poziomie $F_1 = 0{,}55$.

Zniekształcenie anotacji

W celu zmniejszenia jakości przekazywanej informacji zwrotnej referencyjne anotacje zostały zniekształcone tak, by zawierały błędy. Algorytm zniekształcenia referencyjnych miał formę iteracyjnej procedury. W każdej iteracji w sposób losowy wybierane było jedno, niezmodyfikowane wcześniej mikro-zadanie. Dla wszystkich tokenów w wybranym mikro-zadaniu zastosowane zostały dwie reguły wprowadzenia zniekształceń:

- Z prawdopodobieństwem p_d istniejący token zmieniony został na kategorię pustą (brak atrybuty).
- Z prawdopodobieństwem p_2 istniejąca anotacja (pusta lub nie) zastępowana była przez losowo wybraną kategorię.

Proces ten powtarzany był aż do osiągnięcia określonej wartości wybranej metryki jakości (F_1). W przypadku obu wariantów zawierających zniekształcone dane (W_2, W_3) został zastosowany ten sam algorytm różniący się tylko doborem parametrów p_d (średnia jakość: $p_d = 0{,}25$, niska jakość: $p = 0{,}4$), p_r (W_2: $p = 0{,}05$, W_3: $p = 0{,}10$) oraz docelową wartością metryki jakości. Algorytm wprowadzania zniekształceń został opisany w Procedurze 13 zamieszczonej w załączniku (zob. Załącznik B.1).

3.4.3. Zbiór: waga produktów *eBay*

Zbiór danych

Zbiór użyty w tym eksperymencie zawierał informacje na temat wagi[14] produktów zamówionych na platformie *e-commerce eBay*. Jest to autorski zbiór stworzony na potrzebę tego eksperymentu. Surowe dane, które zostały użyte do przygotowania zbioru zostały udostępnione przez firmę *Webinterpret* (zob. Tabela 3.1).

Opis mikro-zadania

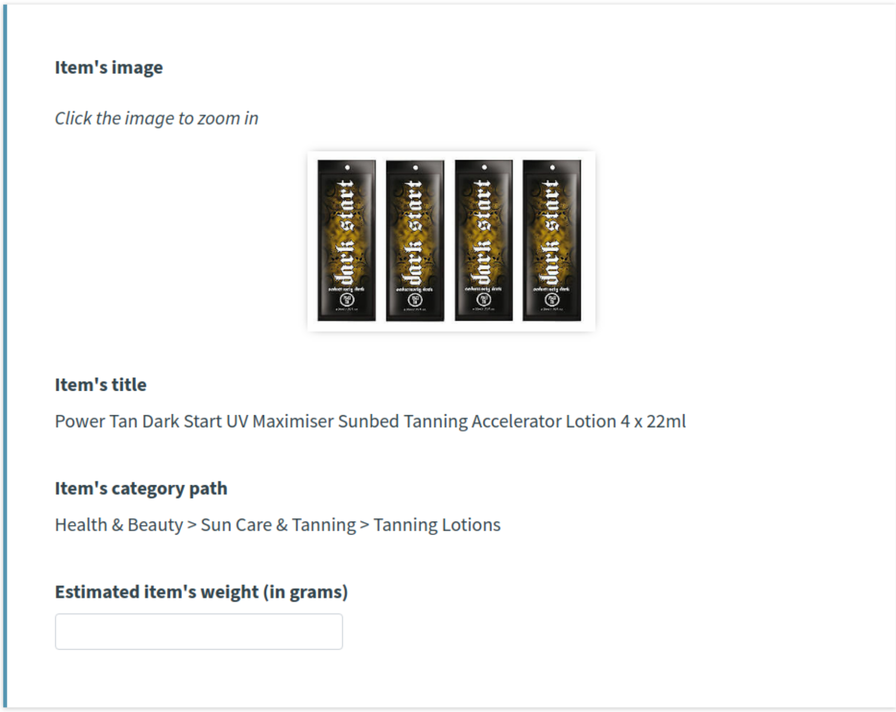
Celem poniższego mikro-zadania było oszacowanie wagi (w gramach) przedstawionego produktu na podstawie dostępnych informacji na temat produktu takie jak zdjęcie czy kategoria przedmiotu. Waga produktów znajdujących się w zbiorze określona została jako liczba całkowita z zakresu od $50g$ do $10000g$.

Mimo że zadanie to dotyczyło wag przedmiotów znanych z życia codziennego, to znajomość precyzyjnej wagi przedmiotów nie należy do wiedzy zdroworozsądkowej. Z tego powodu podczas pracy nad kolejnymi mikro-zadaniami z tego zbioru, anotatorzy nabywali wiedzę domenową dotyczącą wag przedmiotów.

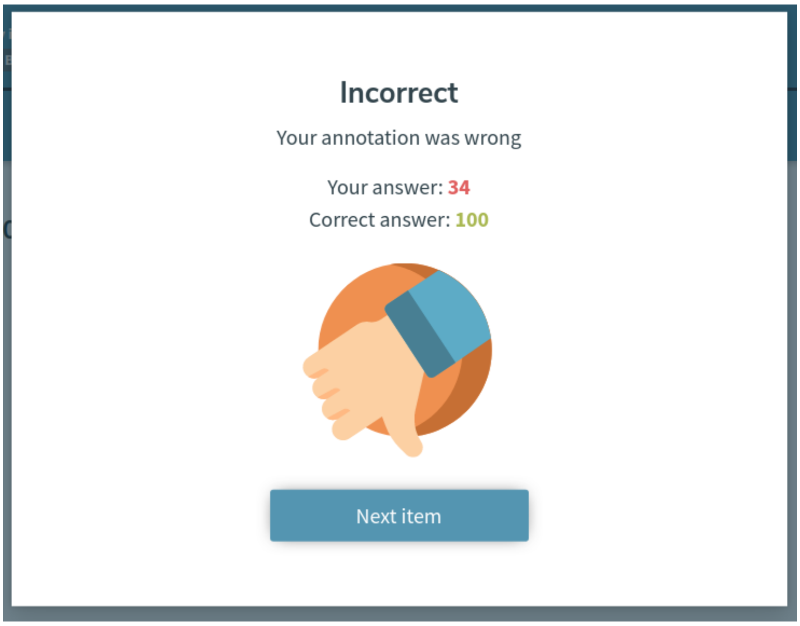
Przed rozpoczęciem pracy nad mikro-zadaniem system wyświetlał protokół anotacyjny. Protokół anotacyjny zawierał techniczną instrukcję opisującą sposób pracy nad zadaniem. W protokole nie zostały zamieszczone przykłady produktów i ich wag. Dokładna treść protokołu anotacyjnego została zamieszczona w załączniku (zob. Dodatek A.3).

Interfejs anotacyjny mikro-zadania zawierał cztery pola: *item's image* (tłum. zdjęcie produktu), *item's title* (tłum. tytuł produktu), *item's category path* (tłum. ścieżka kategorii produktu) oraz *estimated item's weight* (tłum. estymowana waga przedmiotu). Anotator zapoznawał się z treścią mikro-zadania, a następnie wpisywał estymowaną wagę produktu w pole *estimated item's weight* (zob. Rysunek 3.11). Po przesłaniu anotacji system wyświetlał okno zawierające informację zwrotną (zob. Rysunek 3.12).

[14]Wagi produktów znajdujące się w zbiorze zostały zmierzone podczas dostawy produktów do klientów. To oznacza, że waga produktu mierzona była razem z opakowaniem przesyłki. W praktyce różnica pomiędzy wagą produktu a wagą przesyłki może być określona przez błąd $MAE = 250g$. Źródło: konsultacja z firmą *Webinterpret*.



Rysunek 3.11: Interfejs anotacyjny dla zadania „waga produktów *eBay*”



Rysunek 3.12: Okno zawierające informację zwrotną dla zadania „waga produktów *eBay*”

Przygotowanie zbioru danych

Zbiór użyty został w niezmienionej formie. W związku z tym, że zbiór początkowy zawierał jedynie około 1000 elementów, zbiór ten był mniejszy od innych zbiorów, które zawierały około 2000 elementów.

Informacja zwrotna

Treść informacji zwrotnej zależy od wariantu eksperymentu, do którego przypisany został anotator. W każdym wariancie przekazywana jest informacja zwrotna o innej jakości:

- Wysoka jakość (W_1)
  Informacja zwrotna przekazywana anotatorom zawierała anotacje referencyjne;
- Umiarkowana jakość (W_2)
  Informacja zwrotna zawierała losowo zniekształcone anotacje referencyjne. Anotacje zostały zniekształcone tak, by przy porównywaniu z wartościami referencyjnymi otrzymywały jakość mierzoną metryką MAE na poziomie $MAE = 400g$.
- Niska jakość (W_3)
  Informacja zwrotna zawierała losowe zniekształcone anotacje referencyjne. Anotacje zostały zniekształcone tak, by przy porównywaniu z wartościami referencyjnymi otrzymywały jakość mierzoną metryką MAE na poziomie $MAE = 1200g$.

Zniekształcenie anotacji

W celu zmniejszenia jakości przekazywanej informacji zwrotnej referencyjne anotacje zostały zniekształcone tak, by zawierały błędy. Algorytm zniekształcenia referencyjnych miał formę iteracyjnej procedury. W każdej iteracji w sposób losowy wybierane było jedno, niezmodyfikowane wcześniej mikro-zadanie. Dla wybranego mikro-zadania waga referencyjna modyfikowana została poprzez dodanie lub odjęcie losowej wartości szumu. Wartość szumu obliczana była według rozkładu normalnego $\mathcal{N}(0, \sigma^2)$. W przypadku obu wariantów zawierających zniekształcone dane (W_2, W_3) został zastosowany ten sam algorytm różniący się tylko doborem parametru σ (W_2: $\sigma = 400g$, W_3: $\sigma = 1000g$). Proces ten powtarzany był aż do osiągnięcia określonej wartości wybranej metryki jakości (MAE). W przypadku obu wariantów zawierających zniekształcone dane (W_2, W_3) został zastosowany ten sam algorytm różniący się tylko docelową wartością metryki jakości. Algorytm wprowadzania zniekształceń został opisany w Procedurze 14 zamieszczonej w załączniku (zob. Załącznik B.2).

3.4.4. Zbiór: wydźwięk opinii o hotelach

Zbiór danych

Dane użyte w eksperymencie pochodzą ze zbioru *Datafiniti's Business Database*, komercyjnej bazy danych zawierającej dane związane z analityką biznesową. Pełen zbiór danych udostępniany został przez firmę *Datafinity* na jej oficjalnej stronie internetowej. Zbiór użyty w eksperymencie stanowi publicznie dostępny podzbiór pełnego zbioru danych, który została udostępniona na darmowej licencji (zob. Tabela 3.1).

Zbiór ten zawierał dane dotyczące opinii wystawionych przez klientów amerykańskich hoteli. W zbiorze znajdują się informacje hotelu (jego nazwa, lokalizacja) oraz szczegóły opinii (nazwa użytkownika wystawiającego opinię, data jej wystawienia, ocena).

Opis mikro-zadania

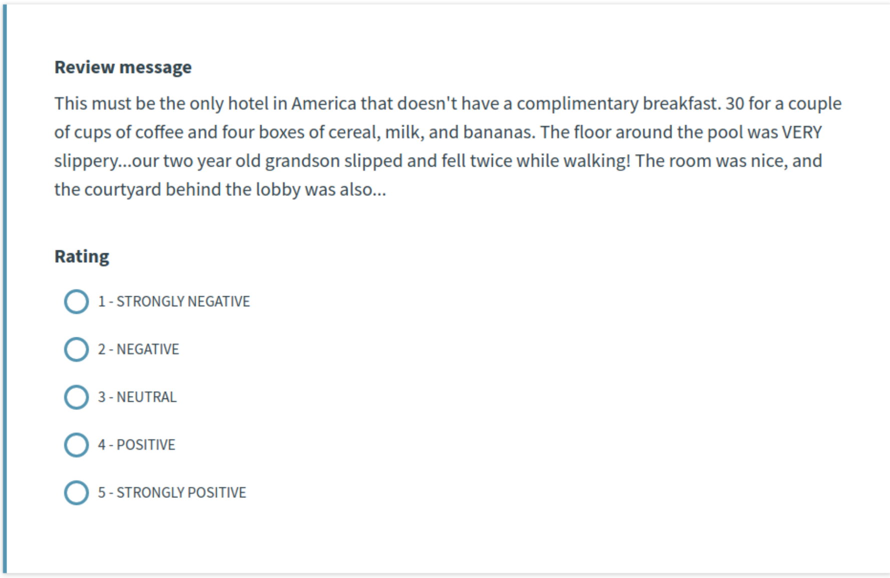
Celem omawianego mikro-zadania było przypisanie oceny wydźwięku tekstowych opinii napisanej przez klienta hoteli. Ocena wybrana przez anotatora powinna odpowiadać nacechowaniu emocjonalnemu oznaczanej opinii. W ramach zadania anotator miał dostęp jedynie do treści opinii, bez dodatkowych informacji o samym hotelu. Ocena odbywała się na pięciostopniowej skali *Likerta*, w której każdej wartości na skali przypisana była inna wartość wydźwięku:

1. *Strongly negative* (tłum. silnie negatywny)
2. *Negative* (tłum. negatywny)
3. *Neutral* (tłum. neutralny)
4. *Positive* (tłum. pozytywny)
5. *Strongly positive* (tłum. silnie pozytywny)

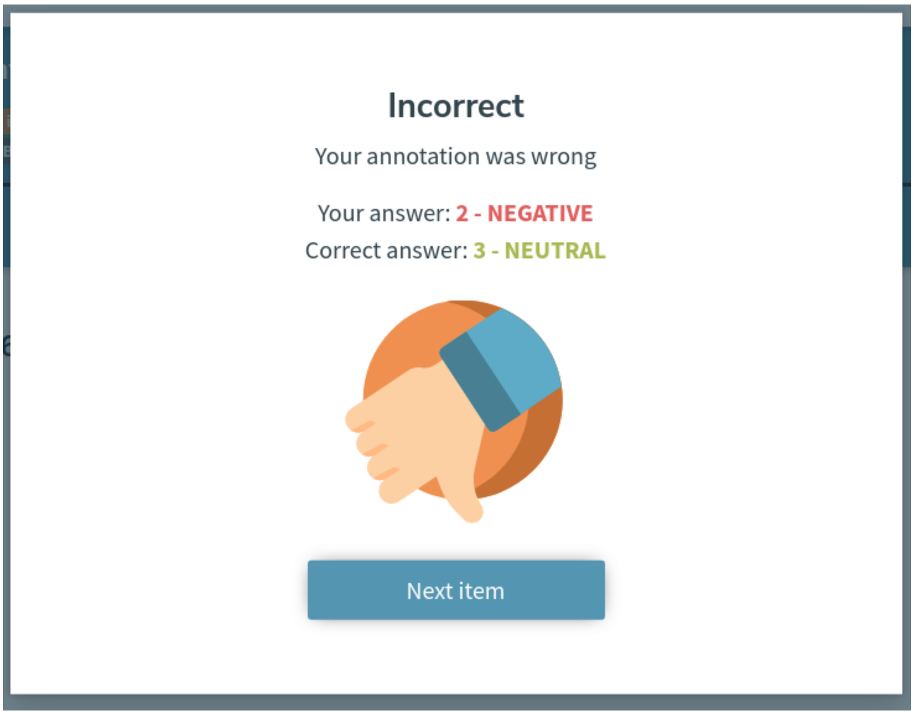
W ramach tego zadania anotator dokonywał oceny wydźwięku danej opinii na podstawie swoich osobistych odczuć. W związku z tym to zadanie to nie wymagało od anotatorów zdobycia dodatkowej wiedzy domenowej, a bazowało jedynie na jego wiedzy zdroworozsądkowej.

Przed rozpoczęciem pracy nad mikro-zadaniem system wyświetlał protokół anotacyjny. Protokół anotacyjny zawierał techniczną instrukcję opisującą sposób pracy nad zadaniem. Dokładna treść protokołu anotacyjnego została zamieszczona w załączniku (zob. Dodatek A.4).

Interfejs anotacyjny mikro-zadania zawierał dwa pola: *review message* (tłum. treść opinii) oraz *rating* (tłum. ocena). Anotator zapoznawał się z treścią mikro-zadania, a następnie uzupełniał wybraną wartość wydźwięku w polu *rating* (zob. Rysunek 3.13). Po przesłaniu anotacji system wyświetlał okno zawierające informację zwrotną (zob. Rysunek 3.14).



Rysunek 3.13: Interfejs anotacyjny dla zadania „wydźwięk opinii o hotelach”



Rysunek 3.14: Okno zawierające informację zwrotną dla zadania „wydźwięk opinii o hotelach”

Przygotowanie zbioru danych

Pełen zbiór danych zawierał szczegółowe informacje dotyczące wystawionych opinii (łącznie 26 kolumn zawierających informacje na temat hotelu oraz użytkownika, który wystawił opinię). Z tego powodu wymagane było dostosowanie zbioru do formy, która mogła być użyta w procesie crowdsourcingu. Dokładne kroki modyfikacji zbioru opracowane zostały w oparciu o własne doświadczenie eksperckie.

1. Wybrane zostały tylko elementy zbioru, które zawierały zarówno treść opinii, jak i ocenę punktową,
2. Wszystkie oceny punktowe (w skali od 1 do 5) zostały zmapowane na pięciostopniową skalę *Likerta* (według skali opisanej powyżej),
3. Dla każdej wartości ze skali ocen wybranych zostało 400 losowych mikro-zadań, którym przypisana została ta ocena,
4. Wszystkie mikro-zadania w zbiorze końcowym (łącznie 2000 elementów) zostały posortowane w sposób losowy.

Informacja zwrotna

Treść informacji zwrotnej zależała od wariantu eksperymentu, do którego przypisany został anotator. W każdym wariancie przekazywana była informacja zwrotna o innej jakości:

- Wysoka jakość (W_1)
  Informacja zwrotna przekazywana anotatorom zawierała anotacje referencyjne;
- Umiarkowana jakość (W_2)
  Informacja zwrotna zawierała losowo zniekształcone anotacje referencyjne. Anotacje zostały zniekształcone tak, by przy porównywaniu z wartościami referencyjnymi otrzymywały dokładność (ang. *accuracy*) na poziomie $ACC = 0{,}75$.
- Niska jakość (W_3)
  Informacja zwrotna zawierała losowe zniekształcone anotacje referencyjne. Anotacje zostały zniekształcone tak, by przy porównywaniu z wartościami referencyjnymi otrzymywały dokładność (ang. *accuracy*) na poziomie $ACC = 0{,}55$.

Zniekształcenie anotacji

W celu zmniejszenia jakości przekazywanej informacji zwrotnej referencyjne anotacje zostały zniekształcone tak, by zawierały błędy. Algorytm zniekształcenia referencyjnych miał formę iteracyjnej procedury. W każdej iteracji w sposób losowy wybierane było jedno, niezmodyfikowane wcześniej mikro-zadanie. Dla wybranego mikro-zadania obecna anotacja referencyjna zastępowana była losową wartością ze skali możliwych odpowiedzi. Proces ten powtarzany był aż do osiągnięcia określonej wartości wybranej metryki jakości (dokładność). W przypadku obu wariantów zawierających zniekształcone dane (W_2, W_3) został zastosowany ten sam algorytm różniący się tylko docelową wartością metryki jakości.

3.4.5. Zbiór: jednostki nazwane

Zbiór danych

Dane użyte w eksperymencie pochodzą ze zbioru *Groningen Meaning Bank* opracowanym przez Uniwersytet w Groningen i dostępny publicznie na oficjalnej stronie projektu[15]. Zbiór ten zawierał tysiące zdań, które zostały m.in. oznaczone tagami, które wskazują występowanie jednostek nazwanych w tekście. W ramach tego zadania została użyta próbka powyższego zbioru[16] (zob. Tabela 3.1).

Opis mikro-zadania

Celem poniższego mikro-zadania było oznaczanie jednostek nazwanych w przedstawionym tekście. Jednostki nazwane to fragmenty tekstu składającego się z jednego lub więcej tokenów, których znaczenie może być przypisane do jednej z wybranych kategorii. W ramach tego mikro-zadania anotatorzy oznaczali jednostki należące do jednej z pięciu kategorii:

– Geographical (tłum. jednostka geograficzna, np. „Tatry”)

– Geopolitical (tłum. jednostka geopolityczna, np. „Polska”)

– Person (tłum. osoba, np. „Jan Nowak”)

– Organization (tłum. organizacja, np. „Google”)

– Time (tłum. czas, np. „1 stycznia”)

Przed rozpoczęciem pracy nad mikro-zadaniem system wyświetlał protokół anotacyjny. Protokół anotacyjny zawierał techniczną instrukcję opisującą sposób pracy nad zadaniem. Dodatkowo w protokole znalazła się tabela zawierającą przykładowe wartości dla każdej z możliwych kategorii. Dokładna treść protokołu anotacyjnego została zamieszczona w załączniku (zob. Dodatek A.5).

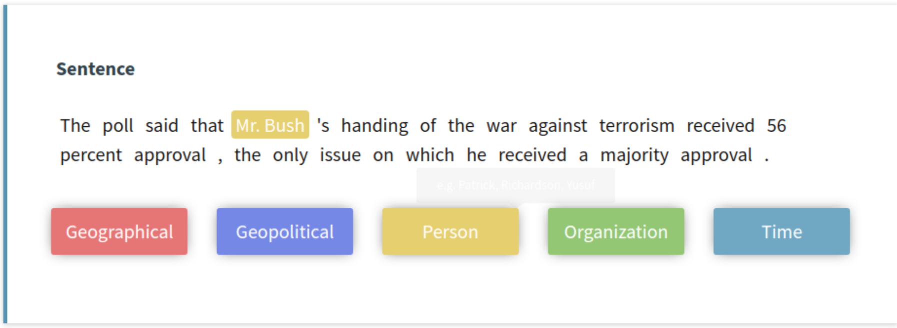
W związku z tym, że zadanie dotyczyło identyfikacji jednostek nazwanych znanych z życia codziennego, część anotatorzy mogła mieć już styczność z niektórymi atrybutami.

[15]`https://gmb.let.rug.nl/`; dostęp: 04.12.2021 r.

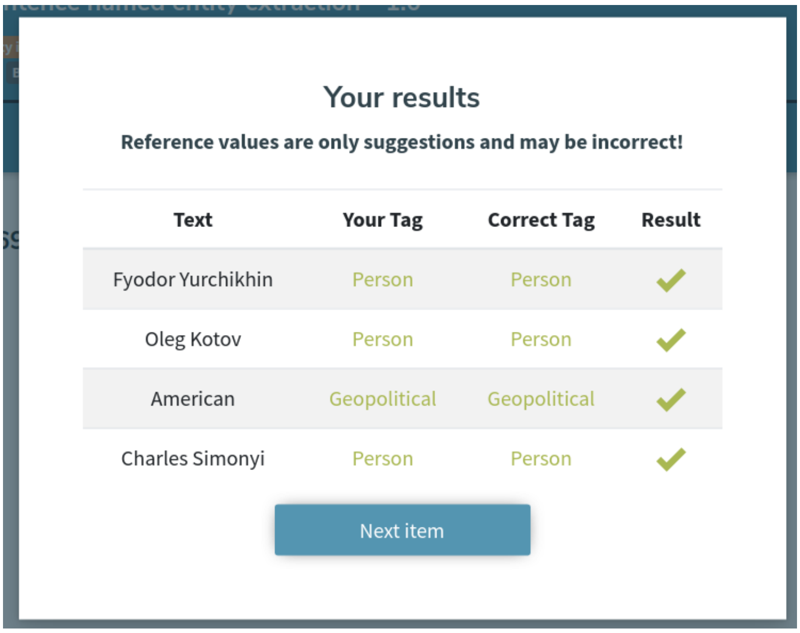
[16]Do zadania został użyty podzbiór pełnego korpusu dostępny pod poniższym linkiem: `https://www.kaggle.com/abhinavwalia95/entity-annotated-corpus`; dostęp: 04.12.2021 r.

Podczas pracy nad kolejnymi mikro-zadaniami z tego zbioru, anotatorzy nabywali wiedzę domenową dotyczącą prawidłowej identyfikacji każdej kategorii jednostek nazwanych, a ich wiedza domenowa była dodatkowo wspierana przez wiedzę zdroworozsądkową zdobytą wcześniej (np. podczas edukacji w szkole).

Interfejs anotacyjny mikro-zadania zawierał jedno pole: *sentence* (tłum. zdanie). Każde mikro-zadanie zawsze zawierało co najmniej jedną jednostkę nazwaną. Anotator oznacza jednostkę nazwaną poprzez wybranie jednego lub więcej tokenów oraz przypisanie ich do wybranej kategorii (zob. Rysunek 3.15). Po przesłaniu anotacji system wyświetla okno zawierające informację zwrotną (zob. Rysunek 3.16).



Rysunek 3.15: Interfejs anotacyjny dla zadania „jednostki nazwane"



Rysunek 3.16: Okno zawierające informację zwrotną dla zadania „jednostki nazwane"

Przygotowanie zbioru danych

Z pełnego zbioru danych został wylosowany podzbiór mikro-zadań. Losowanie przeprowadzone zostało tak, aby dla każdego z pięciu kategorii jednostek nazwanych występowało $n = 400$ mikro-zadań. Łącznie wybrane zostało 2000 elementów.

Informacja zwrotna

Treść informacji zwrotnej zależy od wariantu eksperymentu, do którego przypisany został anotator. W każdym wariancie przekazywana jest informacja zwrotna o innej jakości:

- W_1 – Wysoka jakość
  Informacja zwrotna przekazywana anotatorom zawierała anotacje referencyjne;
- W_2 – Umiarkowana jakość – Informacja zwrotna zawierała losowo zniekształcone anotacje referencyjne. Anotacje zostały zniekształcone tak, by przy porównywaniu z wartościami referencyjnymi otrzymywały jakość mierzoną metryką F_1 na poziomie $F_1 = 0{,}75$.
- W_3 – Niska jakość
  Informacja zwrotna zawierała losowe zniekształcone anotacje referencyjne. Anotacje zostały zniekształcone tak, by przy porównywaniu z wartościami referencyjnymi otrzymywały jakość mierzoną metryką F_1 na poziomie $F_1 = 0{,}55$.

Zniekształcenie anotacji

W celu zmniejszenia jakości przekazywanej informacji zwrotnej referencyjne anotacje zostały zniekształcone tak, by zawierały błędy. Algorytm zniekształcenia referencyjnych miał formę iteracyjnej procedury. W każdej iteracji w sposób losowy wybierane było jedno, niezmodyfikowane wcześniej mikro-zadanie. Dla wszystkich tokenów w wybranym mikrozadaniu zastosowane zostały dwie reguły wprowadzenia zniekształceń:

- Z prawdopodobieństwem p_d istniejący token zmieniony został na kategorię pustą (brak jednostki nazwanej).
- Z prawdopodobieństwem p_2 istniejąca anotacja (pusta lub nie) zastępowana jest przez losowo wybraną kategorię.

Proces ten powtarzany był aż do osiągnięcia określonej wartości wybranej metryki jakości (F_1). W przypadku obu wariantów zawierających zniekształcone dane (W_2, W_3) został zastosowany ten sam algorytm różniący się tylko doborem parametrów p_d (W_2: $p_d = 0{,}25$, W_3: $p = 0{,}4$), p_r (W_2: $p = 0{,}05$, W_3: $p = 0{,}10$) oraz docelową wartością metryki jakości. Algorytm wprowadzania zniekształceń został opisany w Procedurze 15 zamieszczonej w załączniku (zob. Załącznik B.3).

3.4.6. Zbiór: wyrazy bliskoznaczne

Zbiór danych

Autorski zbiór, zawierający pary słów będące wyrazami bliskoznacznymi, oraz przykładowe zdania zawierające te słowa (jedno zdanie dla każdego słowa). Pary słów bliskoznacznych zostały pozyskane z leksykonu *Wordnet*, który jest publicznie dostępny na oficjalnej stronie projektu[17]. Następnie, w celu zwiększenia trudności zadania ze zbioru wybrane zostały pary zawierające mniej popularne słowa. Dodatkowo zbiór par słów został rozszerzony o pary słów, które nie są wyrazami bliskoznacznymi. Dla zebranego zbioru słów wyszukane zostały przykładowe zdania. Do zebrania zdań użyty został zbiór testów dostępnych w ramach projektu *Gutenberg*[18] (zob. Tabela 3.1).

Opis mikro-zadania

Celem poniższego mikro-zadania było dokonanie klasyfikacji par słów na słowa będące wyrazami bliskoznacznymi lub nie. Anotatorzy przypisywali pary słów do jednej z dwóch kategorii:

1. *Yes* (tłum. tak) – słowa bliskoznaczne,
2. *No* (tłum. nie) – słowa nie są bliskoznaczne.

W związku z tym, że zadanie dotyczyło identyfikacji słów bliskoznawczych, to prawidłowe rozwiązywanie mikro-zadań wymagało od anotatorów dobrej znajomości języka angielskiego. Ponieważ w eksperymencie brali udział tylko anotatorzy, który posługiwali się językiem angielskim, to zadanie bazowało na ich wiedzy zdroworozsądkowej i nie wymagało od nich nabycia dodatkowej wiedzy domenowej.

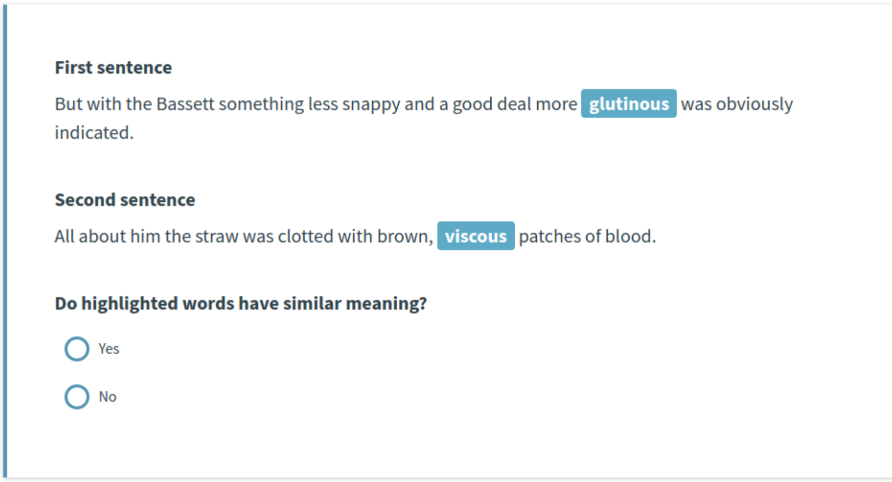
Przed rozpoczęciem pracy nad mikro-zadaniem system wyświetlał protokół anotacyjny. Protokół anotacyjny zawierał techniczną instrukcję opisującą sposób pracy nad zadaniem.

