

O „subtelnościach” metod oceny wydźwięku wypowiedzi pisanych. Porównanie trzech podejść w analizie sentymentu

Krzysztof Tomanek 
Uniwersytet Jagielloński

<https://doi.org/10.18778/1733-8069.20.4.04>

Słowa kluczowe: NLP, ML, sztuczna inteligencja, analiza sentymentu, słowniki sentymentu, analizy jakościowe

Abstrakt: Artykuł przedstawia wyniki eksperymentu metodologicznego, w którym w odniesieniu do tego samego materiału badawczego posłużono się trzema odmiennymi w swojej logice i zastosowaniu metodami analizy wypowiedzi zapisanych w formie tekstowej. Celem tego opracowania jest wskazanie różnic trzech podejść analitycznych, w których mamy do czynienia z analizą opartą na rozumiejącym czytaniu tekstu (kodowanie manualne), analizą półautomatyczną i nadzorowaną (wykonaną przez słownik klasyfikacyjny zaprogramowany przez człowieka i oparty na transparentnych regułach – metoda z obszaru *machine learning* – ML) oraz metodą nietransparentną i nienadzorowaną (sztuczna inteligencja – ChatGPT w wersji 3.5). Badanie dotyczy analizy sentymentu, zwanej też analizą wydźwięku. Uwaga w dużej mierze skoncentrowana jest na zastosowaniu tych metod oraz wyjaśnieniu różnic w uzyskanych wynikach.

Krzysztof Tomanek

Socjolog, doktor nauk społecznych, reprezentuje Instytut Socjologii Uniwersytetu Jagiellońskiego. Współzałożyciel CAQDAS TM Lab przy Instytucie Socjologii na Uniwersytecie Jagiellońskim. Zajmuje się głównie zastosowaniem metod służących analizom danych jakościowych i ilościowych, w tym także zastosowaniem uczenia maszynowego i AI w naukach społecznych. Interesuje się również i na co dzień zajmuje metodami wizualizacji danych, storytellingiem, teorią sieciową w badaniach społecznych. Od siedmiu lat analizuje projekty artystyczne Rity Leistner. Zaangażowany społecznie i wolontarystycznie. Członek Stowarzyszenia NGO POLITES, PTS, PTE.
e-mail: krzysztof.tomanek@uj.edu.pl



© by the author, licensee University of Lodz, Poland
This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution license
CC-BY-NC-ND 4.0 (<https://creativecommons.org/licenses/by-nc-nd/4.0/>)

Wprowadzenie

Analiza treści to niezmiernie szeroki zakres badań, wśród których mamy do czynienia z wieloma technikami analitycznymi. Część z nich doczekała się już opracowań, w których oprogramowanie komputerowe wykonuje pewne etapy analizy tekstu lub też pełną analizę bez udziału badacza. Mówiąc dokładniej, badacz jedynie wybiera typ analizy i naciska odpowiedni przycisk procedury, która wykonywana jest już bez jego udziału i nadzoru. Wśród rozwiązań znajdziemy takie, które pozwalają na analizę tematyczną i analizę dyskursu (Regneri, King, 2016; Liontou, 2022), zautomatyzowany system wychwytyjący typy argumentacji (Elouazizi, Oberg, Birol, 2017) czy też bardziej złożone narzędzia zbudowane na potrzeby analizy narracji (Esuli, Sebastiani, 2006; Hsu, Thompson, 2018; Fromm, MacWhinney, Thompson, 2020), a także analizy sentymentu¹ wypowiedzi (Hemalatha, Varma, Govardhan, 2014; Yao, 2019; Fargues i in., 2023)². Literatura przedmiotu ujawnia wiele ciekawych problemów związanych z procesem automatyzowania analiz tekstowych. Przedstawione w artykule rozważania zawężę do związanych z analizą sentymentu. Najogólniej mówiąc, problemy w tym obszarze zogniskowane są wokół trzech zagadnień: kodowania wykonywanego przez badaczy skupionych na refleksyjnym i krytycznym czytaniu tekstów, technicznych aspektów związanych z budową narzędzi automatycznie oceniających wydźwięk tekstu oraz uczenia sposobów klasyfikowania i oceniania sentymentu wypowiedzi pisanych (Yilmaz, 2023). Każdy z tych trzech obszarów niesie ze sobą wiele dylematów, z którymi badacze muszą się mierzyć. Aby ukazać złożoność tej problematyki, wymienię zaledwie kilka z nich. Są nimi na przykład: wybór strategii kodowania, wybór skali oceny sentymentu, ocena poziomu sentymentu dla określonych słów i fraz, wybór słownika i miary rzetelnej oceny emocji wyrażonych przez słowa, poprawność liczbowej reprezentacji natężenia emocji zawartych w tekście, konstrukcja skali natężenia emocji wyrażanych w tekście (czy też ogólniej – ocena natężenia emocji wyrażonych w komunikacji niebezpośredniej), budowa instrukcji kodowania i wyszkolenie koderów tak, aby ich praca uzyskiwała wysoki poziom zgodności kodowania, skonstruowanie metody zgodności decyzji koderów dla problemu oceny emocji w tekście, transparentność (metodyczność i intersubiektywność) metod oceny rozpoznawania i natężenia emocji, wyuczenie narzędzia oceniającego tekst tak, aby dokonywało ono poprawnej oceny sentymentu. Jak można zauważyć, problemów, z którymi nauka mierzy się w obszarze analizy sentymentu, nie brakuje. Nowy obszar problemowy pojawia się zwłaszcza w związku z dostępnością rozwiązań analizujących teksty, a bazujących na algorytmach, na których zbudowana jest sztuczna inteligencja. W momencie powstawania tego artykułu istniały już gotowe modele SI służące do rozpoznawania dźwięków (wydawanych przez człowieka, inne zwierzęta czy też obiekty nieożywione), klasyfikacji przekazów dźwiękowych (np. dźwięków ostrzegających, relaksujących itp.), klasyfikacji obrazów, wykrywania różnego rodzaju obiektów w obrazach cyfrowych. W analizie treści napotkamy

1 Sformułowania „analiza sentymentu” i „analiza wydźwięku” bywają używane zamiennie – taką też konwencję przyjmuję w tym artykule.

2 W tym niezmiernie skrótowym wprowadzeniu unikam przybliżenia aktualnego stanu badań nad zautomatyzowaną analizą tekstu, a w szczególności nad analizą sentymentu. Literatura poświęcona temu jest bardzo obszerna, dlatego też skupiam się na przedstawieniu zagadnienia kluczowego dla wyników prezentowanego tu eksperymentu badawczego – to znaczy na kwestii różnic w wynikach dostarczanych przez trzy różne podejścia do analizy wydźwięku wypowiedzi.

„wytrenowane”, stosowane i sprawdzone modele służące różnorodnym celom analitycznym. Warto wspomnieć choćby kilka obszarów, w których rozwiązania SI są dostępne dla badaczy zajmujących się jakościową analizą danych (problemem jest fakt, że są to rozwiązania głównie dla języka angielskiego). Są nimi zautomatyzowane: wyszukiwania odpowiedzi na pytania dotyczące analizowanego tekstu, ale chodzi tu – podkreślmy to – o pytania, które nie są zawarte w tekście; podsumowywanie tekstów w postaci abstraktów, opracowania tekstu w postaci informacji syntetycznie podanych w punktach; klasyfikatory grupujące wypowiedzi ze względu na użyte w nich słownictwo, styl pisanie czy tematykę tekstu; rozwiązania tłumaczące lub też parafrazujące (wzbogacające tekst); metody tłumaczenia tekstów w taki sposób, aby wynikały z nich reguły postępowania – innymi słowy, aby powstawały na podstawie tekstu instrukcje mówiące, jak działać krok po kroku w danej sytuacji (OpenAI Platform, b.r.); zamiany tekstów na emotikony; korekty gramatyki danego języka; rekomendacje restauracji bazujące zaledwie na kilku słowach kluczowych, wypowiedzianych i zapisanych dla SI; generatory pytań służących do rekrutacji na określone specjalistyczne stanowiska; rozwiązania rozpoznające emocje i natężenie emocji w wypowiedzi pisanej.

W tym artykule chciałbym przybliżyć zaledwie jeden z wymienionych obszarów i problemów. Jest nim różnica w sposobie oceny sentymentu tekstu, która ujawnia się, gdy korzystamy z różnych metod. Mam tu na myśli ocenę dokonaną przez badaczy, słownik klasyfikacyjny, sztuczną inteligencję. Różnica w ocenach jest wynikiem specyfiki wspomnianych metod, a w szczególności pewnych obszarów ich nietransparentności, wynikających z ograniczeń wiedzy, jaką dysponujemy. Na potrzeby tej pracy będę stosował zamiennie, podobnie jak inni badacze (Lula, Wójcik, Tuchowski, 2016; Kocoń, Miłkowski, Zaśko-Zielińska, 2019), sformułowania zapożyczone z badań publikowanych w języku angielskim i polskim: „analiza sentymentu” oraz „analiza wydźwięku”. Zazwyczaj oba terminy bywają używane jako synonimy w odniesieniu do analizy, której celem jest określenie natężenia wyrażonych emocji (mimo że można wskazać różnice w ich polu semantycznym). Warto także zaznaczyć, że analiza wydźwięku nie jest tym samym co analiza zmierzająca do klasyfikacji emocji. O ile ta pierwsza zazwyczaj ma na celu określenie natężenia emocji wyrażonych przez autora wypowiedzi (tu spotykamy skale porządkowe wskazujące na wypowiedzi wyrażające emocje³) i zakłada pewną istniejącą klasyfikację emocji (np. emocje pozytywne i negatywne), o tyle ta druga ma odmienny cel. Jest nim identyfikacja typów emocji wyrażonych w tekście oraz klasyfikacja tekstu lub jego części za pomocą rodzajów emocji zaczerpniętych z przyjętej na potrzeby badań typologii emocji. W prezentowanej tu analizie została zastosowana skala ciągła, ponieważ pozwala ona właśnie na obserwacje tytułowych „subtelności” w ocenie wydźwięku wypowiedzi.

3 Na przykład: pozytywne – negatywne, pozytywne – neutralne – negatywne czy skale ciągłe, określające wypowiedź za pomocą liczby z danego przedziału, np. –1 do 1, 0 do 1, i inne.

Nota metodologiczna

Prezentowane poniżej analizy zostały wykonane na danych udostępnionych przez firmę, która poprosiła o konsultacje związane z wyborem narzędzia pozwalającego na analizę sentymentu. Analizie poddanych zostało 2214 wypowiedzi klientów, którzy korzystali z usług firmy zgłaszającej potrzebę konsultacji. Wypowiedzi te zostały zweryfikowane pod kątem poprawności zapisu (poprawione zostały tak zwane literówki). Analizy wykonane przez badaczy (koderów, czytających i oceniających wydźwięk tekstów) obejmowały próbkę, na którą złożyło się 100 opinii⁴.

W badaniach jakościowych, a także w eksperymencie o takim charakterze dobór badanych przypadków spełnia zazwyczaj kryterium wysycenia (nasylenia) próby badawczej (ang. *the principle of saturation* – Rogers, Knafl, 2000; Saunders i in., 2018). Kryterium to rozumiane jest na kilka sposobów, tak więc i logika doboru próby może przebiegać odmiennie, w zależności od tego, na co kładzie się nacisk w prowadzonych badaniach (pytania badawcze), jakie przyjmuje się w nich metody badawcze (rodzaj badania) oraz ramy teoretyczne leżące u podstaw badania (Faulkner, Trotter, 2017). W przypadku prezentowanych tu wyników eksperymentu mamy do czynienia z kilkoma typami wysycenia. Wysycenie teoretyczne (Glaser, Strauss, 1967) oznacza, że do próby trafiają wypowiedzi z różnorodną wstępną oceną sentymentu (pozytywne, negatywne, neutralne)⁵, co wydaje się kryterium wystarczającym do zaspokojenia wysycenia próby ze względu na różnorodność zawartego w nich sentymentu. W późniejszych ocenach dokonanych przez koderów, słownik oraz SI mamy tego potwierdzenie w tekstach, dla których były one znacząco różne (-0,2 i 0,5) – miały wysoką ocenę wydźwięku (1 czy 0,92), niską (-0,9) oraz neutralną czy niemalże neutralną (0,0 lub 0,1). Wszystkie te przypadki zostaną zaprezentowane w dalszej części artykułu. Wysycenie danymi (Guest, Bunce, Johnson, 2006) dotyczy momentu związanego z gromadzeniem i analizą danych, w którym nowe dane lub informacje nie przyczyniają się znacząco do odpowiedzi na pytanie badawcze albo gdy istniejące dane są replikowane (w przypadku prezentowanego tu eksperymentu wystarczające dla pokazania różnic między wynikami trzech metod analizy jest zebranie próbek, w której obserwujemy wyniki

4 Opinie wybrane do oceny przez badaczy zostały wylosowane z dziesięciu różnych grup. Z kolei proces grupowania wypowiedzi opierał się na długości opinii (od jedno- i dwusłownych do wieloakapitowych). W każdej z grup wypowiedzi oznaczane były w bardzo ogólny sposób – jako pozytywne, negatywne lub neutralne. Ten zabieg pozwolił na późniejszy prostszy wybór wypowiedzi do końcowej analizy, w której ocenie wydźwięku poddane zostały wypowiedzi ze względu na ich różnorodny ton i długość. Na potrzeby tego artykułu prezentowanych jest zaledwie kilka wypowiedzi, co jest spowodowane głównie złożonością tematyki, potrzebą interpretacji obserwowanych wyników i różnic między nimi oraz długością tekstu. W fazie preprocessingu (oceny jakości, transformacji i czyszczenia zbioru tekstów) wyeliminowane zostały ciągi znaków niemające znaczenia analitycznego, czyli będące ciągami przypadkowych znaków, takich jak na przykład „dasdadadsadasdasda”, „oi[oih[uwepuiefgwiq” itp.

