Politechnika Śląska

Wydział Automatyki, Elektroniki i Informatyki

Rozprawa doktorska

Metody zwiększenia skuteczności wirtualnych konsultantów poprzez minimalizację niepoprawnie rozpoznawanych intencji z wypowiedzi klientów

mgr inż. Łukasz Pawlik

Promotor:

dr hab. inż. Roman Stanisław Deniziak prof. PŚk

Politechnika Świętokrzyska

Gliwice, 2022

Spis treści

1 Wstęp

W dzisiejszych czasach komunikacja klientów z firmą lub instytucją coraz częściej realizowana jest przez telefon, sms, e-mail, czat, a nawet media społecznościowe i rozmowy video. Oprogramowaniem służącym do obsługi masowych kontaktów z klientami są systemy Contact Center (CC). Główną ich rolą jest automatyczna obsługa klienta, kolejkowanie połączeń oraz ich dystrybucja do odpowiednich konsultantów.

W CC obsługiwane są zarówno połączenia telefoniczne przychodzące (ang. inbound) jak i wychodzące (ang. outband). Interakcje przychodzące są inicjowane przez klientów (np. reklamacje), natomiast wychodzące to rozmowy inicjowane przez biuro obsługi klienta (np. sprzedaż). Systemy Contact Center korzystają z integracji telefonii komputerowej CTI (ang. computer telephony integration), interaktywnej odpowiedzi głosowej IVR (ang. interactive voice response) i routingu opartego na umiejętnościach SBR (ang. skills-based routing). Systemy te również archiwizują i przetwarzają dane z różnych kanałów komunikacji dostarczając historię kontaktów z klientem, raporty oraz wskaźniki KPI (ang. key performance indicators) [1].

Contact Center znajduje swoje zastosowanie między innymi w sprzedaży, marketingu, udzielaniu informacji, wsparciu technicznemu, obsłudze reklamacji, przyjmowaniu zgłoszeń, prowadzeniu badań i ankiet oraz windykacji. W ciągu ostatnich dwóch dekad zapewnienie wydajnej i skutecznej obsługi klienta stało się niezbędne dla każdego przedsiębiorstwa. Contact Center jako punkt kontaktu z organizacją, ma znaczący wpływ na pozyskanie, utrzymanie i satysfakcję klienta. Znaczenie CC było w literaturze badane w różnych dziedzinach, ale wraz ze zmianami technologii działanie i struktura tych systemów musi ewoluować, aby sprostać bieżącym wyzwaniom [2].

W Contact Center na przestrzeni ostatnich lat można zaobserwować bardzo dynamiczny rozwój zarówno klasycznych (telefon, sms, e-mail) jak i nowych (czat, media społecznościowe, rozmowy video) kanałów komunikacji [3]. Natomiast ciągle dominującym kanałem komunikacji jest rozmowa telefoniczna. Z kolei czat jest kanałem preferowanym przez osoby młodsze, które często funkcjonują wielozadaniowo i mogą przez to medium komunikacji załatwiać swoje sprawy jednocześnie zajmując się innymi aktywnościami. W tych obszarach rozwój technologiczny jest szczególnie istotny z punktu widzenia podniesienia jakości obsługi klienta jak i obniżenia kosztów firm Contact Center.

Odpowiedzią na te dwa wyzwania jest wprowadzenie wirtualnych konsultantów w postaci voicebotów i chatbotów [4]. W przypadku rozmów telefonicznych dominującą technologią w zakresie automatyzacji obsługi jest IVR [5], który jest skuteczny, ale mało przyjazny dla klientów. Voiceboty znacznie przewyższają IVR w zakresie tak zwanego doświadczenia użytkownika UX (ang. user experience) oferując poziom obsługi zbliżony do kontaktu z człowiekiem. Z kolei w przypadku czatów dotychczasowe formy automatyzacji sprowadzają się do prezentowania klientowi różnego rodzaju list wyboru, które mają na celu skierować go do odpowiedniej grupy specjalistów lub udzielić mu odpowiedzi z bazy wiedzy typu FAQ (ang. frequently asked questions).

W ostatnich latach nastąpił szybki rozwój w obszarze wirtualnych konsultantów, dlatego firmy działające w branży Contact Center coraz odważniej wprowadzają te rozwiązania jako pierwszą linię komunikacji z klientem [6]. Warunkiem podstawowym do wdrożenia takich

inteligentnych rozwiązań jest ich skuteczność, z uwagi na konieczność zachowania ciągłości świadczenia usług. Zrozumienie co wpływa na poprawność działania i tym samym skuteczność voicebotów i chatbotów [7] jest kluczowym zagadnieniem, które decyduje o sukcesie funkcjonowania takich rozwiązań w Contact Center.

System automatycznego rozpoznawania mowy, określany jako ASR (ang. automatic speech recognition) jest podstawowym komponentem, który wpływa na skuteczność głosowego wirtualnego konsultanta w danym modelu językowym [8]. Chociaż skuteczność tych systemów jest bardzo wysoka w warunkach nagrań studyjnych, to w warunkach Contact Center osiągnięcie skuteczności rozpoznawania mowy wystarczającej do stworzenia wirtualnego konsultanta stanowi duże wyzwanie, szczególnie w mniej popularnych modelach językowych, do jakich należy język polski. W zakresie chatbotów zagadnienie automatycznego rozpoznawania mowy nie występuje z uwagi na komunikację tekstową.

Komponentem istotnym zarówno dla voicebotów jak i chatbotów są systemy rozumienia języka naturalnego NLU (ang. natural language understanding) [9], często osadzane w ramach tak zwanych platform NLU. Platformy te są często rozszerzone o dodatkowe elementy takie jak: funkcja wyodrębniania encji, narzędzie do budowania dialogów oraz mechanizmy do testowania bota. Natomiast kluczową ich cechą jest rozpoznawanie intencji na podstawie fraz tekstowych z wypowiedzi klientów. Funkcja rozpoznawania intencji jest wyznaczana poprzez uczenie maszynowe bazujące na zbiorach fraz uczących.

Zarówno w systemach ASR jak i platformach NLU kluczową rolę odgrywają korpusy językowe. Dlatego skuteczność tych rozwiązań może być różna dla poszczególnych modeli językowych, w tym również dla języka polskiego. Producenci tych systemów często dostarczają kompleksowe rozwiązania w postaci zintegrowanego systemu ASR z platformą NLU. Należy jednak pamiętać, że nie zawsze takie rozwiązanie jest optymalne, gdyż to jakie komponenty należy wykorzystać (jakich producentów) może zależeć od warunków w jakich ma pracować voicebot. Dlatego na początku wdrożenia voicebota należy określić z jakim modelem językowym mamy do czynienia, jakiej jakości będzie strumień audio, a nawet jakiej tematyki będzie dotyczyło wdrożenie (np. dla branży medycznej są specjalizowane systemy ASR).