[17] *Wordnet* to słownik semantyczny, w którym znaczenia słów powiązane są relacjami. `https://wordnet.princeton.edu/`; dostęp: 02.03.2022 r.

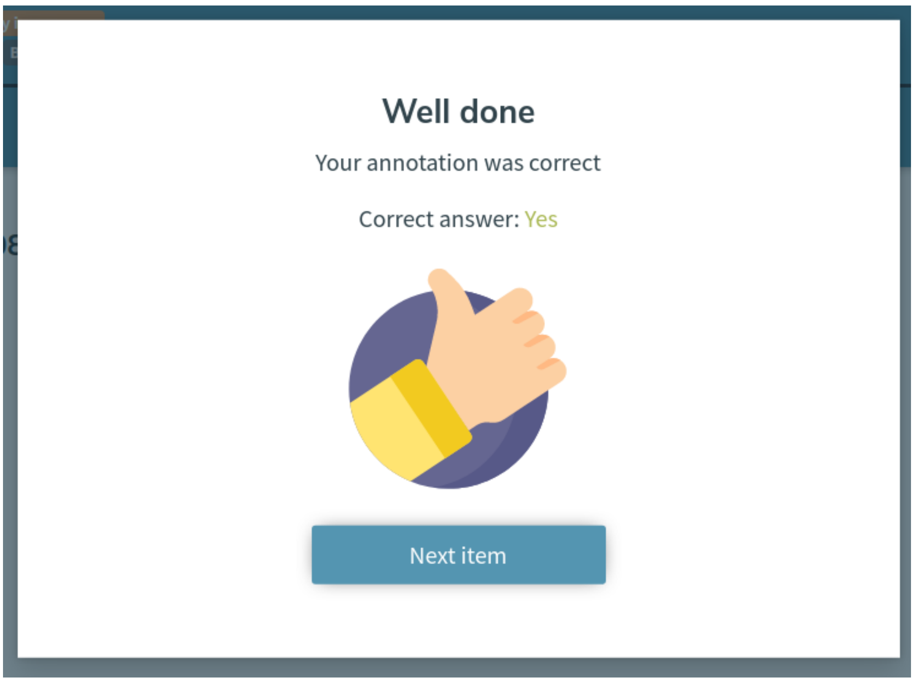
[18] Zbiór ten to kolekcja publicznie dostępnych książek, które udostępnione są na darmowej licencji. `https://www.gutenberg.org/`; dostęp: 02.03.2022. Zbiór danych użyty w eksperymencie pobrany został przy pomocy programu dostępnego pod linkiem: `https://github.com/pgcorpus/gutenberg`; dostęp: 02.03.2022 r.

Dokładna treść protokołu anotacyjnego została zamieszczona w załączniku (zob. Dodatek A.6).

Interfejs anotacyjny mikro-zadania zawierał trzy pola: *first sentence* (tłum. pierwsze zdanie), *second sentence* (tłum. drugie zdanie), *do highlighted words have similar meaning?* (tłum. czy wyróżnione słowa są bliskoznaczne). W dwóch pierwszych polach zawierających zdania zostały wyróżnione słowa, dla których anotator miał udzielić odpowiedzi (zob. Rysunek 3.17). Po zapoznaniu się z treścią zadania anotator dokonywał klasyfikacji poprzez wybranie jednej z dwóch kategorii. Po przesłaniu anotacji system wyświetlał okno zawierające informację zwrotną (zob. Rysunek 3.18).



Rysunek 3.17: Interfejs anotacyjny dla zadania „wyrazy bliskoznaczne”



Rysunek 3.18: Okno zawierające informację zwrotną dla zadania „wyrazy bliskoznaczne”

Przygotowanie zbioru danych

Przygotowany został zbiór losowych par słów będących słowami bliskoznacznymi. Następnie zbiór ten został rozszerzony o pary słów, które nie były bliskoznaczne. W celu ustalenia, czy dwa słowa są bliskoznaczne, użyty został leksykon *WordNet*. Ze wspólnego zbioru usunięte zostały pary słów zawierające popularne zbyt słowa. Popularność danego słow została określona na podstawie częstotliwości jego występowania w zbiorze projektu *Gutenberg*[19]. Dla każdego ze słów wybrane zostało losowe zdanie, które zawierało dane słowo. Zdania zostały pobrane ze zbioru *Gutenberg*. W każdym zdaniu słowo należące do pary słów bliskoznacznych zostało wyróżnione[20]. Z ostatecznego zbioru w sposób losowy wybrane zostało 1000 par, które są bliskoznaczne oraz 1000, które nie są. Wszystkie wiersze (łącznie 2000) w zbiorze danych uporządkowano w losowej kolejności.

Informacja zwrotna

Treść informacji zwrotnej zależała od wariantu eksperymentu, do którego przypisany został anotator. W każdym wariancie przekazywana była informacja zwrotna o innej jakości:

– Wysoka jakość (W_1)
  Informacja zwrotna przekazywana anotatorom zawierała anotacje referencyjne;

– Umiarkowana jakość (W_2)
  Informacja zwrotna zawierała losowo zniekształcone anotacje referencyjne. Anotacje zostały zniekształcone tak, by przy porównywaniu z wartościami referencyjnymi otrzymywały dokładność (ang. *accuracy*) na poziomie $ACC = 0{,}75$.

– Niska jakość (W_3)
  Informacja zwrotna zawierała losowe zniekształcone anotacje referencyjne. Anotacje zostały zniekształcone tak, by przy porównywaniu z wartościami referencyjnymi otrzymywały dokładność (ang. *accuracy*) na poziomie $ACC = 0{,}55$.

Zniekształcenie anotacji

W celu zmniejszenia jakości przekazywanej informacji zwrotnej referencyjne anotacje zostały zniekształcone tak, by zawierały błędy. Algorytm zniekształcenia referencyjnych

[19] W eksperymencie użyte zostały pary, w których częstotliwość pierwszego słowa była niższa 0,001% wszystkich słów (czyli poniżej ok. 40000 wystąpień).

[20] Słowa zostały wyróżnione przy pomocy tagów HTML zmieniających kolor tekstu. W ten sposób wyróżnienie było widoczne w interfejsie anotacyjnym.

miał formę iteracyjnej procedury. W każdej iteracji w sposób losowy wybierane było jedno, niezmodyfikowane wcześniej mikro-zadanie. Dla wybranego mikro-zadania obecna anotacja referencyjna zastępowana była przeciwną kategorią. Proces ten powtarzany był aż do osiągnięcia określonej wartości wybranej metryki jakości (dokładność). W przypadku obu wariantów zawierających zniekształcone dane (W_2, W_3) został zastosowany ten sam algorytm różniący się tylko docelową wartością metryki jakości.

3.5. Analiza wyników

W tej części pracy przedstawiłem analizę wyników przeprowadzonego przeze mnie eksperymentu dla każdego z badanych zbiorów danych. W pierwszej części podrozdziału opisałem metodologię weryfikacji hipotez oraz pytań badawczych, a w dalszej części wyniki oraz wnioski analizy przeprowadzonej zgodnie z opisaną metodologią.

3.5.1. Zebrane anotacje

W ramach przeprowadzonego eksperymentu dla każdego z użytych zbiorów danych wybrane zostało 2000 elementów. Wyjątek stanowił zbiór „waga produktów *eBay*”, w przypadku którego wejściowy zbiór referencyjny zawierał jedynie 1000 elementów. Ostatecznie w eksperymencie wzięło udział 999 anotatorów, którzy łącznie wykonali 56908 oznaczeń. Podczas eksperymentu każde mikro-zadanie oznaczone zostało przez przynajmniej 4 różnych anotatorów. Po zakończeniu eksperymentu wykluczone zostały anotacje, które nie spełniały zdefiniowanych kryteriów minimalnej jakości. Kryteria zastosowane w eksperymencie obejmowały następujące reguły:

1. Wykluczone zostały anotacje anotatorów, którzy oznaczyli mniej niż 10 mikro-zadań.
2. Wykluczone zostały anotacje anotatorów, którzy udzielali dokładniej tej samej odpowiedzi dla wszystkich mikro-zadania w partii.
3. Wykluczone zostały anotacje anotatorów, których średni czas anotacji był mniejszy niż 1 sekunda.

Reguła 1 została zastosowana, aby wykluczyć z eksperymentu anotatorów z pojedynczymi anotacjami[21]. Natomiast reguły 2 oraz 3 zastosowane zostały w celu usunięcia szkodliwych anotatorów, którzy umyślnie wprowadzali złe odpowiedzi w celu oszukania zleceniodawcy.

[21] Było to istotne, ponieważ w eksperymencie porównana była średnia jakość anotatorów. Im wyższa była liczba oznaczonych mikro-zadań, tym bardziej stabilna stawała się jakość anotatorów.

Tabela 3.2: *Zestawienie podstawowych statystyk danych zebranych podczas eksperymentu*

Zbiór danych	Liczba mikro-zadań	Liczba anotatorów	Liczba anotacji	Średnia liczba anotacji dla mikro-zadania	Średnia liczba anotacji dla anotatora
skargi usług bankowych	1986	145	9415	4,74	64,93
atrybuty produktów *eBay*	1871	122	6779	3,62	55,56
waga produktów *eBay*	1017	178	9387	9,23	52,73
wydźwięk opinii o hotelach	2294	182	10378	4,52	57,02
jednostki nazwane	1999	140	8043	4,02	57,45
wyrazy bliskoznaczne	1985	232	12906	6,50	55,63

W związku z tym, że wyżej opisane kryteria zostały zastosowane po zakończeniu eksperymentu, ostatecznie niektóre zbiory zawierają nieznacznie mniej elementów, niż było to początkowo zakładane. Przykładowo zbiór „atrybuty produktów *eBay*" zawierał 1871 elementów, które zostały oznaczone średnio przez 3,62 anotatorów. Zestawienie dokładnej liczby oznaczonych mikro-zadań, liczby anotacji, a także liczby unikalnych anotatorów biorących udział w eksperymencie została przedstawiona w Tabeli 3.2[22].

3.5.2. Metodologia weryfikacji hipotez badawczych

W tej części rozprawy omówiłem metodologię użytą w celu weryfikacji badanych hipotez postawionych na początku tego rozdziału (zob. Paragraf 3.1.1). Dla każdego zbioru danych określony został zbiór metryk jakości. Metryki te obliczone zostały dla grupy anotatorów przypisanej do każdego z wariantów eksperymentu. Weryfikacja hipotez przeprowadzona została osobno dla każdego z badanych zbiorów danych za pomocą tej samej procedury.

Metoda weryfikacji hipotezy H_1

W pierwszej kolejności sprawdziłem, czy zapewnienie synchronicznej informacji zwrotnej w procesie crowdsourcingu w pozytywny sposób przekłada się na jakość pozyskiwanych danych. W związku z tym postawiłem następującą hipotezę:

H_1: Zapewnienie synchronicznej informacji zwrotnej w pozytywny sposób wpływa na jakość pozyskiwanych danych lingwistycznych w procesie crowdsourcingu.

[22] Z powodu błędu w konfiguracji eksperymentu na platformie *MTurk* dla zbiorów „waga produktów *eBay*" oraz „wyrazy bliskoznaczne" zebrane zostało znacznie więcej danych. Nie miało to negatywnego wpływu na przebieg eksperymentu.

W celu weryfikacji powyższej hipotezy przeprowadzona została analiza porównawcza jakości danych pozyskanych dla wariantu, w którym anotatorzy otrzymywali informację zwrotną W_1 oraz wariantu kontrolnego W_1^c. Dla każdego zbioru danych określona została metryka jakości, która została użyta do zamodelowania jakości anotatorów przypisanych do danego wariantu. Jakość całego wariantu obliczana została jako średnia jakości anotatorów z danego wariantu. Wybór konkretnego zbioru metryk uzależniony był od typu mikro-zadania realizowanego w ramach zbioru danych.

W ramach weryfikacji hipotezy H_1 przeprowadzona została analiza rozkładu możliwych różnic jakości pomiędzy wariantami W_1 oraz W_1^c. Różnice te interpretowane były jako obserwowalny wpływ wprowadzenia informacji zwrotnej do procesu pozyskiwania danych. Analiza wykonana została w ramach kilkukrokowej procedury opisanej poniżej (zob. Procedura 10). W celu oszacowania rozkładu różnicy jakości pomiędzy porównywanymi wariantami zastosowana została metoda *bootstrapping*[23]. Oszacowane rozkłady zostały użyte do obliczenia prostych w interpretacji statystyk, które pozwoliły na weryfikację sprawdzanej hipotezy. Obliczenia zostały wykonane za pomocą autorskiej biblioteki *strapping*[24]. Kod użyty do wykonania obliczeń przeprowadzonych w celu porównania wariantów eksperymentu został umieszczony w publicznym repozytorium[25].

Do weryfikacji hipotezy H_1 użyte zostały następujące statystyki opisujące wpływ wprowadzenia informacji zwrotnej na jakość pozyskanych danych:

- CI – przedział ufności; przedstawiony w formie pary $[P05, P95]$ (percentyl 5% i percentyl 95%),
- μ – średnia różnica; średnia wartość obliczona w oparciu o oszacowany rozkład różnic,
- p – prawdopodobieństwo poprawy jakości $P(X > 0)$,
- *d Cohena* – wielkość efektu (ang. *effect size*).

Ponieważ termin „wielkość efektu” może dotyczyć wielu zbliżonych pojęć opisujących

[23] *Bootstrapping* to narzędzie statystyczne służące do przeprowadzania wnioskowania dla testów nieparametrycznych. *Bootstrapping* jest najbardziej przydatny w sytuacji, gdy próbka analizowanych danych pochodzi z rozkładu, który jest nieznany lub zbyt złożony, by użyć standardowych metod testowania statystycznego [Jeremy Orloff, 2014]. Metoda ta służy do estymowania wariancji statystyk, które obliczane są na podstawie danych, takich jak średnia lub odchylenie standardowe. W metodzie tej wejściowy zbiór danych próbkowany jest wielokrotnie (z powtórzeniami) w celu zbudowania wielu, zasymulowanych próbek. Następnie wygenerowane próbki używane są do obliczenia rozkładu wybranych statystyk. Po dokładniejszy opis działania *bootstrappingu* zob. Jeremy Orloff [2014].

[24] `https://pypi.org/project/strapping/`; dostęp: 06.07.2022 r.

[25] `https://github.com/heolin/funcrowd-dataset`; dostęp: 13.07.2022 r.

Procedura 10: Procedura obliczania statystyk używanych w procesie weryfikacji głównych hipotez badawczych

Niech :
X_t – grupa anotatorów z badanego wariantu,
X_c – grupa anotatorów z grupy kontrolnej,
$A = \{A_1, \cdots, A_n\}$ – zbiór zawierający zbiory anotacji wszystkich anotatorów, gdzie A_i to zbiór wszystkich anotacji wykonanych przez anotatora i,
f – funkcja obliczająca wartość wybranej metryki jakości dla danego anotatora[26].

Kroki :

1 Tworzony jest zbiór Q zawierający wartość wybranej metryki jakości dla każdego anotatora i:

$$Q = f(A_i) : i \in A.$$

2 Zbiór Q dzielony jest na dwa podzbiory. Zbiór Q_t zawierający wartości metryki jakości dla anotatorów z grupy X_t przypisanych do badanego wariantu:

$$Q_t = Q_i : i \in X_t$$

oraz analogicznie Q_c zawierający wartości dla anotatorów z grupy X_c.

3 Przy użyciu metody *bootstrapping* oszacowywany jest rozkład różnic R_{tc} pomiędzy średnimi wartościami metryki jakości obu zbiorów:

$$R_{tc} \sim \overline{Q}_t - \overline{Q}_c.$$

W oparciu stworzony rozkład R_{tc} obliczane są opisujące go statystyki: μ, d, p oraz CI.

Wyjście:
μ – średnia różnica,
d – wielkość efektu,
p – prawdopodobieństwo poprawy jakości,
CI – przedział ufności.

Tabela 3.3: *Interpretacja wartości miary d (źródło: [Sawilowsky, 2009])*

Interpretacja efektu	Wartość d Cohena
Bardzo mały	0,01
Mały	0,20
Umiarkowany	0,50
Duży	0,80
Bardzo duży	1,20
Ogromny	2,00

różnicę pomiędzy dwoma zmiennymi, poniżej opisałem, jak ten termin rozumiany jest w niniejszej pracy.

Wielkość efektu

W statystyce wielkość efektu stanowi unormowany sposób zmierzenia różnicy spowodowanej przez zaistnienie danego zjawiska [Kelley & Preacher, 2012]. W ramach przeprowadzonego eksperymentu wielkość efektu reprezentowana była za pomocą miary d Cohena [Cohen, 1988, s. 67], która obliczana została na podstawie poniższej formuły:

$$d = \frac{\overline{X}_a - \overline{X}_b}{s},$$

gdzie

$$s = \sqrt{\frac{\sum_i^{n_a}(X_{a,i} - \overline{X}_a)^2 + \sum_i^{n_b}(X_{b,i} - \overline{X}_b)^2}{n_a + n_b - 2}}.$$

Aby ustandaryzować sposób interpretacji wartości wielkości efektu, wprowadzona została referencyjna klasyfikacja dla przekroju wartości d Cohena. W ramach niniejszej rozprawy użyta została klasyfikacja zaproponowana przez Sawilowsky [2009] (zob. Tabela 3.3).

Metoda weryfikacji hipotezy H_2

Kolejną rzeczą, którą sprawdziłem, było to jak jakość wygenerowanej informacji zwrotnej, przekłada się na jakość pozyskanych danych. A w szczególności czy informacja zwrotna zawierająca błędy może ostatecznie przyczynić się do poprawy jakości danych. W związku z tym postawiłem następującą hipotezę:

H_2: Jakość przekazywanej informacji zwrotnej w procesie crowdsourcingu ma wpływ na jakość pozyskiwanych danych.

W związku z tym, że pozyskiwanie danych dla różnych wariantów eksperymentu rozłożone były w czasie dla każdego z badanych wariantów stworzony został osobny wariant kontrolny (zob. Paragraf 3.3). Z tego powodu nie jest możliwe bezpośrednie porównanie badanych wariantów W_1, W_2 oraz W_3. Dlatego w ramach weryfikacji hipotezy H_2 przeanalizowane zostały różnice pomiędzy znormalizowanymi rozkładami różnic R_{tc}, które zostały obliczone w procesie weryfikacji hipotezy H_1 (zob. Paragraf 3.5.2).

W ramach analizy porównane zostały trzy pary wariantów: W_1 z wariantem W_2, W_1 z wariantem W_3 oraz W_2 z wariantem W_3. Przykładowo podczas porównywania wariantów W_1 i W_2 obliczony został rozkład $\hat{R}_{1,2}$ stanowiący różnicę rozkładów R_{tc} obliczonych dla obu grup:

$$\hat{R}_{1,2} = \hat{R}_{tc}^1 - \hat{R}_{tc}^2,$$

gdzie $\hat{R}_{tc}^i$ to znormalizowany rozkład różnic dla wariantu W_i obliczony poprzez podzielenie rozkładu R_{tc} przez średnią wartość metryki jakości dla anotatorów z grupy kontrolnej:

$$\hat{R}_{tc}^i = \frac{R_{tc}^i}{\overline{X}_c^i}$$

Dla stworzonego rozkładu $\hat{R}_{1,2}$ obliczone zostały te same statystyki opisujące rozkład co w przypadku analizy wykonanej w ramach weryfikacji hipotezy H_1: μ (średnia różnica), d Cohena (wielkość efektu), p (prawdopodobieństwo poprawy jakości) oraz CI (przedział ufności).

3.5.3. Metodologia weryfikacji pytań badawczych

W celu pogłębienia analizy dotyczącą zastosowania informacji zwrotnej w procesie crowdsourcingu analiza wykonana podczas weryfikacji hipotez została rozszerzona o dodatkowe trzy pytania badawcze: P_1, P_2 oraz P_3 (zob. Paragraf 3.1.1). W tej części rozdziału przedstawiłem metodologię przeprowadzenia analizy mającej na celu zapewnienie odpowiedzi na postawione pytania.

Metoda weryfikacji pytania badawczego P_1

Pierwsze pytanie badawcze związane było z eksploracją momentu efektu informacji zwrotnej w procesie crowdsourcingu. W szczególności istotne było sprawdzenie, czy informacja zwrotna ma tylko i wyłącznie efekt długoterminowy na jakość pozyskiwanych danych jak sugeruję literatura przedmiotu (zostało to opisane w pierwszym rozdziale niniejszej pracy, zob. Podrozdział 1.4.1). Z tego powodu zdecydowałem się sformułować następujące pytanie badawcze:

P_1: Czy informacja zwrotna w procesie crowdsourcingu ma efekt długoterminowy (edukacyjny)?

W ramach analizy przeprowadzonej w celu odpowiedzi na pytanie P_1 porównane dane zebrane dla wariantu W_1 z wariantem kontrolnym W_1^c. Według mojej wiedzy istniejąca literatura nie opisuje usystematyzowanego sposobu analizy momentu efektu informacji zwrotnej. Z tego powodu w ramach niniejszej rozprawy zaproponowany został autorski sposób analizy momentu efektu informacji zwrotnej. Szczegóły tego podejścia zostały opisane poniżej.

Analiza momentu efektu informacji zwrotnej

Moment efektu stanowi jeden z wymiarów opisujących informację zwrotną (zob. Paragraf 1.4.1). Wymiar ten określa, w jakim momencie zauważalny jest efekt wprowadzenia informacji zwrotnej. W celu analizy momentu efektu w czasie zaproponowany został model opisujący, jak zmienia się jakość pozyskiwanych danych w czasie po wprowadzeniu informacji zwrotnej. Moment efekt rozumiany jest więc jako funkcja liniowa $\hat{R}$ opisująca przybliżony przebieg zmiany różnicy w średniej jakości pomiędzy wariantami W_1 i W_1^c. Funkcja ta zdefiniowana została w oparciu o dwa parametry, które opisują udział efektu natychmiastowego i efektu długoterminowego:

$$\hat{R} = \alpha x + \beta,$$

gdzie:

- α – opisuje efekt długoterminowy, który stanowi zmianę w różnicy jakości po wykonaniu na samym początku pracy[27],
- β – opisuje efekt natychmiastowy, który stanowi początkową różnicę w jakości.

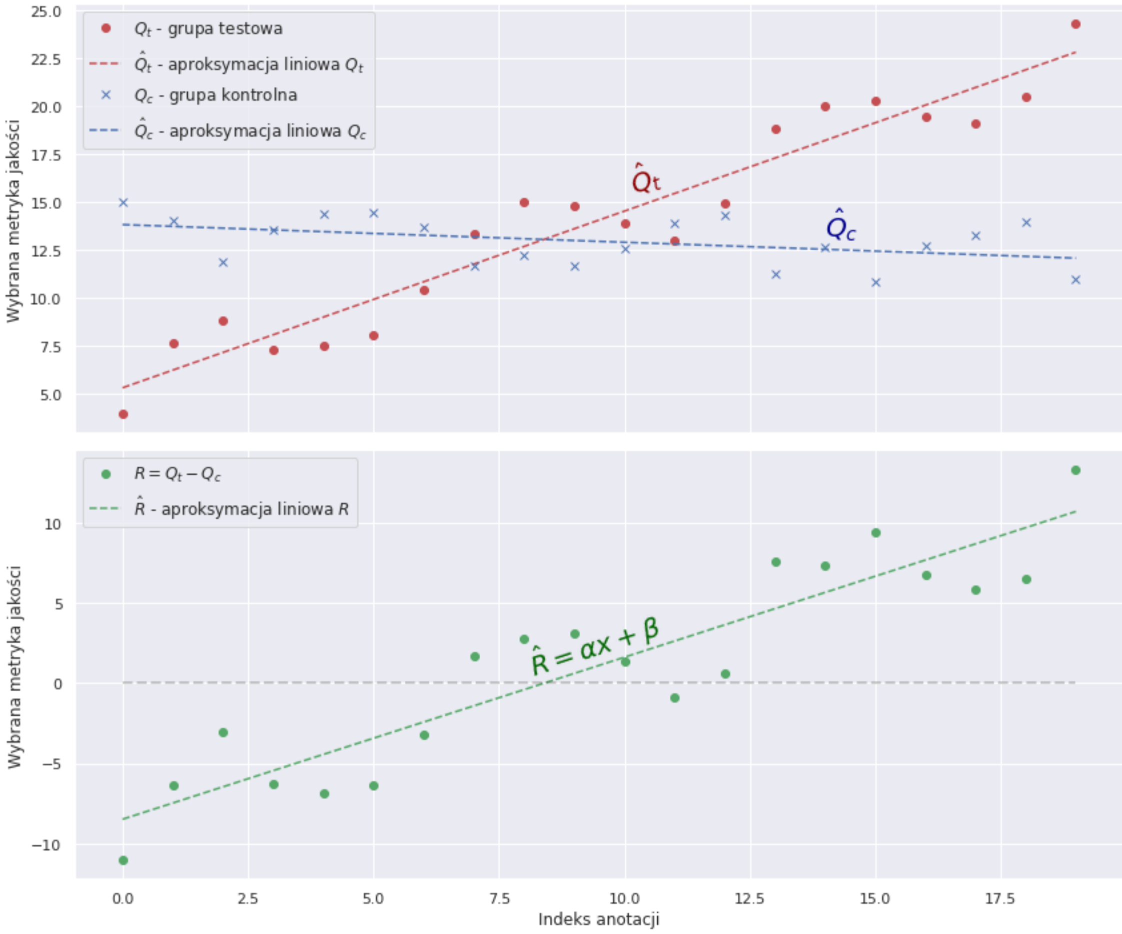
W celu ułatwienia interpretacji wyników analizy wprowadzone zostały dwa dodatkowe parametry opisujące moment efektu w formie procentowej zmiany obliczonej w stosunku do jakości wariantu kontrolnego:

- α_r – opisuje relatywny efekt długoterminowy,

[27] Do wyodrębnienia trendu zmiany jakości użyta została średnia krocząca z oknem $W = 5$. Tak więc efekt natychmiastowy oznacza zmianę jakości po pięciu anotacji.

– β_r – opisuje relatywny efekt natychmiastowy.

Wizualizacja wyżej opisanej funkcji została zamieszczona na Rysunku 3.19. Bardziej szczegółowy opis sposobu obliczenia jej parametrów został przedstawiony w ramach Procedury 11.



Rysunek 3.19: Graficzna interpretacja parametrów modelujących moment efektu informacji zwrotnej

Procedura 11: Procedura obliczania parametrów modelujących moment efektu informacji zwrotnej

Niech:

X_t – grupa anotatorów z wariantu badanego,
X_c – grupa anotatorów z wariantu kontrolnego,
n – liczba anotacji wykonanych przez każdego anotatora,
$A = \{A_1, \cdots, A_m\}$ – zbiór anotacji wszystkich anotatorów
$A_i = (a_1, \cdots, a_n)$ – ciąg anotacji wykonanych przez anotatora i; posortowanych w kolejności
ich wykonywania,
$W = 5$ – wielkość okna średniej kroczącej,
f – funkcja przypisująca ocenę pojedynczej anotacji[a].

Kroki:

Dla każdego anotatora i obliczany jest ciąg zawierający jakość kolejnych, wykonanych przez
niego anotacji:

$$R_i = (f(a_1), \cdots, f(a_n)).$$

Dla każdej z grup obliczany jest ciąg średniej jakości kolejnych anotacji, a następnie w celu wyodrębnienia trendu obliczana jest średnia krocząca z oknem W:

$$Q_t = ma_5(avg(\{R_{i,1} : i \ \in X_t\}), \cdots, avg(\{R_{i,n} : i \ \in X_t\})),$$

oraz analogicznie Q_c powstaje w oparciu o zbiór anotatorów X_c.

Obliczany jest ciąg różnic pomiędzy średnią jakością obu grup:

$$R = (Q_{t,1} - Q_{c,1}, \cdots, Q_{t,n} - Q_{c,n}).$$

Dla obliczonego ciągu różnic R obliczana jest jego aproksymacja $\hat{R}$ metodą regresji liniowej.
Funkcja ta zdefiniowana jest w oparciu parametry α i β:

$$\hat{R}(x) = \alpha x + \beta.$$

oraz analogicznie dla Q_t oraz Q_c obliczane są odpowiednio $\hat{Q}_t$ oraz $\hat{Q}_c$.

Obliczane są dodatkowe parametry α_r oraz β_r relatywne w stosunku do początkowych wartości $\hat{Q}_c$:

$$\alpha_r = \hat{R}(0)/\hat{Q}_c(0), \quad \beta_r = [\hat{R}(n) - \hat{R}(0)]/\hat{Q}_c(0).$$

Wyjście:

α – efekt długoterminowy,
β – efekt długoterminowy,
α_r – relatywny efekt długoterminowy,
β_r – relatywny efekt natychmiastowy.

[a] Przykładowo, w przypadku zadań klasyfikacji może być to funkcja zwracająca 1 dla poprawnej odpowiedzi oraz 0 dla błędnej. Z kolei dla zadań regresji funkcja ta może zwracać absolutną różnicę w stosunku do poprawnej wartości.

Metoda weryfikacji pytania badawczego P_2

Kolejne pytanie badawcze związane z wpływem informacji zwrotnej na szybkość wykonywania mikro-zadań przez anotatorów w procesie crowdsourcingu.

W ramach tego pytania chciałem sprawdzić, czy wprowadzenie informacji zwrotnej spowoduje, że anotatorzy świadomi tego, że są oceniani, będą bardziej skrupulatnie wykonywać swoją pracę, co w konsekwencji wydłuży czas anotacji. Z tego powodu zdecydowałem się sformułować następujące pytanie badawcze:

P_2: Czy informacja zwrotna ma wpływ na szybkość tworzenia anotacji w procesie crowdsourcingu?