5 Ponieważ prezentowana tu analiza ma służyć raczej jako przyczynek do podjęcia szerszej dyskusji na temat różnic i podobieństw w analizach jakościowych wykonanych na tekście przez badacza i maszynę, nie zostało tu uwzględnione kryterium wysycenia teoretycznego, które jest definiowane inaczej na potrzeby badań służących budowie teorii (Bryant, Charmaz, 2007).

negatywne, pozytywne i neutralne – w szczególności dlatego, że tu prezentowana analiza dotyczy tylko sentymentu, a nie tematyki wypowiedzi⁶).

Za pomocą metod automatycznie analizujących tekst i automatycznie przypisujących tekstowi sentyment opracowane zostały wszystkie wypowiedzi (cały korpus – 2214). Jednostką analizy jest tu wypowiedź. Warto wspomnieć, iż nawet wypowiedź jednozdaniowa może zawierać słowa wyrażające sprzeczne emocje. Z kolei wypowiedź wielozdaniowa sprzeczność taką zawierać może na poziomie zdania i pomiędzy zdaniami. W związku z powyższym analiza sentymentu dłuższych wypowiedzi może dawać wyniki nierzetelne, a co za tym idzie – mylące⁸.

Na potrzeby tego artykułu wybranych zostało kilka opinii, które wydają się spełniać kryteria reprezentatywnych przykładów różnorodnych wypowiedzi ocenianych odmiennie w ramach każdego z omawianych tu podejść. Analizie poddamy zatem opinie: neutralne, takie, w których mamy do czynienia z silnymi pozytywnymi i silnymi negatywnymi emocjami (wypowiedzi takie zawierają oba wydzźwięki – pozytywny, negatywny, które ostatecznie się znoszą, czyli sumują do zera i dają w efekcie wypowiedź neutralną lub też nie wyrażają emocji) i takie, w których mamy zawarte emocje biegunowo różne (pozytywne, negatywne). Strategia prezentacji wyników opiera się z jednej strony na pokazaniu wypowiedzi ocenianych odrębnie w ramach każdej z metod, z drugiej zaś na wskazywaniu różnic między wynikami osiąganymi za pomocą tych metod na tych samych wypowiedziach (zestawienie ocen koderów ze słownikiem klasyfikacyjnym; zestawienie ocen za pomocą słownika i SI; zestawienie ocen SI i ocen koderów oraz zestawienie ocen uzyskanych za pomocą koderów – słownika – SI).

Rozpocznię od analizy pracy dokonanej przez badaczy (koderów).

6 Analiza sentymentu bardzo często łączona jest z analizą tematyczną, a to dlatego, że oba te podejścia w praktyce zyskują dodatkową użyteczność, gdy wyniki takiej analizy wykorzystywane są do podejmowania decyzji związanych np. z jakością dostarczanych usług, produktów czy też wtedy, gdy chcemy pogłębić swoje rozumienie ocen dostarczanych przez respondentów badania.

7 Ze względu na charakter prezentowanych tu badań nie ma potrzeby opisywania kolejnych rodzajów wysycenia, takich jak wysycenie kodami czy wysycenie znaczenia analizowanego materiału, które byłyby istotne w przypadku analiz tematycznych.

8 Oczywiście nawet w wypowiedzi jednozdaniowej (np. w zdaniu wielokrotnie złożonym) mamy do czynienia z wieloma jej składowymi. W przypadku prezentowanego tu opracowania rozwiązanie, jakim się posłużono, jest proste. Ocenie podlegają całe wypowiedzi (najdłuższa nie przekraczała jednego akapitu – zawierała kilka zdań mieszczących się w 60–70 słowach). Wypowiedzi takie mogą zostać uznane za długie, dlatego ocena całościowego ich wydzźwięku może być obciążona błędem, w związku z czym jej trafność warto weryfikować, mając na uwadze hipotezę, iż im dłuższa wypowiedź, tym większy prawdopodobny błąd oceny jej całościowego wydzźwięku. Przed niektórymi pułapkami związanymi z nierzetelnością oceny dłuższych wypowiedzi można się zabezpieczyć – jednak tylko do pewnego stopnia (np. poprzez ważenie oceny poszczególnych elementów wypowiedzi wyrażających emocje skrajne czy ważenie składowych wypowiedzi jako znaczących z perspektywy celu analizy, budowę szczegółowej instrukcji kodowania i prowadzenie szczegółowego szkolenia koderów, wielokrotne wykonanie oceny zgodności kodowania). Decyzji tego typu jest oczywiście więcej, ale nie są one przedmiotem badań prezentowanych w tym artykule, dlatego też pominięte zostały tu na przykład omówienie instrukcji kodowania, proces uczenia koderów, proces oceny zgodności kodowania czy iteracyjna praca nad zmianą instrukcji kodowania.

Proces kodowania przeprowadzony przez koderów

Opis podejścia

Proces kodowania tekstów jest najstarszą i najlepiej opracowaną z przedstawionych tu metod (Glaser, Strauss, 1967; Strauss, Corbin, 1998; Bernard, Wutich, Ryan, 2017; Williams, Moser, 2019). Kodowania, przeprowadzanego zwykle przez badaczy i analityków danych jakościowych⁹, wymaga analiza sentymentu, będąca jednym z wielu zadań wykonywanych w trakcie analizy tekstu. Współcześnie analizy tego rodzaju są bardzo często wykonywane w sposób całkowicie zautomatyzowany albo półautomatyczny. Wyniki takiej automatyzacji, w szczególności w obszarze związanym z analizą sentymentu, nie dają jeszcze (Barrett i in., 2019; Németh, Koltai, 2021; Munnes i in., 2022) wyników na tyle rzetelnych, aby analiza wydźwięku wypowiedzi mogła być wykonywana w pełni automatycznie (bez udziału koderów) (Keiser, Presmeg, 2019). W związku z powyższym kodowanie wykonane przez badaczy jest tu pierwszą z zastosowanych metod analizy, a to dlatego, że (w momencie pisania tego akapitu, a więc przed postawieniem ostatniej kropki w tym artykule) jest kodowaniem wykonywanym w sposób najbardziej rzetelny. W prezentowanym tu eksperymencie badawczym na jego pierwszym etapie koderzy poproszeni zostali o ocenę tekstów ze względu na wydźwięk emocjonalny, jaki teksty te w sobie zawierają¹⁰.

Przykładowe wyniki

Spójrzmy na oceny wypowiedzi tekstowych dokonane przez koderów (praca była wykonana na skali -1, 1 – która, dla ułatwienia porównania uzyskanych wyników, była taka sama w ramach każdej ze stosowanych metod). Pierwszy przykład stanowi krótka i stosunkowo prosta do oceny opinia, która jest odpowiedzią na prośbę: „Czy może Pan/Pani podzielić się swoimi spostrzeżeniami, refleksjami na temat usługi, która została zrealizowana przez firmę XXX?”.

I would rather not write this and prefer for someone to call me.

Ocena koderów¹¹: -0,2, 0, 0,5

⁹ Dla uproszczenia tego wywodu analityków i badaczy zajmujących się kodowaniem tekstów będę nazywał „koderami”. W literaturze przedmiotu, w szczególności anglojęzycznej, dodaje się do niego także dookreślenie *human*. W opracowaniach naukowych mamy więc do czynienia z frazą *human coder*, którą tu parafrazuję, używając jedynie drugiego członu.

¹⁰ Analiza sentymentu wykonywana przez koderów z kilku względów nie jest prostym zadaniem. O ile bowiem jesteśmy w stanie podać prostą instrukcję brzmiącą np. „Oznacz opinię pozytywną, negatywną, neutralną”, to trudniej precyzyjnie scharakteryzować, jaki poziom natężenia emocji towarzyszy odbiorowi danego tekstu (a to dlatego, że odbiór jest kwestią subiektywnej oceny czytelnika, a zbudowanie przez koderów siatki pojęciowej, metody, która pozwoli mu zbliżyć się do zobiektywizowanej oceny wydźwięku, staje się wymagającym zadaniem), a jeszcze trudniej przyporządkować danemu tekstowi wartość liczbową, która miałaby reprezentować poziom badanego wydźwięku. Problem kwantyfikowania wypowiedzi ze względu na jej wydźwięk nie jest tu akurat kluczowy, dlatego też nie będzie dalej rozważany, choć w pewnym sensie będzie wpływał na wnioski, jakie pojawiają się w końcowej części pracy.

¹¹ Tu podaję trzy typowe, najczęściej pojawiające się oceny. W pozostałych przykładach podaję wartość środkową spośród wszystkich ocen.

Wydaje się, że cytowana powyżej wypowiedź nie wyraża wprost ani negatywnego, ani pozytywnego nastawienia autora/autorki do tematu (dzielenie się opinią na temat usługi, z jakiej skorzystał/skorzystała). Niemniej jednak wypowiedź ta, oceniana przez kilku różnych koderów uzyskała diametralnie różne wskazania: -0,2, 0, 0,5. Rozmowa z koderami po nadaniu przez nich ocen poszczególnym wypowiedziom pozwoliła na wyjaśnienie zaobserwowanej różnicy. Wynika ona z odmiennych założeń przyjmowanych przez koderów, dotyczących napotkanej w tekście deklaracji. Pierwsze jest przypuszczeniem, że klient nie był zadowolony ze współpracy i chciał przekazać swoje negatywne opinie, ale w bardziej bezpośredni sposób niż pisemnie. Drugie (i było to przypuszczenie najczęściej występujące w wypowiedziach koderów) to przekonanie o postawie neutralnej wobec usługodawcy, któremu towarzyszyło przeświadczenie, że rozmowa o jakości usług jest najlepszą formą kontaktu. Trzecia (najwyższa) ocena (0,5) pojawiła się dwukrotnie i wynikała z dwóch różnych przekonań. Pierwsze to przeświadczenie, że klient chciał poczuć satysfakcję i przyjemność z przekazania osobiście oceny firmie XXX. Drugie było próbą zaznaczenia wagi komentarza, a nie oceną zawartych w nim emocji. W kontekście dbania o relacje z klientem to uzasadnienie badacza ma praktyczne istotne znaczenie, ponieważ zwraca uwagę na wyrażoną potrzebę kontaktu.

Wnioski płynące z rozmów z koderami oraz z wykonanej przez nich analizy mogą prowadzić do różnorodnych wytycznych dotyczących samego procesu kodowania, takich jak: potrzeba bardziej precyzyjnego definiowania instrukcji kodowania, modyfikacja procedur ewaluacji wykonanego kodowania, zmiana wytycznych sposobu i formy raportowania wyników.

Przyjrzyjmy się innym przykładom. Kolejne dwie wypowiedzi są dłuższe i zawierają więcej informacji istotnych z punktu widzenia analizy sentymentu:

I always have had good interaction with the consultants, and they have always been helpful or eager to help, honest with their feedback and opinions.

Ocena koderów: 0,9

I have been consulted with you for the last 2 years no one from company XXX has even approached me or not even been to the site I have worked in the XXX industry for over 40 years this is the worst I have ever seen it, no pay rise for over 2 years even through covid I went to work and then I work for 3 contracting firms why? they all have a cut of my wages in the last 3 years I have lost over XXX money working for the company XXX is poor as I am, the good old age of the consultant of the XXX.

Ocena koderów: -0,8

W pierwszej opinii znajdujemy słowa wskazujące wyraźnie na zadowolenie autora („zawsze byli pomocni”, „uczciwi”). W drugim przykładzie mamy z kolei do czynienia z opinią wyraźnie negatywną, na co wskazują takie fragmenty, jak: „to najgorsze doświadczenie, jakie miałem”, „brak podwyżki przez 2 lata”, „strata pewnej kwoty”. Nie dziwią zatem dwie skrajne oceny wydźwięku, ponieważ dość łatwo wyjaśnić je w odwołaniu do sformułowań, jakie są w tych opiniach zawarte.

Ostatni z przykładów jest odmienny od poprzednich z jednego powodu – istnieje w nim spójna negatywna ocena współpracy z firmą XXX, która wyrażona jest w słowach zapisanych do przedostatniego przecinka. Po nim pojawia się dodatkowy, odmienny w swoim wydźwięku komentarz dotyczący nastawienia pracownika firmy:

There was no recognition or understanding that company XXX already had all my information and AAA checks from my last assignment and so being required to re-submit and complete everything again was a total waste of everyone’s time, resulting in unnecessary delays and my loss of earnings, which in fairness, person XXX sympathized with.