Autor pracy od 20 lat jest związany z branżą Contact Center w zakresie rozwoju i wdrażania rozwiązań informatycznych. Bezpośrednio uczestniczy w tworzeniu systemu CC do wielokanałowej obsługi klienta. Przez ten czas miał okazję obserwować zmiany technologiczne zachodzące w tej branży, które w różnym stopniu przyczyniały się do poprawy jakości obsługi klienta. Nowe technologie często wprowadzały nową jakość w komunikacji m.in.: czat [10], rozmowy video [11], MultiChannel, OmniChannel [12], media społecznościowe [13], a niekiedy okazywało się, że nie przyjęły się one na rynku, czego przykładem jest Interactive Video Response [14], czy Visual IVR [15].

Z doświadczenia autora wynika, że w branży Contact Center, dobór systemu ASR oraz platformy NLU nie jest zagadnieniem prostym, choćby z uwagi na niską jakość rozmów (zarówno sygnału audio, jak i warunki pozatechniczne – dykcja klientów, towarzyszące im emocje czy szumy otoczenia). Często żadne z rozwiązań nie spełnia wymaganych kryteriów, dlatego adaptacja tych systemów dla branży Contact Center powinna nie tylko wiązać się z ich wyborem i integracją, ale również z szukaniem możliwości poprawy ich skuteczności poprzez opracowanie wyspecjalizowanych metod i algorytmów. Celem przeprowadzonych badań była

identyfikacja przyczyn niedoskonałości botów stosowanych w Contact Center, a następnie zaproponowanie metod poprawy ich skuteczności poprzez minimalizację niepoprawnie rozpoznawanych intencji z wypowiedzi klientów. W efekcie opracowano nowatorskie koncepcje, które dzięki swoim zaletom mogą przyczynić się do dalszego rozwoju automatycznych i inteligentnych rozwiązań komunikacji z klientem. Mimo, że badania koncentrowały się na języku polskim, to ich efekty można zaadoptować do innych modeli językowych.

Zakres merytoryczny pracy obejmuje wybrane zagadnienia informatyki oraz lingwistyki. Przeprowadzono analizę rzeczywistych rozmów głosowych i rozmów czat pochodzących z Contact Center funkcjonującego w języku polskim. Współpracowano z psychologami i lingwistami w celu wypracowania skutecznych metod tworzenia voicebotów i chatbotów z uwzględnieniem poprawy transkrypcji automatycznych oraz uwzględnieniem emocji w rozpoznawaniu ukrytych intencji klientów.

Rozdział 2 stanowi przegląd literatury, który przybliża potencjalne zastosowania wirtualnych konsultantów głosowych i tekstowych. Opisano architekturę voicebotów i chatbotów w celu wskazania komponentów, których działanie nie jest wystarczająco efektywne z punktu widzenia branży Contact Center. Rozdział ten zawiera również przegląd literatury dotyczącej systemów automatycznego rozpoznawania mowy oraz platform NLU – uwzględniając badania, które są prowadzone w tych dziedzinach w zakresie ujętym w pracy.

Rozdział 3 zawiera cel i tezę pracy bezpośrednio wynikającą ze zidentyfikowanych braków w zakresie skuteczności dotychczasowych rozwiązań w przeprowadzonym przeglądzie literatury.

W rozdziale 4 zamieszczono definicje oraz formalny model wirtualnego konsultanta. Opisano procedurę uczenia bota oraz proces rozpoznawania intencji i wyboru akcji odpowiedzi bota.

W rozdziale 5 opisano przeprowadzone badania dotyczące wpływu jakość transkrypcji na rozpoznawanie intencji. Zbadano i wprowadzono nowe metody preprocessingu oraz postprocessingu. Oceniono skuteczność tych metod posiłkując się powszechnie stosowanymi metrykami WER (ang. word error rate) oraz LER (ang. letter error rate). Na końcu rozdziału zaprezentowano szczegółowe wyniki przeprowadzonych eksperymentów.

Rozdział 6 poświęcono badaniom wpływu emocji na rozpoznawanie rzeczywistych intencji klientów. Na tej podstawie zaproponowano nową metodę rozpoznawania ukrytych intencji uwzględniającą emocje. Wykorzystano w niej mechanizmy encji dostarczane przez platformy NLU oraz opracowano szereg założeń do tworzenia reguł wnioskowania akcji bota, której celem jest udzielenie jak najbardziej adekwatnej odpowiedzi klientowi. W końcowej części rozdziału przedstawiono wyniki dla kanału głosowego i tekstowego, obrazujące skuteczność nowo opracowanej metody rozpoznawania ukrytych intencji.

W rozdziale 7 przedstawiono system zwiększający skuteczność botów w Contact Center. Zaprezentowano w nim wszystkie moduły tego systemu. Opisano, jak powinien być prowadzony proces przygotowania fraz oraz uczenie modelu dla platformy NLU. Pokazano w jaki sposób została połączona metoda poprawy jakości transkrypcji automatycznej z metodą rozpoznawania ukrytych intencji uwzględniająca emocje klienta. Na końcu rozdziału zaprezentowano końcowe wyniki eksperymentów wpływu nowo opracowanych metod poprawiających skuteczność działania bota.

W podsumowaniu (rozdział 8) zawarty został wykaz najważniejszych osiągnięć pracy, wnioski końcowe oraz propozycje dalszych badań i rozwoju zaprezentowanych w pracy nowatorskich rozwiązań dla Contact Center.

2 Przegląd aktualnych rozwiązań wirtualnych konsultantów w Contact Center

W rozdziale tym przedstawiono rolę i znaczenie wirtualnych konsultantów w komunikacji z klientem implementowanych jako voiceboty i chatboty. Wskazano różne ich zastosowania oraz zaprezentowano ich ogólną architekturę. Następnie przedstawiono przegląd aktualnego stanu wiedzy w zakresie systemów automatycznego rozpoznawania mowy oraz platform NLU. Wskazano również wady i zalety aktualnych rozwiązaniach.

2.1 Kontekst zastosowania wirtualnych konsultantów w Contact Center

Obserwowany na przestrzeni ostatnich lat rozwój systemów z obszaru Contact Center ukierunkowany jest w dużym stopniu na automatyzację wielu zachodzących w nich procesów. Automatyzacja rutynowych czynności z jednej strony pozwala redukować koszty generowane przez systemy CC, z drugiej natomiast stanowi ważny aspekt prospołeczny odciążając agentów od wykonywania żmudnych, stale powtarzających się czynności. Poprawie ulega tym samym komfort pracy agenta, w niemałym stopniu zapobiegając rotacjom na tym trudnym stanowisku pracy [16]. Rozwój technologiczny w kontekście automatyzacji rozmów głosowych oraz rozmów czat jest szczególnie istotny, gdyż w dużej mierze wpływa na jakość obsługi klienta oraz ponoszone koszty. W ostatnich latach dużą popularność w CC zdobywają rozwiązania wirtualnych konsultantów implementowanych w postaci voicebotów dla rozmów telefonicznych oraz chatbotów dla rozmów czat [17].