W celu ustalenia odpowiedzi na powyższe pytanie badawcze przeprowadzona została analiza porównawcza różnicy średniego czasu wykonywania anotacji przez anotatorów w wariancie W_1 oraz w wariancie kontrolnym W_1^c. Analiza przeprowadzona została w sposób analogiczny do metodologii użytej do weryfikacji hipotezy H_1 (zob. Paragraf 3.5.2).

Metoda weryfikacji pytania badawczego P_3

Ostatnie pytanie badawcze eksploruje temat związany z wpływem obecności informacji zwrotnej na zaangażowanie anotatorów w procesie crowdsourcingu. To pytanie badawcze jest bezpośrednio związane z opisanymi w pierwszym rozdziale założeniami teorii „przepływu" (zob. Paragraf 1.4.2). Teoria ta wymienia informację zwrotną jako jeden z podstawowych wariantów wymaganych do zaistnienia stanu „przepływu" mającego pozytywny wpływ na zaangażowanie osoby wykonującej zadanie. Z tego powodu zdecydowałem się sformułować pytanie badawcze:

P_3: Czy informacja zwrotna w procesie crowdsourcingu pozytywnie wpływa na zaangażowanie anotatorów?

Jednak samo zaangażowanie jest trudnym do zdefiniowania, wielowymiarowym konstruktem, które nie jest jednoznacznie zdefiniowane w literaturze[28]. Precyzyjne określenie poziomu zaangażowania uczestnika eksperymentu wybiega poza zakres niniejszej pracy. W związku tym, że wiązałoby się to z wykorzystaniem dedykowanych do tego celu kwestionariuszy. Dlatego w ramach niniejszej pracy przyjęta została uproszczona definicja zaangażowania anotatora rozumiana jako łączna liczba wszystkich oznaczonych mikro-zadań. Zgodnie z tą definicją anotator, który zdecydował się wykonać więcej mikro-zadań

[28] W kontekstach edukacyjnych temat zaangażowania opisany został na przykład przez Reeve [2012].

(np. więcej niż jedną partię mikro-zadań) był bardziej zaangażowany niż anotator, który wykonał mniejszą liczbę mikro-zadań. Głównym czynnikiem wpływającym na motywację anotatorów było otrzymywane wynagrodzone, ale ponieważ anotatorzy we wszystkich wariantach eksperymentu opłacani byli w ten sam sposób nie był to czynnik różnicujący badane grupy.

W celu ustalenia odpowiedzi na powyższe pytanie badawcze przeprowadzona została analiza porównawcza średniej liczby wszystkich wykonanych anotacji przez jednego anotatora w wariancie W_1 oraz w wariancie kontrolnym W_1^c. Analiza przeprowadzona została w sposób analogiczny do metodologii użytej do weryfikacji hipotezy H_1 (zob. Paragraf 3.5.2).

3.5.4. Analiza wyników eksperymentu

W tej części rozprawy omówiłem wyniki analizy przeprowadzonej dla wszystkich badanych zbiorów danych. Dla każdego z omawianych zbiorów zaprezentowane zostały wyniki analizy, której celem była weryfikacja hipotez oraz znalezienie odpowiedzi na pytania badawcze według opisanej powyżej metodologii.

Zbiór: skargi usług bankowych

Oznaczanie zbioru „skargi usług bankowych” związane było z klasyfikacją danych do jednej z pięciu predefiniowanych kategorii (zob. Paragraf 3.4.1). Anotatorzy pracujący nad zadaniem używali wiedzy domenowej, która związana była z poprawną identyfikacją kategorii na podstawie tekstu załączonej skargi. Jakość zebranych danych została oceniona za pomocą metryki F_1.

Poniżej zamieszczone zostały wnioski, które wyciągnąłem na podstawie analizy wykonanej w celu weryfikacji dwóch głównych hipotez eksperymentu (treść hipotez, zob. Paragraf 3.1.1). Wnioski zostały uzupełnione o przykładowe wartości metryk, które obrazują zaobserwowany efekt (lub jego brak). Dla hipotezy H_1 wyliczone zostały metryki opisujące rozkład różnicy pomiędzy wariantem W_1 oraz wariantem kontrolnym W_1^c. W Tabeli 3.4 zamieszczone zostały metryki obliczone dla wszystkich wariantów (W_1–W_3). W celu przedstawienia pełniejszego obrazu przeprowadzonej analizy przygotowany został również wykres przedstawiający rozkład różnic jakości w stosunku do grupy kontrolnej dla każdego z badanych wariantów (zob. Rysunek 3.20). Dla hipotezy H_2 wyliczone zostały metryki opisujące rozkład różnic pomiędzy wariantami W_1 oraz W_2 (opis budowy rozkładu, zob.

Paragraf 3.5.2). W Tabeli 3.4 zamieszczone zostały metryki obliczone dla wszystkich par wariantów.

H_1: Analiza wykazała umiarkowany wzrost jakości danych pozyskiwanych przez anotatorów uzyskujących informację zwrotną ($d = 0{,}41$) oraz bardzo wysokie prawdopodobieństwie zaobserwowania poprawy w jakości danych ($p = 0{,}96$).

H_2: W ramach analizy zaobserwowano duże oraz bardzo duże różnice w jakości pomiędzy porównywanymi wariantami (np. $d = 1{,}73$ przy analizie różnicy $W_1 - W_3$) przy wysokim prawdopodobieństwie zaobserwowania wzrostu jakości ($p = 0{,}88$ dla $W_1 - W_3$). Podczas porównywania wariantów o bardziej zbliżonej jakości ($W_1 - W_2$ oraz $W_2 - W_3$) różnica ta była niższa, lecz wciąż wysoka (np. $d = 0{,}92$ i $p = 0{,}72$ dla $W_1 - W_2$). Obserwowana jakość pozyskiwanych danych spadała wraz ze spadkiem jakości informacji zwrotnej w danym wariantu.

W celu rozszerzenia powyższej analizy przeprowadziłem weryfikację pytań badawczych (treść pytań, zob. Paragraf 3.1.1). Poniżej znajduje się opis wniosków dla każdego z pytań. Dla pytania P_1 wyliczone zostały metryki obrazujące moment efektu informacji zwrotnej. Wartości metryk wraz z graficzną reprezentacją momentu efektu przedstawione zostały na wykresie (zob. Rysunku 3.21). Dla pytania P_2 wyliczone zostały metryki obrazujące rozkład różnicy w średniej szybkości pracy anotatorów dla wariantu W_1 oraz wariantu kontrolnego W_2 (zob. Tabela 3.4). Dla pytania P_3 wyliczone zostały metryki obrazujące rozkład różnicy w liczbie wykonanych anotacji dla wariantu W_1 oraz wariantu kontrolnego W_2 (zob. Tabela 3.4).

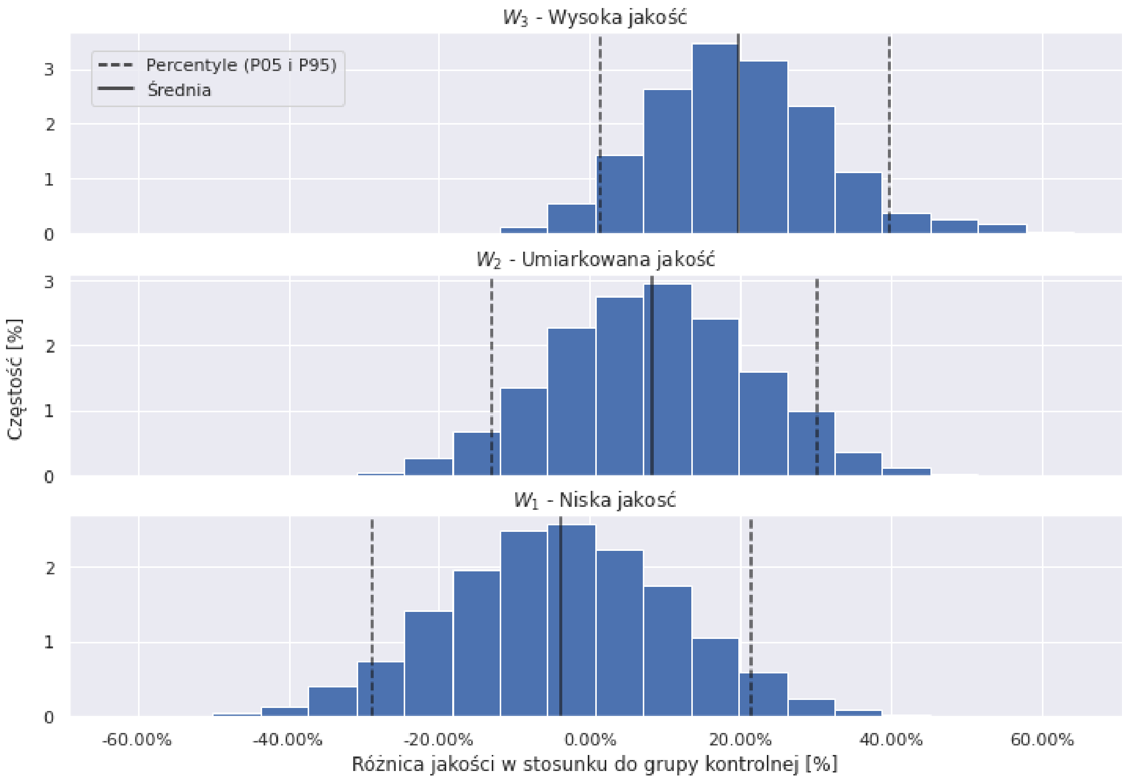
P_1: Zaobserwowano wyraźny pozytywny efekt natychmiastowy wprowadzenia informacji zwrotnej (efekt relatywny $b_r = +21{,}80\%$) oraz ujemny efekt długoterminowy (efekt relatywny $a_r = -10{,}85$).

P_2: Zaobserwowano niewielki ($d = 0{,}1$) wzrost w czasie wykonywania pracy (spadek tempa pracy) przy umiarkowanym prawdopodobieństwie ($p = 0{,}69$).

P_3: Zaobserwowano niewielki wzrost liczby wykonanych zadań ($d = 0{,}13$) przy umiarkowanym prawdopodobieństwie ($p = 0{,}70$).

Tabela 3.4: *Wartości metryk obliczonych w ramach weryfikacji głównych hipotez oraz pytań badawczych dla zbioru „skargi usług bankowych”*

Wariant	P_{05}	μ	P_{95}	p	d Cohena
Hipoteza H_1 — zmiana w metryce F_1					
W_1	+1,42%	+19,72%	+39,81%	0,96	0,41
W_2	-12,95%	+8,34%	+30,21%	0,73	0,20
W_3	-28,93%	-3,89%	+21,32%	0,40	-0,08
Hipoteza H_2 — zmiana w metryce F_1					
$W_1 - W_3$	-9,13%	+23,61%	+57,20%	0,88	1,73
$W_1 - W_2$	-18,29%	+11,38%	+41,56%	0,72	0,92
$W_2 - W_3$	-17,99%	+12,23%	+44,58%	0,73	0,86
Pytanie P_2 — zmiana w szybkości pracy anotatorów					
W_1	-12,74%	+5,04%	+22,26%	0,69	0,10
Pytanie P_3 — zmiana w liczbie wykonanych anotacji					
W_1	-14,44%	+7,99%	+33,41%	0,70	0,13



Rysunek 3.20: Rozkład różnic w jakości pomiędzy wariantami eksperymentu i ich grupami kontrolnymi dla zbioru „skargi usług bankowych”

Metryka	Wartość
α	-0,001
β	0,049
α_r	-10,85%
β_r	+21,80%

Rysunek 3.21: Wizualizacja momentu efektu informacji zwrotnej dla zbioru „skargi usług bankowych”

Zbiór: atrybuty produktów *eBay*

Oznaczanie zbioru „atrybuty produktów *eBay*” związane było ze znajdywaniem fragmentów tekstu, które reprezentują atrybuty produktów sprzedawanych na platformie *eBay* (zob. Paragraf 3.4.2). Każdy z wybranych atrybutów przypisywany był do jednej z sześciu kategorii. Anotatorzy pracujący nad zadaniem używali wiedzy domenowej, która związana była z właściwą identyfikacją każdej z kategorii. Jakość zebranych danych została oceniona za pomocą metryki F_1.

Poniżej zamieszczone zostały wnioski, które wyciągnąłem na podstawie analizy wykonanej w celu weryfikacji dwóch głównych hipotez eksperymentu (treść hipotez, zob. Paragraf 3.1.1). Wnioski zostały uzupełnione o przykładowe wartości metryk, które obrazują zaobserwowany efekt (lub jego brak). Dla hipotezy H_1 wyliczone zostały metryki opisujące rozkład różnicy pomiędzy wariantem W_1 oraz wariantem kontrolnym W_1^c. W Tabeli 3.5 zamieszczone zostały metryki obliczone dla wszystkich wariantów (W_1–W_3). W celu przedstawienia pełniejszego obrazu przeprowadzonej analizy przygotowany został również wykres przedstawiający rozkład różnic jakości w stosunku do grupy kontrolnej dla każdego z badanych wariantów (zob. Rysunek 3.22). Dla hipotezy H_2 wyliczone zostały metryki opisujące rozkład różnic pomiędzy wariantami W_1 oraz W_2 (opis budowy rozkładu, zob. Paragraf 3.5.2). W Tabeli 3.5 zamieszczone zostały metryki obliczone dla wszystkich par wariantów.

H_1: Analiza wykazała umiarkowany wzrost jakości danych pozyskiwanych przez anotatorów uzyskujących informację zwrotną ($d = 0{,}54$) oraz bardzo wysokie prawdopodobieństwie zaobserwowania poprawy w jakości danych ($p = 0{,}99$).

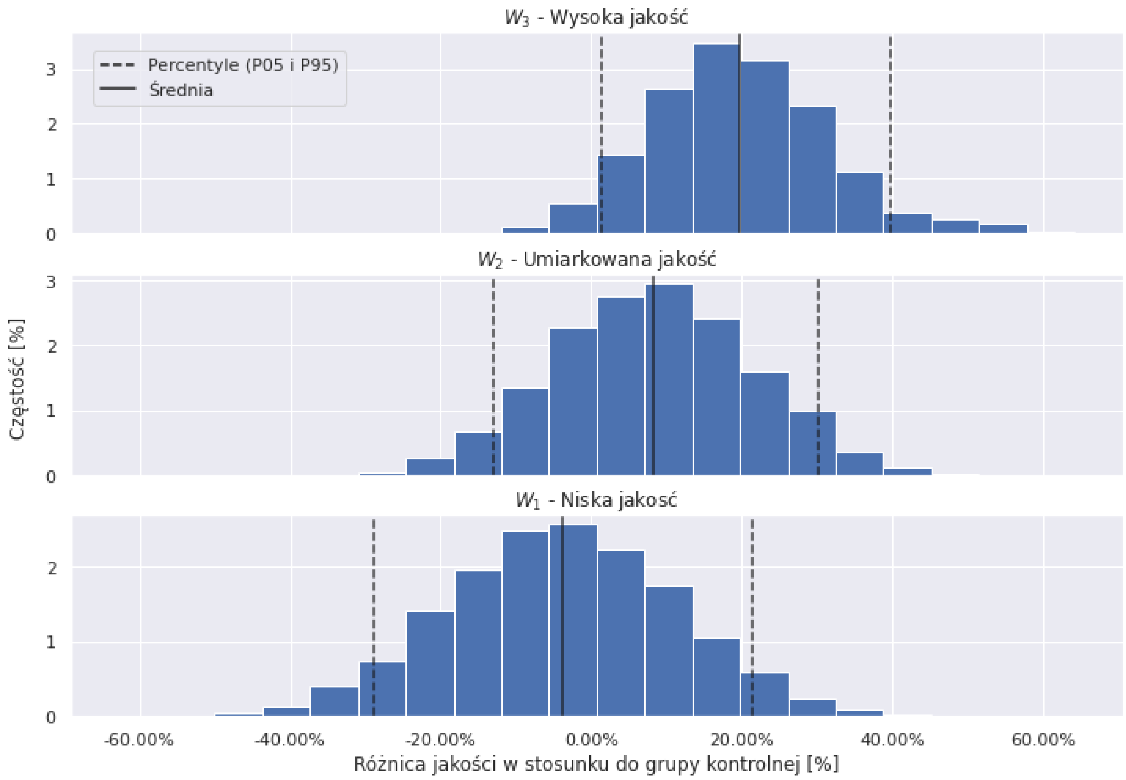
H_2: W ramach analizy zaobserwowano duże i bardzo duże różnice w jakości pomiędzy porównywanymi wariantami (np. $d = 2{,}26$ przy analizie różnicy $W_1 - W_3$) przy wysokim prawdopodobieństwie zaobserwowania wzrostu jakości ($p = 0{,}95$ dla $W_1 - W_3$). Podczas porównywania wariantów o bardziej zbliżonej jakości ($W_1 - W_2$ oraz $W_2 - W_3$) różnica ta była niższa, lecz wciąż bardzo wysoka (np. $d = 1{,}71$ dla $W_1 - W_2$). Obserwowana jakość pozyskiwanych danych spadała wraz ze spadkiem jakości informacji zwrotnej w danym wariantu.

W celu rozszerzenia powyższej analizy przeprowadziłem weryfikację pytań badawczych (treść pytań, zob. Paragraf 3.1.1). Poniżej znajduje się opis wniosków dla każdego z pytań. Dla pytania P_1 wyliczone zostały metryki obrazujące moment efektu informacji zwrotnej. Wartości metryk wraz z graficzną reprezentacją momentu efektu przedstawione zostały na wykresie (zob. Rysunku 3.23). Dla pytania P_2 wyliczone zostały metryki obrazujące rozkład różnicy w średniej szybkości pracy anotatorów dla wariantu W_1 oraz wariantu kontrolnego W_2 (zob. Tabela 3.5). Dla pytania P_3 wyliczone zostały metryki obrazujące rozkład różnicy w liczbie wykonanych anotacji dla wariantu W_1 oraz wariantu kontrolnego W_2 (zob. Tabela 3.5).

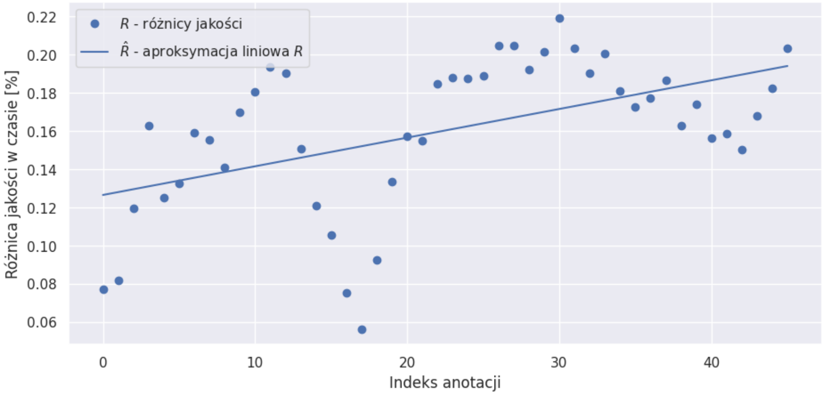
- P_1: Zaobserwowano bardzo wyraźny pozytywny efekt natychmiastowy wprowadzenia informacji zwrotnej (efekt relatywny $b_r = +63{,}32\%$) oraz równie znaczący, pozytywny efekt długoterminowy (efekt relatywny $a_r = +33{,}79\%$).
- P_2: Zaobserwowano umiarkowany ($d = 0{,}4$) wzrost w czasie wykonywania pracy (spadek tempa pracy) przy bardzo wysokim prawdopodobieństwie ($p = 0{,}96$).
- P_3: Zaobserwowano niski ($d = -0{,}23$) spadek liczby wykonanych anotacji przy wysokim prawdopodobieństwie ($1 - p = 0{,}87$).

Tabela 3.5: *Wartości metryk obliczonych w ramach weryfikacji głównych hipotez oraz pytań badawczych dla zbioru „atrybuty produktów eBay”*

Wariant	P_{05}	μ	P_{95}	p	d Cohena
Hipoteza H_1 — zmiana w metryce F_1					
W_1	+21,67%	+64,69%	+107,28%	0,99	0,54
W_2	-29,03%	+18,08%	+65,85%	0,73	0,23
W_3	-46,75%	+2,23%	+49,55%	0,54	0,03
Hipoteza H_2 — zmiana w metryce F_1					
$W_1 - W_3$	+0,13%	+62,46%	+124,18%	0,95	2,26
$W_1 - W_2$	-15,71%	+46,61%	+106,74%	0,89	1,71
$W_2 - W_3$	-52,42%	+15,85%	+85,85%	0,64	0,55
Pytanie P_2 — zmiana w szybkości pracy anotatorów					
W_1	+2,66%	+28,75%	+56,62%	0,96	0,4
Pytanie P_3 — zmiana w liczbie wykonanych anotacji					
W_1	-64,97%	-24,29%	+8,77%	0,13	-0,23



Rysunek 3.22: Rozkład różnic w jakości pomiędzy wariantami eksperymentu i ich grupami kontrolnymi dla zbioru „atrybuty produktów *eBay*”



Metryka	Wartość
α	0,001
β	0,126
α_r	+33,79%
β_r	+63,32%

Rysunek 3.23: Wizualizacja momentu efektu informacji zwrotnej dla zbioru „atrybuty produktów *eBay*"

Zbiór: waga produktów *eBay*

Oznaczanie zbioru „waga produktów *eBay*" związane było z oszacowaniem wagi produktu sprzedawanego na platformie *eBay* na podstawie zdjęcia i opisu przedmiotu (zob. Paragraf 3.4.3). Anotatorzy pracujący nad zadaniem używali wiedzy domenowej, która związana była z precyzyjnym określaniem wagi produktów na podstawie ich opisu i zdjęcia. Jakość zebranych danych została oceniona za pomocą metryki MAE (średni błąd absolutny). Ponieważ metryka ta obrazuje błąd oszacowania, to w przypadku tego zbioru spadek metryki oznaczał poprawę jakości.

Poniżej zamieszczone zostały wnioski, które wyciągnąłem na podstawie analizy wykonanej w celu weryfikacji dwóch głównych hipotez eksperymentu (treść hipotez, zob. Paragraf 3.1.1). Wnioski zostały uzupełnione o przykładowe wartości metryk, które obrazują zaobserwowany efekt (lub jego brak). Dla hipotezy H_1 wyliczone zostały metryki opisujące rozkład różnicy pomiędzy wariantem W_1 oraz wariantem kontrolnym W_1^c. W Tabeli 3.6 zamieszczone zostały metryki obliczone dla wszystkich wariantów (W_1–W_3). W celu przedstawienia pełniejszego obrazu przeprowadzonej analizy przygotowany został również wykres przedstawiający rozkład różnic jakości w stosunku do grupy kontrolnej dla każdego z badanych wariantów (zob. Rysunek 3.24). Dla hipotezy H_2 wyliczone zostały metryki opisujące rozkład różnic pomiędzy wariantami W_1 oraz W_2 (opis budowy rozkładu, zob. Paragraf 3.5.2). W Tabeli 3.6 zamieszczone zostały metryki obliczone dla wszystkich par wariantów.

H_1: Analiza wykazała umiarkowany wzrost jakości danych pozyskanych przez anotatorów uzyskujących informację zwrotną ($d = -0{,}5$) oraz bardzo wysokie prawdopodobieństwo zaobserwowania poprawy w jakości danych ($1 - p = 1{,}0$).

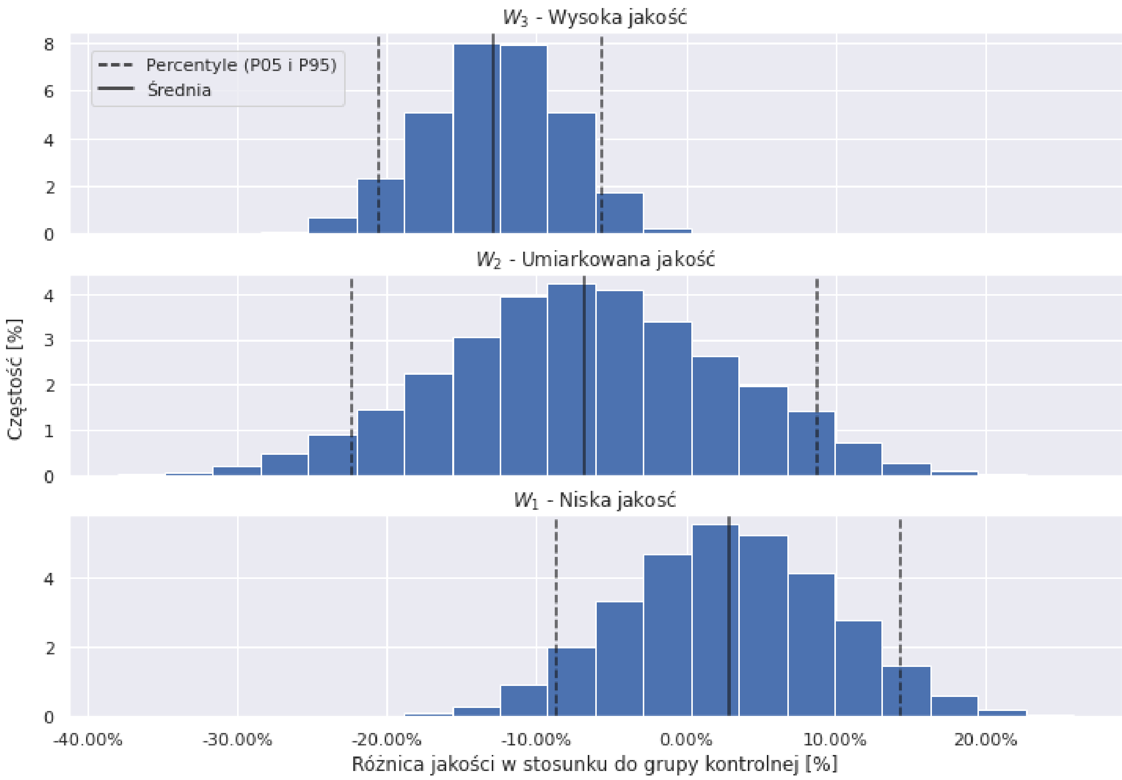
- H_2: W ramach analizy zaobserwowano bardzo duże różnice pomiędzy porównywanymi wariantami (np. $d = -2{,}66$ przy analizie $W_1 - W_3$) przy wysokim prawdopodobieństwie zaobserwowania wzrostu ($p = 0{,}95$ dla $W_1 - W_3$). Podczas porównywania wariantów o bardziej zbliżonej jakości ($W_1 - W_2$ oraz $W_2 - W_3$) różnica ta była niższa, lecz wciąż wysoka (np. $d = -0{,}82$ dla $W_1 - W_2$). Obserwowana jakość pozyskiwanych danych spadała wraz ze spadkiem jakości informacji zwrotnej w danym wariantu.

W celu rozszerzenia powyższej analizy przeprowadziłem weryfikację pytań badawczych (treść pytań, zob. Paragraf 3.1.1). Poniżej znajduje się opis wniosków dla każdego z pytań. Dla pytania P_1 wyliczone zostały metryki obrazujące moment efektu informacji zwrotnej. Wartości metryk wraz z graficzną reprezentacją momentu efektu przedstawione zostały na wykresie (zob. Rysunku 3.25). Dla pytania P_2 wyliczone zostały metryki obrazujące rozkład różnicy w średniej szybkości pracy anotatorów dla wariantu W_1 oraz wariantu kontrolnego W_2 (zob. Tabela 3.6). Dla pytania P_3 wyliczone zostały metryki obrazujące rozkład różnicy w liczbie wykonanych anotacji dla wariantu W_1 oraz wariantu kontrolnego W_2 (zob. Tabela 3.6).

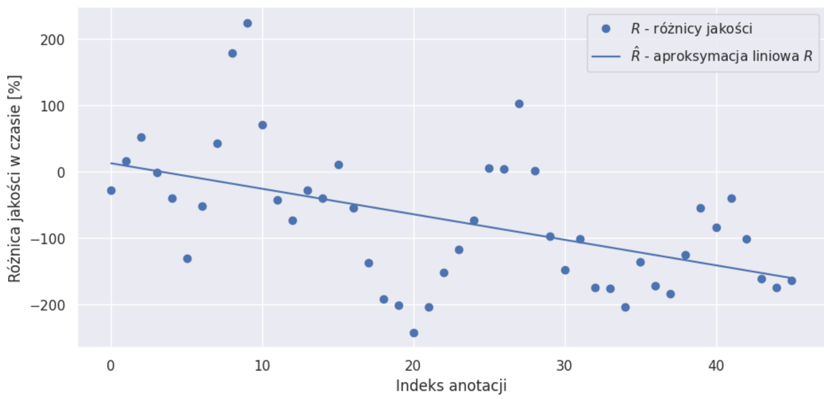
- P_1: Nie zaobserwowano wyraźnego efektu natychmiastowy wprowadzenia informacji zwrotnej (efekt relatywny $b_r = +1{,}31\%$) natomiast zaobserwowano znaczący, pozytywny efekt długoterminowy (efekt relatywny $a_r = -18{,}04\%$ dla błędu anotacji).
- P_2: Zaobserwowano niewielki ($d = -0{,}25$) spadek czasu wykonywania pracy (wzrost tempa pracy) przy wysokim prawdopodobieństwie ($1 - p = 0{,}90$).
- P_3: Zaobserwowano niski ($d = 0{,}17$) wzrost liczby wykonanych anotacji przy umiarkowanym prawdopodobieństwie ($p = 0{,}80$).