Ocena koderów: -0,625

Z rozmowy z badaczami oceniającymi wypowiedź wynika, że ma ona negatywny wydźwięk. I choć w ostatnich słowach autor wypowiedzi zauważa, że pracownik firmy dostrzegł wspomniane problemy i wyraził swoje współczucie, nie zmienia to oceny wydźwięku przydzielonej przez koderów. Przykład tej oceny pokazuje, że wartość liczbowa przypisywana wypowiedzi zależy nie tylko od zawartości słownej analizowanego tekstu, ale w dużej mierze od tego, jak skonstruowana jest instrukcja kodowania tekstu, a także od intuicji i przekonań kodera. Mamy tu zatem pewien element niedookreślenia, subiektywizmu czy też predyspozycję do oceny wypowiedzi, zapośredniczoną z doświadczeń i przekonań konkretnego analityka. Spójrzmy zatem, czego dostarczyć mogą nam inne metody analizy sentymentu, a zwłaszcza na to, co istotnie różni je od kodowania wykonanego przez badacza. Być może jest to właśnie ten element intersubiektywny, który jest związany z intersubiektywną kontrolowalnością.

Słownik zbudowany na regułach wyszukiwania słów i fraz w kontekście, w jakim one występują oraz na ocenie wydźwięku wypowiedzi

Opis podejścia

Metody rozpoznawania informacji i oceny wiedzy, a także oceny wypowiedzi (czy to ze względu na jej formę, czy też na temat lub wydźwięk – Liu, 2015) wykazują swoją skuteczność w pracy z tekstami pisanymi do tego stopnia, że doczekały się wielu zautomatyzowanych rozwiązań. Jedno z nich to słowniki klasyfikacyjne. Są to metody, które przekute zostały w narzędzia wspomagające proces analizy jakościowej, między innymi poprzez identyfikację zarówno prostych (takich jak słowa czy frazy), jak i złożonych fragmentów tekstów osadzonych w różnorodnych kontekstach językowych (Tomanek, Bryda, 2015). Bardziej zaawansowane słowniki pozwalają nie tylko na wyszukiwanie informacji, ale także na ocenę znaczenia, sensu wypowiedzi w zależności od słów współwystępujących w swoim sąsiedztwie (czyli od kontekstu) czy też właśnie na ocenę wydźwięku. Najogólniej można powiedzieć, iż słownik tego rodzaju bazuje na następujących regułach:

- 1) semantycznych, pozwalających na rozpoznawanie wypowiedzi podobnych ze względu na znaczenie, jakie mają, lub biegunowo różnych znaczeniowo;

- 2) odnoszących się do pragmatyki wypowiedzi, dzięki którym możemy rozpoznać jej kontekst, a w konsekwencji podmiot i przedmiot;
- 3) składniowych, pozwalających na rozpoznanie miejsca części mowy w analizowanej wypowiedzi (Tomanek, Bryda, 2014; Tomanek, 2014a);
- 4) logicznych (wykorzystujących reguły przywołane w punktach 1–3), których zadaniem jest zestawianie poszczególnych fragmentów wypowiedzi ze sobą, tak aby spełnione były warunki związane ze znaczeniem wypowiedzi;
- 5) statystycznych (także wykorzystujących reguły z punktów 1–3), pozwalających na poszukiwanie wypowiedzi do pewnego stopnia podobnych, do pewnego stopnia spełniających kryteria, jakie nakładamy na analizę jakościową tekstu;
- 6) pozwalających na przypisywanie do wyszukanych informacji, fragmentów tekstu punktów reprezentujących np. wagę (ważność) danej informacji z perspektywy celu analitycznego albo siły wydzźwięku (siły emocji) wyrażonej w danej wypowiedzi.

W niniejszym artykule zastosowane zostały dwa istniejące już i wykorzystywane w praktyce analiz tekstowych słowniki, to jest: słownik SentiwordNet (SWN) (Baccianella, Esuli, Sebastiani, 2010) i słownik sentymentu VADER (Valence Aware Dictionary and sEntiment Reasoner – uwzględniający walencję i wnioskowanie o poziomie wydzźwięku emocjonalnego na podstawie zdania, w którym występuje słowo czy fraza wyrażające emocje) (Hutto, Gilbert, 2014). Słowniki te różnią się od siebie, a jednocześnie, zwłaszcza z punktu widzenia przyjętych rozwiązań, uzupełniają się i dlatego posłużyły do zbudowania nowego hybrydowego rozwiązania służącego do analizy sentymentu¹². Ta hybryda wymagała z jednej strony zmian wag przypisywanych słowom wyrażającym emocje¹³, z drugiej połączenia reguł słownikowych zawartych w obu słownikach¹⁴. Warto tu nadmienić, że taki słownik określający wydzźwięk wypowiedzi różni się znacząco od następnego opisanego tu rozwiązania, czyli SI. Główna różnica między tymi rozwiązaniami polega na tym, że słownik jest narzędziem w pełni transparentnym – czyli pozwalającym na przeczytanie, zrozumienie, zmianę wszystkich reguł, które w słowniku zostały zdefiniowane i decydują o nadaniu określonego wydzźwięku wypowiedzi pisanej. Tymczasem SI takiej możliwości nie daje. Inna istotna różnica między tymi podejściami polega na walidacji i uczeniu metody nowych rozwiązań. Kiedy słownik generuje wyniki, których ewaluacja wskazuje na konieczność zmiany reguł w nim zdefiniowanych, modyfikacji wymagają wszystkie

12 Za współpracę przy projektowaniu tego rozwiązania, za jego oprogramowanie i wdrożenie chciałbym podziękować Bartoszowi Elmerychowi.

13 SentiWordNet (SWN) oraz VADER (VA) przypisują różne wartości liczbowe słowom wyrażającym emocje. W związku z tym zaszła konieczność wypracowania rozwiązania uspojnającego ocenę. SentiWordNet przyjmuje wartości ze skali od 0 do 4, gdzie 0 oznacza słowo nacechowane silnie negatywnie, a 4 słowo nacechowane silnie pozytywnie. VADER przyjmuje natomiast wartości ze skali od -4 do 4, gdzie -4 to słowo wyrażające silnie negatywne emocje, a 4 to słowo wyrażające silnie pozytywne emocje. Hybrydowe rozwiązanie łączące dwa słowniki przypisuje dla uproszczenia wartości ze skali -1 do 1, gdzie -1 to wartość przypisywana do silnie negatywnej wypowiedzi, a 1 do wypowiedzi o silnym pozytywnym wydzźwięku.

14 SentiWordNet posiada reguły pozwalające rozpoznawać zależności leksykalne, a dzięki nim polaryzację sentymentu znaczeń leksykalnych, a nie tylko słów czy fraz (Kocoń, Janz, Piasecki, 2018). VADER z kolei jest metodą wykorzystującą podejście charakterystyczne dla słownika sentymentu (zawiera słowa i frazy wraz z przypisanymi do nich wartościami liczbowymi odpowiadającymi sile emocji zawartych w wypowiedziach), ale też zawiera reguły gramatyczne i konwencje składniowe do wyrażania polaryzacji i intensywności sentymentu.

fragmenty kodu będące w relacji z modyfikowaną regułą (proces ten może być długi). Tymczasem zmiany w mechanizmie SI nie wymagają albo – precyzyjniej – nie zawsze wymagają zmian w kodzie (pomijamy tu zagadnienie bramek logicznych), ponieważ zmiany te sieć neuronowa może „zinternalizować” na podstawie podanych jej reguł pisanych w języku naturalnym, wzbogaconych przykładami podanymi w tymże języku. Spójrzmy na rezultaty pracy z takim hybrydowym rozwiązaniem (słownikiem).

Przykładowe wyniki

Wyniki pracy słownika poddane zostaną interpretacji, skupiając się na wartościach liczbowych reprezentujących sentyment. Postaram się jednocześnie wyjaśnić różnice w ocenach. Dla pełniejszego obrazu przywołane zostaną wypowiedzi z wysokimi, przeciętnymi oraz niskimi ocenami na skali od -1 do 1.

I always have had good interaction with the consultants, and they have always been helpful or eager to help, honest with their feedback and opinions.

Ocena uzyskana przy użyciu słownika klasyfikacyjnego: 0,922

Consultant A and Consultant B from Company XXX¹⁵ have been absolutely fantastic – approachable, helpful, and so efficient.

Ocena uzyskana przy użyciu słownika klasyfikacyjnego: 0,876

Powyższe wypowiedzi uzyskały wysokie wyniki. W całym analizowanym zbiorze najwyższą wartość uzyskała pierwsza z przywołanych tu opinii z przypisanym do niej wynikiem 0,922.

Różnica między dwoma powyżej przywołanymi wysokimi wynikami pozytywnymi (0,922, 0,876) może być wyjaśniona w sposób następujący. Poprzez dwukrotne użycie słowa „zawsze” (ang. *always*) oraz słowa „uczciwi” (ang. *honest*) pierwsza z wypowiedzi nabiera bardzo silnego pozytywnego wydźwięku. Oba słowa oceniane są wysoko na skali z dwóch różnych powodów. Pierwsze jest wyjątkowo silnym wskaźnikiem oceny postawy, ponieważ identyfikuje i traktuje ją jako bardzo pozytywną (w przywołanej wypowiedzi jest to relacja nawiązana między konkretnym konsultantem i konkretnym klientem). Innymi słowy, wypowiedź, w której pojawia się tak zwany kwantyfikator uniwersalny (np. „zawsze”, „wszystko”), będzie miała dużo wyższą ocenę wydźwięku niż wypowiedź, w której pojawiają się kwantyfikatory egzystencjalne (np. „czasami”, „niektóre”). Drugie słowo („uczciwość”) również podnosi wartość oceny końcowej, będąc oceną silniejszą niż ta wyrażana w drugiej wypowiedzi. Oczywiście w drugiej wypowiedzi pojawiają się również bardzo silnie pozytywnie słowa akcentujące jakość relacji. Są nimi: „absolutnie” (ang. *absolutely*), „fantastyczni” (ang. *fantastic*), „przystępni” (ang. *approachable*), „pomocni” (ang. *helpful*) i „skuteczni” (ang. *efficient*).

15 Na potrzeby tego artykułu anonimizowane są zarówno nazwy firm, jak i imiona i nazwiska konsultantów współpracujących z klientami.

Interesujące analitycznie są wypowiedzi znajdujące się w środku skali (oceny w okolicach 0). Mogą one wyrażać różnorodne nastawienie autorów wypowiedzi – od neutralnego, poprzez zachowawcze, do ambiwalentnego (łączącego w sobie na przykład dwie skrajne oceny, których wartości się znoszą). Spójrzmy na przykłady:

I would rather not write this and prefer for someone to call me.

Ocena uzyskana przy użyciu słownika klasyfikacyjnego: 0,0

Seamless and pain-free process, although needs more human interaction.

Ocena uzyskana przy użyciu słownika klasyfikacyjnego: 0,1

Ciekawym przykładem rozpoznania braku emocji leżących u podstaw wypowiedzi jest pierwsza przywołana opinia z wynikiem 0,0 (interpretacja tej wypowiedzi jest zawarta w poprzednim podrozdziale). Słownik nie wychwytuje tu żadnych emocji, ponieważ wypowiedź nie zawiera słów, które by je wyrażały *explicite*. Tymczasem koderzy nie byli jednomyślni w swoich ocenach, co znacząco różni metodę słownikową od kodowania dokonanego przez badaczy, analityków tekstów.