W obecnych czasach obserwowany jest dynamiczny rozwój firm teleinformatycznych wdrażających systemy z obszaru Contact Center. Spowodowane jest to między innymi szerzącą się pandemią COVID-19, która w sposób zdecydowany wpływa na zmianę dotychczas praktykowanego modelu komunikacji międzyludzkiej. Znaczna część kontaktów zarówno prywatnych, jak i biznesowych odbywa się obecnie w sposób zdalny. Z dużym prawdopodobieństwem można zakładać, że ten wymuszony trend zwłaszcza w relacjach biznesowych już z nami pozostanie. W związku z tym, w wielu branżach coraz częściej wykorzystywane są rozwiązania systemów infolinii CC. Systemy te obecnie budowane są z coraz szerszym wykorzystaniem algorytmów sztucznej inteligencji AI (ang. artificial intelligence), w tym głównie metod uczenia maszynowego ML (ang. machine learning) oraz metod przetwarzania języka naturalnego NLP (ang. natural language processing) i analityki Big Data. Implementacja rozwiązań wspieranych algorytmami AI ma szczególne znaczenie w budowie chatbotów oraz voicebotów [18].

Obecnie jednym z najczęstszych zastosowań voicebota i chatbota jest obsługa rutynowych i powtarzalnych połączeń, w których można w niemal 100% wyeliminować czynnik ludzki. Boty potrafią samodzielnie przeprowadzać rozmowy telefoniczne i czat z klientem, całkowicie rozwiązując niektóre sprawy [6]. Dzisiaj w typowym Contact Center co najmniej 25% obsługiwanych połączeń jest klasyfikowanych jako ściśle rutynowe [19]. Innymi słowy, jedna czwarta pracowników centrum musi poświęcać czas na rozmowy, które można łatwo zautomatyzować za pomocą botów głosowych i tekstowych. Agenci, którzy spędzają czas na rutynowych rozmowach, mogą szybko stracić motywację do rozwiązywania problemów, a nawet wypalić się ze względu na powtarzalny i monotonny charakter rozmów. Z drugiej strony, boty są przeznaczone do powtarzalnych zadań. Ponadto ludzie są bardziej szczerzy w stosunku do botów głosowych i tekstowych niż w stosunku do realnej osoby. To sprawia, że

boty nadają się do wszelkiego rodzaju ankiet i połączeń ankietowych, które mogą być nie tylko w 100% zautomatyzowane, ale także dostarczać bardziej trafnych odpowiedzi [19].

Kolejnym zastosowaniem jest użycie voicebota zamiast systemu IVR. Często klienci w IVR wybierają błędną ścieżkę i wówczas rola agenta w zasadzie sprowadza się do wysłuchania klienta i przełączenie go do innego zespołu specjalistów. Znane są przypadki, gdzie liczba błędnie skierowanych rozmów stanowi nawet 15% spośród wszystkich rozmów [20]. W takiej sytuacji użycie voicebota, który prawidłowo rozpozna intencje klienta i skieruje go do prawidłowej grupy specjalistów znacznie zwiększa efektywność Contact Center.

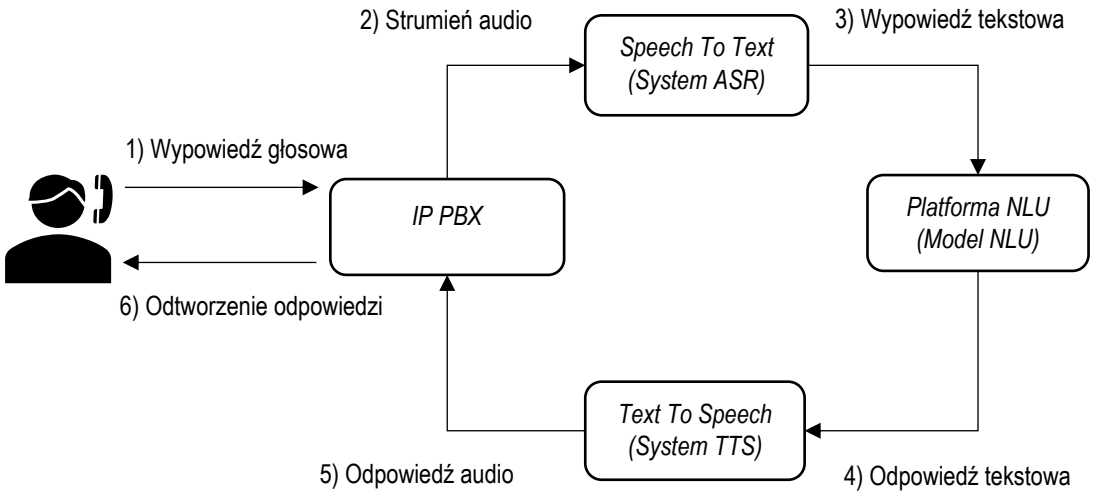
W dzisiejszych czasach systemy IVR są często stosowane, aby umożliwić klientom sprecyzowanie swojego żądania, otrzymać podstawową odpowiedź lub przekierować do odpowiedniego agenta. Ale systemy IVR są nieintuicyjne i mogą być bardzo mylące. Mogą zaoszczędzić czas niektórych pracowników Contact Center, kierując do nich odpowiednią część dzwoniących, ale wielu klientów i tak wybiera ogólne żądanie „Połącz z agentem" i trafia do nieodpowiedniej grupy specjalistów. W przeciwieństwie do IVR zastosowanie voicebota zapewnia klientom bardziej naturalną, opartą na sztucznej inteligencji rozmowę. Zastosowana technologia przetwarzania języka naturalnego (NLP) rozumie intencję rozmowy, a w połączeniu z technologiami zamiany mowy na tekst i zamiany tekstu na mowę tworzy solidną podstawę dla responsywnych, konwersacyjnych botów głosowych wykorzystujących AI. Zamiast wybierać konkretny numer na klawiaturze lub odpowiadać na IVR prostymi słowami, klienci mogą komunikować się z voicebotami w taki sam sposób, jakby rozmawiali z rzeczywistą osobą [19].

W artykule [21] przedstawiono badania wykorzystania voicebota jako alternatywnego rozwiązania do klasycznego IVR'a. Porównano obsługę połączeń z użyciem IVR i z użyciem voicebota. Zastosowanie voicebota spowodowało, że liczba klientów, którzy rozłączają się przed przełączeniem do agenta, spadła o 5%, a liczba klientów, którzy rozłączają się już po przełączeniu się do agenta, zmniejszyła się o 6%. W efekcie przy zastosowaniu voicebota w porównaniu do IVR liczba klientów przełączonych i obsłużonych przez agenta wzrosła aż o 10%.

2.2 Architektura wirtualnych konsultantów

Na rysunku 2.1 przedstawiono ogólną architekturę voicebota. Podstawowymi komponentami voicebota są:

- *IP PBX* [22] to platforma realizująca połączenia telefoniczne z wykorzystaniem technologii VoIP [23].
- *Speech To Text (System ASR)* [24] to platforma realizująca usługę zamiany mowy na tekst.
- *Text To Speech (System TTS)* [25] to platforma realizująca usługę zamiany tekstu na mowę.
- *Platforma NLU* [26] to platforma uczenia maszynowego dla zautomatyzowanych rozmów tekstowych, odpowiedzialna między innymi za rozpoznawanie intencji.
- *Model NLU* to zapis cyfrowy procesu uczenia *Platformy NLU*, najczęściej przechowywany na tej platformie w postaci pliku binarnego.