Tabela 3.6: *Wartości metryk obliczonych w ramach weryfikacji głównych hipotez oraz pytań badawczych dla zbioru „waga produktów eBay”*

Wariant	$\mathbf{P_{05}}$	μ	$\mathbf{P_{95}}$	**p**	**d Cohena**
Hipoteza $\mathbf{H_1}$ — zmiana w metryce MAE					
W_1	-20,56%	-12,83%	-5,65%	0,0	-0,5
W_2	-22,31%	-6,75%	+8,75%	0,24	-0,28
W_3	-8,70%	+2,85%	+14,38%	0,65	0,12
Hipoteza $\mathbf{H_2}$ — zmiana w metryce MAE					
$W_1 - W_3$	-29,79%	-15,68%	-1,47%	0,03	-2,66
$W_1 - W_2$	-22,44%	-6,08%	+11,51%	0,27	-0,83
$W_2 - W_3$	-28,52%	-9,60%	+9,52%	0,20	-1,16
Pytanie $\mathbf{P_2}$ — zmiana w szybkości pracy anotatorów					
W_1	-31,29%	-13,89%	+4,05%	0,10	-0,25
Pytanie $\mathbf{P_3}$ — zmiana w liczbie wykonanych anotacji					
W_1	-5,29%	+6,79%	+20,20%	0,80	0,17



Rysunek 3.24: Rozkład różnic w jakości pomiędzy wariantami eksperymentu i ich grupami kontrolnymi dla zbioru „waga produktów *eBay*”



Metryka	Wartość
α	-3,838
β	12,561
α_r	-18,04%
β_r	+1,31%

Rysunek 3.25: Wizualizacja momentu efektu informacji zwrotnej dla zbioru „waga produktów *eBay*"

Zbiór: wydźwięk opinii o hotelach

Oznaczanie zbioru „wydźwięk opinii o hotelach" związane było z przypisywaniem oceny w pięciostopniowej skali *Likerta* do opinii napisanej przez klienta hotelu (zob. Paragraf 3.4.4). Anotatorzy pracujący nad zadaniem posługiwali się wiedzą zdroworozsądkową, ponieważ dokonywali oni wyboru oceny na podstawie swoich osobistych odczucia na temat danej opinii. Jakość zebranych danych została oceniona za pomocą metryki MAE (średni błąd absolutny). Ponieważ metryka ta obrazuje błąd oszacowania, to w przypadku tego zbioru spadek metryki oznaczał poprawę jakości.

Poniżej zamieszczone zostały wnioski, które wyciągnąłem na podstawie analizy wykonanej w celu weryfikacji dwóch głównych hipotez eksperymentu (treść hipotez, zob. Paragraf 3.1.1). Wnioski zostały uzupełnione o przykładowe wartości metryk, które obrazują zaobserwowany efekt (lub jego brak). Dla hipotezy H_1 wyliczone zostały metryki opisujące rozkład różnicy pomiędzy wariantem W_1 oraz wariantem kontrolnym W_1^c. W Tabeli 3.7 zamieszczone zostały metryki obliczone dla wszystkich wariantów (W_1–W_3). W celu przedstawienia pełniejszego obrazu przeprowadzonej analizy przygotowany został również wykres przedstawiający rozkład różnic jakości w stosunku do grupy kontrolnej dla każdego z badanych wariantów (zob. Rysunek 3.26). Dla hipotezy H_2 wyliczone zostały metryki opisujące rozkład różnic pomiędzy wariantami W_1 oraz W_2 (opis budowy rozkładu, zob. Paragraf 3.5.2). W Tabeli 3.7 zamieszczone zostały metryki obliczone dla wszystkich par wariantów.

H_1: Analiza nie wykazała różnic w jakości danych pozyskanych przez anotatorów uzyskujących informację zwrotną w porównaniu do grupy kontrolnej. Zaobserwowany efekt był bliski zeru ($d = 0{,}01$), a prawdopodobieństwo wzrostu jakości danych bliskie 50% ($p = 0{,}5213$).

H_2: Podczas porównywania wariantów zaobserwowano pomiędzy nimi wyraźne różnice. W wariancie W_2 zaobserwowano umiarkowany spadek jakości danych ($d = -0{,}32$), a dla wariantu W_2 duży wzrost ($d = 0{,}87$). Analiza porównywania wariantów $W_1 – W_3$ wykazało bardzo duży ujemny efekt ($d = -3{,}02$) przy wysokim prawdopodobieństwie spadku jakości ($1 - p = 0{,}9871$). Porównywanie pozostałych wariantów wielkość efektu również utrzymuje się na wysokich wartościach ($d = 1{,}12$ dla $W_1 - W_2$ oraz $d = -3{,}6$ dla $W_2 - W3$).

W celu rozszerzenia powyższej analizy przeprowadziłem weryfikację pytań badawczych (treść pytań, zob. Paragraf 3.1.1). Poniżej znajduje się opis wniosków dla każdego z pytań. Dla pytania P_1 wyliczone zostały metryki obrazujące moment efektu informacji zwrotnej. Wartości metryk wraz z graficzną reprezentacją momentu efektu przedstawione zostały na wykresie (zob. Rysunku 3.27). Dla pytania P_2 wyliczone zostały metryki obrazujące rozkład różnicy w średniej szybkości pracy anotatorów dla wariantu W_1 oraz wariantu kontrolnego W_2 (zob. Tabela 3.7). Dla pytania P_3 wyliczone zostały metryki obrazujące rozkład różnicy w liczbie wykonanych anotacji dla wariantu W_1 oraz wariantu kontrolnego W_2 (zob. Tabela 3.7).

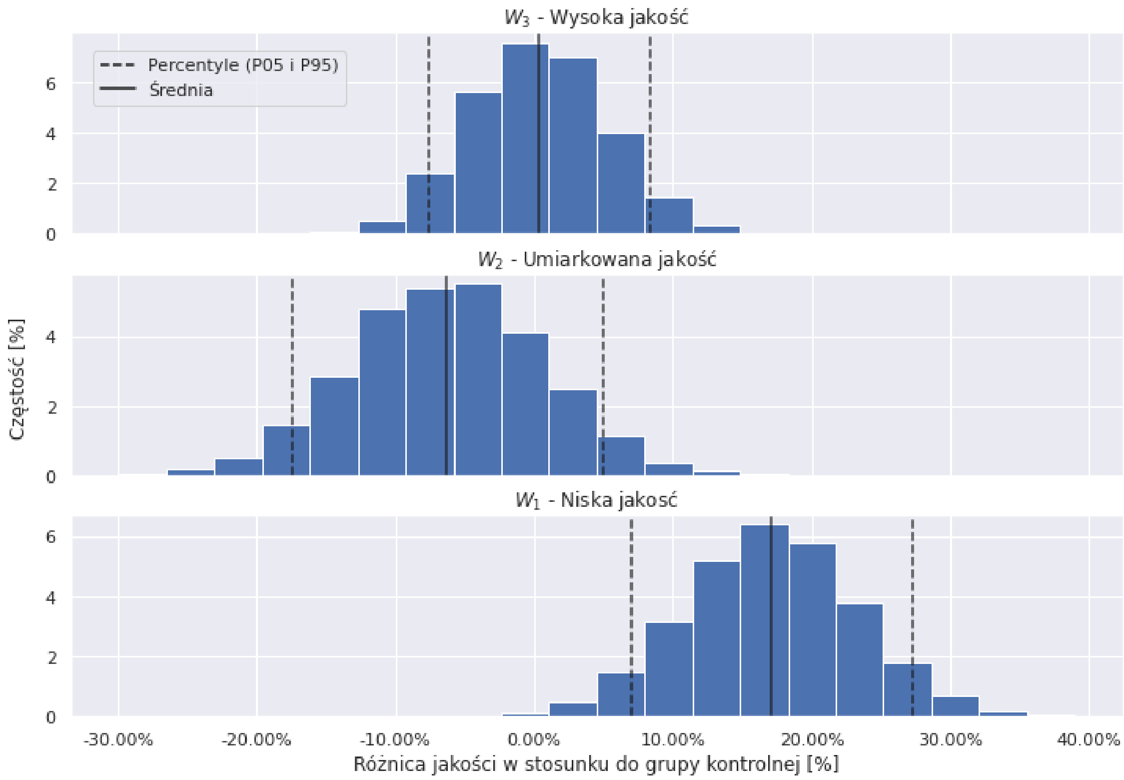
P_1: Zaobserwowano niski efekt natychmiastowy wprowadzenia informacji zwrotnej (efekt relatywny $b_r = +12{,}44\%$) oraz zaobserwowano znaczący, negatywny efekt długoterminowy (efekt relatywny $a_r = -21{,}62\%$ dla błędu anotacji). Oznacza to, że wraz z liczbą wykonanych zadań obniżała się jakość pracy danego anotatora.

P_2: Zaobserwowano niski ($d = 0{,}26$) wzrost czasu wykonywania pracy (spadek tempa pracy) przy wysokim prawdopodobieństwie ($p = 0{,}94$).

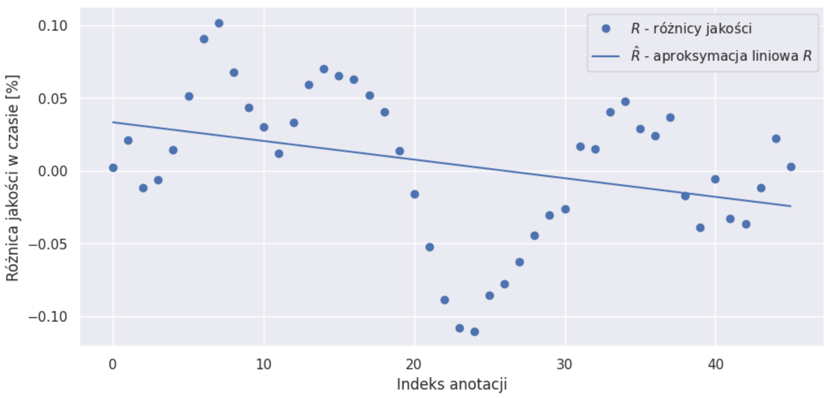
P_3: Zaobserwowano niski ($d = -0{,}27$) spadek liczby wykonanych zadań przy wysokim prawdopodobieństwie ($1 - p = 0{,}94$).

Tabela 3.7: *Wartości metryk obliczonych w ramach weryfikacji głównych hipotez oraz pytań badawczych dla zbioru „wydźwięk opinii o hotelach”*

Wariant	P_{05}	μ	P_{95}	p	d Cohena
Hipoteza H_1 — zmiana w metryce MAE					
W_1	-7,60%	+0,33%	+8,36%	0,52	0,01
W_2	-17,46%	-6,30%	+4,96%	0,18	-0,32
W_3	+6,99%	+17,07%	+27,26%	0,99	0,87
Hipoteza H_2 — zmiana w metryce MAE					
$W_1 - W_3$	-30,31%	-16,74%	-4,16%	0,01	-3,02
$W_1 - W_2$	-6,40%	+6,63%	+20,20%	0,79	1,12
$W_2 - W_3$	-38,64%	-23,37%	-8,02%	0,00	-3,6
Pytanie P_2 — zmiana w szybkości pracy anotatorów					
W_1	-0,85%	+20,33%	+42,17%	0,94	0,26
Pytanie P_3 — zmiana w liczbie wykonanych anotacji					
W_1	-19,53%	-9,19%	+0,32%	0,06	-0,27



Rysunek 3.26: Rozkład różnic w jakości pomiędzy wariantami eksperymentu i ich grupami kontrolnymi dla zbioru „wydźwięk opinii o hotelach”



Metryka	Wartość
α	-0,001
β	0,033
α_r	-21,62%
β_r	+12,44%

Rysunek 3.27: Wizualizacja momentu efektu informacji zwrotnej dla zbioru „wydźwięk opinii o hotelach”

Zbiór: jednostki nazwane

Oznaczanie zbioru „jednostki nazwane” związane było ze znajdywaniem fragmentów tekstu, które oznaczały jednostki nazwane (zob. Paragraf 3.4.5). Każda z wybranych jednostek przypisywany był do jednej z sześciu kategorii. Anotatorzy pracujący nad zadaniem używali wiedzy domenowej, która związana była z właściwą identyfikacją każdej z kategorii. Jakość zebranych danych została oceniona za pomocą metryki F_1.

Poniżej zamieszczone zostały wnioski, które wyciągnąłem na podstawie analizy wykonanej w celu weryfikacji dwóch głównych hipotez eksperymentu (treść hipotez, zob. Paragraf 3.1.1). Wnioski zostały uzupełnione o przykładowe wartości metryk, które obrazują zaobserwowany efekt (lub jego brak). Dla hipotezy H_1 wyliczone zostały metryki opisujące rozkład różnicy pomiędzy wariantem W_1 oraz wariantem kontrolnym W_1^c. W Tabeli 3.8 zamieszczone zostały metryki obliczone dla wszystkich wariantów (W_1–W_3). W celu przedstawienia pełniejszego obrazu przeprowadzonej analizy przygotowany został również wykres przedstawiający rozkład różnic jakości w stosunku do grupy kontrolnej dla każdego z badanych wariantów (zob. Rysunek 3.28). Dla hipotezy H_2 wyliczone zostały metryki opisujące rozkład różnic pomiędzy wariantami W_1 oraz W_2 (opis budowy rozkładu, zob. Paragraf 3.5.2). W Tabeli 3.8 zamieszczone zostały metryki obliczone dla wszystkich par wariantów.

H_1: Analiza wykazała umiarkowany wzrost jakości danych pozyskanych przez anotatorów uzyskujących informację zwrotną ($d = 0{,}43$) przy bardzo wysokim prawdopodobieństwie wzrostu jakości ($p = 0{,}96$).

H_2: Podczas porównywania wariantów zaobserwowano pomiędzy nimi wyraźne różnice. Dla wariantów W_2 oraz W_3 efekt poprawy jakości danych był bliski zeru. Analiza różnicy $W_1 - W_3$ wykazała bardzo duży wzrost ($d = 1{,}45$) oraz umiarkowane prawdopodobieństwo wzrostu ($d = 0{,}85$). Analogicznie w przypadku porównywania wariantów $W_1 - W_2$. Natomiast w przypadku różnicy $W_2 - W_3$ efekt był bliski zeru ($d = 0{,}05$).

W celu rozszerzenia powyższej analizy przeprowadziłem weryfikację pytań badawczych (treść pytań, zob. Paragraf 3.1.1). Poniżej znajduje się opis wniosków dla każdego z pytań. Dla pytania P_1 wyliczone zostały metryki obrazujące moment efektu informacji zwrotnej. Wartości metryk wraz z graficzną reprezentacją momentu efektu przedstawione zostały na wykresie (zob. Rysunku 3.29). Dla pytania P_2 wyliczone zostały metryki obrazujące rozkład różnicy w średniej szybkości pracy anotatorów dla wariantu W_1 oraz wariantu kontrolnego W_2 (zob. Tabela 3.8). Dla pytania P_3 wyliczone zostały metryki obrazujące rozkład różnicy w liczbie wykonanych anotacji dla wariantu W_1 oraz wariantu kontrolnego W_2 (zob. Tabela 3.8).

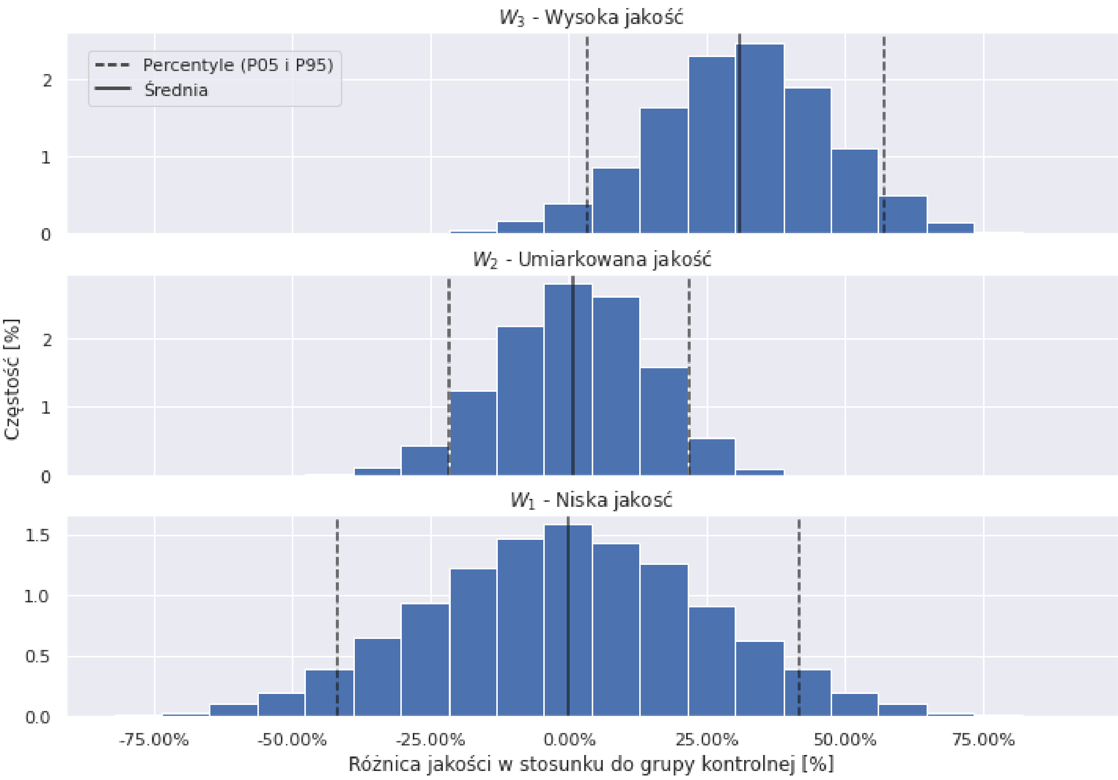
P_1: Zaobserwowano wyraźny pozytywny efekt natychmiastowy wprowadzenia informacji zwrotnej (efekt relatywny $b_r = +19{,}46\%$) oraz równie znaczący, pozytywny efekt długoterminowy (efekt relatywny $a_r = +25{,}63\%$).

P_2: Nie zaobserwowano znaczącej różnicy pomiędzy wariantem W_1 a wariantem kontrolnym ($d = -0{,}05$ przy $p = 0{,}40$).

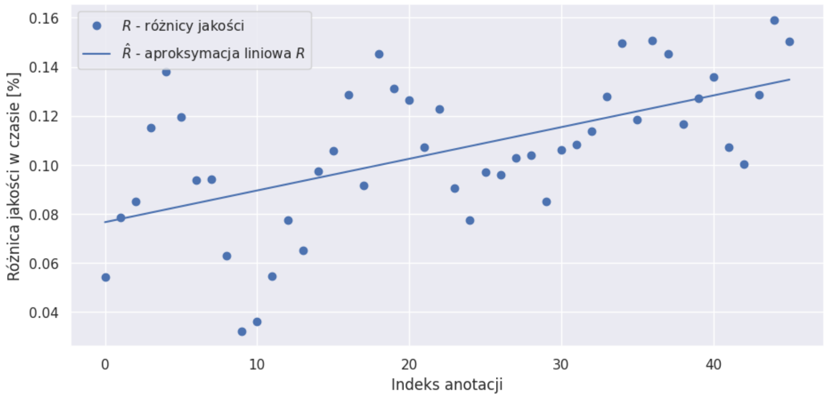
P_3: Nie zaobserwowano znaczącej różnicy pomiędzy wariantem W_1 a wariantem kontrolnym ($d = 0{,}02$ przy $p = 0{,}53$.

Tabela 3.8: *Wartości metryk obliczonych w ramach weryfikacji głównych hipotez oraz pytań badawczych dla zbioru „jednostki nazwane”*

Wariant	$\mathbf{P_{05}}$	μ	$\mathbf{P_{95}}$	**p**	**d Cohena**
Hipoteza $\mathbf{H_1}$ — zmiana w metryce $\mathbf{F_1}$					
W_1	+3,31%	+30,87%	+57,05%	0,96	0,43
W_2	-21,53%	+0,76%	+21,75%	0,53	0,02
W_3	-41,89%	-0,16%	+41,63%	0,49	-0,0
Hipoteza $\mathbf{H_2}$ — zmiana w metryce $\mathbf{F_1}$					
$W_1 - W_3$	-17,67%	+31,03%	+78,35%	0,85	1,45
$W_1 - W_2$	-3,11%	+30,11%	+62,39%	0,93	2,04
$W_2 - W_3$	-48,26%	+0,92%	+47,65%	0,52	0,05
Pytanie $\mathbf{P_2}$ — zmiana w szybkości pracy anotatorów					
W_1	-19,60%	-2,38%	+15,31%	0,40	-0,05
Pytanie $\mathbf{P_3}$ — zmiana w liczbie wykonanych anotacji					
W_1	-20,71%	+1,15%	+22,93%	0,53	0,02



Rysunek 3.28: Rozkład różnic w jakości pomiędzy wariantami eksperymentu i ich grupami kontrolnymi dla zbioru „jednostki nazwane”



Metryka	Wartość
α	0,001
β	0,077
α_r	+19,46%
β_r	+25,63%

Rysunek 3.29: Wizualizacja momentu efektu informacji zwrotnej dla zbioru „jednostki nazwane”

Zbiór: wyrazy bliskoznaczne

Oznaczanie zbioru „wyrazy bliskoznaczne” związane podjęciem binarnej decyzji czy para podanych słów ma podobne znaczenie, czy też nie. W związku z tym, że anotatorami były osoby, które znały język angielski, to pracujący nad zadaniem posługiwali się wiedzą zdroworozsądkową (nabytą podczas nauki języka). Jakość zebranych danych została oceniona za pomocą metryki dokładność (ang. *accuracy*).

Poniżej zamieszczone zostały wnioski, które wyciągnąłem na podstawie analizy wykonanej w celu weryfikacji dwóch głównych hipotez eksperymentu (treść hipotez, zob. Paragraf 3.1.1). Wnioski zostały uzupełnione o przykładowe wartości metryk, które obrazują zaobserwowany efekt (lub jego brak). Dla hipotezy H_1 wyliczone zostały metryki opisujące rozkład różnicy pomiędzy wariantem W_1 oraz wariantem kontrolnym W_1^c. W Tabeli 3.9 zamieszczone zostały metryki obliczone dla wszystkich wariantów (W_1–W_3). W celu przedstawienia pełniejszego obrazu przeprowadzonej analizy przygotowany został również wykres przedstawiający rozkład różnic jakości w stosunku do grupy kontrolnej dla każdego z badanych wariantów (zob. Rysunek 3.30). Dla hipotezy H_2 wyliczone zostały metryki opisujące rozkład różnic pomiędzy wariantami W_1 oraz W_2 (opis budowy rozkładu, zob. Paragraf 3.5.2). W Tabeli 3.9 zamieszczone zostały metryki obliczone dla wszystkich par wariantów.

H_1: Analiza wykonała niski wzrost jakości danych pozyskanych przez anotatorów uzyskujących informację zwrotną ($d = 0{,}26$) oraz wysokie prawdopodobieństwo wzrostu jakości ($d = 0{,}92$).

H_2: Podczas porównywania wariantów zaobserwowano pomiędzy nimi znaczące różnice. Dla wszystkich wariantów wzrost jakości był niski. Obserwowana jakość pozyskiwanych danych spadała wraz ze spadkiem jakości informacji zwrotnej w danym wariantu. W przypadku. Analiza różnicy $W_1 - W_3$ wykazała bardzo duży efekt ($d = 1{,}03$) przy umiarkowanym prawdopodobieństwie wzrostu ($p = 0{,}78$). W przypadku pozostałych par obserwowalny efekt był mniejszy.

W celu rozszerzenia powyższej analizy przeprowadziłem weryfikację pytań badawczych (treść pytań, zob. Paragraf 3.1.1). Poniżej znajduje się opis wniosków dla każdego z pytań. Dla pytania P_1 wyliczone zostały metryki obrazujące moment efektu informacji zwrotnej. Wartości metryk wraz z graficzną reprezentacją momentu efektu przedstawione zostały na wykresie (zob. Rysunku 3.31). Dla pytania P_2 wyliczone zostały metryki obrazujące rozkład różnicy w średniej szybkości pracy anotatorów dla wariantu W_1 oraz wariantu kontrolnego W_2 (zob. Tabela 3.9). Dla pytania P_3 wyliczone zostały metryki obrazujące rozkład różnicy w liczbie wykonanych anotacji dla wariantu W_1 oraz wariantu kontrolnego W_2 (zob. Tabela 3.9).

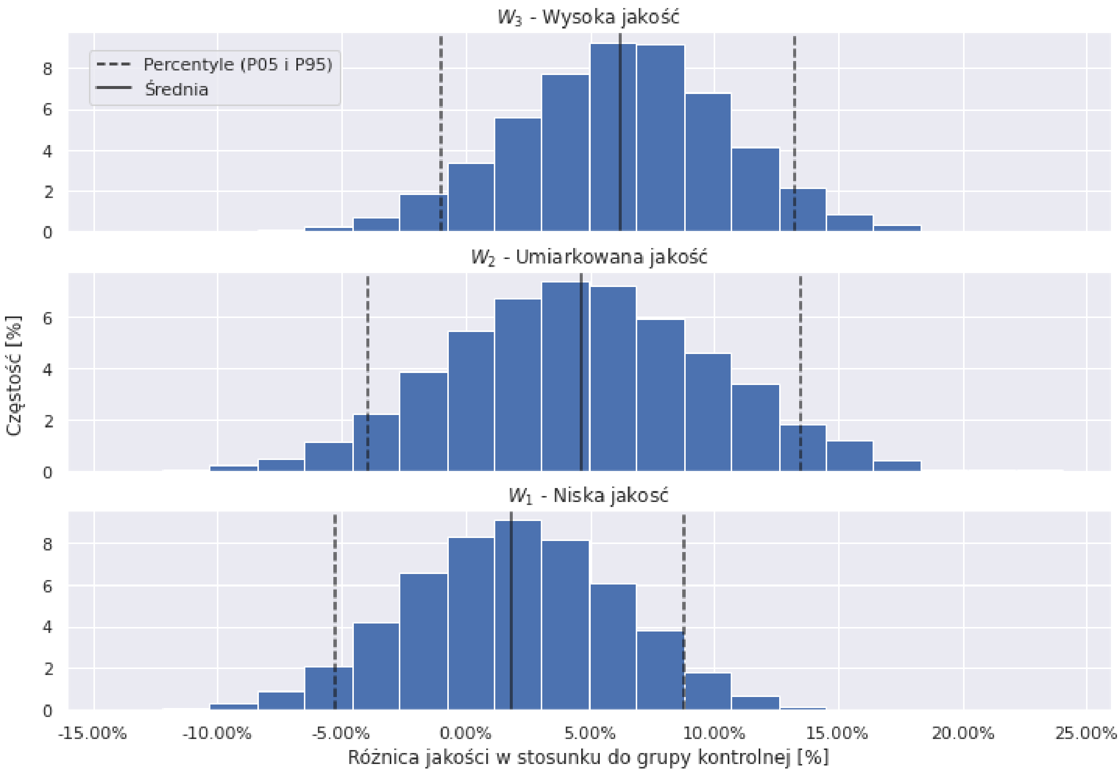
P_1: Zaobserwowano niski pozytywny efekt natychmiastowy wprowadzenia informacji zwrotnej (efekt relatywny $b_r = +4{,}51\%$) oraz niski pozytywny efekt długoterminowy (efekt relatywny $a_r = +2{,}37\%$).

P_2: Zaobserwowano niski wzrost ($d = 0{,}27$) w czasie wykonywania pracy (spadek tempa pracy) przy wysokim prawdopodobieństwie ($p = 0{,}93$).

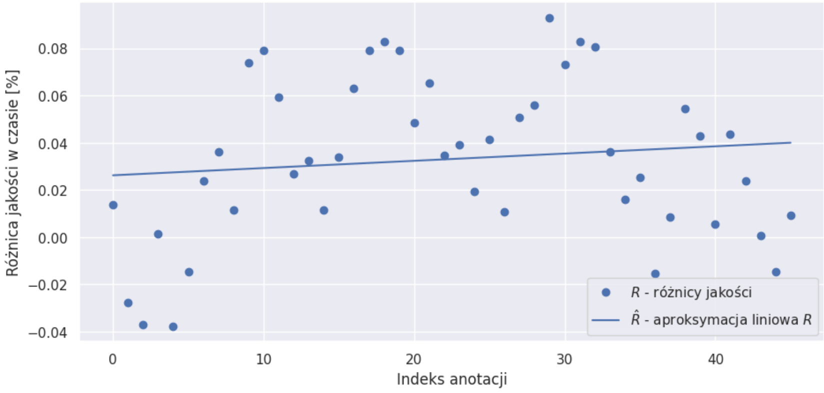
P_3: Zaobserwowano niewielki wzrost liczby wykonanych zadań ($d = 0{,}13$) przy umiarkowanym prawdopodobieństwie ($p = 0{,}70$).

Tabela 3.9: *Wartości metryk obliczonych w ramach weryfikacji głównych hipotez oraz pytań badawczych dla zbioru „wyrazy bliskoznaczne”*

Wariant	$\mathbf{P_{05}}$	μ	$\mathbf{P_{95}}$	**p**	**d Cohena**
Hipoteza $\mathbf{H_1}$ — zmiana w metryce *dokładność*					
W_1	-0,97%	+6,23%	+13,24%	0,92	0,26
W_2	-3,93%	+4,68%	+13,50%	0,81	0,26
W_3	-5,26%	+1,84%	+8,81%	0,67	0,11
Hipoteza $\mathbf{H_2}$ — zmiana w metryce *dokładność*					
$W_1 - W_3$	-5,71%	+4,38%	+13,97%	0,78	1,03
$W_1 - W_2$	-9,28%	+1,55%	+12,20%	0,59	0,32
$W_2 - W_3$	-7,28%	+2,84%	+14,24%	0,65	0,59
Pytanie $\mathbf{P_2}$ — zmiana w szybkości pracy anotatorów					
W_1	-2,88%	+24,21%	+48,99%	0,93	0,27
Pytanie $\mathbf{P_3}$ — zmiana w liczbie wykonanych anotacji					
W_1	-5,06%	+5,36%	+19,59%	0,72	0,1



Rysunek 3.30: Rozkład różnic w jakości pomiędzy wariantami eksperymentu i ich grupami kontrolnymi dla zbioru wyrazy bliskoznaczne



Metryka	Wartość
α	0,0
β	0,026
α_r	+2,37%
β_r	+4,51%

Rysunek 3.31: Wizualizacja momentu efektu informacji zwrotnej dla zbioru wyrazy bliskoznaczne

3.5.5. Podsumowanie eksperymentu

Poniżej zmieszczone zostało krótkie podsumowanie wszystkich przeprowadzonych eksperymentów, a także wnioski płynące z otrzymanych wyników.