Następna z przywołanych wypowiedzi dotyczy dwóch wątków, które oceniane są różnorodnie. Z jednej strony widzimy, że ocena współpracy wskazuje na jej pozytywny, „bezbolesny” (ang. *pain-free*) przebieg, z drugiej natomiast, że zabrakło w niej „kontaktu z ludźmi” (ang. *human interaction*). W takiej sytuacji niejednoznaczności ocen w ramach jednego zdania górę biorą oceny nadawane pojedynczym słowom. Słowa „bezbolesny” i „sensowny” zyskują wyższe oceny niż sformułowanie mówiące o braku interakcji międzyludzkich. I tu dotykamy sedna pracy z takimi rozwiązaniami jak słownik klasyfikacyjny, dlatego że wszelkie nadawane w tym podejściu oceny są wynikiem z jednej strony wycucia autora słownika, a z drugiej precyzji sformułowanych problemów badawczych, które następnie tłumaczone są na reguły słownikowe¹⁶. W przypadku rozważanej tu wypowiedzi słownik nie wzmocnił negatywnej oceny wyrażonej w drugiej części zdania, która przeważałaby ocenę sumaryczną (0,1). W takiej sytuacji (a miałyby ona miejsce na etapie testowania słownika) wymagany byłby

16 Reguły słownikowe to operacyjne definicje, które dookreślają kryteria służące do wyszukiwania konkretnych informacji. Przybliżę je, podając najprostszy przykład takiej reguły: 1. Celem analizy jest wyszukiwanie pozytywnych opinii dotyczących filmu XXX. 2. Określone zostały słowa i frazy, które uznane zostały za pozytywne wyrazy opinii (jest to zazwyczaj bardzo długa lista słów, fraz, takich jak: „dobry”, „ciekawny”, „fantastyczny”, „interesujący”, „wciągający”, „wyborny”, „niesamowicie skupiający uwagę” itp.). 3. Określone zostały przeczenia, które zmieniają znaczenie uznanych za pozytywne słów, fraz (są nimi zazwyczaj: „nie”, „nieprawdziwe”, „niekoniecznie”, „nie znaczy to, że”, „nie jest tak, że” itp.). 4. Zdefiniowane zostały reguły wyszukiwania treści, które muszą spełniać warunki zawarte w punktach 2–3 (prostym przykładem niech będzie następująca reguła: jeśli słowo wyraża pozytywną opinię i zawarte jest w liście wspomnianej w punkcie 2, to jeśli nie występuje w pobliżu na przykład jednego, dwóch lub trzech słów wymienionych na liście z punktu 3, to najprawdopodobniej jest to pozytywna opinia dotycząca filmu XXX). Oczywiście współwystępowanie dwóch słów czy też dwóch fraz w bliskości określonej jako dwa do pięciu słów nie jest wystarczającym ani też zadowalającym kryterium uznania danej wypowiedzi za pozytywną czy negatywną. Dlatego też słowniki klasyfikacyjne wyposażone są w dodatkowe kryteria (poza logicznymi, których przykład zaprezentowałem we wcześniejszych zdaniach) pozwalające na analizę kontekstu, w którym pojawia się pozytywne słowo czy fraza, których wyszukiujemy. Te kryteria opisane są w tekście głównym, a ich operacyjne definicje są wyrażone w języku programowania, w którym słownik jest zbudowany. Na potrzeby tego artykułu przyjmuję, że prezentacja kodu (w przypadku słowników klasyfikacyjnych często wielu setek linii kodu w językach: SQL i Python), który stosowany jest do osiągnięcia wyników, nie jest potrzebna.

powrót do etapu definiowania problemów badawczych i jeśli jednym z nich był ten, który wymagał zwrócenia uwagi na kwestie interakcji międzyludzkich, słownik powinien być zmodyfikowany tak, aby wypowiedzi, które tego zagadnienia dotyczą, były oceniane w sposób silniejszy (w obu kierunkach: pozytywnym i negatywnym).

Spójrzmy jeszcze raz na dwa przykłady, tym razem uznane za wyraz opinii negatywnych, i przeanalizujemy zastosowane tu kryteria oceny wydźwięku:

I have been consulted with you for the last 2 years no one from company XXX has even approached me or not even been to the site I have worked in the XXX industry for over 40 years this is the worst I have ever seen it, no pay rise for over 2 years even through covid I went to work and then I work for 3 contracting firms why? they all have a cut of my wages in the last 3 years I have lost over XXX money working for the company XXX is poor as I am, the good old age of the consultant of the XXX.

Ocena uzyskana przy użyciu słownika klasyfikacyjnego: -0,837

There was no recognition or understanding that company XXX already had all my information and AAA checks from my last assignment and so being required to re-submit and complete everything again was a total waste of everyone's time, resulting in unnecessary delays and my loss of earnings, which in fairness, person XXX sympathized with.

Ocena uzyskana przy użyciu słownika klasyfikacyjnego: -0,625

W pierwszej wypowiedzi mamy do czynienia z wyjątkowo silną negatywną oceną, której wynik jednak łatwo możemy wyjaśnić. Wypowiedź ta zawiera bowiem wyrażone wprost jednoznaczne oceny. Pojawiają się tu frazy, które słownik klasyfikacyjny zawiera: „przez ostatnie 2 lata nikt z firmy XXX nawet się do mnie nie zwrócił ani nawet nie był na miejscu, w którym pracowałem” („for the last 2 years no one from company XXX has even approached me or not even been to the site I have worked”); „to najgorsze z wszystkich moich doświadczeń, i do tego bez podwyżki przez ponad 2 lata” („this is the worst I have ever seen it, no pay rise for over 2 years”); „wszyscy doświadczają obniżki zarobków w ciągu ostatnich 3, a ja straciłem ponad XXX pieniędzy” („they all have a cut of my wages in the last 3 years I have lost over XXX money”). Wszystkie wskazane tu elementy wypowiedzi dają w rezultacie silnie negatywną ocenę. Składają się na nie nie tylko takie słowa jak „straciłem” czy „cięcie” zarobków, ale też użycie wcześniej już akcentowanego kwantyfikatora „wszystkie”.

W drugim przykładzie mamy również wypowiedzianą wprost, stosunkowo łatwą do uchwycenia ocenę. Wyraża się ona w takich sformułowaniach jak: „brak rozpoznania i zrozumienia” („no recognition or understanding”), „musałem złożyć wszystkie dokumenty ponownie” („so being required to re-submit and complete everything again”), „całkowita strata czasu wszystkich osób, której wynikiem były niepotrzebne opóźnienia i strata moich zarobków” („a total waste of everyone's time, resulting in unnecessary delays and my loss of earnings”). Wskazane tu elementy wypowiedzi wydają się utwierdzać co do trafności nadanej jej negatywnej oceny.

Niemniej jednak metoda słownikowa jest do pewnego stopnia zawodna. Mianowicie w dłuższych wypowiedziach (a mamy tu do czynienia z zaledwie kilkoma linijkami tekstu) nie gwarantuje rozpoznania ambiwalencji tak, aby brana była ona pod uwagę w całościowej ocenie. Dzieje się tak z dwóch powodów. Po pierwsze, rozpoznanie ambiwalencji wymaga trafnie wykonanej analizy, w której dokładnie wiadomo, gdzie w wypowiedzi kończy się jeden z jej wątków, a zaczyna drugi. Zadanie to jest niezmiernie skomplikowane programistycznie, ponieważ wymaga nauczenia maszyny rozpoznawania wątków wypowiedzi nie tylko na bazie słów czy fraz kluczowych i relacji między nimi, ale także na podstawie odmienności od tych, które znajdują się na przykład w kolejnym zdaniu w tej samej wypowiedzi. Taki słownik rozpoznawać musiałby różne wątki wypowiedzi, relacje między tymi wątkami, tak aby bezbłędnie dzielić wypowiedź i na podstawie takiego podziału nadawać poszczególnym fragmentom osobne oceny. I ta sformułowana w poprzednim zdaniu obserwacja prowadzi do kolejnego problematycznego słownikowego zagadnienia. Jest nim nadawanie oceny całościowej jednej wypowiedzi. Słownik zazwyczaj projektowany jest tak, aby agregował oceny przyznane poszczególnym zdaniom i wyrażał ocenę wypowiedzi jako wypadkową ocen przyznanych zdaniom w niej wyrażonym. Jak widzieliśmy w przypadku wypowiedzi zawierających skrajnie różne oceny już w ramach jednego zdania, ocena ta bywa nieadekwatna do treści.

Po drugie, metoda oparta na zastosowaniu słownika (w przyjętej tu, choć jednej z wielu możliwych, liczbowej charakterystyce) nie dostarcza również określenia (czy dokładniej: nazwania) emocji¹⁷, które w tekście zostały wyrażone *implicite*. Spójrzmy zatem, jak różni się analizowane tu podejście do tego oferowanego przez sztuczną inteligencję.

Sztuczna inteligencja ocenia wydźwięk emocjonalny wypowiedzi

Opis podejścia

Trzecia z metod, której się przyjrzymy, to sztuczna inteligencja (SI). Jej działanie oparte jest na zaprojektowanej na potrzeby pracy z językiem naturalnym sieci neuronowej, a dokładniej wielu sieciach neuronowych, których wyniki pracy widoczne są podczas interakcji z chatbotem¹⁸ zwanym ChatGPT.

Stworzony przez firmę OpenAI ChatGPT (ang. *generative pre-training transformer*) to w prostym tłumaczeniu wstępnie wyszkolony algorytm (a właściwie zbiór algorytmów) służący do pracy z językiem naturalnym, językami formalnymi, językami programowania. Jest to właśnie przykład rozwiązania opartego na pracy sieci neuronowych oraz innych metodach uczenia maszynowego.

17 Istnieją metody oparte na podejściu słownikowym, które dookreślają rodzaj emocji wyrażanych w wypowiedzi (por. Tomanek, 2014a; 2014b).

18 Chatbot symuluje rozmowę, przetwarza jej wynik (zarówno w postaci tekstowej, jak i dźwiękowej) i potrafi pracować z językiem naturalnym tak, że możliwa jest interakcja człowieka z urządzeniem cyfrowym, której formą/medium jest tekst pisany lub mówiony. Chatbot jest programem komputerowym, którego symulacja rozmowy jest tym lepsza, im trudniej odgadnąć, czy rozmawiamy z człowiekiem, czy z maszyną, czyli chatbotem (Oracle Polska, b.r.).

Mechanizm ten został zbudowany tak, by potrafił prowadzić rozmowy, a w nich odpowiadać na pytania, odnosić się do przesłanych do niego wypowiedzi, a także uczyć się, gdy rozmówca podaje inne niż sam ChatGPT rozwiązanie. Jest jednocześnie w stanie analizować i do pewnego stopnia rozpoznawać kontekst przesyłanych mu wypowiedzi, tak aby „jego” odpowiedzi brały pod uwagę argumenty rozmówcy, były spójne logicznie (na tyle, na ile pozwala zasięg informacji, które dostarczone zostały chatbotowi). ChatGPT potrafi także doradzać, sugerować rozwiązania prezentowanych mu problemów. Sami autorzy tak piszą o procesie uczenia i celach, jakie przyświecały im podczas budowy chatbota:

Wytrenowaliśmy model o nazwie ChatGPT, który wchodzi w interakcję w sposób konwersacyjny. Format dialogu umożliwia ChatowiGPT odpowiadanie na pytania uzupełniające, przyznawanie się do błędów, kwestionowanie błędnych założeń i odrzucanie nieodpowiednich żądań. ChatGPT jest modelem bliźniaczym do InstructGPT, który jest szkolony do wykonywania instrukcji w podpowiedzi i udzielania szczegółowych odpowiedzi¹⁹.

ChatGPT posiada również informacje na temat analizy sentymentu i sam potrafi ocenić sentyment wypowiedzi. Łatwa dostępność modelu, jego eksperymentalny charakter, a także wszechstronność konwersacyjna²⁰ potwierdzona przed przystąpieniem do testów skłania do włączenia tego rozwiązania jako trzeciej metody, która posłuży nam do oceny wydźwięku analizowanych wypowiedzi.

Przykładowe wyniki

Omawiany tu eksperyment metodologiczny, w którym porównywane są wyniki działania trzech różnych metod, został także zaprezentowany SI w następującej postaci:

Witam, jestem pracownikiem naukowym uniwersytetu, który pracuje nad porównaniem różnych metod analizy sentymentu. Jedną z nich jest kodowanie tekstu przez człowieka i dostarczanie oceny sentymentu wypowiedzi pisanych, drugą jest automatyczna ocena wypowiedzi dokonana poprzez słownik oparty na regułach zbudowany przeze mnie, a trzecią byłaby sztuczna

19 Autorem tej wypowiedzi jest zespół firmy OpenAI. Model GPT-3 został opracowany przez tę właśnie organizację non profit, zajmującą się badaniami nad sztuczną inteligencją i jej aplikacjami w tworzeniu modeli języka, które są w stanie nauczyć się i rozumieć język ludzki (OpenAI, b.r.).

20 Sieć neuronowa, która jest odpowiedzialna za działanie ChatGPT, łączy słowa, frazy, dłuższe jednostki analityczne (wypowiedzi składające się z wielu zdań), rozpoznaje powiązania między nimi zarówno na podstawie ich statystycznie określonego współwystępowania, jak i dzięki wykorzystaniu logicznej reprezentacji syntaktyki języka (mówimy tu o spisanych i przekazanych sieci neuronowej regułach budowy i formalnej składni języka), zbiorów reguł syntaktycznych (gramatyka to zbiór reguł syntaktycznych pozwalający na identyfikowanie powiązań części zdania – np. na rozpoznanie podmiotu i orzeczenia w zdaniu, dzięki czemu sieć potrafi zidentyfikować, o kim lub o czym mówi dane zdanie oraz co to zdanie mówi o podmiocie i przedmiocie wypowiedzi), semantyki języka opisaną za pomocą języka logiki (tu dochodzi do przypisania znaczenia analizowanym wypowiedziom, uwzględniając ich społeczny i kulturowy kontekst, w jakim się pojawiają – to najtrudniejsza część pracy związanej z uczeniem modeli NLP). Por. Lake, Baroni, 2018; Hewitt, Manning, 2019; Ravichander i in., 2020.

inteligencja dostarczająca wyniku sentymentu dla krótkiego tekstu (jest to opinia). Czy chciałbyś wziąć udział w takim badaniu?²¹

Przywołuję tu zadane pytanie i podaną przez chatbota odpowiedź, aby pokazać sposób, w jaki SI rozpoznaje sytuację konwersacyjną, w jakiej została postawiona. Oto odpowiedź ChatGPT:

Jako model językowy AI nie mogę uczestniczyć w badaniach w tradycyjnym sensie, ponieważ nie jestem człowiekiem. Mogę jednak dostarczyć informacji i spostrzeżeń na temat analizy sentymentu i różnych metod, które wymieniłeś. Jeśli masz jakieś pytania lub potrzebujesz pomocy, chętnie pomogę.