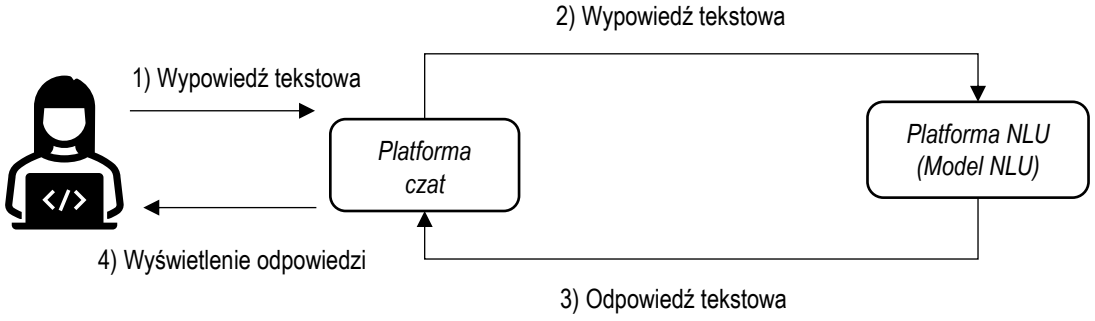


Rysunek 2.1. Architektura voicebota

Zarówno serwer *IP PBX* jak i serwer *Text To Speech* to rozwiązania stosowane już od wielu lat i realizujące swoje funkcje niemal bezbłędnie, stąd również w voicebotcie są to komponenty, które nie wpływają na jego skuteczność.

Na rysunku 2.2 przedstawiono ogólną architekturę chatbota. Podstawowymi komponentami chatbota są:

- *Platforma czat* to platforma realizująca usługę konwersacji tekstowej.
- *Platforma NLU* to platforma uczenia maszynowego dla zautomatyzowanych rozmów tekstowych, odpowiedzialna między innymi za rozpoznawanie intencji.
- *Model NLU* to zapis cyfrowy procesu uczenia *Platformy NLU*, najczęściej przechowywany na tej platformie w postaci pliku binarnego.



Rysunek 2.2. Architektura chatbota

Platforma czat to serwer, którego działanie nie ma istotnego wpływu na skuteczność chatbota, gdyż jego rola sprowadza się do wymiany komunikatów tekstowych między klientem a botem.

Skuteczność voicebota i chatbota w największym stopniu zależy od *Platformy NLU*, od jej poprawności działania, ale również od jakości *Modelu NLU*, który powstaje w wyniku procesu

uczenia. W przypadku chatbota, poza *Platformą NLU* i *Modelem NLU* nie ma innych istotnych elementów z punktu widzenia jego skuteczności. Natomiast dla voicebota istotną rolę dodatkowo odgrywa usługa *Speech To Text (System ASR)*, gdyż od jakości transkrypcji wypowiedzi klienta zależy bezpośrednio jakość fraz wejściowych przekazywanych do *Platformy NLU*.

Analiza architektury voicebota i chatbota wskazuje, na które komponenty należy zwracać uwagę pod kątem budowania skutecznych botów dla Contact Center. Istotny jest wybór zarówno *Platformy NLU* jak i usługi *Speech To Text*, a także przebadanie ich pod kątem określonego modelu językowego oraz jakości rozmów pojawiających się w danym obszarze zastosowania. Szczegółowa analiza powinna wskazać, które funkcje w tych komponentach wymagają poprawy pod kątem zastosowania ich w Contact Center.

2.3 Systemy automatycznego rozpoznawania mowy

Do niedawna rozpoznawanie mowy stanowiło jedno z największych wyzwań w zakresie uczenia maszynowego, z uwagi na nieograniczoną naturę i wewnętrzną zmienność języka mówionego. Połączenie głębokich sieci neuronowych ze starszymi koncepcjami modelowania neuronowego przyniosły postęp zarówno w modelowaniu akustycznym, jak i językowym. Systemy te zazwyczaj wykorzystują architektury konwolucyjnych sieci neuronowych w modelowaniu akustycznym oraz wielowarstwowe sieci rekurencyjne z bramkowaną pamięcią zarówno do modelowania akustycznego, jak i językowego [27].

W ciągu ostatnich kilku lat wydajność systemów automatycznego rozpoznawania mowy znacznie się poprawiła. W rezultacie zastosowanie ASR jest coraz bardziej powszechne i liczba wdrożeń znacznie się zwiększyła [28]. Mimo to nadal większość systemów ASR posiada ograniczenia w zakresie częstotliwość próbkowania, szumu, hałasu, głośności wypowiedzi, modeli językowych oraz zastosowania w różnych dziedzinach począwszy od wyszukiwania głosowego, automatycznych napisów do filmów, aż po Contact Center [29].

Opisywana w literaturze skuteczność rozpoznawania mowy przez znane systemy ASR sięga nawet 95% [30]. Wartość ta osiągana jest jednak zwykle w przypadku nagrań o jakości studyjnej (np. w filmach, w audycjach radiowych). Sygnały dźwiękowe typowo występujące w systemach CC są niezwykle zróżnicowane oraz często zakłócane. Sygnał taki cechuje zwykle częstotliwość próbkowania zaledwie na poziomie 8 kHz, 16 bitowa rozdzielczość oraz format PCM (ang. pulse-code modulation). Ponadto, rozmowy realizowane są często z wykorzystaniem technologii VoIP (ang. Voice over Internet Protocol) zapisywane zwykle w formacie mono bez wydzielonych osobno kanałów klienta i agenta, przez co wypowiedzi rozmówców nakładają się na siebie. Kanały te charakteryzują się zwykle różnymi parametrami, np. różną amplitudą, czy też różnym poziomem szumów pochodzących z otoczenia rozmówców [31]. Dla wspomnianych parametrów sygnału dźwiękowego w modelu języka angielskiego osiągana skuteczność automatycznego rozpoznawania mowy oceniona została na poziomie 86% [32]. Wartość ta jest na granicy możliwości efektywnego wykorzystywania otrzymywanych transkrypcji automatycznych w systemach CC. Systemy transkrypcji automatycznych są trudniejsze w realizacji i jeszcze mniej efektywne dla mało popularnych modeli językowych, między innymi dla języka polskiego.

Każdy system ASR należy oceniać stosując kilka kryteriów: poziom wsparcia danego modelu języka, mechanizmy tworzenia niestandardowych modeli danego języka, wydajność i skuteczność czy radzenie sobie ze słabym sygnałem dźwiękowym. Głównym komponentem,

który wykorzystywany jest w budowie voicebotów są metody rozpoznawania intencji. Metody te budowane są z kolei z wykorzystaniem systemów ASR [33]. Obecnie na rynku istnieje wiele różnych systemów ASR dostępnych zarówno nieodpłatnie jak i na zasadach komercyjnych. Systemy te cechują się wieloma różnymi parametrami, od których zależą możliwości ich wykorzystania w budowie metod rozpoznawania intencji. Wykonano przegląd 10 popularnych systemów rozpoznawania mowy pod kątem potencjalnego użycia ich w Contact Center. W tabeli 2.1 zestawione zostały popularne systemy ASR wraz z parametrami istotnymi z punktu widzenia możliwości ich potencjalnych zastosowań w systemach CC.