Tabela 3.10: *Zestawienie wyników analiz jakości dla przeprowadzonego eksperymentu (kolor zielony – wartości dodatnie, kolor czerwony – wartości ujemne)*

Zbiór danych	Metryka	$\mathbf{W_1}$			$\mathbf{W_2}$			$\mathbf{W_3}$		
		$\mathbf{P_{05}}$	$\boldsymbol{\mu}$	$\mathbf{P_{95}}$	$\mathbf{P_{05}}$	$\boldsymbol{\mu}$	$\mathbf{P_{95}}$	$\mathbf{P_{05}}$	$\boldsymbol{\mu}$	$\mathbf{P_{95}}$
skargi usług bankowych	F_1	+1,42%	+19,72%	+39,81%	-12,95%	+8,34%	+30,21%	-28,93%	-3,89%	+21,32%
atrybuty produktów *eBay*	F_1	+21,67%	+64,69%	+107,28%	-29,03%	+18,08%	+65,85%	-46,75%	+2,23%	+49,55%
waga produktów *eBay*	MAE	-20,56%	-12,83%	-5,65%	-22,31%	-6,75%	+8,75%	-8,70%	+2,85%	+14,38%
wydźwięk opinii o hotelach	MAE	-7,60%	+0,33%	+8,36%	-17,46%	-6,30%	+4,96%	+6,99%	+17,07%	+27,26%
jednostki nazwane	F_1	+3,31%	+30,87%	+57,05%	-27,53%	+0,76%	+21,75%	-41,89%	-0,16%	+41,63%
wyrazy bliskoznaczne	*dokładność*	-0,97%	+6,23%	+13,24%	-3,93%	+4,68%	+13,50%	-5,26%	+1,84%	+8,81%

Podsumowanie analizy hipotez

Analiza wykonana w celu weryfikacji hipotezy H_1 wykazała istotny pozytywny wpływ informacji zwrotnej na jakość pozyskiwanych danych dla czterech z sześciu badanych zbiorów danych (zob. Tabela 3.10). W przypadku zbioru „wydźwięk opinii o hotelach" oraz „wyrazy bliskoznaczne" informacja zwrotna nie miała znaczącego wpływu jakość danych. Warto zauważyć, że w przypadku obu wspomnianych zbiorów danych zadanie odwoływało się do wiedzy zdroworozsądkowej anotatorów. Natomiast najsilniejszy efekt widoczny był dla zadań, które wymagały od anotatora przyswojenia wiedzy domenowej, czyli dla zbiorów: „atrybuty produktów *eBay*" oraz „jednostki nazwane". Można przypuszczać, że jest to jeden z czynników decydujących o skuteczności informacji zwrotnej. Siła efektu przekazywanej informacji zwrotnej różniła się dla każdego z badanych zbiorów. Ostateczne zestawienie zmiany jakości dla każdego z badanych zbiorów danych we wszystkich trzech wariantach zostało zamieszczone w Tabeli 3.10. W związku, z tym że dla dwóch badanych zbiorów wpływ informacji zwrotnej nie był zauważalny, to przed zastosowaniem tego mechanizmu w zadaniu crowdsourcingowym zalecane jest przeprowadzenie testów w celu weryfikacji skuteczności tego podejścia dla oznaczanego zbioru danych[29].

Wyniki analizy przeprowadzonej w celu weryfikacji hipotezy H_2 ujawniły znaczące różnice w jakości danych pozyskiwanych w różnych wariantach eksperymentu. Efekt ten widoczny był dla wszystkich badanych zbiorów danych. W przypadku większości badanych zbiorów obserwowana różnica jakości była mniejsza podczas porównywania wariantów o zbliżonej jakości informacji zwrotnej ($W_1 - W_2$ oraz $W_2 - W_3$) niż w przypadku porównywania wariantów skrajnych ($W_1 - W_3$). W przypadku zbiorów, dla których informacja zwrotna dała pozytywny efekt, obniżenie jakości informacji zwrotnej skutkowało obniżeniem jakości pozyskiwanych danych. Mimo że najkorzystniejszy efekt obserwowany był dla informacji zwrotnej o najwyższej jakości (wariant W_1), to pozytywny efekt wprowadzenia tego mechanizmu widoczny był również w przypadku informacji zwrotnej o obniżonej jakości (wariant W_2). Oznacza to, że zastosowanie informacji zwrotnej w celu podwyższenia jakości danych pozyskiwanych w procesie crowdsourcingu możliwe jest również w przypadku, gdy nie pochodzi ona ze zbioru referencyjnego lub nie jest tworzona przez grupę ekspertów, a jest ona np. wygenerowana w sposób automatyczny przez wybrany algorytm.

[29]Przykładowo, możliwe jest przeprowadzenie testu porównawczego (tzw. „testu A/B") w analogiczny wariantów eksperymentu opisanych w ramach niniejszej rozprawy. W teście tego typu system crowdsourcingowy udostępniałby interfejs w dwóch wariantach: z informacją zwrotną oraz bez.

Podsumowanie analizy pytań badawczych

Analiza wykonana dla pytania badawczego P_1 potwierdziła, że pozytywny efekt informacji zwrotnej zachodzi nie tylko dla procesu długoterminowego (rozumianego jako uczenie się anotatorów), ale również w procesie natychmiastowym. Wyraźny, pozytywny efekt natychmiastowy widoczny był w 4 z 6 analizowanych zbiorów (zob. Tabela 3.10). Oznacza to, że w przypadku obecności informacji zwrotnej jakość oznaczanych danych wzrastała już po oznaczeniu kilku pierwszych anotacji. Najbardziej wyraźny efekt długoterminowy zauważony został dla zadań, które wymagały od anotatorów przyswojenia najbardziej złożonej wiedzy domenowej: „atrybuty produktów *eBay*", „jednostki nazwane" oraz „waga produktów *eBay*". Natomiast w przypadku zbioru „wydźwięk opinii o hotelach" zauważony został wyraźny ujemny efekt długotrwały, który minimalizował wpływ pozytywnego efektu natychmiastowego.

Analiza pytania P_2 nie wykazała systematycznego wpływu przekazywania informacji zwrotnej na czas wykonywania zadań. Wyniki różniły się dla każdego ze zbiorów danych, ale dla żadnego ze zbiorów obserwowany efekt nie był duży. Największy (umiarkowany) efekt obserwowany został dla zbioru „atrybuty produktów *eBay*", w którym przekazywanie informacji zwrotnej spowodowało pogorszenie tempa pracy anotatorów. Wyniki te nie dały podstawy, by stwierdzić, że informacja zwrotna wpływa na szybkość pracy w procesie crowdsourcingu.

W ramach analizy pytania P_3 niewykazane zostały znaczące różnice w liczbie wykonanych zadań przez osoby otrzymujące informację zwrotną w stosunku do wariantu kontrolnego. Obserwowany efekt jest niski bądź bliski zeru. Kierunek efektu różni się też pomiędzy zbiorami danych: dla części efekt jest pozytywny, a dla części – negatywny. Wyniki te nie dały jednoznacznej odpowiedzi na pytanie, czy informacja zwrotna wpływa na motywację anotatorów w procesie crowdsourcingu.

ROZDZIAŁ 4

Model *Dynamicznej Informacji Zwrotnej*

Wyniki eksperymentu opisanego w Rozdziale 3 wykazały istotny (dla czterech z sześciu badanych zbiorów), pozytywny wpływ informacji zwrotnej na jakość danych pozyskiwanych w procesie crowdsourcingu (hipoteza H_1, zob. Paragraf 3.5.5). Omówione wyniki eksperymentu potwierdzają również zależność pomiędzy jakością informacji zwrotnej a jakością pozyskanych danych (hipoteza H_2, zob. Paragraf 3.5.5). Ponieważ wyższa jakość informacji zwrotnej bezpośrednio przekłada się na wzrost jakości danych, zasadne jest opracowanie dedykowanych rozwiązań poprawiających jakość informacji zwrotnej.

W ramach niniejszego rozdziału zaprezentowany został autorski model *Dynamicznej Informacji Zwrotnej* (*DIZ*). Model ten wykorzystywał komponenty używane w procesie nauczania maszynowego (takie jak model ucznia; zob. Rozdział 2.2) oraz aktualne oznaczenia anotatora w celu wygenerowania informacji zwrotnej.

W pierwszej części rozdziału przedstawiony został ogólny opis działania modelu *DIZ* oraz jego proces trenowania. Następnie omówiony został eksperyment, który przeprowadziłem w celu porównania skuteczności modelu *DIZ*, z modelem, który nie wykorzystuje aktualnych oznaczeń anotatora do wygenerowania informacji zwrotnej. Działanie modelu przetestowane zostało dla danych rzeczywistych, a także danych symulowanych (powstałych w procesie modyfikacji danych rzeczywistych).

4.1. Model *Dynamicznej Informacji Zwrotnej*

W tej części niniejszego rozdziału opisana została definicja modelu *Dynamicznej Informacji Zwrotnej*, jego podstawowe założenia, a także zastosowanie modelu w procesie nauczania maszynowego.

4.1.1. Budowa modelu *Dynamicznej Informacji Zwrotnej*

Rozważmy problem oznaczania zbioru danych $\mathcal{D} = \{x_1, \cdots, x_n\}$ przy pomocy metody crowdsourcing, w ramach której dodatkowo zaimplementowany został proces nauczania maszynowego. Prawidłowe rozwiązania dla mikro-zadań należących do zbioru $\mathcal{D}$ danych zdefiniowane są przez nieznaną funkcję docelową f, która dla treści mikro-zadania x_i zwraca wartość prawidłowego rozwiązania y.

W ramach procesu nauczania maszynowego informacja zwrotna przekazywana jest przez model nauczyciela, który odpowiedzialny jest za przygotowanie sygnałów nauczających (zob. Paragraf 2.1.1). Rozważmy sytuację, w której treść informacji zwrotnej generowana jest w sposób automatyczny przez model M_g i przekazywana synchronicznie (zob. klasyfikacja informacji zwrotnej, Paragraf 1.4.1). Model M_g tworzony jest przy pomocy wybranego algorytmu uczenia maszynowego oraz dostępnego zbioru przykładów uczących $\mathcal{D}_u$. Na początku procesu oznaczania danych zbiór ten składa się z elementów pochodzących z dostępnego zbioru referencyjnego. W sytuacji, gdy zebrana została już część oznaczeń anotatorów, zbiór $\mathcal{D}_u$ może zostać rozszerzony o dane powstałe po zagregowaniu zbioru zebranych dotychczas anotacji $\mathcal{D}_a$ (np. przy pomocy algorytmu głosowania większościowego, zob. Paragraf 2.3.2)[1].

W trakcie procesu crowdsourcingu, po oznaczeniu mikro-zadania x_i, anotator j przesyła swoje rozwiązanie y_i^j. Następnie model M_g używany jest do wygenerowania informacji zwrotnej:

$$y_i^* = M_g(x_i; \theta_g), \tag{4.1}$$

gdzie θ_g to parametry modelu M_g.

Wartość przekazywanej informacji zwrotnej y_i^* jest taka sama bez względu na wartość anotacji y_i^j. W związku z tym, że estymacje modelu M_g są jednakowe dla wszystkich anotatorów bez względu na ich anotacje, z tego powodu model ten nazywany będzie „modelem globalnym". Warto zauważyć, że w sytuacji, gdy jakość anotatora j przewyższa jakość modelu globalnego M_g, przekazywana informacja zwrotna może negatywnie wpływać na proces nauczania (poprawy stanu wiedzy anotatora j).

W ramach niniejszej rozprawy zaproponowałem nowe rozwiązanie dla powyższego problemu – Model *Dynamicznej Informacji Zwrotnej* (*DIZ*) M_{DIZ}. Główne założenia modelu zostały opisane w Definicji 10:

Definicja 10 (Model Dynamicznej Informacji Zwrotnej)

Niech:

X *– zbiór treści mikro-zadań,*

Y *– zbiór możliwych wartości wyjściowych.*

[1] Podejście to opisane zostało np. w ramach pracy Sheng [2011].

Rozważamy mikro-zadanie $x_i \in X$

$y_i^j \in Y$ *– anotacja stworzona przez anotatora* j *dla mikro-zadania* x_i,

$M_u^j \in \mathbb{R}^n$ *– model ucznia dla anotatora* j, *określone w formie* n*-wymiarowego wektora,*

$y_i^* = M_g(x_i); y_i^* \in Y$ *– anotacja stworzona przez model globalny* M_g.

Model Dynamicznej Informacji Zwrotnej (DIZ) zdefiniowany jest jako funkcja

$$M_{DIZ} : X \times Y \times Y \times \mathbb{R}^n \rightarrow Y,$$

która oblicza wartość $\hat{y}_i$ *dla czterech zmiennych: treści mikro-zadania* x_i, *anotacji* y_i^*, *anotacji* y_i^j, *modelu ucznia* M_u^j *oraz parametrów modelu* θ_{DIZ}. *Funkcja ta ma postać:*

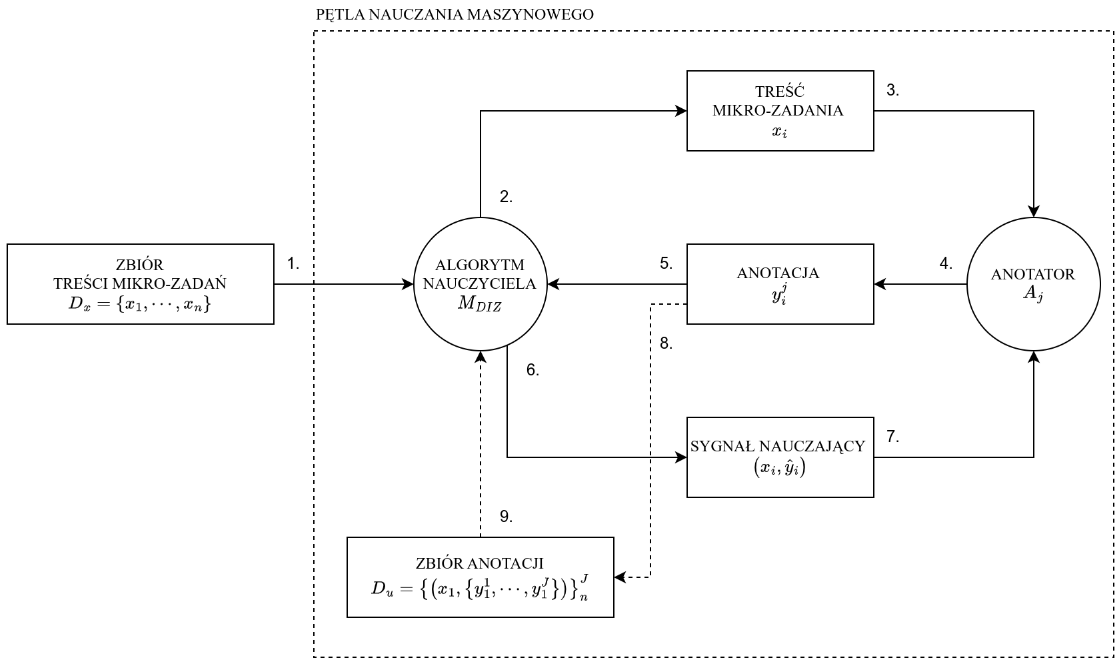
$$\hat{y}_i = M_{DIZ}(x_i, y_i^*, y_i^j, M_u^j; \theta_{DIZ}). \tag{4.2}$$

W modelu M_{DIZ} informacja zwrotna generowana jest poprzez połączenie anotacji wygenerowanej przez model globalny M_g, anotacji y_i^j stworzonej przez anotatora j oraz modelu ucznia M_u^j reprezentującego jego obecny stan wiedzy. W sytuacji, gdy model M_{DIZ} wykryje, że anotacja y_i^j jest lepsza niż ta pochodząca z modelu M_g, to odpowiedź anotatora zostanie przekazana w formie informacji zwrotnej, w przeciwnym wypadku, użyje anotacji wygenerowanej przez model globalny. W ten sposób informacja zwrotna generowana przez model M_{DIZ} może korzystnie wpływać na proces nauczania anotatorów o gorszej (niż model M_g) jakości oznaczeń oraz nie pogarszać skuteczności nauczania anotatorów o wyższej jakości oznaczeń.

4.1.2. Proces działania modelu *Dynamicznej Informacji Zwrotnej*

Model *DIZ* nie zakłóca podstawowej pętli nauczania i może on zostać z łatwością zaimplementowany w procesie nauczania maszynowego. Proces nauczania maszynowego przy użyciu modelu nauczyciela implementującego model *DIZ* przedstawiony został w Procedurze 12 oraz w formie graficznej na Diagramie 4.1[2].

[2] Przedstawiony tu opis procesu nauczania maszynowego skupia się przede wszystkim na wymianie danych pomiędzy komponentami procesu, oraz na udziale modelu *DIZ* w tym procesie. Opis procesu nauczania maszynowego zawarty w Rozdziale 2 (zob. Procedura 4) opisywał sposób tworzenia modelu ucznia. W swoich podstawowych założeniach obie procedury opisują ten sam proces.



Rysunek 4.1: Diagram procesu działania modelu *DIZ*

Procedura 12: Process działania modelu *DIZ* w procesie nauczania maszynowego

Niech :

$\mathcal{D}_x = \{x_1, \cdots, x_n\}$ – zbiór treści mikro-zadań, gdzie $x_i \in X$,

M_{DIZ} – algorytm nauczyciela implementujący model DIZ,

$A = \{A_1, \cdots, A_J\}$ – zbiór wszystkich anotatorów,

$\mathcal{D}_a = \emptyset$ – zbiór anotacji.

Kroki :

1 Algorytm nauczyciela implementujący M_{DIZ} przyjmuje pełen zbiór treści mikro-zadań.

2 Algorytm M_{DIZ} wybiera treść mikro-zadania x_i.

3 Algorytm M_{DIZ} przekazuje treść mikro-zadania x_i do anotatora A_j.

4 Anotator A_j tworzy anotację y_i^j dla mikro-zadania x_i.

5 Anotacja y_i^j przekazywana jest do algorytmu M_{DIZ}.

6 Algorytm M_{DIZ} generuję informację zwrotną i przekazuje ją w formie sygnału nauczającego $(x_i, \hat{y}_i)$,

7 Anotator A_j aktualizuje swój obecny stan wiedzy na podstawie przekazanego sygnału nauczającego.

8 Anotacja y_i^j dodawana jest do zbioru anotacji $\mathcal{D}_a$.

9 Algorytm nauczyciela aktualizuje model M_{DIZ} na podstawie aktualnego zbioru anotacji $\mathcal{D}_a$.

Kroki od 2 do 9 powtarzane są aż do oznaczenia całego zbioru $\mathcal{D}_x$.

Wyjście:

$\mathcal{D}_u = \{(x_i, \{y_i^1, \cdots, y_i^J\})\}_{i=1}^n$ – zbiór zebranych anotacji.

4.2. Badanie skuteczności modelu *Dynamicznej Informacji Zwrotnej*

W tym podrozdziale omówiony został eksperyment przeprowadzony w celu zweryfikowania skuteczności modelu *DIZ*. W pierwszej części omówione zostały komponenty wykorzystane do implementacji modelu *DIZ* dla każdego z sześciu badanych zbiorów danych. Następnie przedstawiony został ogólny przebieg procesu, który został użyty do trenowania i testowania modelu *DIZ*. W ostatniej części niniejszego rozdziału przedstawione zostały wyniki eksperymentu oraz omówione wyciągnięte wnioski.

4.2.1. Problem badawczy

Celem eksperymentu było zweryfikowanie, czy zaimplementowanie modelu *DIZ* w procesie crowdsourcingu wpłynie pozytywnie na jakości generowanej informacji zwrotnej. W eksperymencie model *DIZ* został porównany z modelem referencyjnym, który nie wykorzystuje aktualnych oznaczeń anotatora do wygenerowania informacji zwrotnej. Eksperyment przeprowadzony został na zbiorach danych, które zostały również wykorzystane do wcześniej wykonanego eksperymentu omawianego w Rozdziale 3. W ramach analizy wyników eksperymentu zweryfikowane zostały poniższe pytania badawcze:

P_1: Czy zastosowanie modelu *DIZ* pozwala na wygenerowanie informacji zwrotnej o wyższej jakości w porównaniu do modelu referencyjnego?

P_2: Czy jakość oznaczeń anotatorów wpływa na jakość informacji zwrotnej generowanej przez model *DIZ*?

Dla każdego zbioru danych rozpatrzone zostały pytania pomocnicze oznaczone odpowiednio P_1^i oraz P_2^i, gdzie i to numer porządkowy danego zbioru danych.

4.2.2. Warianty eksperymentu

Eksperyment został wykonany w dwóch różnych wariantach: wariant bazowy, w którym do generowania informacji zwrotnej użyty został model *DIZ* oraz wariant kontrolny. W ramach wariantu kontrolnego model *DIZ* zastąpiony został modelem referencyjnym, w którym oznaczenia użytkowników y_i^j oraz model ucznia M_g^j nie są używane.

- Użycie modelu *DIZ* (W_{DIZ})
  Informacja zwrotna generowana jest przy użyciu modelu *DIZ*.

- Wariant kontrolny (W_C)
  Informacja zwrotna generowana jest tylko przy pomocy modelu referencyjnego.

4.2.3. Implementacja modeli *Dynamicznej Informacji Zwrotnej*

W ramach przeprowadzonego eksperymentu modele *DIZ* zostały zaimplementowane dla każdego z sześciu badanych zbiorów danych. Modele te zostały zaimplementowaniu przy użyciu sieci neuronowych. W związku z tym, że zbiory danych wykorzystane w eksperymencie dotyczą różnych typów zadań i różnią się formatem danych, dla każdego z nich stworzona została osobna sieć neuronowa. Każdy z modeli został stworzony przy pomocy tego samego zbioru elementów. Poniżej opisane zostały trzy grupy elementów użytych do zbudowania modeli *DIZ* wykorzystanych w opisywanym eksperymencie: *reprezentacja danych*, *reprezentacja modeli ucznia* oraz *budowa sieci neuronowej*.

Reprezentacja danych

Pierwszym elementem była odpowiednia reprezentacja danych. Treść mikro-zadań oraz anotacji zostały dostosowane do formy wektorów, które mogą zostać użyte przez modele sieci neuronowych. W eksperymencie zostały użyte dwa typy wektorowej reprezentacji danych:

– **Wektor *one-hot***

  Wektor *one-hot* stanowi jeden ze sposobów wektorowej reprezentacji danych kategorycznych. W przypadku gdy K jest zbiorem możliwych kategorii, to pojedyncza kategoria może być reprezentowana w postaci $|K|$-wymiarowego wektora zawierającego dokładnie jedną wartość 1 w pozycji odpowiadającej indeksowi danej kategorii, a wszystkie pozostałe wartości wektora są równe 0 [Jurafsky & Martin, 2021, s. 148].

  Przykładowo, rozważmy zbiór $K = \{a, b, c\}$. Kategorie z tego zbioru mają następująca reprezentację w formie wektora *one-hot*:

$$\begin{aligned} V_a &= [1, 0, 0] \\ V_b &= [0, 1, 0] \\ V_c &= [0, 0, 1] \end{aligned} \tag{4.3}$$

– **Wektor ważony metodą *TF-IDF***

  Wektor ważony metodą *TF-IDF* to sposób wektorowej reprezentacji danych tekstowych stanowiący modyfikację metody *one-hot*[3]. Wektory *TF-IDF* dla dokumentów tekstowych są tworzone przez przypisanie wagi każdemu słowu w dokumencie, proporcjonalnej do częstości występowania słowa w dokumencie tf, ale odwrotnie

[3]Dane tekstowe mogą być traktowane jako dane kategoryczne, w którym każde słowo reprezentowane jest jako osobna kategoria, a zbiór wszystkich kategorii stanowi słownik możliwych słów.

proporcjonalnej do częstości występowania słowa we wszystkich dokumentach idf. Waga $w_{t,d}$ dla słowa t w dokumencie d liczona jest według formuły:

$$w_{t,d} = tf_{t,d} \times idf_t, \tag{4.4}$$

gdzie

$$\begin{aligned} df_{t,d} &= log_{10}(count(t,d)+1) \\ idf_t &= log_{10}(\frac{N}{df_t}) \end{aligned}, \tag{4.5}$$

funkcja $count(t,d)$ określa liczbę wystąpień termu t w dokumencie d, a df_t to liczba dokumentów zawierających term t.

W ten sposób słowa, które występują często w jednym dokumencie, ale rzadko w innych, otrzymują wysoką wagę, a słowa, które występują w każdym dokumencie, otrzymują niską wagę. Dokładny opis tej metody został opisany m.in. w ramach pracy [Jurafsky & Martin, 2021, s. 113-116].

Inną, bardziej rozbudowaną metodą wektorowej reprezentacji tekstu jest *word embedding*. Jednak metoda ta nie jest skuteczna w przypadku przetwarzania danych tekstowych w domenie *e-commerce*[4]. Ponieważ w badaniu analizowane były dwa zbiory związane z domeną *e-commerce*: „atrybuty produktów *eBay*" oraz „waga produktów *eBay*", zdecydowałem się na użycie prostszej, lecz bardziej uniwersalnej metody reprezentacji danych tekstowych: wektory ważone metodą *TF-IDF*.

Reprezentacja modeli ucznia

Ponieważ mikro-zadania we wszystkich badanych zbiorów danych związane były z interpretacją języka naturalnego, zdecydowałem się na użycie prostszej wersji modeli reprezentacji wiedzy ucznia – modeli jakości ucznia[5]. W badaniu użyte zostały dwa typy modeli jakości ucznia. Dla zadań związanych z klasyfikacją użyta została *tablica pomyłek*, a w przypadku zadań związanych z estymacją wartości liczonych użyty został zbiór metryk:

– **Tablica pomyłek**

Tablica pomyłek (zob. Paragraf 2.2.2) została przygotowana poprzez porównanie

[4] Wniosek ten został wyciągnięty na podstawie własnego doświadczenia pracy jako analityk danych w firmie *Webinterpret*. Istnieją również badania raportujące ten problem np. praca Mudgal et al. [2018].

[5] Podejście to sprawdza się lepiej w przypadku danych reprezentowanych przez wektory o dużym wymiarze, zob. Paragraf 2.2.

odpowiedzi anotatora udzielonymi dla zbioru treningowego z wartościami referencyjnymi. Następnie tablica pomyłek została przetransformowana do postaci jednowymiarowego wektora, w którym wszystkie wiersze macierzy zostały skonkatenowane jeden za drugim:

$$M_u = [Q_{1,1}, \cdots, Q_{1,n}, Q_{2,1}, \cdots, Q_{n,n}]. \tag{4.6}$$

– **Zbiór metryk**

Zbiór metryk użytych do zdefiniowania modelu jakości ucznia zawiera metryki opisujące błąd przesunięcia μ i odchylenia standardowego σ^2 (zob. Paragraf 2.2.2). Dodatkowo zbiór został rozszerzony o dwie metryki używane do oceny jakości estymacji wartości liczbowych: MAE oraz MSE (zob. Paragraf 2.2.2). Metryki zostały obliczone poprzez porównanie estymacji anotatora dla zbioru treningowego z wartościami referencyjnymi. Zbiór metryk został przedstawiony w formie wektora:

$$M_u = [\mu, \sigma^2, MAE, MSE]. \tag{4.7}$$

Budowa sieci neuronowej

Sieci neuronowe (ang. *neural networks*) stanowią grupę algorytmów uczenia maszynowego, które zainspirowane zostały przez biologiczny proces uczenia zachodzący w mózgu [Goodfellow et al., 2016, s. 13]. Dokładniejszy opis działania sieci neuronowych można znaleźć np. w [Goodfellow et al., 2016, s. 168-177]. W ramach niniejszej rozprawy przedstawiony został uproszczony opis wybranych elementów sieci neuronowej, które zostały wykorzystane do implementacji w ramach opisywanego eksperymentu modeli DIZ.

– **Perceptron wielowarstwowy (ang. *multilayer perceptron*)**

Perceptron wielowarstwowy stanowi podstawowy model sieci neuronowych. Model ten składa się z co najmniej trzech warstw neuronów: wejściowej, ukrytej i wyjściowej. Sygnały wejściowe są wprowadzane do sieci przez neurony z warstwy wejściowej, która przekazuje sygnały do warstwy ukrytej. W warstwie ukrytej sygnały są przetwarzane i przekazywane do kolejnej warstwy ukrytej bądź warstwy wyjściowej. Działanie pojedynczej warstwy ukrytej perceptronu wielowarstwowego zdefiniowane jest według następującego wzoru:

$$f(x; w,t) = g(x^\top w + b), \tag{4.8}$$

gdzie g to wybrana funkcja aktywacji, w to macierz parametrów, a b to wektor przesunięcia (ang. *bias*). W ramach opisywanego eksperymentu użyte zostały dwie funkcje aktywacji: *ReLU* oraz *softmax*.

– **Funkcja aktywacji *ReLU***

  Funkcja aktywacji *ReLU* (ang. *Rectified Linear Unit*) stanowi jedną z podstawowych funkcji aktywacji. Funkcja ta zwraca wartość wejściową, jeśli ta jest dodatnia, lub zero w przeciwnym wypadku:

$$ReLU(x) = max(0, x) \tag{4.9}$$

  Funkcja *ReLU* jest powszechnie stosowana dla warstw ukrytych sieci neuronowej. Może być również używana w warstwach wyjściowych dla problemów związanych z estymacją wartości liczbowych.