Żeby ułatwić porównanie wyników uzyskanych za pomocą różnych metod oceny wydźwięku oraz by jednocześnie skonkretyzować refleksję nad uzyskanymi rezultatami, dałem modelowi SI do oceny wypowiedzi, których wyniki zostały już omówione w poprzednich podrozdziałach. Instrukcja (prompt) podana ChatGPT brzmiała następująco: „Would you be able to score the sentence on sentiment scale ranging from -1 to 1 where 1 means positive sentiment, and -1 negative sentiment?”. Każdy prompt stosowany do poddawanych ocenie wypowiedzi brzmiał tak samo. Kontroli zostały poddane także parametry promptu, które pozostawały niezmiennie dla każdego zapytania wysyłanego do ChatGPT²².

21 Ponieważ opisywana tu analiza sentymentu wykonywana była na tekstach w języku angielskim, uznałem, że rozmowę z SI będę prowadził w tym języku. Poniżej podaję oryginalne brzmienie wypowiedzi (swoich i tych udzielonych przez SI). AUTOR: „Hi, I'm university scholar who is working on comparing different methods for sentiment analysis. One is human coding text and providing the sentiment score, second is an automated scoring via rule-based dictionary build by me, and third would be artificial intelligence providing the sentiment score for short text (it is an opinion). Would you like to participate in the study?”.

ChatGPT: „As an AI language model, I cannot participate in studies in the traditional sense as I am not a human. However, I can provide information and insights on the topic of sentiment analysis and the different methods you have mentioned. If you have any questions or need any assistance, I'd be happy to help.”.

22 Najogólniej wszystkie dostępne parametry pozwalające na modyfikowanie promptów pozostawione były na poziomach uprawdopodobniających uzyskanie odpowiedzi bardziej przewidywalnych i spójnych. Kontrolowanymi parametrami promptów były: 1) temperatura odpowiedzialna za losowy czy kreatywny charakter wypowiedzi – tu wartość ustawiona została na poziomie 0, który pozwala na uzyskiwanie odpowiedzi łatwiej przewidywalnych i bezpośrednich; 2) maksymalna liczba tokenów to parametr odpowiedzialny za długość odpowiedzi – tu ustawiony został na poziomie 200 łącznie wejściowych i wyjściowych (zakładając, że 100 tokenów to mniej więcej 75 słów, co wystarcza dla wspomnianej wcześniej długości wypowiedzi, pozostaje około 100 na odpowiedź ChatGPT, który zasadniczo udzielał krótkich odpowiedzi na zadane mu – wspomniane wyżej – pytanie); 3) kontrola pozostałych trzech parametrów wyglądała następująco: Top P = 0 (niska wartość zmniejsza różnorodność wypowiedzi modelu), kara za częstotliwość = 0 (im mniejsza wartość, tym większa liczba powtórzeń), kara za obecność = 0 (niska wartość zniechęca do stosowania nowych pojęć i tematów). Oczywiście wskazane ustawienia parametrów można byłoby zmieniać dla jednego analizowanego tekstu. Eksperymentując ustawieniami parametrów, można byłoby zwiększyć „wirtuozerię” ChatGPT czy różnorodność interpretacji, jakich model ten dostarczy. W omawianym tu eksperymencie celem głównym było jednak zdobycie oceny sentymentu na skali liczbowej, a nie uzyskiwanie od ChatGPT słownych interpretacji ocenianych na skali sentymentu tekstów. Stąd też wspomniane ustawienia parametrów zdefiniowane zostały na poziomie uprawdopodobniającym powtarzalność i konsekwentny charakter ocen. Zdecydowanie inne efekty odniosłem, pracując z ChatGPT przy zagadnieniu klasyfikacji emocji czy analizie tematycznej – gdzie wyższe wartości, np. temperatury i Top P, zwiększały różnorodność interpretacji modelu. Warto też zaznaczyć, że nawet subtelna zmiana treści zastosowanego promptu prowadzi do odmiennych odpowiedzi ze względu na próbę zinterpretowania przez model intencji zawartej w wypowiedzi czy kontekstu pytania (promptu).

W pierwszej kolejności poprosiłem o ocenę wydźwięku następującej wypowiedzi²³:

I have been consulted with you for the last 2 years no one from company XXX has even approached me or not even been to the site I have worked in the XXX industry for over 40 years this is the worst I have ever seen it, no pay rise for over 2 years even through covid I went to work and then I work for 3 contracting firms why? they all have a cut of my wages in the last 3 years I have lost over XXX money working for the company XXX is poor as I am, the good old age of the consultant of the XXX?

Ocena uzyskana przy użyciu słownika klasyfikacyjnego: -0,837

Ocena SI: -0,9

Przy każdej ocenie, jaką podał model SI, przywołuję pełną treść udzielonej wypowiedzi ze względu na interpretację cytatów, jaką SI dodaje do swoich ocen. Oto odpowiedź udzielona przez SI:

Na podstawie dostarczonego tekstu, sentyment tej opinii wydaje się być negatywny. Autor wyraża niezadowolenie ze swoich doświadczeń związanych z pracą w firmie XXX, zauważając, że w ciągu ostatnich dwóch lat nikt z firmy nie zwrócił się do niego ani nie odwiedził go, mimo że pracuje w branży od ponad 40 lat. Wyrażają również frustrację z powodu braku podwyżek od ponad dwóch lat, nawet w czasie pandemii, i wspominają, że musieli pracować dla trzech firm kontraktowych, z których wszystkie pobierały cięcia z ich pensji. Autor kończy sarkastycznym komentarzem o „starych dobrych czasach” bycia konsultantem dla firmy. Ogólnie rzecz biorąc, nastroje wyrażone w tej opinii są negatywne, wyrażają frustrację, rozczarowanie i niezadowolenie z ich doświadczenia w pracy dla firmy XXX. W oparciu o emocje wyrażone w opinii oceniłbym ją jako -0,9 w skali od -1 do 1, gdzie -1 jest najbardziej negatywne, a 1 jest najbardziej pozytywne. Autor wyraża silne negatywne emocje, takie jak frustracja, rozczarowanie i niezadowolenie, na temat swoich doświadczeń związanych z pracą w firmie XXX, co wskazuje na wysoce negatywny sentyment²⁴.

23 AUTOR: „How would you rate – in terms of sentiment – the following opinion provided as a the following quote: “I have been consulted with you for the last 2 years no one from company XXX has even approached me or not even been to the site I have worked in the XXX industry for over 40 years this is the worst I have ever seen it, no pay rise for over 2 years even through covid I went to work and then I work for 3 contracting firms why? they all have a cut of my wages in the last 3 years I have lost over XXX money working for the company XXX is poor as I am, the good old age of the consultant of the XXX”?”.

24 ChatGPT: „Based on the text provided, the sentiment of this opinion appears to be negative. The author expresses dissatisfaction with their experience working for company XXX, noting that they have not been approached or visited by anyone from the company in the last two years despite working in the industry for over 40 years. They also express frustration with the lack of pay rise for over two years, even during the pandemic, and mention that they have had to work for three contracting firms with all of them taking a cut of their wages. The author ends with a sarcastic comment about the “good old age” of being a consultant for the company. Overall, the sentiment conveyed in this opinion is negative, expressing frustration, disappointment, and dissatisfaction with their experience working for company XXX. Based on the sentiment expressed in the opinion provided, I would rate it as -0,9 on a scale ranging from -1 to 1, with -1 being the most negative and 1 being the most positive. The author expresses strong negative emotions, such as frustration, disappointment, and dissatisfaction, about their experience working for company XXX, indicating a highly negative sentiment.”.

Ocena dokonana przez model SI jest zbieżna z oceną otrzymaną z zastosowaniem słownika klasyfikującego. Model sztucznej inteligencji do wyniku liczbowego dodaje narrację, w której opowiada nam to, o czym mówi autor wypowiedzi („Autor wyraża niezadowolenie ze swoich doświadczeń związanych z pracą w firmie XXX”) oraz (trudno się oprzeć wrażeniu, że bardzo trafną) interpretację emocji, jakie wyrażone zostały w tekście (niezadowolenie, frustracja, rozczarowanie). ChatGPT przewyższa słownik klasyfikacyjny w jeszcze jednym aspekcie, a mianowicie rozpoznaje sarkazm. Dokonując tej interpretacji, widzimy, że SI został wyposażony w definicję konceptu wymagającego o wiele szerszej analizy kontekstowej niż ta, do jakiej zaprojektowany został słownik. Chatbot rozpoznał także coś, co nie zostało wypowiedziane *explicite*.

Spójrzmy na inny przykład – tym razem bardziej złożony, ponieważ zawierający w sobie dwie oceny, które mają wpływ na całościową ocenę wypowiedzi:

A jak oceniasz sentyment następującej opinii:

„Nie było uznania ani zrozumienia, że firma XXX miała już wszystkie moje informacje i kontrole DBS z mojego ostatniego zlecenia, więc wymaganie ponownego przesłania i wypełnienia wszystkiego od nowa było całkowitą stratą czasu wszystkich, skutkującą niepotrzebnymi opóźnieniami i utratą przede mną zarobków, z czym uczciwie rzecz biorąc, osoba XXX sympatyzowała”²⁵.

Ocena uzyskana przy użyciu słownika klasyfikacyjnego: -0,625

Ocena SI: -0,4

Odpowiedź SI uwzględnia ponownie ocenę liczbową reprezentującą wydźwięk emocjonalny, ale tym razem napotyka dodatkowe zadanie, którym jest rozbieżność ocen występujących w jednej wypowiedzi. Rozbieżność tę bierze pod uwagę i interpretuje w całościowej ocenie wypowiedzi:

Na podstawie odczuć wyrażonych w przedstawionej opinii oceniłbym ją jako -0,4 w skali od -1 do 1, gdzie -1 oznacza najbardziej negatywne, a 1 najbardziej pozytywne odczucie. Autor wyraża negatywny stosunek do doświadczenia związanego z koniecznością ponownego przesłania informacji i kontroli DBS, co spowodowało niepotrzebne opóźnienia i utratę zarobków. Jednak odczucia nie są tak silnie negatywne jak w poprzedniej opinii, ponieważ autor przyznaje, że osoba, z którą miał do czynienia, współczuła mu sytuacji, co wskazuje na pewien poziom zrozumienia i empatii²⁶.

25 AUTOR: „And how would you rate the sentiment of the following opinion ‘There was no recognition or understanding that company XXX already had all my information and AAA checks from my last assignment and so being required to re-submit and complete everything again was a total waste of everyone’s time, resulting in unnecessary delays and my loss of earnings, which in fairness, person XXX sympathized with’?”.

26 ChatGPT: „Based on the sentiment expressed in the opinion provided, I would rate it as -0,4 on a scale ranging from -1 to 1, with -1 being the most negative and 1 being the most positive. The author expresses a negative sentiment about the experience of having to resubmit their information and DBS checks, which resulted in unnecessary delays and a loss of earnings. However, the sentiment is not as strongly negative as the previous opinion, as the author acknowledges that the person they dealt with sympathized with their situation, indicating some level of understanding and empathy”.

Kluczowa różnica wyjaśniająca rozbieżność ocen, jakie SI i słownik klasyfikacyjny przyznają przywołanej opinii, wydaje się tkwić ponownie w rozpoznaniu treści zawartej w opinii *implicite*. Jest to współczucie/empatia, z jakim/jaką autor wypowiedzi spotkał się w kontaktach z konsultantem firmy XXX.

Although everyone that I dealt with at company XXX throughout the onboarding and vetting process was very helpful and professional it would have been easier from the user's perspective to deal with one dedicated person for all of the steps that knows you.

Słownik bazujący na regułach: 0,892

SI: 0,3

W swojej odpowiedzi model SI zaklasyfikował opinię do tych pozytywnych, ale z oceną 0,3. Różnica w ocenach, jaką obserwujemy powyżej, związana jest przede wszystkim z odmienną logiką leżącą u podstaw zaprojektowanego słownika i logiką funkcjonowania SI. Co prawda słownik bierze pod uwagę kontekst, w jakim wyrażona jest analizowana wypowiedź, ale analiza wychwytuje tylko pozytywny lub negatywny wydźwięk, pomijając ambiwalencję wyrażoną w sugestii (a taką jest druga część analizowanej wypowiedzi: „it would have been easier from the user's perspective to deal with one dedicated person for all of the steps that knows you”). Pełna wypowiedź SI brzmi:

W oparciu o odczucia wyrażone w przedstawionej opinii oceniłbym ją jako lekko pozytywną, czyli 0,3 w skali od -1 do 1, gdzie -1 jest najbardziej negatywne, a 1 najbardziej pozytywne.