Tabela 2.1. Wybrane systemy automatycznego rozpoznawania mowy

Lp.	System ASR	Model języka polskiego	Licencja	Strumieniowanie w czasie rzeczywistym	API	Model użycia oprogramowania	Raporty i rozliczenia	WER literatura	WER eksperymenty
1	Microsoft Speech to Text [34]	Tak	Komercyjna	Tak	http	SaaS	Tak	10,98% [35]	16,07%
2	Nuance ASR [36]	Tak	Komercyjna	Tak	http	On Premises	Nie	58,00% [37]	40,25%
3	Google Speech-to-Text [38]	Tak	Komercyjna	Tak	gRPC	SaaS	Tak	11,87% [35]	17,79%
4	VoiceLab ASR [39]	Tak	Komercyjna	Tak	http	On Premises	Nie	-	17,89%
5	Techmo ASR [40]	Tak	Komercyjna	Tak	gRPC	On Premises	Nie	-	25,25%
6	PrimeSpeech [41]	Tak	Komercyjna	Tak	MRCP	On Premises	Nie	-	-
7	Mozilla Deep Speech [42]	Częściowo	Open Source	Tak	Natywni klienci	On Premises	Nie	65,00% [37]	-
8	Amazon Transcribe [43]	Nie	Komercyjna	Tak	http	SaaS	Tak	10,27% [35]	-
9	Facebook wav2letter [44]	Nie	Open Source	Nie	Native	On Premises	Nie	-	-
10	IBM Watson Speech to Text [45]	Nie	Komercyjna	Tak	http	SaaS	Tak	20,79% [35]	-

Artykuł [46] wymienia trzech światowych gigantów technologicznych jako czołówkę w zakresie systemów ASR: Google, Microsoft i IBM. Google Speech-to-Text (GST) jest zdecydowanym liderem wśród innych usług ASR pod względem dokładności i obsługiwanych języków (w tym języka polskiego). Możliwe jest w nim bezpośrednie przesyłanie strumieniowanych plików dźwiękowych. Obsługuje między innymi formaty kodowania dźwięku takie jak MP3, FLAC, LINEAR16, MULAW, AMR, AMR WB, OGG OPUS, WEBM OPUS. Rozpoznawanie mowy można dostosować do konkretnego kontekstu, dodając do słownictwa zbiór podpowiedzi. Microsoft Speech to Text (MST) obsługuje język polski i strumieniowanie audio poprzez interfejs API REST. Domyślny format przesyłania strumieniowego audio to WAV (częstotliwość próbkowania 16 kHz lub 8 kHz, 16 bitowa rozdzielczość), są obsługiwane również inne formaty: MP3, FLAC, OGG OPUS, ALAW, MULAW. MST wyróżnia się tym, że usługa umożliwia użytkownikom tworzenie własnych modeli językowych na bazie modeli podstawowych (danych akustycznych i językowych). Usługa IBM Watson Speech to Text to system ASR, który wykorzystuje inteligencję maszynową do łączenia informacji o gramatyce i strukturze języka z wiedzą o składowych

sygnału audio w celu dokładnej transkrypcji ludzkiego głosu. Podczas przetwarzania kolejnych fragmentów mowy, system potrafi wykonać wsteczną korektę transkrypcji. System obsługuje strumienie audio z częstotliwością próbkowania 16 kHz i 8 kHz między innymi w formatach MP3, MPEG, WAV, FLAC, OPUS. Niestety IBM Watson nie wspiera natywnie modelu języka polskiego.

Według badań [35], dla głównych trzech platform (MST, GST, IBM Watson) zamiany mowy na tekst dla języka angielskiego, średni współczynnik błędów słów (WER) wyniósł nawet 10,98%. Niemniej jednak przeprowadzone przez autora tej pracy wstępne eksperymenty wykazywały, że dla języka polskiego zarówno dla MST jak i GST średni WER przekroczył wartość 16% (szczegółowe dane pokazano w tabeli 2.1).

W artykule [47] zestawiono między innymi Amazon Transcribe z Nuance ASR. Amazon Transcribe charakteryzuje się wysoką dokładnością transkrypcji. AWS od 12 lat jest najbardziej wszechstronną i powszechnie stosowaną platformą chmurową na świecie. To doświadczenie można zobaczyć w wynikach dokładności. W przeciwieństwie do innych usług transkrypcji, platforma transkrybowania Amazon produkuje teksty gotowe do użycia, bez konieczności dalszej ich edycji. Aby to osiągnąć, AWS Transcribe zwraca szczególną uwagę na interpunkcję, wynik zaufania, możliwe alternatywy, generowanie znacznika czasu, niestandardowe słownictwo. Z kolei Nuance Transcription Engine jest uznawany za najlepsze oprogramowanie do transkrypcji medycznej na świecie. Niestety oznacza to, że w innych branżach, dokładność transkrypcji nie będzie tak dobra, jak transkrypcji medycznych. Obsługuje strumieniowane audio z częstotliwością próbkowania od 8 kHz (dla telefonii) do 99 kHz. Obsługuje formaty audio: WAVE PCM, MS ADPCM, IMA ADPCM, ALAW, MULAW, VOX, MP3, WAV, WMA, DSS, DS2 i M4A. Zawiera zbiór podstawowych modeli akustycznych, które zapewniają niezwykle dokładne wyniki rozpoznawania dla przeszkolonych profili użytkowników [48].

Według badań [35], dla platformy Amazon, średni WER zamiany mowy na tekst dla języka angielskiego wyniósł 10,27% i był najlepszy spośród wszystkich transkryptorów zestawionych w teście. Natomiast z punktu widzenia prowadzonych badań elementem eliminującym Amazon Transcribe był brak wsparcia dla języka polskiego.

Mozilla Deep Speech [49] to system zamiany mowy na tekst o otwartym kodzie źródłowym, wykorzystujący model wyszkolony za pomocą technik uczenia maszynowego i oparty na artykule badawczym Baidu Deep Speech [50]. DeepSpeech korzysta z TensorFlow firmy Google.

Jeżeli chodzi o Nuance ASR i Mozilla Deep Speech dostępne są testy porównawcze przeprowadzone w 2018 roku dla rozmów dotyczących opieki zdrowotnej [37], które pokazują, że Mozilla Deep Speech osiągnęła WER na poziomie 65%, podczas gdy Nuance uzyskał 58%. W testach były również uwzględnione ASR Google 44%, IBM Watson 38% i Microsoft 35%, które wypadły lepiej od Mozilla Deep Speech oraz Nuance. Ponadto model języka polskiego w Mozilla Deep Speech należałoby zbudować samodzielnie co wykluczyło go z dalszych badań.

Wav2Letter to system rozpoznawania mowy, który został opracowany przez Facebook AI. Ostatnio Wav2Letter przekształcił się w zestaw narzędzi zorientowany na wydajność, umożliwiający wybór spośród różnych architektur sieciowych, funkcji strat i danych

wejściowych. Oprogramowanie jest we wczesnej fazie rozwojowej i aktualnie nie nadaje się do zastosowania w Contact Center [49].