– **Funkcja aktywacji *softmax***

  Funkcja aktywacji *softmax* używana jest w celu przekształcenia wektora wartości na wektor prawdopodobieństw. Wartość funkcji *softmax* dla i-tej wartości wektora obliczana jest według następującego wzoru:

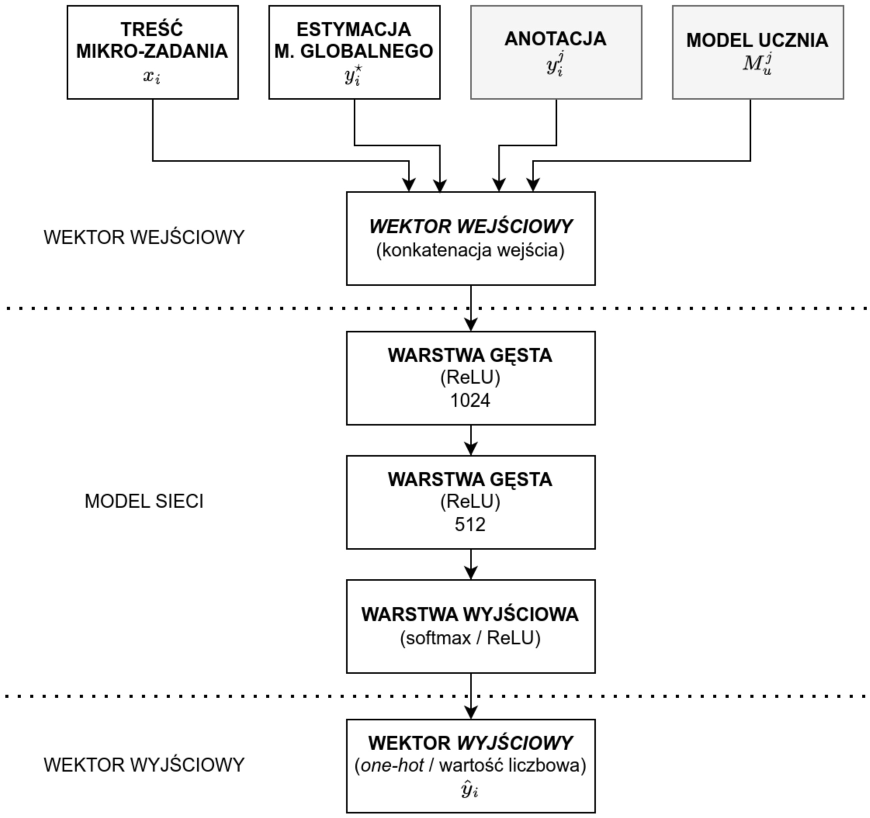
$$softmax(x)_i = \frac{exp(x_i)}{\sum_{j=1}^{n} exp(x_j)} \tag{4.10}$$

  Wynikowy wektor zawiera wartości z zakresu od 0 do 1, które sumują się do 1, co umożliwia interpretację wyników jako prawdopodobieństw. Funkcja *softmax* jest stosowana jako funkcja aktywacji w warstwie wyjściowej sieci neuronowej stworzonej dla problemów klasyfikacji wieloklasowej.

Architektura implementacji modeli *DIZ*

W celu ujednolicenia eksperymentów i ułatwienia interpretacji wyników zaimplementowane modele korzystały jedynie z podstawowych elementów sieci neuronowej, ponieważ celem eksperymentu było zweryfikowanie korzyści z użycia aktualnej anotacji do wygenerowania informacji zwrotnej, a nie stworzenie jak najlepszych modeli rozwiązujących dany problem.

Wszystkie modele *DIZ* zaimplementowane w ramach eksperymentu, jak i odpowiadające im modele referencyjne stosowane w wariantach kontrolnych, zostały stworzone na podstawie tej samej, ogólnej architektury sieci neuronowej (zob. Rysunek 4.2). Budowa tej architektury podzielona została na trzy części:



Rysunek 4.2: Ogólna architektura sieci neuronowej użytej do implementacji modelu *DIZ* oraz modelu referencyjnego (pomijane są bloki szare)

1. **Wektor wejściowy**

   Wektor wejściowy zawierał dane, które przekazywane są do sieci neuronowej. Dla każdego mikro-zadania x_i, wektor wejściowy stworzony został poprzez konkatenację elementów:

   (a) **Treść mikro-zadania** x_i – wektor zawierający treść mikro-zadania. Dokładny format wektora różnił się dla każdego zbioru danych.

   (b) **Estymacja modelu globalnego** y_i^* – wektor zawierający wartości wyjściowe wygenerowane przez model globalny M_g. W przypadku zadań klasyfikacji był to wektor w formacie *one-hot*, a w przypadku zadań związanych z estymacją wartości liczbowych był to wektor jednoelementowy, zawierający liczbę rzeczywistą.

   (c) **Anotacja** y_i^j – wektor zawierający anotację stworzoną przez anotatora j. Format wektora był taki sam jak w przypadku wektora y_i^*. Element ten nie był używany w przypadku implementacji modelu referencyjnego.

   (d) **Model ucznia** M_u^j – wektor zawierający wartości parametrów modelu ucznia. W przypadku zadań klasyfikacji był to wektor zawierający wartości z tablicy pomyłek, a w przypadku zadań związanych z estymacją wartości liczbowych był to wektor zawierający zbiór metryk. Element ten nie był używany w przypadku implementacji modelu referencyjnego.

2. **Model sieci**
   Model sieci składał się z trzech warstw: dwóch warstw gęstych (zawierających odpowiednio 1024 i 512 neuronów) z funkcją aktywacją $ReLU$ oraz jednej warstwy wyjściowej, której funkcja aktywacji zależała od typu zadania. W przypadku zadań klasyfikacji była to funkcja *softmax*, a w przypadku zadań związanych z estymacją wartości liczbowych była to funkcja $ReLU$.

3. **Wektor wyjściowy** $\hat{y}_i$
   Wektor wyjściowy zawierał estymowaną odpowiedź modelu. Format wektora wyjściowego był taki sam jak w przypadku wektorów y_i^* oraz y_i^j.

4.2.4. Przebieg eksperymentu

W ramach przeprowadzonego eksperymentu porównana została jakość informacji zwrotnej wygenerowanej w wariancie W_{DIZ} oraz W_C. Dodatkowo w eksperymencie zostały zdefiniowane dwa parametry definiujące jego przebieg:

– **Liczba treningowych mikro-zadań** (S) – proporcja pomiędzy wielkością zbioru treningowego i zbioru testowego określona w formie procentowej wartości.

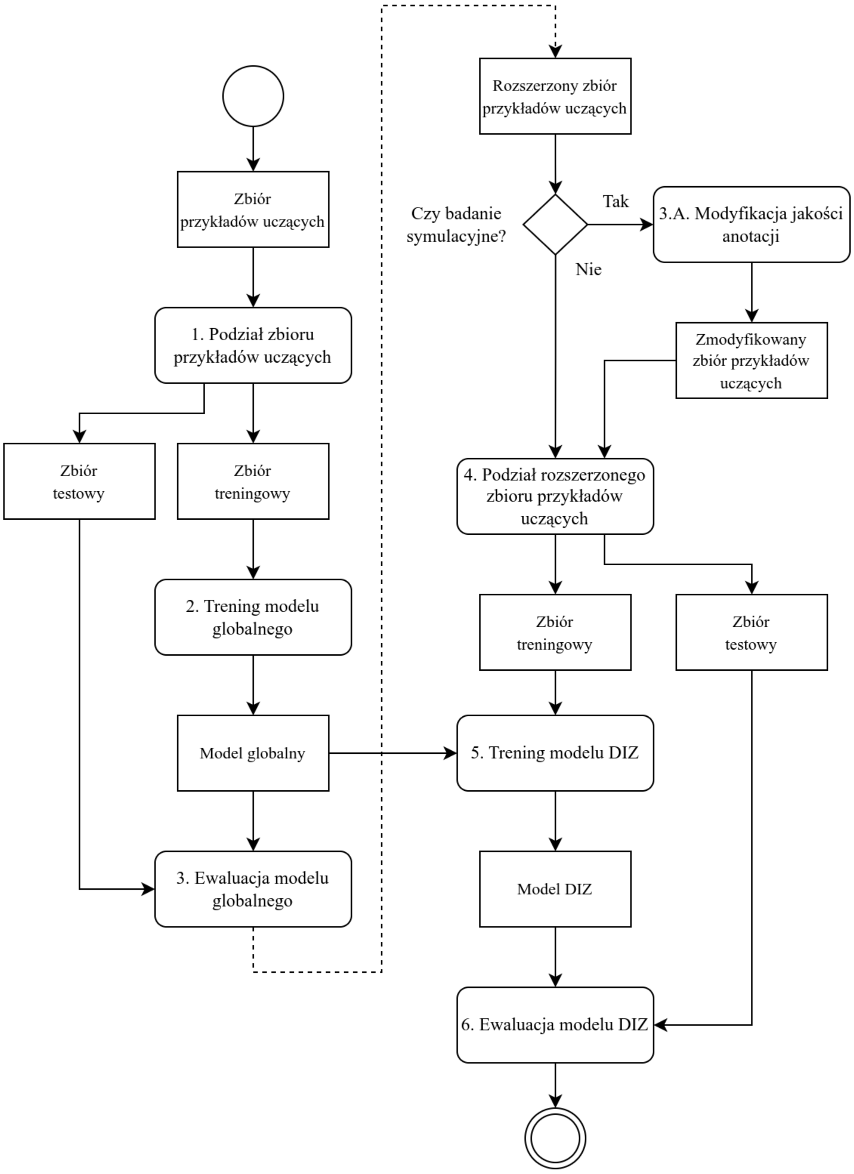
– **Jakość anotacji** (q) – średnia jakość anotacji w zbiorze przykładów uczących.

Eksperyment został powtórzony dla każdego z sześciu badanych zbiorów danych oraz dla wszystkich możliwych pary wartości dwóch parametrów, które definiowały konkretną konfigurację jednego przebiegu eksperymentu.

W celu uzyskania zbioru przykładów uczących o odpowiedniej jakości, w czasie każdego przebiegu eksperymentu, wejściowych zbiór danych został automatycznie zmodyfikowany tak, by jego jakość odpowiadała wartości określonej przez parametr q.

W celu zwiększenia rzetelności oszacowania wydajności modelu *DIZ* w eksperymencie użyta została metoda walidacji krzyżowej[6]. Dla każdej pary parametrów eksperyment został potworzony pięć razy dla wariantu bazowego W_{DIZ} oraz pięć razy dla wariantu kontrolnego W_C. W ramach każdego powtórzenia na nowo losowany był podział zbioru

[6] Metoda walidacji krzyżowej opisana została np. w pracy [Jurafsky & Martin, 2021, s. 70]



Rysunek 4.3: Diagram przebiegu procesu uczenia modelu *DIZ*

przykładów uczących na zbiór treningowy i testowy. Kroki pojedynczego przebiegu eksperymentu przedstawione zostały w formie graficznej na Rysunku 4.3, a także w formie listy kroków zamieszczonych poniżej:

1. **Podział zbioru przykładów uczących**

   Zbiór przykładów uczących $\mathcal{D}_u = \{(x_1, y_1), \cdots, (x_n, y_n)\}$ został podzielony na dwa podzbiory:

   – **zbiór treningowy** ($\mathcal{D}^{train}$) – zbiór używany podczas procesu trenowania modelu uczenia maszynowego,

– **zbiór testowy** ($\mathcal{D}^{test}$) – zbiór używany podczas ewaluacji jakości modelu uczenia maszynowego.

Wielkość obu zbiorów definiowana została przez wartość parametru S. Z każdej partii wybierane było S mikro-zadań (np. 30% całego zbioru), które użyte były do stworzenia zbioru treningowego $\mathcal{D}_u^{train}$. Pozostałe mikro-zadania trafiają do zbioru testowego $\mathcal{D}_u^{test}$.

2. **Trening modelu globalnego**

   Stworzony został model globalny M_g na podstawie zbioru treningowego $\mathcal{D}_u^{train}$.

3. **Ewaluacja modelu globalnego**

   Przeprowadzona została ewaluacja modelu globalnego M_g przy użyciu zbioru testowego $\mathcal{D}_u^{test}$ oraz wybranej metryki jakości.

4. **Modyfikacja jakości anotacji**

   Jakość anotacji została zmodyfikowana w sposób sztuczny tak, by średnia jakość całego zbioru anotacji osiągnęła wartość parametru q według wybranej metryki jakości.

5. **Podział rozszerzonego zbioru przykładów uczących**

   Określony został rozszerzony zbiór przykładów uczących:
   $\mathcal{D}_a = \{(\hat{x}_{1,1}, y_1), (\hat{x}_{1,m}, y_1), \cdots (\hat{x}_{n,m} y_n)\}$, gdzie $\hat{x}_{i,j}$ zdefiniowany został jako para składająca się z danych wejściowych dla i-tego mikro-zadania oraz anotacji j-tego anotatora dla tego mikro-zadania: $\hat{x}_{i,j} = (x_i, y_{i,j})$.

   Podział rozszerzonego zbioru $\mathcal{D}_a$ na zbiory $\mathcal{D}_a^{train}$ oraz $\mathcal{D}_a^{test}$ odbywał się w taki sposób, aby podział mikro-zadań w zbiorze $\mathcal{D}_a$ odpowiadał podziałowi w zbiorze $\mathcal{D}_u$.

6. **Trening modelu *DIZ***

   Stworzony został model Dynamicznej Informacji Zwrotnej M_{DIZ} na podstawie zbioru treningowego $\mathcal{D}_a^{train}$.

7. **Ewaluacja modelu *DIZ***

   Przeprowadzona została ewaluacja modelu M_{DIZ} przy użyciu zbioru testowego $\mathcal{D}_a^{test}$ oraz wybranej metryk jakości.

4.2.5. Wyniki eksperymentu

W tej części niniejszej rozprawy przedstawione zostały wyniki analizy przeprowadzonej w celu zweryfikowania skuteczności modelu *DIZ*. Ewaluacja skuteczności modelu została przeprowadzona poprzez porównanie wybranej metryki jakości obliczonej dla oznaczeń wygenerowanych przy pomocy modelu *DIZ* oraz tych, które zostały wygenerowane przy użyciu modelu referencyjnego. Proces ewaluacji przeprowadzony został dla różnych konfiguracji modelu *DIZ* i przebiegał zgodnie z procedurą opisaną w poprzedniej części rozprawy (zob. Paragraf 4.2.4). Wyniki porównania modeli przedstawione zostały w postaci różnicy metryki jakości uzyskanej dla obu modeli. Wyniki analizy zaprezentowane zostały dla dwóch typów danych:

– **dane empiryczne** – dane zebrane podczas eksperymentu użyte w niezmodyfikowanej formie,

– **dane symulacyjne** – dane wygenerowane poprzez modyfikację danych empirycznych w celu uzyskaniu danych o różnym poziomie jakości.

Celem zastosowania danych symulacyjnych było zbadanie, jak zmiana jakości anotacji przekłada się na jakość informacji zwrotnej generowanej przez model *DIZ*. Dokładny sposób wprowadzania modyfikacji zbioru danych różnił się dla każdego ze zbiorów danych użytych w eksperymencie. W sytuacji, gdy jakość danych danego zbioru miała zostać obniżona, stosowany był algorytm zniekształcenia anotacji. Algorytmy zniekształcania anotacji dla każdego z badanych zbiorów danych opisane zostały w Podrozdziale 3.4. W sytuacji, gdy jakość danych miała zostać zwiększona, użyty został iteracyjny algorytm, w którym w każdym kroku losowa anotacja zamieniana była na poprawną. Algorytm ten powtarzany był aż do momentu osiągnięcia określonego poziomu jakości.

Poniżej omówione zostały wyniki eksperymentu dla każdego z sześciu wykorzystanych zbiorów danych.

Zbiór: skargi usług bankowych

Zbiór „skargi usług bankowych” związany był z zadaniem klasyfikacji tekstu do jednej z pięciu zdefiniowanych kategorii. Dokładniejszy opis samego zadania został przedstawiony w Rozdziale 3 (zob. Paragraf 3.4.1). Jakość modeli stworzonych w ramach eksperymentu oceniana została według metryki F_1.

Na Rysunku 4.4 przedstawione zostało zestawienie procentowej różnicy jakości pomiędzy wariantami W_{DIZ} i W_C dla wszystkich przebiegów eksperymentu przeprowadzonych

dla tego zbioru danych. Na osi X umieszczona została wielkości zbioru treningowego, a na osi Y – jakość anotacji użytych w eksperymencie (zarówno w zbiorze treningowym, jak i testowym) wyrażona przy pomocy metryki F_1. Dokładna wielkość zbiorów treningowego oraz testowego (użytych do trenowania modelu globalnego) oraz zbiorów rozszerzonych, zawierających anotacje (użytych do trenowania modelu *DIZ* oraz modelu referencyjnego) przedstawiona została w Tabeli 4.1.

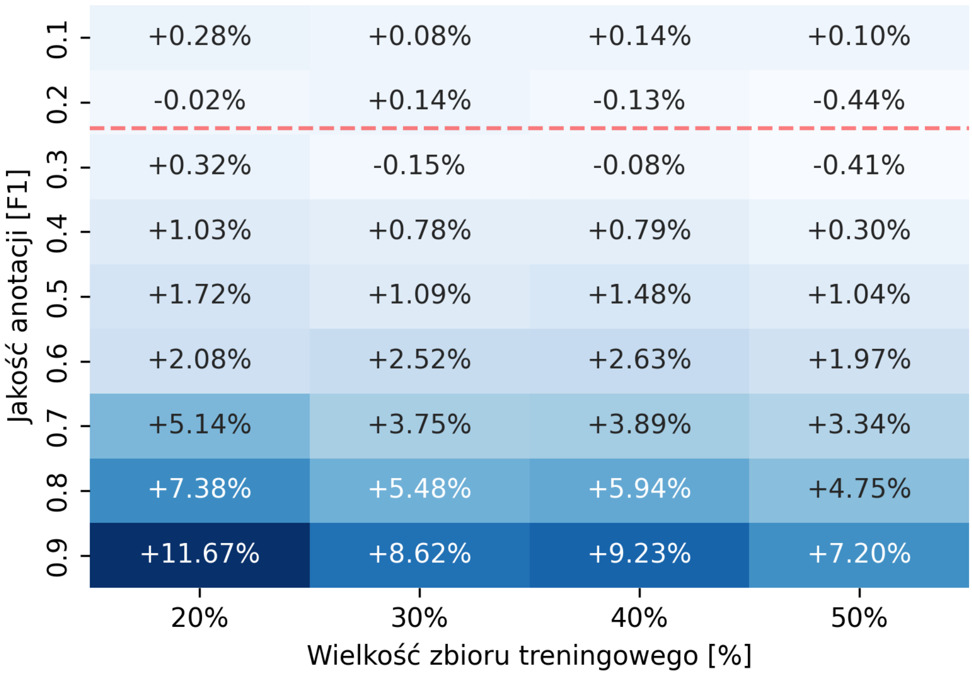
W ramach analizy wyników eksperymentów dokonałem weryfikacji dwóch pytań badawczych opisanych w Paragrafie 4.2.1:

P_1^1: Jakość dla zbioru danych empirycznych obserwowana została na poziomie $q = F_1 = 0{,}24$. Dla tej wartości nie zaobserwowano zauważalnej różnicy pomiędzy wariantami W_{DIZ} i W_C. W przypadku danych symulacyjnych obserwowana była wyraźna poprawa jakości w sytuacji, gdy parametr q był na poziomie 0,4 lub wyższym.

P_2^1: Obserwowana różnica pomiędzy W_{DIZ} i W_C była bezpośrednio zależna od obu parametrów (q i S) zdefiniowanych w ramach eksperymentu:

Wzrost jakości anotacji q miał pozytywny wpływ na jakość modelu *DIZ* (zwiększenie pozytywnej różnicy pomiędzy wariantami). Wpływ jakości anotacji na jakość modelu *DIZ* była nieliniowy – jakość modelu rosła szybciej niż jakość anotacji.

Zwiększenie liczby treningowych mikro-zadań (parametr S) przełożyło się na poprawę jakości modelu globalnego, a tym samym spadek różnicy pomiędzy wariantami. Najwyższa różnica jakości na poziomie +11,67% osiągnięta została dla $q = 0{,}9$ oraz $S = 20\%$. Zwiększenie wielkości zbioru treningowego do $S = 50\%$ spowodowało spadek różnicy do +7,20%. Podobna tendencja widoczna była dla wszystkich wartości parametru q.



Rysunek 4.4: Jakość modelu *DIZ* w zależności od wartości parametrów symulacji dla zbioru „skargi usług bankowych" (jakość anotacji ze zbioru empirycznego oznaczona czerwoną linią)

Tabela 4.1: *Wielkość zbioru danych użytego w symulacji modelu DIZ dla zbioru „skargi usług bankowych"*

	Mikro-zadania		**Anotacje**	
Wielkość zbioru treningowego [%]	**Treningowe**	**Testowe**	**Treningowe**	**Testowe**
20%	394	1564	1121	4441
30%	590	1368	1680	3882
40%	785	1173	2240	3322
50%	980	978	2787	2775

Zbiór: atrybuty produktów eBay

Zbiór „atrybuty produktów eBay" związany był z zadaniem oznaczania w tekście atrybutów należących do jednej z sześciu zdefiniowanych kategorii. Dokładniejszy opis samego zadania został przedstawiony w Rozdziale 3 (zob. Paragraf 3.4.2). Jakość modeli stworzonych w ramach eksperymentu oceniana została według metryki F_1.

Na Rysunku 4.5 przedstawione zostało zestawienie procentowej różnicy jakości pomiędzy wariantami W_{DIZ} i W_C dla wszystkich przebiegów eksperymentu przeprowadzonych dla tego zbioru danych. Na osi X umieszczona została wielkości zbioru treningowego, a na osi Y – jakość anotacji użytych w eksperymencie (zarówno w zbiorze treningowym, jak i testowym) wyrażona przy pomocy metryki F_1. Dokładna wielkość zbiorów treningowego oraz testowego (użytych do trenowania modelu globalnego) oraz zbiorów rozszerzonych, zawierających anotacje (użytych do trenowania modelu *DIZ* oraz modelu referencyjnego) przedstawiona została w Tabeli 4.2.

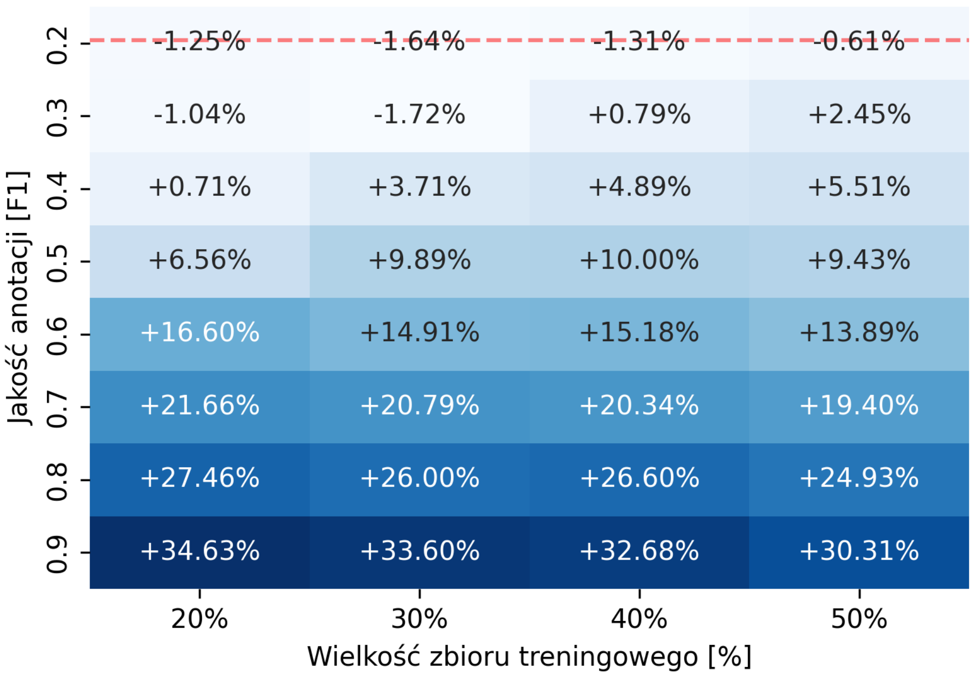
W ramach analizy wyników eksperymentów dokonałem weryfikacji dwóch pytań badawczych opisanych w Paragrafie 4.2.1:

P_1^2: Jakość dla zbioru danych empirycznych obserwowana została na poziomie $q = F_1 = 0{,}19$. Dla tej wartości zaobserwowano negatywny wpływ wprowadzenia modelu *DIZ* (wariant C_{DIZ}). W przypadku danych symulacyjnych obserwowana była wyraźna poprawa jakości w sytuacji, gdy parametr q był na poziomie 0,4 lub wyższym.

P_2^2: Obserwowana różnica pomiędzy W_{DIZ} i W_C była bezpośrednio zależna od obu parametrów (q i S) zdefiniowanych w ramach eksperymentu:

Wzrost jakości anotacji q miał pozytywny wpływ na jakość modelu *DIZ* (zwiększenie pozytywnej różnicy pomiędzy wariantami). Wpływ jakości anotacji na jakość modelu *DIZ* była nieliniowy – jakość modelu rosła szybciej niż jakość anotacji.

Zwiększenie liczby treningowych mikro-zadań (parametr S) przełożyło się na poprawę jakości modelu globalnego, a tym samym spadek różnicy pomiędzy wariantami. Najwyższa różnica jakości na poziomie +34,63% osiągnięta została dla $q = 0{,}9$ oraz $S = 20\%$. Zwiększenie wielkości zbioru treningowego do $S = 50\%$ spowodowało spadek różnicy do +30,31%. Podobna tendencja widoczna była dla wszystkich wartości parametru q.



Rysunek 4.5: Jakość modelu *DIZ* w zależności od wartości parametrów symulacji dla zbioru „atrybuty produktów eBay" (jakość anotacji ze zbioru empirycznego oznaczona czerwoną linią)

Tabela 4.2: *Wielkość zbioru danych użytego w symulacji modelu DIZ dla zbioru „atrybuty produktów eBay"*

	Mikro-zadania		**Anotacje**	
Wielkość zbioru treningowego [%]	**Treningowe**	**Testowe**	**Treningowe**	**Testowe**
20%	677	1811	1807	5025
30%	901	1587	2440	4392
40%	1060	1428	2894	3938
50%	1250	1238	3440	3392

Zbiór: waga produktów eBay

Zbiór „waga produktów eBay" związany był z zadaniem estymacji wagi produktu w gramach. Dokładniejszy opis samego zadania został przedstawiony w Rozdziale 3 (zob. Paragraf 3.4.3). Jakość modeli stworzonych w ramach eksperymentu oceniana została według metryki MAE. Metryka ta służy do opisania błędu estymacji, oznacza to, że im niższa wartość metryki, tym wyższa jakość modelu.

Na Rysunku 4.6 przedstawione zostało zestawienie procentowej różnicy jakości pomiędzy wariantami W_{DIZ} i W_C dla wszystkich przebiegów eksperymentu przeprowadzonych

dla tego zbioru danych. Na osi X umieszczona została wielkości zbioru treningowego, a na osi Y – jakość anotacji użytych w eksperymencie (zarówno w zbiorze treningowym, jak i testowym) wyrażona przy pomocy metryki MAE. Dokładna wielkość zbiorów treningowego oraz testowego (użytych do trenowania modelu globalnego) oraz zbiorów rozszerzonych, zawierających anotacje (użytych do trenowania modelu *DIZ* oraz modelu referencyjnego) przedstawiona została w Tabeli 4.3.

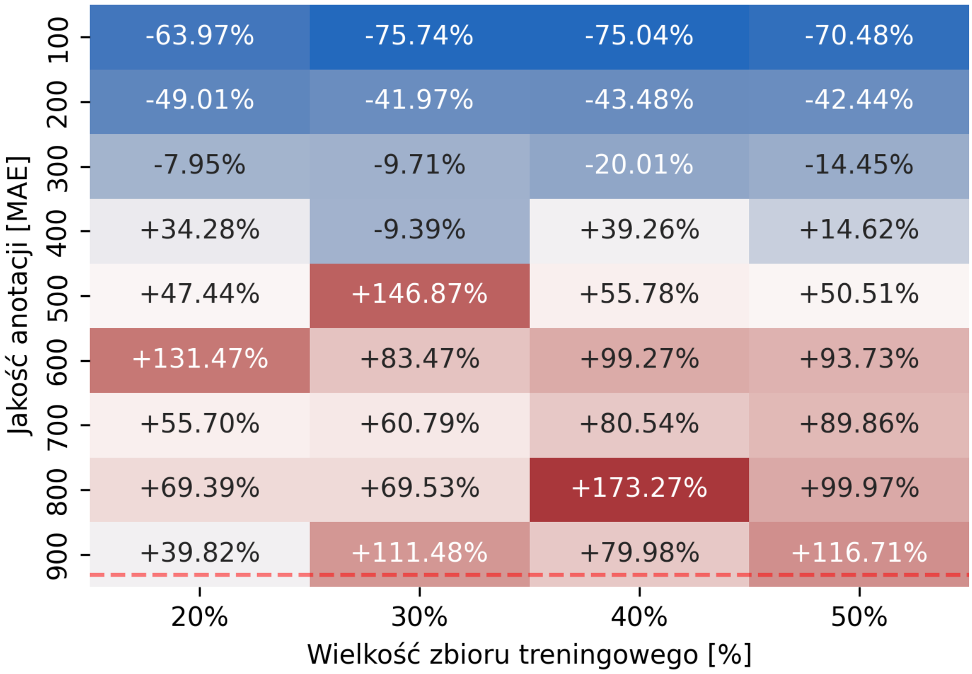
W ramach analizy wyników eksperymentów dokonałem weryfikacji dwóch pytań badawczych opisanych w Paragrafie 4.2.1:

P_1^3: Jakość dla zbioru danych empirycznych obserwowana została na poziomie $q = MAE = 931{,}22$. Dla tej wartości zaobserwowano negatywny wpływ wprowadzenia modelu *DIZ* (wariant C_{DIZ}). W przypadku danych symulacyjnych obserwowana była wyraźna poprawa jakości w sytuacji, gdy parametr q był na poziomie 300 lub niższa.

P_2^3: Obserwowana różnica pomiędzy W_{DIZ} i W_C była bezpośrednio zależna od obu parametrów (q i S) zdefiniowanych w ramach eksperymentu:

Wzrost jakości anotacji q miał pozytywny wpływ na jakość modelu *DIZ* (zwiększenie pozytywnej różnicy pomiędzy wariantami). Wpływ jakości anotacji na jakość modelu *DIZ* nie był tak stabilny (wzrost jakości nie był monotoniczny), jak w przypadku innych zbiorów danych i obserwowane było wiele wartości odstających (np. w przypadku $S = 20\%$ jakość modelu *DIZ* była znacząco gorsza dla $q = 600$ – 954,77, niż jakości dla $q = 700$ – 408,06).