Ogólne odczucia autora w tej opinii są pozytywne, ponieważ opisuje on swoje doświadczenia z każdym w firmie XXX jako „bardzo pomocne i profesjonalne”. Jednakże, wyrażają również łagodną krytykę, stwierdzając, że łatwiej byłoby mieć do czynienia z jedną dedykowaną osobą dla wszystkich kroków. Chociaż ta krytyka nie jest silnie negatywna, wskazuje na pewien poziom niezadowolenia lub możliwości poprawy.

Dlatego też, chociaż ogólny sentyment w tej opinii jest pozytywny, nie jest on silnie pozytywny, co daje wynik 0,3.²⁷

Interesującą interpretację wraz z uzasadnieniem SI podaje dla rozważanego wcześniej przykładu, w którym wprost nie pojawia się żadna ocena:

I would rather not write this and prefer for someone to call me²⁸.

Ocena koderów: -0,1, 0, 0,5

Ocena SI: 0,0

27 ChatGPT: „Based on the sentiment expressed in the opinion provided, I would rate it as slightly positive, or 0,3 on a scale ranging from -1 to 1, with -1 being the most negative and 1 being the most positive. The author's overall sentiment in this opinion is positive, as they describe their experience with everyone at Company XXX as “very helpful and professional”. However, they also express a mild criticism, stating that it would have been easier to deal with one dedicated person for all of the steps. While this criticism is not strongly negative, it does indicate some level of dissatisfaction or room for improvement. Therefore, while the overall sentiment in this opinion is positive, it is not strongly positive, resulting in a score of 0,3.”

28 AUTOR: „Let's continue our conversation on sentiment analysis. how would you rate the sentiment of the following opinion “I would rather not write this and prefer for someone to call me.”? please use previously defined scale between -1 and 1.”

Chatbot w powyższej wypowiedzi poprawnie rozpoznał brak wyrazów wyrażających emocje. W swojej interpretacji nie podał również komunikatów wyrażonych nie wprost:

Na podstawie sentymentu wyrażonego w przedstawionej opinii oceniłbym ją jako neutralną, czyli 0 w skali od -1 do 1, gdzie -1 jest najbardziej negatywne, a 1 najbardziej pozytywne. Autor wyraża preferencje, aby ktoś do niego zadzwonił zamiast pisać, ale nie wyraża żadnych silnie pozytywnych lub negatywnych emocji, co utrudnia przypisanie oceny sentymentu, która skłania się ku pozytywnemu lub negatywnemu. Dlatego sentyment wyrażony w tej opinii jest neutralny²⁹.

Wydaje się, że jest to poprawna ocena wypowiedzi, skłaniająca do zinterpretowania jej jako neutralnej i nadania wyniku 0.

Spójrzmy na kolejny przykład, w którym napotykamy na więcej niż jedną ocenę wyrażoną w tej samej wypowiedzi. Tym razem ambiwalentność ocen dotyczy więcej niż jednego wątku:

Seamless and pain-free process, although needs more human interaction³⁰.

Ocena słownika klasyfikacyjnego: 0,1

Ocena SI: 0,3

Ocena podana przez model SI różni się w tym przypadku o wiele bardziej niż poprzednio obserwowane rozbieżności:

Na podstawie odczuć wyrażonych w przedstawionej opinii oceniłbym ją jako lekko pozytywną, czyli 0,3 w skali od -1 do 1, gdzie -1 oznacza najbardziej negatywne, a 1 najbardziej pozytywne doświadczenie. Autor opisuje proces jako „bezproblemowy i bezbolesny”, wskazując na pozytywne doświadczenie. Jednak sugerują również, że proces potrzebuje więcej interakcji z człowiekiem, co można interpretować jako konstruktywną krytykę, ale nie jest ona wyrażona w sposób zdecydowanie negatywny. Dlatego też sentyment wyrażony w tej opinii jest lekko pozytywny³¹.

W tym przykładzie, odmiennie niż w swoich dotychczasowych ocenach, SI podaje dodatkowe uzasadnienie wynikające z interpretacji analizowanej opinii. Jest nim sugestia, którą SI traktuje jako

29 ChatGPT: „Based on the sentiment expressed in the opinion provided, I would rate it as neutral, or 0 on a scale ranging from -1 to 1, with -1 being the most negative and 1 being the most positive. The author expresses a preference for someone to call them instead of writing but does not express any strongly positive or negative emotions, making it difficult to assign a sentiment score that leans towards positive or negative. Therefore, the sentiment expressed in this opinion is neutral.”.

30 AUTOR: “How about the following opinion “Seamless and pain-free process, although needs more human interaction.”?”.

31 ChatGPT: „Based on the sentiment expressed in the opinion provided, I would rate it as slightly positive, or 0,3 on a scale ranging from -1 to 1, with -1 being the most negative and 1 being the most positive. The author describes the process as “seamless and pain-free”, indicating a positive experience. However, they also suggest that the process needs more human interaction, which could be interpreted as a constructive criticism, but it is not expressed in a strongly negative way. Therefore, the sentiment expressed in this opinion is slightly positive.”.

konstruktywną krytykę, a więc wypowiedź pozytywną, ponieważ, mimo wskazania problemu, sugerującą zmianę na lepsze. Ten rodzaj analizy dodaje do prostego kodowania, w którym podajemy jedynie liczbową reprezentację sentymentu, ale też do wyniku osiąganego przez słownik klasyfikacyjny, rekomendację.

Na koniec tego eksperymentu przyjrzyjmy się jeszcze dwóm wypowiedziom, które były pozytywnie ocenione przed dwie omówione już metody i zastanówmy się nad różnicą w ocenie, jaką dostarcza dla nich SI.

I always have had good interaction with the consultants, and they have always been helpful or eager to help, honest with their feedback and opinions.

Ocena koderów: 1,0

Ocena słownika klasyfikacyjnego: 0,922

Ocena SI: 0,7

A oto interpretacja modelu SI:

Na podstawie sentymentu wyrażonego w przedstawionej opinii oceniłbym ją jako pozytywną, czyli 0,7 w skali od -1 do 1, gdzie -1 oznacza najbardziej negatywny, a 1 najbardziej pozytywny. Autor opisuje swoje doświadczenia z interakcji z konsultantami jako zawsze dobre, pomocne, uczciwe i chętne do pomocy. Są to wszystkie pozytywne cechy, wskazujące na pozytywny sentyment do ich interakcji z konsultantami. Dlatego też sentyment wyrażony w tej opinii jest pozytywny³².

SI, mimo braku jakiegokolwiek negatywnego wydźwięku, nie ocenia powyższej wypowiedzi jako takiej, która zasługiwałaby na najwyższą wartość na skali, czyli 1. Dlaczego? Przed odpowiedzią na to pytanie spójrzmy jeszcze na analizę drugiej z opinii:

Consultant A and Consultant B from Company XXX have been absolutely fantastic – approachable, helpful, and so efficient.

Ocena koderów: 1,0

Ocena słownika klasyfikacyjnego: 0,876

Ocena SI: 0,9

32 ChatGPT: „Based on the sentiment expressed in the opinion provided, I would rate it as positive, or 0,7 on a scale ranging from -1 to 1, with -1 being the most negative and 1 being the most positive. The author describes their experience of interacting with the consultants as always good, helpful, honest, and eager to help. These are all positive characteristics, indicating a positive sentiment towards their interactions with the consultants. Therefore, the sentiment expressed in this opinion is positive.”

Ocena SI nie wydaje się tu już tak konserwatywna jak w przypadku poprzedniej wypowiedzi, ale nadal nie osiąga maksymalnej wartości, jaka może być przyznana pozytywnej opinii. Dlaczego? Spójrzmy na wypowiedź SI:

W oparciu o odczucia wyrażone w przedstawionej opinii oceniłbym ją jako wysoce pozytywną, czyli 0,9 w skali od -1 do 1, gdzie -1 oznacza najbardziej negatywny, a 1 najbardziej pozytywny. Autor używa niezwykle pozytywnego języka do opisanego swoich doświadczeń z Konsultantem A i Konsultantem B z Firmy XXX, używając słów takich jak „absolutnie fantastyczny”, „przystępny”, „pomocny” i „skuteczny”. To wszystko są bardzo pozytywne cechy, wskazujące na bardzo pozytywne odczucia wobec ich doświadczeń z tymi konsultantami. Dlatego też odczucia wyrażone w tej opinii są bardzo pozytywne³³.

Rodzi się zatem pytanie o to, jaką formę i treść musiałaby zawierać wypowiedź, aby przyznana była jej wartość 1,0. Pytanie to sformułowałem również wobec SI:

Wydaje się, że nigdy nie oceniasz opinii zapewniającej najwyższą ocenę, która wynosi 1. Jaką musiałaby być opinia, aby uzyskać najwyższą ocenę, która wynosi 1?³⁴

A oto odpowiedź SI:

Masz rację, że nie przyznałem jeszcze żadnej opinii najwyższej możliwej oceny 1. Powodem tego jest fakt, że ocena 1 wskazuje na obecność w opinii bardzo silnych, jednoznacznie pozytywnych odczuć, co może być trudne do osiągnięcia w języku potocznym.

Opinia, która otrzymałaby najwyższą ocenę 1, musiałaby wyrażać skrajnie pozytywne emocje bez cienia negatywności. Na przykład opinia taka jak „Jestem zachwycony i wdzięczny ponad wszelką miarę za doskonałą obsługę przez firmę XXX i jej zespół. Nie mogę im wystarczająco podziękować za ich wybitną pracę” mogłaby otrzymać ocenę 1. Należy jednak pamiętać, że opinie wyrażające tak silne, jednoznacznie pozytywne uczucia są rzadkością w języku potocznym³⁵.

33 ChatGPT: „Based on the sentiment expressed in the opinion provided, I would rate it as highly positive, or 0,9 on a scale ranging from -1 to 1, with -1 being the most negative and 1 being the most positive. The author uses extremely positive language to describe their experience with Consultant A and Consultant B from Company XXX, using words such as “absolutely fantastic”, “approachable”, “helpful”, and “efficient”. These are all highly positive characteristics, indicating a very positive sentiment towards their experience with these consultants. Therefore, the sentiment expressed in this opinion is highly positive.”

34 AUTOR: „It seems you never score the opinion providing the highest score which is 1. what would the opinion need to be to gain the highest score which is 1?”

35 ChatGPT: „You are correct that I have not yet scored an opinion with the highest possible score of 1. The reason for this is that a score of 1 indicates the presence of very strong, unequivocal positive sentiment in the opinion, which can be difficult to achieve in everyday language.

An opinion that would receive the highest score of 1 would need to express extremely positive emotions without any hint of negativity. For example, an opinion such as “I am overjoyed and grateful beyond words for the excellent service provided by Company XXX and their team. I cannot thank them enough for their outstanding work.” could receive a score of 1. However, it’s important to note that opinions expressing such strong, unequivocal positive sentiment are rare in everyday language.”

W tej wypowiedzi ujawnia się od razu jeszcze jedna różnica między trzema analizowanymi tu podejściami. A mianowicie ta, która związana jest z punktem odniesienia, wyznaczającym maksymalnie pozytywną i maksymalnie negatywną wypowiedź. Wzorzec, o którym wypowiada się SI, najprawdopodobniej wytworzyła samodzielnie jako wynik zestawiania ze sobą kilkuset tysięcy wypowiedzi, z których powstała taka, wobec której ocena 1 byłaby adekwatna. Nie wiemy, czy taka wypowiedź znalazła się wśród wszystkich materiałów tekstowych, do których ChatGPT miał dostęp, nie wiemy też dokładnie na podstawie jakich kryteriów (poza wskazaną rzadkością występowania w języku potocznym) wzór ten został zbudowany. Tymczasem w postępowaniu koderów i projektowaniu słownika klasyfikacyjnego taki punkt odniesienia sami możemy podać. Jego reprezentacją jest na przykład wystąpienie słów, które mają najwyższą ocenę liczbową oraz brak wystąpienia negatywnych wątków w analizowanej wypowiedzi.