Z uwagi, że w prowadzonych badaniach kluczowa była obsługa języka polskiego, rozpatrywano również polskich producentów ASR: VoiceLab, PrimeSpeech oraz Techmo. Pozyskano wersje testowe od VoiceLab oraz Techmo. Każdy z tych dwóch producentów dostarcza bardzo skuteczny system ASR, który cechuje między innymi możliwość transkrypcji nagrań lub transkrypcji strumienia audio w czasie rzeczywistym, wsparcie dla nagrań o częstotliwości próbkowania 16 kHz oraz 8 kHz, wsparcie różnych formatów nagrań: WAV, RAW, FLAC. ASR'y te dysponują intuicyjnym API programistycznym HTTP (VoiceLab) lub gRTP (Techmo). Jedną z kluczowych ich cech jest to, że obsługują język polski i do tworzenia jego modelu przywiązują bardzo dużą uwagę, gdyż jest to dla nich podstawowy język.

Na podstawie oceny zestawionych w tabeli 2.1 parametrów określono, które z systemów ASR mogą potencjalnie zostać wykorzystane na potrzeby budowy metody transkrypcji dedykowanej dla systemów CC obsługujących język polski. Kryteria jakie były analizowane to przede wszystkim: natywna obsługa modelu języka polskiego, sposób licencjonowania, możliwości pracy w czasie rzeczywistym, dostępne API programistyczne, model dostarczania oprogramowania SaaS (ang. Software as a Service – w chmurze) lub On Premises (oprogramowanie instalowane na własnej infrastrukturze) oraz dostęp do zaawansowanych, przejrzystych narzędzi służących do monitorowania, raportowania i rozliczania oferowanych usług (co jest bardzo istotne z punktu widzenia zastosowań biznesowych). Ponieważ, prezentowane w pracy rozwiązanie dedykowane jest do transkrypcji rozmów w systemach CC prowadzonych w języku polskim, dlatego naturalnym kryterium wyboru była konieczność natywnej obsługi modelu języka polskiego. Z punktu widzenia docelowych zastosowań w CC wsparcia tego języka nie posiadają lub posiadają na poziomie niewystarczającym następujące rozwiązania: Amazon Transcribe, Facebook wav2letter, IBM Watson Speech to Text oraz Mozilla Deep Speech. Dlatego też, wyżej wspomniane systemy nie były uwzględniane podczas badań prezentowanych w dalszej części niniejszej pracy. Innym istotnym kryterium, z punktu widzenia docelowego zastosowania opracowywanego rozwiązania była konieczność uruchomienia go w bezpiecznym oraz popularnym środowisku chmurowym typu Microsoft Azure, Google Cloud, Amazon AWS, IBM Cloud lub Oracle Cloud Infrastructure [51]. Dodatkowym aspektem było zapewnienie przez system ASR możliwości douczania modelu językowego poprzez wykorzystanie dostępnych wbudowanych wewnętrznie mechanizmów uczących.

Poza wyżej opisanymi kryteriami zdecydowano również o samodzielnej weryfikacji 5 transkryptorów: MST, GST, VoiceLab, Techmo, Nuance. Celem był wybór dwóch z nich do przeprowadzania docelowych badań. Z uwagi na to, że dysponowano już pierwszym zbiorem nagrań rzeczywistych rozmów przeprowadzonych w kanałach dźwiękowych systemu Contact Center w języku polskim, postanowiono przeprowadzić własne testy transkrypcji automatycznych dla 50 nagrań, których wyniki WER zaprezentowano w tabeli 2.1. Na podstawie analizy powyższych kryteriów oraz wstępnych testów określono, że w dalszych badaniach będą brane pod uwagę 2 rozwiązania: Microsoft Speech to Text oraz Google Speech-to-Text.

Z uwagi, że rozmowy CC są niskiej jakości to ich transkrypcja automatyczna wytworzona nawet przez najlepsze systemy ASR jest niewystarczająca do efektywnego zastosowania jej

w voicebotach. Słabej jakości transkrypcja przekłada się na obniżoną skuteczność rozpoznawania intencji. Intencje są co do zasady trenowane prawidłowymi zdaniami z danego modelu językowego, więc nieuzasadnione jest trenowanie ich zdaniami zawierającymi błędy tylko dlatego, że transkrypcja jest nieprawidłowa. Dlatego skuteczne rozpoznawanie mowy w Contact Center stanowi istotne wyzwanie, aby kolejne komponenty składające się na voicebota mogły prawidłowo funkcjonować.

Nawet najlepsze systemy ASR mają niedoskonałości z punktu widzenia ich zastosowań w Contact Center. System Microsoft Speech to Text nie radzi sobie z nagraniami stereo w zakresie dzielenia kanałów rozmówców. Z kolei Google ASR w zakresie douczania posiada tylko mechanizm podpowiedzi, który nie jest wystarczający w zastosowaniach Contact Center, a jedynie sprawdza się przy tworzeniu rozwiązań do rozpoznawania poleceń głosowych.

W systemach ASR brakuje mechanizmów, które pozwalałyby na ignorowanie dźwięków nieartykułowanych (np. „hmm", „aha", „eee") pojawiających się bardzo często w rozmowach Contact Center. Systemy ASR słabo sobie radzą z wyrazami podobnie brzmiącymi np. „IMEI" versus „email". Wynika to po części z tego, że klienci wymawiają takie wyrazy niewyraźnie, mając świadomość, że taki wyraz osadzony w kontekście będzie zrozumiały. O ile jest to prawdziwym twierdzeniem w odniesieniu do człowieka to nie sprawdza się to w przypadku wirtualnych konsultantów. Ponadto obcojęzyczne słowa są wypowiadane w różny sposób (np. „ajfon", „ifon"), a systemy ASR albo zapisują je w sposób fonetyczny, albo próbują dopasować je do innych słów mieszczących się w ramach modelu danego języka. W efekcie powstaje duża paleta wyrazów posiadających to samo znaczenie, skutkiem czego bardzo trudno jest dla nich zbudować frazy uczące dla danej intencji. Skomplikowana odmiana wyrazów języka polskiego, powoduje, że automatyczne rozpoznawanie mowy często nie jest wystarczająco skuteczne. Istnieje więc potrzeba przeprowadzania dokładnej analizy czy sprowadzenie wyrazów do bezokoliczników będzie skuteczne w przypadku zastosowania ich w voicebotach. W systemach ASR brakuje mechanizmów normalizacji zapisów, między innymi zamiany do małych liter, normalizacji liczb, kwot, dat oraz godzin. To nie tylko negatywnie wpływa na platformy NLU, ale również utrudnia skuteczne mierzenie jakości transkrypcji z wykorzystaniem metryk WER oraz LER.

Należy też podkreślić, że w trakcie przeglądu literatury nie znaleziono prac badawczych dotyczących poprawy jakości transkrypcji automatycznych dedykowanych dla systemów CC opartych na modelu języka polskiego. Wskazuje to na konieczność przeprowadzenia badań ukierunkowanych na rozwiązanie zidentyfikowanych problemów. Zasadnym jest zbadanie czy korekcja wejściowego strumienia audio może zwiększyć skuteczność transkrypcji, czy systemy ASR posiadają mechanizmy douczania, które sprawdzą się w rozmowach CC, czy każdy system ASR prawidłowo radzi sobie z rozdzieleniem kanałów klienta i agenta. Kolejna faza badań powinna koncentrować się na poprawie jakości wygenerowanej transkrypcji automatycznej poprzez detekcję w jej tekście powtarzających się schematów błędów i zaproponowanie metod ich eliminacji. Kryterium decydującym o przydatności danej metody powinny być wyniki metryk WER oraz LER na wydzielonym walidacyjnym zbiorze danych, który nie był wykorzystywany w fazie analizy i przygotowywania metod.