Zwiększenie liczby treningowych mikro-zadań (parametr S) przełożyło się na poprawę jakości modelu globalnego, a tym samym spadek różnicy pomiędzy wariantami. Najwyższa różnica jakości na poziomie $-462{,}91$ osiągnięta została dla $q = 100$ oraz $S = 20\%$. Zwiększenie wielkości zbioru treningowego do $S = 50\%$ spowodowało spadek różnicy do $-319{,}15$. Podobna tendencja widoczna była dla wszystkich wartości parametru q.



Rysunek 4.6: Jakość modelu *DIZ* w zależności od wartości parametrów symulacji dla zbioru „waga produktów eBay" (jakość anotacji ze zbioru empirycznego oznaczona czerwoną linią)

Tabela 4.3: *Wielkość zbioru danych użytego w symulacji modelu DIZ dla zbioru „waga produktów eBay"*

	Mikro-zadania		**Anotacje**	
Wielkość zbioru treningowego [%]	**Treningowe**	**Testowe**	**Treningowe**	**Testowe**
20%	530	2120	1877	7510
30%	795	1855	2825	6562
40%	1060	1590	3761	5626
50%	1325	1325	4703	4684

Zbiór: wydźwięk opinii o hotelach

Zbiór „wydźwięk opinii o hotelach" związany był z analizą wydźwięku tekstu opinii o hotelach, w ramach której dokonywano wyboru wartości w skali od 1 do 5. Dokładniejszy opis samego zadania został przedstawiony w Rozdziale 3 (zob. Paragraf 3.4.4). Jakość modeli stworzonych w ramach eksperymentu oceniana została według metryki MAE. Metryka ta służy do opisania błędu estymacji, oznacza to, że im niższa wartość metryki, tym wyższa jakość modelu.

Na Rysunku 4.7 przedstawione zostało zestawienie procentowej różnicy jakości pomiędzy wariantami W_{DIZ} i W_C dla wszystkich przebiegów eksperymentu przeprowadzonych dla tego zbioru danych. Na osi X umieszczona została wielkości zbioru treningowego, a na osi Y – jakość anotacji użytych w eksperymencie (zarówno w zbiorze treningowym, jak i testowym) wyrażona przy pomocy metryki MAE. Dokładna wielkość zbiorów treningowego oraz testowego (użytych do trenowania modelu globalnego) oraz zbiorów rozszerzonych, zawierających anotacje (użytych do trenowania modelu *DIZ* oraz modelu referencyjnego) przedstawiona została w Tabeli 4.4.

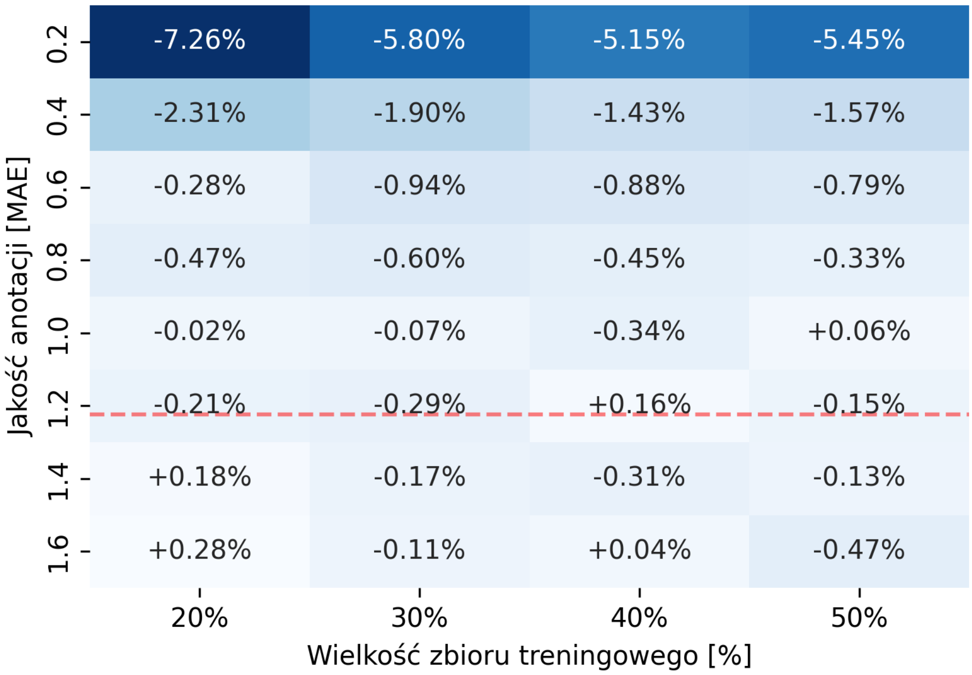
W ramach analizy wyników eksperymentów dokonałem weryfikacji dwóch pytań badawczych opisanych w Paragrafie 4.2.1:

P_1^4: Jakość dla zbioru danych empirycznych obserwowana została na poziomie $q = MAE = 1{,}22$. Dla tej wartości nie zaobserwowano zauważalnej różnicy pomiędzy wariantami W_{DIZ} i W_C. W przypadku danych symulacyjnych obserwowana była wyraźna poprawa jakości w sytuacji, gdy parametr q był na poziomie 0,4 lub niższa.

P_2^4: Obserwowana różnica pomiędzy W_{DIZ} i W_C była bezpośrednio zależna od obu parametrów (q i S) zdefiniowanych w ramach eksperymentu:

Wzrost jakości anotacji q miał pozytywny wpływ na jakość modelu *DIZ* (zwiększenie pozytywnej różnicy pomiędzy wariantami). Jednak wpływ jakości anotacji na jakość modelu *DIZ* był widoczny przede wszystkim dla anotacji o najwyższej jakości: $q = 0{,}4$ lub niższe. Dla wyższych wartości parametru q różnice pomiędzy wariantami W_{DIZ} i W_C były nieznaczne.

Zwiększenie liczby treningowych mikro-zadań (parametr S) przełożyło się na poprawę jakości modelu globalnego, a tym samym spadek różnicy pomiędzy wariantami. Najwyższa różnica jakości na poziomie $-7{,}26\%$ osiągnięta została dla $q = 0{,}2$ oraz $S = 20\%$. Zwiększenie wielkości zbioru treningowego do $S = 50\%$ spowodowało spadek różnicy do $-5{,}45\%$.



Rysunek 4.7: Jakość modelu *DIZ* w zależności od wartości parametrów symulacji dla zbioru „wydźwięk opinii o hotelach” (jakość anotacji ze zbioru empirycznego oznaczona czerwoną linią)

Tabela 4.4: *Wielkość zbioru danych użytego w symulacji modelu DIZ dla zbioru „wydźwięk opinii o hotelach”*

	Mikro-zadania		**Anotacje**	
Wielkość zbioru treningowego [%]	**Treningowe**	**Testowe**	**Treningowe**	**Testowe**
20%	390	1560	1298	5220
30%	585	1365	1952	4566
40%	780	1170	2605	3913
50%	975	975	3255	3263

Zbiór: jednostki nazwane

Zbiór „jednostki nazwane” związany był z zadaniem oznaczania w tekście jednostek nazwanych (ang. *named entities*) należących do jednej z pięciu zdefiniowanych kategorii. Dokładniejszy opis samego zadania został przedstawiony w Rozdziale 3 (zob. Paragraf 3.4.5). Jakość modeli stworzonych w ramach eksperymentu oceniana została według metryki F_1.

Na Rysunku 4.8 przedstawione zostało zestawienie procentowej różnicy jakości pomiędzy wariantami W_{DIZ} i W_C dla wszystkich przebiegów eksperymentu przeprowadzonych

dla tego zbioru danych. Na osi X umieszczona została wielkości zbioru treningowego, a na osi Y – jakość anotacji użytych w eksperymencie (zarówno w zbiorze treningowym, jak i testowym) wyrażona przy pomocy metryki F_1. Dokładna wielkość zbiorów treningowego oraz testowego (użytych do trenowania modelu globalnego) oraz zbiorów rozszerzonych, zawierających anotacje (użytych do trenowania modelu *DIZ* oraz modelu referencyjnego) przedstawiona została w Tabeli 4.5.

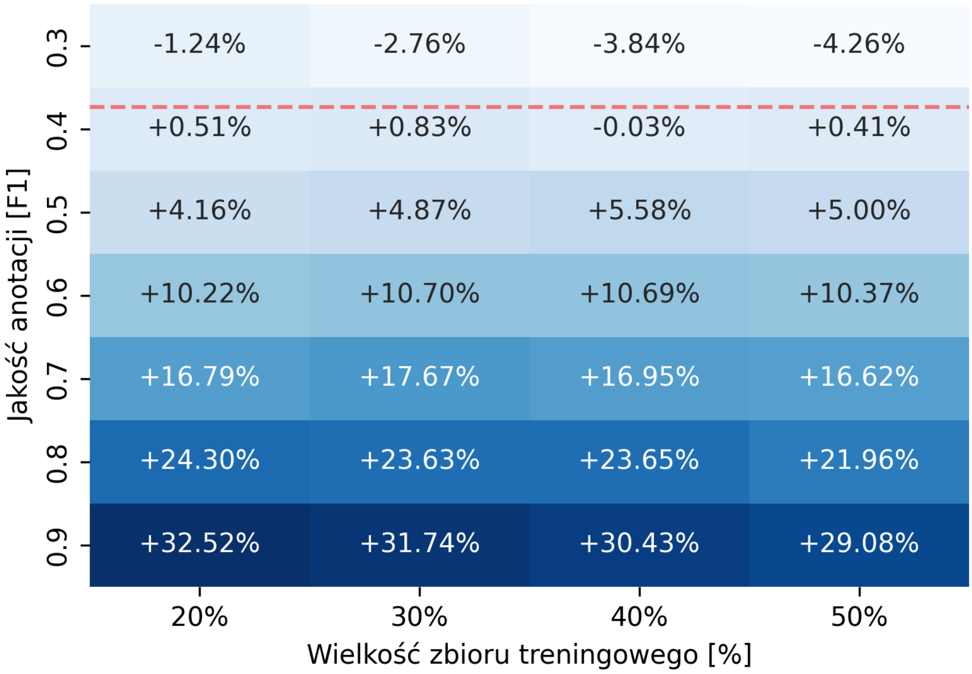
W ramach analizy wyników eksperymentów dokonałem weryfikacji dwóch pytań badawczych opisanych w Paragrafie 4.2.1:

P_1^5: Jakość dla zbioru danych empirycznych obserwowana została na poziomie $q = F_1 = 0{,}37$. Dla tej wartości zaobserwowano negatywny wpływ wprowadzenia modelu *DIZ* (wariant C_{DIZ}). W przypadku danych symulacyjnych obserwowana była wyraźna poprawa jakości w sytuacji, gdy parametr q był na poziomie 0,5 lub wyższym.

P_2^5: Obserwowana różnica pomiędzy W_{DIZ} i W_C była bezpośrednio zależna od obu parametrów (q i S) zdefiniowanych w ramach eksperymentu:

Wzrost jakości anotacji q miał pozytywny wpływ na jakość modelu *DIZ* (zwiększenie pozytywnej różnicy pomiędzy wariantami). Wpływ jakości anotacji na jakość modelu *DIZ* była nieliniowy – jakość modelu rosła szybciej niż jakość anotacji.

Zwiększenie liczby treningowych mikro-zadań (parametr S) przełożyło się na poprawę jakości modelu globalnego, a tym samym spadek różnicy pomiędzy wariantami. Najwyższa różnica jakości na poziomie +32,52% osiągnięta została dla $q = 0{,}9$ oraz $S = 20\%$. Zwiększenie wielkości zbioru treningowego do $S = 50\%$ spowodowało spadek różnicy do +29,08%. Podobna tendencja widoczna była dla wszystkich wartości parametru q.



Rysunek 4.8: Jakość modelu *DIZ* w zależności od wartości parametrów symulacji dla zbioru „jednostki nazwane” (jakość anotacji ze zbioru empirycznego oznaczona czerwoną linią)

Tabela 4.5: *Wielkość zbioru danych użytego w symulacji modelu DIZ dla zbioru „jednostki nazwane”*

	Mikro-zadania		**Anotacje**	
Wielkość zbioru treningowego [%]	**Treningowe**	**Testowe**	**Treningowe**	**Testowe**
20%	580	2320	1614	6429
30%	870	2030	2418	5625
40%	1160	1740	3227	4816
50%	1450	1450	4035	4008

Zbiór: wyrazy bliskoznaczne

Zbiór „wyrazy bliskoznaczne” związany był z zadaniem klasyfikacji binarnej, w ramach której dokonywana była weryfikacja czy dwa słowa są wyrazami bliskoznacznymi. Dokładniejszy opis samego zadania został przedstawiony w Rozdziale 3 (zob. Paragraf 3.4.6). Jakość modeli stworzonych w ramach eksperymentu oceniana została według metryki *dokładność ACC* (ang. *accuracy*).

Na Rysunku 4.9 przedstawione zostało zestawienie procentowej różnicy jakości pomiędzy wariantami W_{DIZ} i W_C dla wszystkich przebiegów eksperymentu przeprowadzonych

dla tego zbioru danych. Na osi X umieszczona została wielkości zbioru treningowego, a na osi Y – jakość anotacji użytych w eksperymencie (zarówno w zbiorze treningowym, jak i testowym) wyrażona przy pomocy metryki ACC. Dokładna wielkość zbiorów treningowego oraz testowego (użytych do trenowania modelu globalnego) oraz zbiorów rozszerzonych, zawierających anotacje (użytych do trenowania modelu *DIZ* oraz modelu referencyjnego) przedstawiona została w Tabeli 4.6.

W ramach analizy wyników eksperymentów dokonałem weryfikacji dwóch pytań badawczych opisanych w Paragrafie 4.2.1:

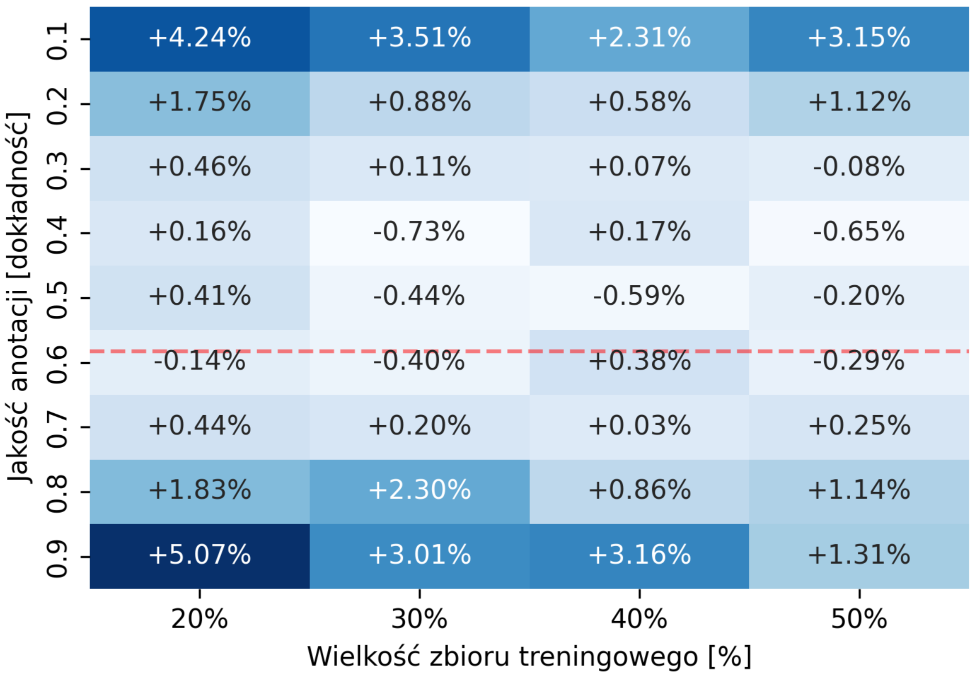
P_1^6: Jakość dla zbioru danych empirycznych obserwowana została na poziomie $q = ACC = 0{,}58$. Dla tej wartości zaobserwowano negatywny wpływ wprowadzenia modelu *DIZ* (wariant C_{DIZ}). W przypadku danych symulacyjnych obserwowana była wyraźna poprawa jakości w sytuacji, gdy parametr q był na poziomie 0,8 lub wyższym, a także 0,2 lub niższym[7] .

P_2^6: Obserwowana różnica pomiędzy W_{DIZ} i W_C była bezpośrednio zależna od obu parametrów (q i S) zdefiniowanych w ramach eksperymentu:

Osiągnięcie skrajnych wartości dla jakości anotacji q (wartości bliskie 0 i bliskie 1) miał pozytywny wpływ na jakość modelu *DIZ* (zwiększenie pozytywnej różnicy pomiędzy wariantami). Wpływ jakości anotacji na jakość modelu *DIZ* była nieliniowy – jakość modelu rosła szybciej niż zmieniała się jakość anotacji.

Zwiększenie liczby treningowych mikro-zadań (parametr S) przełożyło się na poprawę jakości modelu globalnego, a tym samym spadek różnicy pomiędzy wariantami. Najwyższa różnica jakości na poziomie +5,07% osiągnięta została dla $q = 0{,}9$ oraz $S = 20\%$. Zwiększenie wielkości zbioru treningowego do $S = 50\%$ spowodowało spadek różnicy do +1,31%. Podobna tendencja widoczna była dla wszystkich wartości parametru q w których różnica pomiędzy wariantami była duża.

[7] Ponieważ zadanie związane jest z klasyfikacją binarną, bardzo niska jakość anotacji pozytywnie wpływa na jakość modelu, który uczy się wybierać odpowiedź przeciwną od tej wybranej w ramach anotacji.



Rysunek 4.9: Jakość modelu *DIZ* w zależności od wartości parametrów symulacji dla zbioru „wyrazy bliskoznaczne” (jakość anotacji ze zbioru empirycznego oznaczona czerwoną linią)

Tabela 4.6: *Wielkość zbioru danych użytego w symulacji modelu DIZ dla zbioru „wyrazy bliskoznaczne”*

	Mikro-zadania		**Anotacje**	
Wielkość zbioru treningowego [%]	**Treningowe**	**Testowe**	**Treningowe**	**Testowe**
20%	390	1560	1298	5220
30%	585	1365	1952	4566
40%	780	1170	2605	3913
50%	975	975	3255	3263

4.2.6. Podsumowanie eksperymentu

W ramach niniejszego rozdziału zaprezentowany został autorski model *Dynamicznej Informacji Zwrotnej* (*DIZ*). Główną cechą odróżniają model *DIZ* od istniejących rozwiązań jest zastosowanie w nim modelu ucznia oraz aktualnej anotacji w procesie generowania informacji zwrotnej. W celu weryfikacji skuteczności modelu *DIZ* przeprowadziłem eksperyment, w ramach którego przetestowane zostało sześć przykładowych implementacji modelu *DIZ*. Każda z implementacji związana była z innym zbiorem danych. Zbiory danych użyte w eksperymencie zostały zebrane podczas wcześniej przeprowadzonego eksperymentu opisanego w Rozdziale 3. Ewaluacja skuteczności modelu został przeprowadzony poprzez porównanie wybranej metryki jakości (wybór metryki zależał od typu zadania zdefiniowanego dla każdego z badanych zbiorów danych) obliczonej dla oznaczeń wygenerowanych przy pomocy modelu *DIZ* oraz tych, które zostały wygenerowane przy użyciu modelu referencyjnego. Jako model referencyjny użyty został model, który do estymacji nie używał ani modelu ucznia, ani aktualnych anotacji. Analiza wyników eksperymentu przeprowadzona została dla danych empirycznych (pochodzących z niezmodyfikowanego zbioru danych) oraz danych symulacyjnych (zmodyfikowanych tak, by średnia jakość anotacji osiągnęła odpowiedni poziom).

Podsumowanie analizy pytań badawczych

Celem analizy wyników przeprowadzonego eksperymentu było określenie odpowiedzi na dwa pytania badawcze:

P_1: Czy zastosowanie modelu *DIZ* pozwala na wygenerowanie informacji zwrotnej o wyższej jakości w porównaniu do modelu referencyjnego?

P_2: Czy jakość oznaczeń anotatorów wpływa na jakość informacji zwrotnej generowanej przez model *DIZ*?

Analiza wykonana na wynikach eksperymentu przeprowadzonego na danych empirycznych nie wykazała skuteczności modelu *DIZ* w poprawie jakości generowanej informacji zwrotnej. Natomiast, analiza wyników dla danych symulacyjnych wykazała skuteczność modelu *DIZ*, zwłaszcza w sytuacji, gdy jakość anotacji jest wysoka. Analiza ta wykazała również proporcjonalną zależność pomiędzy jakością anotacji a jakością samego modelu. Oznacza to, że jakość generowanej informacji zwrotnej wzrastała wraz ze wzrostem jakości samych anotacji. Jednak jakość modelu *DIZ* (w porównaniu do modelu referencyjnego) różniła się dla każdego z analizowanych zbiorów danych.

Ponieważ skuteczność modelu *DIZ* zależy od typu zbioru danych, a także od jakości oznaczeń tworzonych przez anotatorów, implementacja modelu *DIZ* w procesie nauczania maszynowego powinna być poprzedzona odpowiednimi testami potwierdzającymi skuteczność tego rozwiązania dla danego problemu. Dodatkowym rozwiązaniem, które mogłoby korzystnie wpłynąć na jakość generowanej informacji zwrotnej, jest warunkowe stosowanie modelu *DIZ* tylko dla anotatorów tworzących oznaczenia o wyższej jakości. Alternatywnym podejściem byłoby stworzenie modelu, w którym decyzja odbywałaby się w sposób automatyczny. W ramach eksperymentu implementacja modeli *DIZ* wykonana została przy pomocy sieci neuronowych o prostej architekturze. Możliwe jest, że skuteczność tego modelu byłaby większa w przypadku użycia bardziej złożonej architektury, która w sposób automatyczny odrzuci anotacje o niskiej jakości.

ROZDZIAŁ 5

Podsumowanie

Celem mojej rozprawy doktorskiej było potwierdzenie skuteczności użycia informacji zwrotnej w procesie crowdsourcingu jako mechanizmu poprawy jakości danych dla zadań zawierających dane lingwistyczne. Aby zrealizować wspomniany cel, przeprowadziłem eksperyment, w którym zweryfikowałem skuteczność mechanizmu informacji zwrotnej dla różnych zbiorów danych, które dotyczyły zadań związanych z przetwarzaniem języka naturalnego.

Przeprowadzony przeze mnie przegląd literatury związanej z zastosowaniem mechanizmu informacji zwrotnej w metodzie crowdsourcingu wykazał niewielką liczbę badań, które dotyczyłyby skuteczności tego mechanizmu dla zadań zawierających dane lingwistyczne (zob. Paragraf 1.4.2). Badania opisane w literaturze przedmiotu potwierdzają skuteczność tego mechanizmu, ale zakres prowadzonych prac jest ograniczony i pozostawia pole do dalszych badań. Opisywane w literaturze eksperymenty dotyczyły przede wszystkim pisania dłuższej (zazwyczaj kilkuzdaniowej) formy tekstowej (np. tworzenia podsumowań tekstów lub recenzji produktów). Mechanizm informacji zwrotnej nie został dokładnie przebadany dla zadań przetwarzania języka naturalnego takich jak klasyfikacja, analiza wydźwięku czy oznaczanie jednostek nazwanych. Przeanalizowana przeze mnie literatura wskazywała również brak badań, które opisywałyby sposób, w jaki jakość przekazywanej informacji zwrotnej przekłada się na jakość tworzonych przez anotatorów oznaczeń. Co więcej, brak jest również badań, które porównywałby skuteczność informacji zwrotnej dla różnych typów mikro-zadań.

Na podstawie przedstawionego powyżej tła teoretycznego sformułowałem dwie główne hipotezy badawcze:

H_1: Zapewnienie synchronicznej informacji zwrotnej w pozytywny sposób wpływa na jakość pozyskiwanych danych lingwistycznych w procesie crowdsourcingu.

H_2: Jakość przekazywanej informacji zwrotnej w procesie crowdsourcingu ma pozytywny wpływ na jakość pozyskiwanych danych.

W celu zweryfikowania postawionych hipotez przeprowadziłem eksperyment, w ramach którego przetestowałem skuteczność mechanizmu informacji zwrotnej dla sześciu różnych zbiorów danych: „skargi usług bankowych”, „atrybuty produktów *eBay*”, „waga produktów *eBay*”, „jednostki nazwane”, „wydźwięk opinii o hotelach”, „wyrazy bliskoznaczne” (zob. Rozdział 3). Każdy ze zbiorów dotyczył innego typu zadania, ale wszystkie związane były z przetwarzaniem języka naturalnego. Badanie przeprowadzone zostało w czterech wariantach, które zostały podzielone na dwie grupy. W grupie pierwszej znalazły się warianty, w których anotatorzy otrzymywali informację zwrotną dotyczącą jakości ich oznaczenia. W ramach tej grupy wyróżnione zostały trzy warianty, które określały jakość przekazywanej informacji zwrotnej: wysoka jakość, umiarkowana jakość oraz niska jakość. Druga grupa zawierała wariant kontrolny, w którym mechanizm informacji zwrotnej nie był stosowany.

Dostępne narzędzia do zbierania danych w metodzie crowdsourcingu nie posiadają funkcjonalności, które umożliwiłyby wdrożenie i przetestowanie skuteczności mechanizmu informacji zwrotnej. Z tego powodu eksperyment został przeprowadzony przy użyciu dwóch systemów: platformy *Amazon Mechanical Turk* (*MTurk*), która została użyta do rekrutacji uczestników badania oraz autorskiego systemu *Funcrowd*, w którym znajdował się interfejs anotacyjny obsługujący informację zwrotną, a także interfejs pozwalający na automatyczną integrację z platformą *MTurk*. Kod systemu *Funcrowd* został udostępniony w formie otwartoźródłowej na otwartej licencji i może być stosowany do przeprowadzenia kolejnych eksperymentów.

W ramach przeprowadzonego eksperymentu, dla każdego z użytych zbiorów danych wybranych zostało 2000 elementów[1]. Ostatecznie w eksperymencie wzięło udział 999 anotatorów, którzy łącznie wykonali 56908 oznaczeń. Wyniki przeprowadzonego eksperymentu wykazały istotny, pozytywny wpływ informacji zwrotnej na jakość danych pozyskiwanych w procesie crowdsourcingu dla czterech z sześciu badanych zbiorów (hipoteza H_1). W ramach eksperymentu potwierdzona została również zależność pomiędzy jakością informacji zwrotnej a jakością pozyskanych danych (hipoteza H_2). Efekt ten widoczny był dla wszystkich badanych zbiorów danych. Mimo że najkorzystniejszy wpływ badanego mechanizmu obserwowany był dla informacji zwrotnej o najwyższej jakości, to pozytywny efekt wprowadzenia informacji zwrotnej widoczny był również w przypadku informacji zwrotnej o obniżonej jakości. Oznacza to, że zastosowanie informacji zwrotnej w celu podwyższenia jakości danych pozyskiwanych w procesie crowdsourcingu możliwe jest również w przypadku, gdy nie pochodzi ona ze zbioru referencyjnego lub nie jest tworzona przez grupę ekspertów, a jest ona np. wygenerowana w sposób automatyczny przez wybrany algorytm.

[1] Wyjątek stanowił zbiór „waga produktów *eBay*”, w przypadku którego wejściowy zbiór referencyjny zawierał jedynie 1000 elementów.

Przeprowadzona analiza literatury przedmiotu wykazała również, że źródłem treści informacji zwrotnej w większości istniejących systemów jest ręcznie tworzona ocena ekspertów lub predefiniowany zbiór referencyjny. Z tego powodu w ramach niniejszej rozprawy zaproponowałem również autorski model Dynamicznej Informacji Zwrotnej (*DIZ*), którego zadaniem było generowanie informacji zwrotnej w sposób automatyczny. Główną cechą odróżniają model *DIZ* od istniejących rozwiązań jest zastosowanie w nim modelu ucznia oraz danych z aktualnej anotacji w procesie generowania informacji zwrotnej. W celu zweryfikowania skuteczności stworzonego przeze mnie rozwiązania przeprowadziłem eksperyment, w którym działanie modelu zostało sprawdzone dla danych empirycznych oraz danych symulacyjnych. Zbiory danych użyte w eksperymencie zostały zebrane podczas wcześniej przeprowadzonego eksperymentu, który opisany został w Rozdziale 3. Ewaluacja skuteczności modelu została przeprowadzona poprzez porównanie wybranej metryki jakości obliczonej dla oznaczeń wygenerowanych przy pomocy modelu *DIZ* oraz tych, które wygenerowane zostały przy użyciu modelu referencyjnego. Jako model referencyjny użyty został model, który do estymacji nie używał ani modelu ucznia, ani aktualnych anotacji.

Analiza wyników eksperymentu przeprowadzonego dla danych empirycznych nie wykazała skuteczności modelu *DIZ* dla poprawy jakości generowanej informacji zwrotnej. Natomiast, analiza wyników dla danych symulacyjnych wykazała skuteczność modelu *DIZ*, zwłaszcza w sytuacji, w której jakość anotacji jest wysoka. Analiza ta wykazała również proporcjonalną zależność między jakością anotacji a jakością samego modelu. Oznacza to, że jakość generowanej informacji zwrotnej wzrastała wraz ze wzrostem jakości samych anotacji. Jednak jakość modelu *DIZ* (w porównaniu do modelu referencyjnego) różniła się dla każdego z analizowanych zbiorów danych.

Dalsze prace związane z tematem badań opisywanych w ramach niniejszej rozprawy planuję skierować na rozwój systemu *Funcrowd* w celu stworzenia uniwersalnej, łatwej we wdrożeniu platformy crowdsourcingowej, która pozwoli na zastosowanie mechanizmu informacji zwrotnej. System jest cały czas przeze mnie rozwijany i został już wdrożony m.in. w projekcie *Sprawdzamy Jak Jest*, w którym anotatorzy nieodpłatnie weryfikują dokumenty przesłane przez polskie instytucje publiczne.