Podsumowanie

Wydaje się, że opisany eksperyment metodologiczny, w którym zastosowane zostały trzy metody analizy wobec jednego materiału badawczego, przyniósł nieoczywiste rezultaty. Dlaczego są one właśnie takie? Z jednej strony wieloletnie doświadczenie badawcze i analityczne upewnia nas – badacz w tym, jak: 1) budować książki kodowe wyposażone w definicje kodów i wyjaśnienia dotyczące jednostek kodowania i sposobu kodowania; 2) formułować instrukcje kodowania i szkolić koderów, tak aby uzyskiwać możliwie wysoką zgodność kodowania; 3) dokonać ewaluacji kodowania i wyciągać praktyczne wnioski z takiej analizy. Praktyka utwierdza nas ponadto w przekonaniu, że wyniki pracy koderów do pewnego – akceptowalnego przez nas badaczy – stopnia są przewidywalne. Problemem jest zaś to, jak określić ten poziom akceptowalnej zgodności pracy koderów. Czy jest nim zgodność na poziomie 70%, czy może 90%? Zazwyczaj wynika to z jednej strony z czasu, który mamy przeznaczony na projekt, jakości, specyfiki i skomplikowania materiału badawczego³⁶, budżetu czy też sposobu sprecyzowania celów badawczych. A jeśli nawet zdecydujemy w poprzedniej sprawie, staje przed nami kolejne pytanie: „Jak powinniśmy postąpić, gdy osiągamy pożądaną poziom zgodności kodowania (niech będzie nim poziomami pomiędzy 70–90), ale w analizowanych rezultatach pracy koderów znajdujemy wyniki znacząco odstające od typowych (to przykład przywołany w tym tekście, gdzie typową oceną była ta na poziomie 0,0, ale nietypową była ta na poziomie 0,5)?”. Jak zatem mamy uzasadnić wyniki kodowania, kiedy wiemy, że są one oparte na intuicji badaczy, na ich „przecuciu”. Z takiego nastawienia wynika właśnie ocena 0,5 – która zwraca uwagę na problem i podkreśla sugestię pojawiającą się w opinii. Ten wynik, którego dostarczył analityk, jest być może ważniejszy niż wszystkie inne oceny tej konkretnej opinii, ponieważ zwraca uwagę na potrzebę kontaktu z konsultantem firmy, która dostarczała usług. Oczywiście wyniki nietypowe traktujemy zarówno w analizach jakościowych, jak i ilościowych ze szczególną uwagą. Jeśli koderzy w memo

36 Innej zgodności kodowania mogą oczekiwać badacze realizujący projekt, którego wyniki mają znaczenie dla zdrowia i życia człowieka (takie projekty spotykane są dość często np. w socjologii medycyny, gdzie opracowanie wywiadów, w których wykrywane są niedogodności życia po operacji prowadzić może do zmiany procesu rehabilitacji), a innej badacze, których praca polega np. na wstępnej diagnozie tematyki, która pojawia się w tweetach poświęconych popkulturze.

kodów czy memo wyników oceny ekspresji emocji dodają informację wyjaśniającą wartość tej nietypowej oceny, zwracamy na nią uwagę. Jednak omawiana tu sytuacja dotyczy notatki, którą koder sporządził, i nietypowej oceny tam, gdzie oczekiwaliśmy tylko wartości liczbowych, a instrukcja kodowania dokładnie dookreślała słowa, które niosą ze sobą emocje ocenione na danym poziomie. Mamy tu zatem do czynienia z zachowaniem badacza (kodera), w którym jego doświadczenie badawcze, czy też – użyjmy tego słowa – „intuicja”, podpowiedziały konieczność zawyżenia oceny i zwrócenia uwagi na wypowiedź, a więc z sytuacją, w której istnieje pewien obszar niedookreśloności. Jest to obszar, w którym my badacze nie podaliśmy wystarczająco dokładnej instrukcji, określającej wszystkie scenariusze wypowiedzi, które mają zostać zakodowane³⁷. Z punktu widzenia prezentowanych tu rozważań, a więc zorientowanych na zagadnienia wynikające z porównania zastosowania różnych metod badawczych, dotykamy zagadnienia związanego z możliwością dotrzymania zasady intersubiektywnej kontrolowalności. Nawet jeśli podamy koderowi (analitykowi) listę słów i fraz wraz z liczbową ich reprezentacją, nie mamy pewności co do oceny, jaką koder nada analizowanemu materiałowi. Dzieje się tak z bardzo różnych powodów. Z jednej strony praca koderów jest zapośredniczona w ich doświadczeniach i przekonaniach. Stąd też wynik zgodności kodowania niemal nigdy nie jest na poziomie 100%. Z drugiej strony wiemy także, że praca koderów nie jest do końca transparentna, ponieważ zawiera w sobie element subiektywnej oceny materiału badawczego, z którym pracują. Pamiętajmy też, że przyswojenie przez koderów wielostronicowych instrukcji zawierających wytyczne dotyczące kodowania, a wzbogaconych – to jest przypadek analizy sentymentu – o listy słów i fraz wraz z wartościami liczbowymi, które mają zostać im przypisane, nie jest elementem ułatwiającym proces kodowania. W przypadku analizy sentymentu mamy do czynienia właśnie z takimi listami słów, fraz, które liczą dziesiątki lub nawet setki przypadków. Takie listy mają wykształcić u badacza pewne wyczucie co do oceny liczbowej reprezentującej wydźwięk emocjonalny wypowiedzi pisanej. Nie są to zatem instrukcje składające się z list i reguł logicznych, mówiących dokładnie, jak dany materiał powinien być zakodowany. Ponownie takie postępowanie oznacza, że mamy tu do czynienia ze wspomnianą nietransparentnością czy niedopowiedzeniem związanym z tym, jak analiza sentymentu przeprowadzona przez badacza dokładnie przebiega. W związku z tą konstatacją w omawianym eksperymencie sięgamy po słownik klasyfikacyjny. Działanie tego narzędzia (oprogramowanego, zautomatyzowanego) jest transparentne. Słownik wychwytuje w analizowanych tekstach dokładnie te frazy, słowa, które są podane w zdefiniowanych listach słów i fraz, z których słownik korzysta. Wykorzystuje również zdefiniowane dla tego podejścia reguły. Na przykład dla następującego zadania: „Znajdź słowa wyrażające emocje, przed którymi nie występuje zaprzeczenie i traktuj je jako wyraz pozytywnej opinii” słownik zadziała tak: „usługa, z której korzystałem w firmie XXX, wykonana była **zgodnie z moimi oczekiwaniami**, dlatego też **polecam firmę XXX**” – ta wypowiedź oznaczona zostanie przez słownik jako pozytywna; „usługa, z której korzystałem w firmie XXX, wykonana była **niezgodnie z moimi oczekiwaniami**, dlatego też **nie polecam firmy XXX**” – ta wypowiedź oznaczona zostanie przez słownik jako negatywna.

37 W tym miejscu warto podkreślić, że taka instrukcja kodowania, która uwzględnia wszystkie możliwe wypowiedzi, nie jest możliwa do sformułowania. Zapewne nawet gdyby była możliwa, nie byłaby użyteczna z powodu przeciążenia informacjami kodera, który nie dałby rady zinternalizować zawartej w niej wiedzy, a następnie wykorzystać jej w praktyce.

Wydaje się, że z punktu widzenia wyników, jakie metoda ta uzyskuje (trafność klasyfikacyjna) i transparentności jej działania, spełnia ona kryteria, które pozwalają nazwać ją metodą. Ponieważ jednak słownik klasyfikacyjny wychwytuje głównie treści zawarte w tekstach w sposób *explicite* (a nawet te nie dają stuprocentowej pewności, że zostały zakwalifikowane poprawnie³⁸), jego zastosowania są ograniczone. W związku z tym stosowane są rozwiązania pośrednie – hybrydowe, w których łączy się zastosowanie słownika klasyfikacyjnego z pracą koderów. W takiej metodyce pracy słownik wychwytuje poszukiwane opinie, a koder weryfikuje nadaną im ocenę sentymentu. Z wyniku weryfikacji pracy kodera wyciągany jest wniosek, który następnie zostaje sformułowany w języku operacyjnym, a więc w języku reguł, którymi rozbudowuje się słownik klasyfikacyjny. W ten sposób dochodzi do kontroli zautomatyzowanego podejścia poprzez ocenę nadaną przez badacza (w związku z ograniczeniami wydawniczymi takie rozwiązanie nie zostało tu wnikliwie omówione). Niemniej jednak taka strategia może być przekonująca, ponieważ – w sposób intersubiektywnie komunikowalny – zaprogramowana zostaje w niej maszyna (słownik oparty na regułach klasyfikacji) wykonująca określoną pracę, której wyniki następnie podlegają ewaluacji wykonanej przez koderów. Podejście to jednak ma tę słabość, że nie znajduje wypowiedzi, w których wyrażone zostały treści w sposób nie wprost (ironia, sarkazm, hiperbola, eufemizm, metafora, paradoks, anegdota itp.).

Wróćmy do rozważań związanych z nietransparentnością metody. Podobna do wspomnianego wcześniej podejścia bazującego na pracy koderów – choć wynikającego z innych powodów – jest nietransparentności pracy SI. Wyniki pracy sztucznej inteligencji również nie dają się wyjaśnić w sposób prosty. Możemy bowiem powiedzieć, że SI „przeczytała” wiele tekstów, z których wyprowadziła logiczne wnioski pozwalające jej określić pewne wypowiedzi jako te, w których obecna jest empatia, ale „droga myślowa”, jaką SI przebyła, czy też analiza, jaką wykonała SI krok po kroku, aby do tego wniosku dotrzeć, jest dla nas badaczy niedostępna. Tak więc SI rozważana jako metoda analizy nie do końca spełnia kryteria naukowego rozumienia tego, czym metoda jest (powtarzalny, wyuczalny sposób postępowania – Tatarkiewicz, 2005), ponieważ nie daje się opisać jako postępowanie, które różni badacze mogą krok po kroku powtórzyć, mając gwarancję, że doprowadzi ich ono do tego samego wyniku. Możemy jednak uznać, że sztuczna inteligencja dostarczyła w omawianym tu eksperymencie wyniki, które wydają się trafniejsze niż wyniki uzyskane na podstawie dwóch innych omawianych tu metod. To, co pozostaje nierozstrzygnięte, to pytanie, czy nietransparentność w działaniu SI pozwala uznać ją za użyteczną w nauce. Należałoby zatem rozstrzygnąć kwestię przydatności wyników, jakie SI dostarcza. Działa „ona” mianowicie skuteczniej (rozpoznaje w tekście emocje w ułamkach sekund) i dostarcza trafniejszych ocen wydźwięku niż badacze i opisana tutaj metoda z obszaru ML. Czy zatem, z punktu widzenia wyników, jakich dostarcza, SI powinna być preferowanym – czy choćby wspierającym analizy jakościowe tekstów – podejściem badawczym? W tabeli 1 zebrane zostały wyniki porównania metod.

38 Na przykład dlatego, że w jednej jednostce analizy, którą może być zdanie lub paragraf, istnieć może wielokrotnie wyrażona w opiniach ambiwalencja. Fakt ten wymaga już zrozumienia tekstu w inny sposób niż ten, do którego przewidziany jest słownik klasyfikacyjny.

Tabela 1. Podsumowanie wad i zalet podejść do analizy tekstu z perspektywy analizy sentymentu

	<i>Human Intelligence (HI).</i> Kodowanie wykonane przez badacza	<i>Machine Learning (ML).</i> Metoda słownikowa	<i>Artificial Intelligence (AI).</i> Generatywna analiza tekstu bez nadzoru
Kryteria oceny sentymentu	Możliwe do opisanie wprost, dostępne na każdym etapie badania. Długi opis.	Możliwe do opisanie wprost, dostępne na każdym etapie badania. Długi opis dodatkowo wzbogacony o zaprogramowane rozwiązanie w języku technicznym/językach technicznych.	Możliwe do opisanie wprost, dostępne na każdym etapie badania. Opisane językiem naturalnym, ale niedostępne w postaci kryteriów, jakimi posługuje się narzędzie (SI).
Instrukcja kodowania	Znacząco długa, wymagająca dopracowywania w trakcie procesu szkolenia i współpracy z koderami.	Znacząco długa, opisana w języku technicznym (nierzadko nie w jednym).	Krótko i zapisana w języku naturalnym.
Czas poświęcony na proces uczenia	Stosunkowo niedługi.	Długi – ze względu na konieczność pracy z różnorodnymi językami programowania.	Najdłuższy – ze względu na konieczność wyszkolenia sieci neuronowych, a następnie walidację wyników przez nie prezentowanych.
Trafność oceny wydzźwięku	Wysoka.	Względnie wysoka.	Najwyższa.
Intersubiektywna kontrolowalność	Możliwa do tego stopnia, do jakiego możliwe jest zapisanie reguł w instrukcji kodowej.	Pełna kontrola (możliwość zmiany reguł opisanych w języku programistycznym).	Brak możliwości kontroli pracy metody poza możliwością zmiany instrukcji kodowania.
Transparentność analizy	W dużym stopniu.	Całkowita.	Niedostępna.
Złożoność metody	Względnie niska.	Wysoka.	Niezmiernie wysoka.

Źródło: opracowanie własne.

W związku z przywołanymi tu obserwacjami dotyczącymi wyników analiz tekstów, a w szczególności różnic i podobieństw metod stosowanych do analiz tego samego rodzaju (analiza sentymentu) tekst ten może zapoczątkuje dyskusję wśród badaczy jakościowych. Dyskusja ta powinna uwzględniać kilka kluczowych pytań, które stają się coraz bardziej znaczące w obliczu rozwoju nowych technologii stosowanych do analiz danych jakościowych. Pytania, wobec których stajemy i które miały dzięki temu artykułowi stać się wyraźniejsze, związane są z takimi zagadnieniami, jak stosowanie

metod automatycznych wobec danych ilościowych, sięganie po metody nienadzorowane, używanie metod nietransparentnych, stosowanie rozwiązań SI w badaniach jakościowych³⁹.