2.4 Platformy NLU wykorzystujące mechanizmy AI/ML

Rozpoznawanie głosu (znane również jako rozpoznawanie mowy lub zamiana mowy na tekst) umożliwia transkrypcję głosu na tekst. Komputer przechwytuje głos za pomocą mikrofonu i dostarcza tekstową transkrypcję wypowiedzianych słów. Korzystając z prostego poziomu przetwarzania tekstu, można opracować funkcję sterowania głosowego za pomocą prostych poleceń, takich jak „skręć w lewo" lub „zadzwoń do Jana". Aby osiągnąć wyższy poziom zrozumienia, poza prostymi poleceniami należy uwzględnić warstwę NLU. NLU spełnia zadanie czytania ze zrozumieniem. Aby zbudować dobrą warstwę NLU, która może zrozumieć wypowiedzi człowieka, trzeba zapewnić szeroki i kompleksowy zestaw przykładowych pojęć i kategorii w obszarze tematycznym lub domenie. Tym samym należy dostarczyć listę powiązanych próbek, lub jeszcze lepiej zbiór możliwych zdań dla każdej pojedynczej intencji, którą klient może aktywować w voicebocie [29]. Dokładność poziomu zrozumienia zależy od liczby wstępnie skonfigurowanych próbek w voicebocie. Te próbki to zdania wypowiedziane przez klienta, które reprezentują jego intencję. Voicebot następnie przekłada je na działanie lub odpowiedź.

W zakresie budowy voicebotów i chatbotów podstawowym wyzwaniem jest poprawne rozpoznawanie intencji, jakie występują w wypowiedziach klientów w dyskursie Contact Center. Obecnie, na rynku istnieje wiele różnych metod rozpoznawania intencji klientów, stanowiących doskonałe platformy wsparcia do budowy wirtualnych konsultantów działających zarówno w kanałach głosowych, jak i tekstowych [52]. Są one udostępniane nieodpłatnie lub na zasadach komercyjnych, najczęściej w modelu SaaS, czasem w modelu On Premises. Rozwiązania te dostarczają funkcje zapewniające możliwości rozpoznawania intencji dla różnych modeli językowych. Niektóre z rozwiązań [53], [54] umożliwią również budowę złożonych scenariuszy rozmowy oraz dostarczają API programistyczne i mechanizmy tworzenia własnych wtyczek w różnych językach programowania. W tabeli 2.2 zestawiono najbardziej znane platformy NLU umożliwiające rozpoznawanie intencji.

Tabela 2.2. Wybrane systemy rozpoznawania intencji

Lp.	**Platforma NLU**	**Język polski**	**Budowanie scenariuszy rozmowy**	**Rodzaj licencji**	**API**	**Model użycia oprogramowania**		**Senty-ment**	***F1-score***	
						On Premises	**SaaS**		[55]	[56]
1	Microsoft LUIS [57]	Nie	Tak	Komercyjna	Tak	Nie	Tak	Nie	0,71	0,35
2	Google Dialogflow [58]	Tak	Tak	Komercyjna	Tak	Nie	Tak	Tak	0,78	0,82
3	Facebook Wit.ai [59]	Tak	Nie	Komercyjna	Tak	Nie	Tak	Nie	-	0,58
4	IBM Watson Assistant [60]	Nie	Tak	Komercyjna	Tak	Tak	Tak	Nie	0,84	0,82
5	VoiceLab Intent Recognition [61]	Tak	Nie	Komercyjna	Tak	Tak	Tak	Nie	-	-
6	Rasa [62]	Tak	Tak	Open Source	Tak	Tak	Nie	Nie	0,75	0,62

W artykule [55] porównano cztery platformy NLU: LUIS, Dialogflow, IBM Watson i Rasa. Badania przeprowadzono na rzeczywistych danych tekstowych pochodzących z popularnej strony internetowej *stackoverflow.com* [63] zawierającej pytania i odpowiedzi w zakresie języków programowania. Wyniki skuteczność tych badań w postaci metryki *F1-score* [64] zaprezentowano w tabeli 2.2 w kolumnie „[55]". W badaniach uwzględniono jeszcze jeden korpus *Repository*. Dla tych dwóch korpusów oceniono trzy kryteria: klasyfikację intencji, ocenę ufności, wyodrębnianie encji. W efekcie końcowym stworzono ranking, w którym kolejność pozycji począwszy od najlepszej platformy NLU była następująca:

1. IBM Watson,
2. Rasa,
3. Dialogflow,
4. LUIS.

Z kolei w artykule [56] zestawiono platformy NLU zasilane danymi uczącymi i wiadomościami wejściowymi z chatbota uniwersyteckiego używanego przez studentów. Wyniki skuteczności jako metrykę *F1-score* zaprezentowano w tabeli 2.2 w kolumnie „[56]". Warto podkreślić, że w tych badaniach ponownie Dialogflow, IBM Watson oraz Rasa osiągnęły najlepsze wyniki. W zestawieniu pojawił się też Wit.ai oraz Amazon LEX, ale obie platformy nie wspierają modelu języka polskiego.

Na podstawie oceny zestawionych w tabeli 2.2 parametrów określono, które z platform NLU mogą potencjalnie zostać wykorzystane na potrzeby budowy metody rozpoznawania intencji uwzględniającej emocje dla systemów CC obsługujących język polski. Do oceny wybrano 6 platform NLU opracowanych przez różnych producentów. Główne kryteria, jakie były brane pod uwagę to: konieczność natywnej, wysokopoziomowej obsługi modelu języka polskiego oraz konieczność zapewnienia możliwości budowy własnych scenariuszy rozmów. Kryterium językowe nie zostało spełnione w stopniu wystarczającym przez rozwiązania Microsoft LUIS oraz IBM Watson Assistant. Z kolei rozwiązania Facebook Wit.ai oraz VoiceLab Intent Recognition nie udostępniają możliwości tworzenia własnych scenariuszy rozmów.

W wyniku przeglądu literatury, analizy powyższych kryteriów oraz wstępnych testów określono, że w dalszych badaniach będą brane pod uwagę dwa rozwiązania: Google Dialogflow oraz Rasa. Dialogflow to platforma komercyjna działająca w modelu SaaS, natomiast Rasa jest platformą Open Source udostępniającą oprogramowanie w modelu On Premises.