Bibliografia

Abu-Mostafa, Y. S., Magdon-Ismail, M., & Lin, H.-T. (2012). *Learning from data.* AMLBook.

Albarqouni, S., Baur, C., Achilles, F., Belagiannis, V., Demirci, S., & Navab, N. (2016). Aggnet: Deep learning from crowds for mitosis detection in breast cancer histology images. *IEEE Trans. Med. Imaging*, *35*(5), 1313-1321.

Alonso, O., Rose, D. E., & Stewart, B. (2008). Crowdsourcing for relevance evaluation. *SIGIR Forum*, *42*(2).

Bauer, A., & Popović, Z. (2017, 12). Collaborative problem solving in an open-ended scientific discovery game. *Proc. ACM Hum.-Comput. Interact.*, *1*(CSCW).

Bernstein, M. S., Little, G., Miller, R. C., Hartmann, B., Ackerman, M. S., Karger, D. R., ... Panovich, K. (2010). Soylent: A word processor with a crowd inside. In *Proceedings of the 23nd annual acm symposium on user interface software and technology* (pp. 313–322). ACM.

Boim, R., Greenshpan, O., Milo, T., Novgorodov, S., Polyzotis, N., & Tan, W. (2012). Asking the right questions in crowd data sourcing. In *2012 ieee 28th international conference on data engineering* (p. 1261-1264).

Boutsis, I., & Kalogeraki, V. (2014). On task assignment for real-time reliable crowdsourcing. In *2014 ieee 34th international conference on distributed computing systems* (p. 1-10).

Bu, Q., Simperl, E., Chapman, A., & Maddalena, E. (2019, Jan 01). Quality assessment in crowdsourced classification tasks. *International Journal of Crowd Science*, *3*(3), 222-248.

Cao, C. C., She, J., Tong, Y., & Chen, L. (2012). Whom to ask? jury selection for decision making tasks on micro-blog services. *CoRR*, *abs/1208.0273*.

Chan, J., Dang, S., & Dow, S. P. (2016). Improving crowd innovation with expert facilitation. In *Proceedings of the 19th acm conference on computer-supported cooperative work & social computing* (p. 1223–1235). New York, NY, USA: Association for Computing Machinery.

Chilton, L. B., Horton, J. J., Miller, R. C., & Azenkot, S. (2010). Task search in a human computation market. In *Proceedings of the acm sigkdd workshop on human computation* (p. 1–9). New York, NY, USA: Association for Computing Machinery.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences.* Lawrence Erlbaum Associates.

Csikszentmihalyi, M. (1991). *Flow: The psychology of optimal experience.* New York, NY: Harper Perennial.

Dalvi, N., Dasgupta, A., Kumar, R., & Rastogi, V. (2013). Aggregating crowdsourced binary ratings. In *Proceedings of the 22nd international conference on world wide web* (p. 285–294). New York, NY, USA: Association for Computing Machinery.

Daniel, F., Kucherbaev, P., Cappiello, C., Benatallah, B., & Allahbakhsh, M. (2018). Quality control in crowdsourcing: A survey of quality attributes, assessment techniques, and assurance actions. *ACM Comput. Surv.*, *51*(1).

Dawid, P., Skene, A. M., Dawidt, A. P., & Skene, A. M. (1979). Maximum likelihood estimation of observer error-rates using the em algorithm. *Applied Statistics*, 20–28.

Dolgui, A., & Proth, J.-M. (2013, 11). Outsourcing: Definitions and analysis. *International Journal of Production Research*, *51*.

Donmez, P., Carbonell, J. G., & Schneider, J. (2009). Efficiently learning the accuracy of labeling sources for selective sampling. In *Proceedings of the 15th acm sigkdd international conference on knowledge discovery and data mining* (p. 259–268). New York, NY, USA: Association for Computing Machinery.

Dontcheva, M., Morris, R. R., Brandt, J. R., & Gerber, E. M. (2014). Combining crowdsourcing and learning to improve engagement and performance. In *Proceedings of the sigchi conference on human factors in computing systems* (p. 3379–3388). New York, NY, USA: Association for Computing Machinery.

Doroudi, S., Kamar, E., Brunskill, E., & Horvitz, E. (2016). Toward a learning science for complex crowdsourcing tasks. In *Proceedings of the 2016 chi conference on human factors in computing systems* (p. 2623–2634). New York, NY, USA: Association for Computing Machinery.

Dow, S., Kulkarni, A., Klemmer, S., & Hartmann, B. (2012). Shepherding the crowd yields better work. In *Proceedings of the acm 2012 conference on computer supported cooperative work* (pp. 1013–1022). ACM.

Dudley, J. J., Jacques, J. T., & Kristensson, P. O. (2019). Crowdsourcing interface feature design with bayesian optimization. In *Proceedings of the 2019 chi conference on human factors in computing systems* (p. 1–12). New York, NY, USA: Association for Computing Machinery.

Dziedzic, D. (2016). Use of the free to play model in games with a purpose: the robocorp game case study. *Bio-Algorithms and Med-Systems*, *12*(4), 187–197.

Eickhoff, C., Harris, C. G., de Vries, A. P., & Srinivasan, P. (2012). Quality through flow and immersion: Gamifying crowdsourced relevance assessments. In *Proceedings of the 35th international acm sigir conference on research and development in information retrieval* (p. 871–880). New York, NY, USA: Association for Computing Machinery.

Faridani, S., Hartmann, B., & Ipeirotis, P. G. (2011). What's the right price? pricing tasks for finishing on time. In *Proceedings of the 11th aaai conference on human computation* (p. 26–31). AAAI Press.

Fleiss, J., Cohen, J., & Everitt, B. (1969). Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin*, *72*, 323-327.

Fornaciari, T., Cagnina, L. C., Rosso, P., & Poesio, M. (2020). Fake opinion detection: how similar are crowdsourced datasets to real data? *Language Resources and Evaluation*, 1-40.

Franklin, M. J., Trushkowsky, B., Sarkar, P., & Kraska, T. (2013). Crowdsourced enumeration queries. In *Proceedings of the 2013 ieee international conference on data engineering (icde 2013)* (p. 673–684). USA: IEEE Computer Society.

Fromreide, H., Hovy, D., & Søgaard, A. (2014, 5). Crowdsourcing and annotating NER for Twitter #drift. In *Proceedings of the ninth international conference on language resources and evaluation (LREC'14)* (pp. 2544–2547). Reykjavik, Iceland: European Language Resources Association (ELRA).

Gaikwad, S. N. S., Morina, D., Ginzberg, A., Mullings, C. A., Goyal, S., Gamage, D., ... Bernstein, M. S. (2019). Boomerang: Rebounding the consequences of reputation feedback on crowdsourcing platforms. *CoRR*, *abs/1904.06722*.

Geéron, A. (2017). *Hands-on machine learning with scikit-learn and tensorflow : concepts, tools, and techniques to build intelligent systems.* Sebastopol, CA: O'Reilly Media.

Goldman, S., & Kearns, M. (1995, February). On the complexity of teaching. *J. Comput. Syst. Sci.*, *50*(1), 20–31.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning.* MIT Press.

Grier, D. A. (2005). *When computers were humans.* Princeton: Princeton University Press.

Ho, C.-J., Jabbari, S., & Vaughan, J. W. (2013). Adaptive task assignment for crowdsourced classification. In (p. I–534–I–542). JMLR.org.

Hong, J., Lee, K., Xu, J., & Kacorri, H. (2020). Crowdsourcing the perception of machine teaching. *CoRR*, *abs/2002.01618*.

Horton, J. J. (2010). Employer expectations, peer effects and productivity: Evidence from a series of field experiments. *CoRR*, *abs/1008.2437*.

Howe, J. (2006a). *Crowdsourcing: A definition.* Retrieved from `https://crowdsourcing.typepad.com/cs/2006/06/crowdsourcing_a.html`

Howe, J. (2006b, 06). The rise of crowdsourcing. *Wired Magazine*, *14*(6). Retrieved from `http://www.wired.com/wired/archive/14.06/crowds.html`

Huang, S.-W., & Fu, W.-T. (2013). Enhancing reliability using peer consistency evaluation in human computation. In *Proceedings of the 2013 conference on computer supported cooperative work* (p. 639–648). New York, NY, USA: Association for Computing Machinery.

Ipeirotis, P. G., Provost, F., & Wang, J. (2010). Quality management on amazon mechanical turk. In *Proceedings of the acm sigkdd workshop on human computation* (p. 64–67). New York, NY, USA: Association for Computing Machinery.

Jeremy Orloff, J. B. (2014). Reading for 24: Bootstrap confidence intervals. In *Introduction to probability and statistics – mit course no. 18.05.* Cambridge MA. Retrieved from `https://ocw.mit.edu/courses/18-05-introduction-to-probability-and-statistics-spring-2014/`

Jiang, Z., & Huang, Y. (2016, 01). The role of feedback in dynamic crowdsourcing contests: A structural empirical analysis. *SSRN Electronic Journal*.

Jurafsky, D., & Martin, J. H. (2021). *Speech and language processing : an introduction to natural language processing, computational linguistics, and speech recognition. 3rd edition draft.* Retrieved from `https://web.stanford.edu/~jurafsky/slp3/ed3book_jan122022.pdf`

Kazai, G., Kamps, J., Koolen, M., & Milic-Frayling, N. (2011). Crowdsourcing for book search evaluation: Impact of hit design on comparative system ranking. In *Proceedings of the 34th international acm sigir conference on research and development in information retrieval* (p. 205–214). New York, NY, USA: Association for Computing Machinery.

Kazai, G., Kamps, J., & Milic-Frayling, N. (2012). The face of quality in crowdsourcing relevance labels: Demographics, personality and labeling accuracy. In *Proceedings of the 21st acm international conference on information and knowledge management* (p. 2583–2586). New York, NY, USA: Association for Computing Machinery.

Kecman, V. (2005, 05). Support vector machines – an introduction. In (Vol. 177, p. 605-605).

Kelley, K., & Preacher, K. (2012, 04). On effect size. *Psychological Methods*, *17*, 137–152.

Kittur, A., Smus, B., Khamkar, S., & Kraut, R. E. (2011). Crowdforge: Crowdsourcing complex work. In *Proceedings of the 24th annual acm symposium on user interface software and technology* (p. 43–52). New York, NY, USA: Association for Computing Machinery.

Kuncheva, L. I., Whitaker, C. J., & Duin, R. P. W. (2003). Limits on the majority vote accuracy in classifier fusion. *Pattern Analysis and Applications*, *6*, 22–31.

Lasecki, W. S., Rzeszotarski, J. M., Marcus, A., & Bigham, J. P. (2015). The effects of sequence and delay on crowd work. In *Proceedings of the 33rd annual acm conference on human factors in computing systems* (p. 1375–1378). New York, NY, USA: Association for Computing Machinery.

Lave, J., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation.* Cambridge University Press.

Li, G., Wang, J., Zheng, Y., & Franklin, M. J. (2016). Crowdsourced data management: A survey. *IEEE Transactions on Knowledge & Data Engineering*, *28*(9), 2296–2319.

Littlestone, N., & Warmuth, M. (1994, 05). The weighted majority algorithm. *Information and Computation*, *108*.

Mamykina, L., Smyth, T. N., Dimond, J. P., & Gajos, K. Z. (2016). Learning from the crowd: Observational learning in crowdsourcing communities. In *Proceedings of the 2016 chi conference on human factors in computing systems* (p. 2635–2644). New York, NY, USA: Association for Computing Machinery.

Marcus, A., Karger, D., Madden, S., Miller, R., & Oh, S. (2012). Counting with the crowd. *Proc. VLDB Endow.*, *6*(2), 109–120.

Mudgal, S., Li, H., Rekatsinas, T., Doan, A., Park, Y., Krishnan, G., ... Raghavendra, V. (2018). Deep learning for entity matching: A design space exploration. In *Proceedings of the 2018 international conference on management of data* (p. 19–34). New York, NY, USA: Association for Computing Machinery.

Newell, E., & Ruths, D. (2016). How one microtask affects another. In *Proceedings of the 2016 chi conference on human factors in computing systems* (p. 3155–3166). New York, NY, USA: Association for Computing Machinery.

Nguyen, C., Oh, O., Alothaim, A., de Vreede, T., & de Vreede, G.-J. (2015, 11). Engaging with online crowd: A flow theory approach. In (p. 175-189).

Nguyen, T. T. D. T., Garncarz, T., Ng, F., Dabbish, L. A., & Dow, S. P. (2017). Fruitful feedback: Positive affective language and source anonymity improve critique reception and work outcomes. In *Proceedings of the 2017 acm conference on computer supported cooperative work and social computing* (p. 1024–1034). New York, NY, USA: Association for Computing Machinery.

Patil, K. R., Zhu, J., Kopeć, L. u., & Love, B. C. (2014). Optimal teaching for limited-capacity human learners. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems* (Vol. 27). Curran Associates, Inc.

Poesio, M., Chamberlain, J., Kruschwitz, U., Robaldo, L., & Ducceshi, L. (2013, 04). Phrase detectives. *ACM Transactions on Interactive Intelligent Systems*, *3*, 1-44.

Polikar, R. (2006). Ensemble based systems in decision making. *IEEE Circuits and Systems Magazine*, *6*(3), 21-45.

Prestopnik, N. R., & Crowston, K. (2012). Citizen science system assemblages: Understanding the technologies that support crowdsourced science. In *Proceedings of the 2012 iconference* (p. 168–176). New York, NY, USA: Association for Computing Machinery.

Rafferty, A., Brunskill, E., Griffiths, T., & Shafto, P. (2011, 06). Faster teaching by pomdp planning. In (Vol. 6738, p. 280-287).

Raykar, V. C., & Yu, S. (2012). Eliminating spammers and ranking annotators for crowdsourced labeling tasks. *Journal of Machine Learning Research*, *13*(16), 491-518.

Raykar, V. C., Yu, S., Zhao, L. H., Valadez, G. H., Florin, C., Bogoni, L., & Moy, L. (2010). Learning from crowds. *J. Mach. Learn. Res.*, *11*, 1297–1322.

Reeve, J. (2012). A Self-determination Theory Perspective on Student Engagement. In S. Christenson, A. Reschly, & C. Wylie (Eds.), *Handbook of research on student engagement* (pp. 149–172). Springer, Boston, MA: Springer International Publishing.

Rogstadius, J., Kostakos, V., Kittur, A., Smus, B., Laredo, J., & Vukovic, M. (2011, 7). An assessment of intrinsic and extrinsic motivation on task performance in crowdsourcing markets. *Proceedings of the International AAAI Conference on Web and Social Media*, *5*(1).

Rzeszotarski, J. M., Chi, E. H., Paritosh, P., & Dai, P. (2013). Inserting micro-breaks into crowdsourcing workflows. In *Human computation and crowdsourcing: Works in progress and demonstration abstracts, an adjunct to the proceedings of the first AAAI conference on human computation and crowdsourcing, november 7-9, 2013, palm springs, ca, USA.*

Rzeszotarski, J. M., & Kittur, A. (2011). Instrumenting the crowd: Using implicit behavioral measures to predict task performance. In *Proceedings of the 24th annual acm symposium on user interface software and technology* (p. 13–22). New York, NY, USA: Association for Computing Machinery.

Sakurai, Y., Okimoto, T., Oka, M., Shinoda, M., & Yokoo, M. (2013). Ability grouping of crowd workers via reward discrimination. In *Hcomp.*

Sawilowsky, S. S. (2009). New effect size rules of thumb. *Journal of Modern Applied Statistical Methods*, *8*, 26.

Sheng, V. S. (2011). Simple multiple noisy label utilization strategies. In *2011 ieee 11th international conference on data mining* (p. 635-644).

Shinohara, A., & Miyano, S. (1991, February). Teachability in computational learning. *New Gen. Comput.*, *8*(4), 337–347.

Singla, A., Bogunovic, I., Bartók, G., Karbasi, A., & Krause, A. (2014). Near-optimally teaching the crowd to classify. *CoRR*, *abs/1402.2092*.

Sokolova, M., & Lapalme, G. (2009, 07). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, *45*, 427-437.

Su, H., Deng, J., & Fei-Fei, L. (2012). Crowdsourcing annotations for visual object detection. In *Workshops at the twenty-sixth aaai conference on artificial intelligence.*

Toxtli, C., Richmond-Fuller, A., & Savage, S. (2020). Reputation agent: Prompting fair reviews in gig markets. *CoRR*, *abs/2005.06022*.

Vapnik, V. N. (1998). *The nature of statistical learning theory.* John Wiley & Sons.

Waggoner, B., & Chen, Y. (2014). Output agreement mechanisms and common knowledge. In *Hcomp.*

Wang, G., Hao, J., Ma, J., & Jiang, H. (2011). A comparative assessment of ensemble learning for credit scoring. *Expert Systems with Applications*, *38*(1), 223-230.

Wang, P., Nagrecha, K., & Vasconcelos, N. (2021a). Gradient-based algorithms for machine teaching. In *2021 ieee/cvf conference on computer vision and pattern recognition (cvpr)* (p. 1387-1396).

Wang, P., Nagrecha, K., & Vasconcelos, N. (2021b, June). Gradient-based algorithms for machine teaching. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition (cvpr)* (p. 1387-1396).

Wang, S., Xiao, X., & Lee, C.-H. (2015). Crowd-based deduplication: An adaptive approach. In *Proceedings of the 2015 acm sigmod international conference on management of data* (p. 1263–1277). New York, NY, USA: Association for Computing Machinery.

Wiggins, G. (1998). *Educative assessment. designing assessments to inform and improve student quality.* Jossey-Bass Publishers.

Willett, K. W., Lintott, C. J., Bamford, S. P., Masters, K. L., Simmons, B. D., Casteels, K. R. V., ... et al. (2013, Sep). Galaxy zoo 2: detailed morphological classifications for 304 122 galaxies from the sloan digital sky survey. *Monthly Notices of the Royal Astronomical Society*, *435*(4), 2835–2860.

Willett, W., Heer, J., & Agrawala, M. (2012). Strategies for crowdsourcing social data analysis. In *Proceedings of the sigchi conference on human factors in computing systems* (p. 227–236). New York, NY, USA: Association for Computing Machinery.

Yang, R., Xue, Y., & Gomes, C. (2017, 09). Pedagogical value-aligned crowdsourcing: Inspiring the wisdom of crowds via interactive teaching..

Yu, L., André, P., Kittur, A., & Kraut, R. (2014). A comparison of social, learning, and financial strategies on crowd engagement and output quality. In *Proceedings of the 17th acm conference on computer supported cooperative work & social computing* (p. 967–978). New York, NY, USA: Association for Computing Machinery.

Zaidan, O. F., & Callison-Burch, C. (2011, 6). Crowdsourcing translation: Professional quality from non-professionals. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies* (pp. 1220–1229). Portland, Oregon, USA: Association for Computational Linguistics.

Zhao, Z., Yan, D., Ng, W., & Gao, S. (2013). A transfer learning based framework of crowd-selection on twitter. In *Proceedings of the 19th acm sigkdd international conference on knowledge discovery and data mining* (p. 1514–1517). New York, NY, USA: Association for Computing Machinery.

Zheng, Y., Cheng, R., Maniu, S., & Mo, L. (2015, 03). On optimality of jury selection in crowdsourcing..

Zhou, Y., Nelakurthi, A. R., & He, J. (2018). Unlearn what you have learned: Adaptive crowd teaching with exponentially decayed memory learners. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining, KDD 2018, london, uk, august 19-23, 2018* (pp. 2817–2826).

Zhu, H., Dow, S. P., Kraut, R. E., & Kittur, A. (2014). Reviewing versus doing: Learning and performance in crowd assessment. In *Proceedings of the 17th acm conference on computer supported cooperative work & social computing* (p. 1445–1455). New York, NY, USA: Association for Computing Machinery.

Zhu, X., Singla, A., Zilles, S., & Rafferty, A. N. (2018). An overview of machine teaching. *CoRR*, *abs/1801.05927*.

Spis tabel

Spis rysunków

DODATEK A

Treść protokołów anotacyjnych

Dodatek ten zawiera treść protokołów anotacyjnych użytych w eksperymencie opisywanym w ramach niniejszej pracy. Ponieważ eksperyment przeprowadzany został dla odbiorców posługujących się językiem angielskim, wszystkie protokoły napisane zostały w tym języku. Protokoły zamieszczone poniżej pozostały w niezmienionej formie.

A.1. Protokół anotacyjny: skargi usług bankowych

General overview

In this task you will see consumer complaints on financial products and company responses. All data are gathered for banks from the United States.

Your task is to assign each of those complaints to one of the selected categories.

Task procedure

1. Read **Complaint massage**

2. Select the right category in the **Complaint category** field.

Categories descriptions

Category	**Description**
Debt collection	Issues related to debt collection, attempting to collect debt not owed, notification about the debt, medical debt
Mortgage	Cases around mortgages, reversing mortgage, loan modification, closing the mortgage
Credit reporting	Problems with incorrect information on a credit report, fraud investigation, getting credit report score
Credit card	Checking credit card limit, late fee, problems related to identity theft or fraud, bankruptcy

A.2. Protokół anotacyjny: atrybuty produktów *eBay*

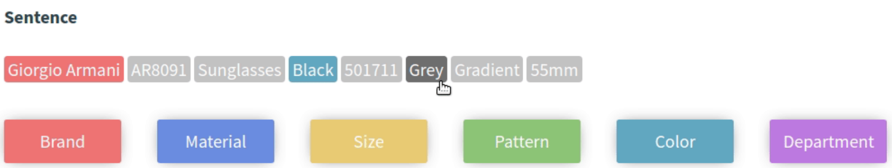
General overview

In this task, you will see a text from the item’s title or description.

Your task is to extract an a phrase (one word or multiple) that represents an item’s attribute.

Task procedure

1. Select the text you want and click the button with the correct attribute. **Sometimes a text may not include any tags, in such case you can submit an empty text**.



Attribute categories with examples

Category	Examples					
Brand	’Gucci’	’Pyle’	’Adidas’	’Nautica’	’Athleta’	’SMC Networks’
Material	’Nylon’	’Tulle’	’Cotton Blend’	’Polyester’	’Cotton’	’Gold’
Size	’2-3’	’XXS’	’47’	’$1\frac{1}{2}$’	’56’	’Plus Size’
Pattern	’Plaid’	’Floral’	’Leopard’	’Camouflage’	’Polka Dot’	’Paisley’
Color	’Silver’	’White’	’Mahogany’	’Gray’	’Brown’	’Pink’
Department	’Girls’	’Boys’	’Men’	’Women’	’Kids’	’Unisex’

2. You can clear your selection by clicking it.

3. After you’re done with your work click **Submit** button.

A.3. Protokół anotacyjny: waga produktów *eBay*

General overview

Your task is to estimate a weight of the displayed item.

All shown items are products sold in the eCommerce store.

Task procedure

1. Take a look on the **Item's image**.

   This field should contain an example photo of the item.

2. Read **Item's title**

   Check if the title contains information about the number of items in the package.

3. Read **Item's category path**

   This field contains a path of categories to which item was assigned in the store. Categories can help you identify the type of the item you are annotating.

4. Try to estimate weight of the item in **grams** and write your guess into **Estimated item's weight** field.

   Max weight for items in this task is 10 000 grams.

A.4. Protokół anotacyjny: wydźwięk opinii o hotelach

General overview

Your task is to assign a right sentiment score based on the Hotel's review message.

Task procedure

1. Read the **Review message**

   This field contains a full review message which was sent by one of the Hotel's customer.

2. Decide if review was **positive**, **neutral** or **negative**.

   High score means the review was postivie, low score – negative.

3. Assign right sentiment score in the **Rating** field.

A.5. Protokół anotacyjny: jednostki nazwane

General overview

In this task, you will see one sentence.

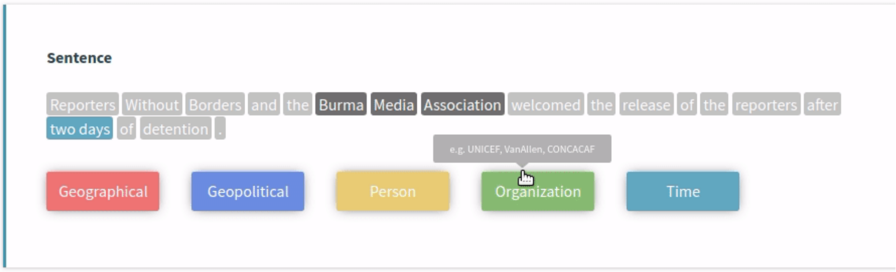
Your task is to select one or multiple words that represents one named entitity.

Named enitity are phrases that can be assigned to one from following category:

Geographical, Geopolitical, Person, Organization, Time

Task procedure

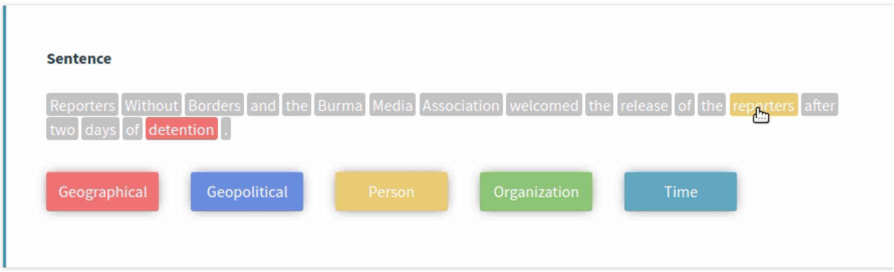
1. Select words you want and click the button with the correct tag.



Named enitity tags with examples

Category	Examples		
Geographical	'Tennesse'	'Ganges'	'Brookly'
Geopolitical	'Italy'	'Italian'	'Argentina'
Person	'Patrick'	'Richardson'	'Yusuf'
Organization	'UNICEF'	'VanAllen'	'CONCACAF'
Time	'1960'	'present-day'	'two days'

2. You can clear your selection by clicking it.



3. After you're done with your work click **Submit** button.

A.6. Protokół anotacyjny: wyrazy bliskoznaczne

General overview

Your task is to decide if two highlighted words have a similar meaning.

Highlighted words will look like this.

Notice, we are not looking only for direct synonyms, but more general similarity,

like: happy contented joyful glad overjoyed successful

Task procedure

1. Read the **First sentence**.
2. Read the **Second sentence**.
3. Compare highlighted words and decide if they have a similar meaning.
4. Select option **Yes** if you think both words have a similar meaning, and bf otherwise.

DODATEK B

Algorytmy zniekształcenia anotacji referencyjnych

Załącznik ten zawiera rozszerzony opis algorytmów użytych w celu zniekształcenia anotacji referencyjnych użytych w eksperymencie przedstawionym w Rozdziale 3. W załączniku zawarta została treść algorytmów trzech zbiorów danych: „atrybuty produktów *eBay*”, „waga produktów *eBay*” oraz „jednostki nazwane”. Ze względu na swoją prostotę algorytmy użyte dla pozostałych trzech zbiorów nie wymagały dodatkowego opisu.

B.1. Zbiór: atrybuty produktów *eBay*

Procedura 13: Algorytm zniekształcania anotacji referencyjnych w zbiorze „atrybuty produktów *eBay*”

Niech :

T – zbiór tokenów;

J – zbiór kategorii atrybutów, gdzie $J = \{0, 1, ..., n\}$, gdzie 0 to kategoria „pusta”, a $n = 6$;

A – zbiór anotacji, gdzie a_i to kategoria przypisana do tokenu t_i;

p_d – prawdopodobieństwo usunięcia anotacji (W_2: $p_d = 0.25$, W_3: $p_d = 0.4$);

p_r – prawdopodobieństwo zamiany anotacji (W_2: $p_r = 0.05$, W_3: $p_r = 0.10$);

Kroki :

1 Dla każdego tokenu $t_i \in T$:

- Zastąp anotację a_j kategorią pustą $x = 0$ z prawdopodobieństwem p_d:

$$a_i := x, \text{ gdzie } P(X = x) = \begin{cases} 1 - p_d, & a_i \\ p_d, & 0 \end{cases}$$

- Zastąp anotację a_j kategorią losową z prawdopodobieństwem p_r:

$$a_i := x, \text{ gdzie } P(X = x) = \begin{cases} 1 - p_r, & a_i \\ p_r, & rand(1, n) \end{cases}$$

Wyjście:

A – Finalny zbiór anotacji dla wszystkich mikro-zadań

B.2. Zbiór: waga produktów *eBay*

Procedura 14: Algorytm zniekształcania anotacji referencyjnych w zbiorze „atrybuty produktów *eBay*"

Niech :

a – referencyjna wartość wagi danego produktu

σ^2 – odchylenie standardowe wprowadzanego szumu (W_2: $\sigma = 400g$, W_3: $\sigma = 1000g$);

Kroki :

1 Dodaj losową wartość szumu do referencyjnej wagi a:

$$a := a + \mathcal{N}(0, \sigma^2)$$

Wyjście:

A – Finalny zbiór anotacji dla wszystkich mikro-zadań

B.3. Zbiór: jednostki nazwane

Procedura 15: Algorytm zniekształcania anotacji referencyjnych w zbiorze „jednostki nazwane”

Niech :

T – zbiór tokenów;

J – zbiór kategorii jednostek nazwanych, gdzie $J = \{0, 1, ..., n\}$ $(n = 5)$, gdzie 0 to kategoria „pusta”;

A – zbiór anotacji, gdzie a_i to kategoria przypisana do tokenu t_i;

p_d – prawdopodobieństwo usunięcia anotacji (W_2: $p_d = 0{,}25$, W_3: $p_d = 0.4$);

p_r – prawdopodobieństwo zamiany anotacji (W_2: $p_r = 0{,}05$, W_3: $p_r = 0{,}10$);

Kroki :

1 Dla każdego tokenu $t_i \in T$:

- Zastąp anotację a_j kategorią pustą $x = 0$ z prawdopodobieństwem p_d:

$$a_i := x, \text{ gdzie } P(X = x) = \begin{cases} 1 - p_d, & a_i \\ p_d, & 0 \end{cases}$$

- Zastąp anotację a_j kategorią losową z prawdopodobieństwem p_r:

$$a_i := x, \text{ gdzie } P(X = x) = \begin{cases} 1 - p_r, & a_i \\ p_r, & rand(1, n) \end{cases}$$

Wyjście:

A – Finalny zbiór anotacji dla danego mikro-zadania