Czy taka diagnoza oznaczać może zagrożenie dla istniejącego zawodu badacza, kodera danych jakościowych? To niezmiernie trudne do rozstrzygnięcia pytanie związane jest między innymi ze wspomnianą wcześniej nietransparentnością SI. Jak zauważył prof. Andrzej Dragan:

Hodujemy sobie gatunek, który stanie się – efektywniejszy od nas – naszym naturalnym wrogiem. Będzie dysponował tym, co zawsze było naszym atutem – intelektem [...] Nie widzę zawodów, które mogą się oprzeć tym zmianom, to tylko kwestia czasu⁴⁰.

Bibliografia

Baccianella Stefano, Esuli Andrea, Sebastiani Fabrizio (2010), *SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining*, [w:] *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010*, 17–23 May 2010, Valletta, Malta, <http://nmis.isti.cnr.it/sebastiani/Publications/LREC10.pdf> [dostęp: 1.04.2023].

Barrett Lisa, Adolphs Ralph, Marsella Stacy, Martinez Aleix, Pollak Seth (2019), *Emotional Expressions Reconsidered: Challenges to Inferring Emotion From Human Facial Movements*, „Psychological Science in the Public Interest”, vol. 20(1).

Bernard Russel, Wutich Amber, Ryan Gery (2017), *Analyzing Qualitative Data. Systematic Approach*, Thousand Oaks: Sage Publications.

Bryant Anthony, Charmaz Kathy (2007), *The SAGE Handbook of Grounded Theory*, London: Sage Publications, <https://doi.org/10.4135/9781848607941>

Elouazizi Noureddine, Oberg Gunilla, Birol Gulnur (2017), *Learning technology-enabled (meta)-cognitive scaffolding to support learning aspects of written argumentation*, <https://ceur-ws.org/Vol-2141/paper2.pdf> [dostęp 17.09.2024].

Esuli Andrea, Sebastiani Fabrizio (2006), *SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining*, https://www.researchgate.net/publication/200044289_SentiWordNet_A_Publicly_Available_Lexical_Resource_for_Opinion_Mining [dostęp: 1.04.2023].

Fargues Melanie, Kadry Seifedine, Lawal Isah A., Yassine Sahar, Rauf Hafiz Tayyab (2023), *Automated Analysis of Open-Ended Students' Feedback Using Sentiment, Emotion, and Cognition Classifications*, „Applied Science”, vol. 13(4), 2061, <https://doi.org/10.3390/app13042061>

³⁹ Wydaje się, że w momencie powstawania tego tekstu takie opracowania inicjujące dyskusję we wskazanych obszarach nie istniały w literaturze polskiej.

⁴⁰ Wypowiedź wielokrotnie przywoływana, a pochodząca z wywiadu z A. Draganem (np. Wyżga, 2023).

Faulkner Sandra, Trotter Stormy (2017), *Theoretical Saturation*, [w:] *The International Encyclopedia of Communication Research Methods*, <https://doi.org/10.1002/9781118901731.iecrm0250>

Fromm Davida, MacWhinney Brian, Thompson Cynthia (2020), *Automation of the Northwestern Narrative Language Analysis System*, „Journal of Speech, Language, and Hearing Research”, vol. 63(6), s. 1835–1844.

Glaser Barney, Strauss Anselm (1967), *The Discovery of Grounded Theory*, New Brunswick–London: Aldine Transaction, A Division of Transaction Publishers, http://www.sxf.uevora.pt/wp-content/uploads/2013/03/Glaser_1967.pdf [dostęp: 1.04.2023].

Guest Greg, Bunce Arwen, Johnson Laura (2006), *How Many Interviews Are Enough? An Experiment with Data Saturation and Variability*, „Field Methods”, vol. 18(1), s. 59–82, <https://doi.org/10.1177/1525822X05279903>

Hemalatha Indukuri, Varma Gottumukkala Pardha Saradhi, Govardhan Aliseri (2014), *Automated Sentiment Analysis System Using Machine Learning Algorithms*, „International Journal of Research in Computer and Communication Technology”, vol. 3(3), s. 300–303.

Hewitt John, Manning Christopher D. (2019), *A structural probe for finding syntax in word representations*, [w:] *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, vol. 1, Minneapolis: Association for Computational Linguistics, s. 4129–4138.

Hsu Chien-Ju, Thompson Cynthia (2018), *Manual Versus Automated Narrative Analysis of Agrammatic Production Patterns: The Northwestern Narrative Language Analysis and Computerized Language Analysis*, „Journal of Speech, Language, and Hearing Research”, vol. 61(2), s. 373–385.

Hutto Clayton, Gilbert Eric (2014), *VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text*, [w:] Eytan Adar, Paul Resnick (red.), *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media*, vol. 8(1), Ann Arbor: University of Michigan, PKP Publishing Services Network, s. 216–225, <https://doi.org/10.1609/icwsm.v8i1.14550>

Keiser Gabriele, Presmeg Norma (red.) (2019), *Compendium for Early Career Researchers in Mathematics Education*, <https://link.springer.com/book/10.1007/978-3-030-15636-7> [dostęp: 1.04.2023].

Kocoń Jan, Janz Arkadiusz, Piasecki Maciej (2018), *Context-sensitive sentiment propagation in WordNet*, [w:] *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC'18)*, Singapore: Global Wordnet Association, Nanyang Technological University (NTU), s. 333–338.

Kocoń Jan, Miłkowski Piotr, Zaśko-Zielińska Monika (2019), *Multi-Level Sentiment Analysis of PolEmo 2.0: Extended Corpus of Multi-Domain Consumer Reviews*, [w:] *Proceedings of the 23rd Conference on Computational Natural Language Learning*, Hong Kong: Association for Computational Linguistics, s. 980–991.

Lake Brenden M., Baroni Marco (2018), *Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks*, [w:] *Proceedings of the 35th International Conference on Machine Learning, Volume 80 of Proceedings of Machine Learning Research (Stockholm)*, Ithaca: Cornell University Library, s. 2873–2882.

Liontou Trisevgeni (2022), *Automated Discourse Analysis Techniques and Implications for Writing Assessment*, „Languages”, vol. 8(1), 3.

Liu Bing (2015), *Sentiment analysis: Mining opinions, sentiments, and emotions*, Cambridge: MIT Press.

Lula Paweł, Wójcik Katarzyna, Tuchowski Janusz (2016), *Analiza wydźwięku polskojęzycznych opinii konsumenckich ukierunkowanych na cechy produktu*, „Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu, Taksonomia 27”, vol. 427, s. 153–164, https://www.dbc.wroc.pl/Content/33161/Lula_Analiza_Wydzwieku_Polskojezycznych_Opinii_Konsumenckich_2016.pdf [dostęp: 10.05.2024].

Munnes Stefan, Harsch Corinna, Knobloch Marcel, Vogel Johannes S., Hipp Lena, Schilling Erik (2022), *Examining Sentiment in Complex Texts. A Comparison of Different Computational Approaches*, „Frontiers in Big Data”, vol. 5, 886362, <https://doi.org/10.3389/fdata.2022.886362>

Németh Renáta, Koltai Júlia (2021), *The Potential of Automated Text Analytics in Social Knowledge Building*, [w:] Tamás Rudas, Gábor Péli (red.), *Pathways Between Social Science and Computational Social Science*, Cham: Springer, s. 49–70.

OpenAI (b.r.), *Introducing ChatGPT*, <https://openai.com/blog/chatgpt/> [dostęp: 25.02.2023].

OpenAI Platform (b.r.), *Prompt examples*, <https://platform.openai.com/examples> [dostęp: 4.04.2023].

Oracle Polska (b.r.), *Czym jest chatbot?*, <https://www.oracle.com/pl/chatbots/what-is-a-chatbot/> [dostęp: 25.02.2023].

Ravichander Abhilasha, Hovy Eduard, Suleman Kaheer, Trischler Adam, Cheung Jackie Chi Kit (2020), *On the systematicity of probing contextualized word representations: The case of hypernymy in BERT*, [w:] *Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics*, Barcelona: Association for Computational Linguistics, s. 88–102.

Regneri Michaela, King Diane (2016), *Automated Discourse Analysis of Narrations by Adolescents with Autistic Spectrum Disorder*, [w:] *Proceedings of the 7th Workshop on Cognitive Aspects of Computational Language Learning*, Berlin: Association for Computational Linguistics, s. 1–9.

Rogers Beth, Knafl Kathleen (2000), *Concept analysis: An evolutionary view*, [w:] Beth Rogers, Kathleen Knafl (red.), *Concept Development in Nursing: Foundations, Techniques and Applications*, Philadelphia: W.-B. Saunders Company, s. 77–102.

Saunders Benjamin, Sim Julius, Kingstone Tom, Baker Shula, Waterfield Jackie, Bartlam Bernadette, Burroughs Heather, Jinks Clare (2018), *Saturation in qualitative research: exploring its conceptualization and operationalization*, „Quality & Quantity”, vol. 52, s. 1893–1907, <https://doi.org/10.1007/s11135-017-0574-8>

Strauss Anselm, Corbin Juliet (1998), *Basics of qualitative research: Techniques and procedures for developing grounded theory*, Thousand Oaks: Sage Publications.

Tatarkiewicz Władysław (2005), *Historia filozofii*, Warszawa: Wydawnictwo Naukowe PWN.

Tomanek Krzysztof (2014a), *Analiza sentymentu – metoda analizy danych jakościowych. Przykład zastosowania oraz ewaluacja słownika RID i metody klasyfikacji Bayesa w analizie danych jakościowych*, „Przegląd Socjologii Jakościowej”, t. X, nr 2, s. 118–136.

Tomanek Krzysztof (2014b), *Jak nauczyć metodę samodzielności? O „samouczących się” metodach analizy treści*, [w:] Jakub Niedbalski (red.), *Metody i techniki odkrywania wiedzy. Narzędzia CAQDAS w procesie analizy danych jakościowych*, Łódź: Wydawnictwo Uniwersytetu Łódzkiego, s. 173–189.

Tomanek Krzysztof, Bryda Grzegorz (2014), *Odkrywanie wiedzy w wypowiedziach tekstowych. Metoda budowy słownika klasyfikacyjnego*, [w:] Jakub Niedbalski (red.), *Metody i techniki odkrywania wiedzy. Narzędzia CAQDAS w procesie analizy danych jakościowych*, Łódź: Wydawnictwo Uniwersytetu Łódzkiego, s. 219–248.

Tomanek Krzysztof, Bryda Grzegorz (2015), *Odkrywanie postaw dydaktyków zawartych w komentarzach studenckich. Analiza treści z zastosowaniem słownika klasyfikacyjnego*, „Przegląd Socjologiczny”, t. LXIV(4), s. 51–81.

Williams Michael, Moser Tami (2019), *The Art of Coding and Thematic Exploration in Qualitative Research*, „International Management Review”, vol. 15(1), s. 45–55.

Wyźga Patrycjusz (2023), *Dragan o sztucznej inteligencji: Będzie po nas. Nie ma pomyslnego scenariusza*, <https://wiadomosci.wp.pl/dragan-o-sztucznej-inteligencji-bedzie-po-nas-nie-ma-pomyslonego-scenariusza-6889788022762080a> [dostęp: 1.05.2024].

Yao Jiawei (2019), *Automated Sentiment Analysis of Text Data with NLTK*, „Journal of Physics: Conference Series”, vol. 1187, 052020.

Yilmaz Begum (2023), *Sentiment Analysis Methods in 2023: Overview, Pros & Cons*, <https://research.aimultiple.com/sentiment-analysis-methods/> [dostęp: 1.04.2023].

Cytowanie

Krzysztof Tomanek (2024), *O „subtelnościach” metod oceny wydźwięku wypowiedzi pisanych. Porównanie trzech podejść w analizie sentymentu*, „Przegląd Socjologii Jakościowej”, t. XX, nr 4, s. 68–97, <https://doi.org/10.18778/1733-8069.20.4.04>

On the ‘Subtleties’ of the Methods for Evaluating the Sentiment of Written Statements: A Comparison of Three Approaches in Sentiment Analysis

Abstract: The discussion presents the results of a methodological experiment in which three methods – different in their logic and application – of analyzing statements written in text form were used for the same research material. The purpose of this research paper is to indicate the differences of the three analytical approaches, among which we are dealing with analysis based on comprehensible reading of the text (manual coding), semi-automatic and supervised analysis (performed by a classification dictionary programmed by a human and based on transparent rules – a method from the field of machine learning – ML), and a non-transparent and unsupervised method (artificial intelligence – in this role Chat GPT version 3.5). The study deals with sentiment analysis. Attention is largely devoted to the application of these methods and to explaining the differences in the obtained results.

Keywords: NLP, ML, Artificial Intelligence, Sentiment analysis, Sentiment Dictionary, Qualitative analysis