Jak wskazano w tabeli 2.2, niektóre metody rozpoznawania intencji mają też zaimplementowane mechanizmy oceniające sentyment [65]. Jednak w istniejących rozwiązaniach, określenie sentymentu nie uwzględnia języka polskiego, jest możliwe tylko dla języka angielskiego [66]. Dodatkowo sentyment rozpoznawany jest jedynie w oparciu o tekst [65] i nie uwzględnia parametrów sygnału dźwiękowego. W tym przypadku wyniki rozpoznania sentymentu w rozmowie są znormalizowane i zawierają się w zakresie od -1.0 do 1.0, pozwalając na określenie afektu negatywnego, pozytywnego lub zachowania neutralnego. Z punktu widzenia budowy inteligentnych botów, prowadzących w sposób automatyczny rozmowy na infoliniach CC, dostarczany mechanizm rozpoznawania sentymentu nie jest wystarczający. Znacznie korzystniejsza byłaby możliwość określania konkretnych rodzajów emocji występujących w rozmowach CC.

Emocjonalne wypowiedzi klientów zawierają często ukryte intencje, których boty nie są w stanie rozpoznać, ponieważ analizują tylko treść wypowiedzi bez uwzględnienia emocji, co jest niewystarczające, aby udzielić klientowi prawidłowej odpowiedzi.

Funkcjonalność właściwej obsługi rozpoznawanych emocji może być kluczową cechą przyszłych wirtualnych konsultantów. W ostatnich latach w coraz większym stopniu korzysta się z voicebotów i chatbotów opartych o sztuczną inteligencję (AI). Stały się one potężnym narzędziem komunikacji między ludźmi a maszynami. Jednak opracowanie lepszych interfejsów głosowych i tekstowych w celu prowadzenia jeszcze bardziej naturalnej rozmowy

z botem wymaga uwzględnienia emocji, ponieważ jest to istotny aspekt komunikacji między ludźmi. Emocje można wykrywać w oparciu o standardowe metody rozpoznawania tekstu i rozumienia języka naturalnego (NLU), a także analizowanie głosu użytkownika [67]. Podejmowane są również próby wbudowania empatii w mechanizmy wirtualnych konsultantów [68], [69] czy usprawnienie interfejsu człowiek-komputer [70].

W 2000 roku, gdy systemy rozpoznawania intencji nie były jeszcze tak rozwinięte jak w obecnych czasach, podjęto próbę zbudowania „Interfejsu człowiek-komputer wrażliwego na emocje" [71]. Został wówczas zaimplementowany moduł zarządzania dialogami jako sieć przejść stanów (ang. state transition network). Eksperymenty opierały się na zbiorze kilku tysięcy smutnych, gniewnych i neutralnych fragmentów mowy z angielskich filmów. Opisano w jaki sposób moduł zarządzania dialogami dostosowywał interakcję z użytkownikiem (podpowiedzi, informacje zwrotne, przepływ dialogu) po rozpoznaniu jego emocji. W artykule [71] wskazano, że poziom wykrycia wyrażanych emocji w wypowiedzi osiągnął około 64%. Nie zaprezentowano jednak algorytmów uczenia i trenowania modułu dialogów, ani wyników skuteczności rozmów człowiek-komputer. W podsumowaniu artykułu wspomniano również, że dodatkowe kategorie emocji oraz poprawa interakcji z użytkownikiem wymaga dalszych badań. Jednak w kolejnych artykułach tych samych autorów [72], [73] nie znaleziono istotnych rozwinięć tematyki budowania dialogów uwzględniających emocje. Artykuł [71] jest na pewno interesującym wstępem w zakresie interakcji człowiek-komputer z uwzględnieniem emocji, ale z uwagi, że powstał w 2000 roku nie obejmuje tematyki platform NLU opartych o sztuczną inteligencję. Ponadto jego mankamentem jest bazowanie na fragmentach mowy z angielskich filmów, a nie na rzeczywistych rozmowach Contact Center, które cechują się określoną specyfiką i jakością. Nie uwzględnia on też specyfiki języka polskiego. Natomiast interesujące jest, że już wtedy autorzy dostrzegli, że specyficzne dla emocji dostosowania interakcji człowiek-komputer powinny być wyraźnie określone w fazie projektowania dialogu i różnią się w zależności od konkretnego zastosowania.

W artykule [67] opisano badania, które miały w zamierzeniu ocenić, czy użytkownicy mogą oszukiwać voicebota reagującego na emocje. Poproszono niewielką liczbę uczestników o naśladowanie pięciu podstawowych emocji (Radość, Smutek, Złość, Strach, Neutralność) i użyto ogólnie dostępnego detektora emocji w głosie OpenVokaturi [74], aby uzyskać informacje zwrotne na temat tego, czy ich emocje zostały rozpoznane jako zamierzone (rzeczywiste). Badania te są na wczesnym etapie, ale mają być kontynuowane w celu opracowania metod rozróżniania między prawdziwymi i fałszywymi emocjami w voicebotach.

W ramach projektu przedstawionego w pracy [75] opracowano chatbota, który podczas konwersacji pozwala klientom wyrażać emocje poprzez użycie emotikon. Założono, że klienci mogą wybierać emotikony z podanej listy. Wprowadzono w nim cztery stany określające stan klienta: zadowolony, zły, neutralny i ten, który nie używa emotikon. Wykorzystano teorię kontroli afektów (ang. affect control theory) do wnioskowania i śledzenia stanów emocjonalnych osób podczas przesyłania wiadomości online. W ocenie autorów wyniki tych badań mogą poprawić projektowanie chatbotów i sprawić, że ich komunikacja będzie bardziej naturalna. Podkreślają również, że zrozumienie afektywnego znaczenia emotikon pomoże maszynom prowadzić znacznie wydajniejsze rozmowy z ludźmi.

Praca [76] bada skuteczność mechanizmu pamięci długoterminowej LSTM (ang. long short-term memory) opartej na głębokim uczeniu się (ang. deep learning) w identyfikacji emocji

tekstowych. Badanie przeprowadzono na zbiorze danych z sześcioma grupami emocjonalnymi. Wyniki eksperymentalne dowiodły, że klasyfikacja emocji oparta na LSTM zapewnia stosunkowo większą dokładność w porównaniu z istniejącymi metodami uczenia się. Badania skupiają się na interakcjach z chatbotem. Autorzy zauważyli, że klienci często wykazują emocje, takie jak zdenerwowanie, desperacja, przygnębienie itp. Podkreślają, że w takich okolicznościach różnice między oprogramowaniem, a człowiekiem znikają i klienci oczekują, że oprogramowanie dokładnie zinterpretuje zachowanie i uczucia ludzi. Zrozumienie tych emocji i dostosowanie odpowiedzi do klienta umożliwia podtrzymywanie głębszych relacji z klientami i poznanie ich emocjonalnych potrzeb.

Automatyczne rozpoznawanie intencji rozmówcy jest przedmiotem badań związanych z opracowywaniem botów, systemów rekomendacyjnych, systemów obsługi klienta. W systemie AntProphet [77] intencje klienta są określane metodą uczenia maszynowego na podstawie jego dotychczasowego zachowania. Z kolei system SBotScope [78] został opracowany w celu identyfikacji intencji zapytań kierowanych do wyszukiwarek internetowych. Głównym celem było wykrywanie złośliwych zachowań typu wykrywanie luk w systemach czy zbieranie adresów e-mailowych. Również inteligentny Search Bot [79] został opracowany w celu rozpoznawania intencji zapytań. Jednak w tym przypadku celem jest nauczenie się rozpoznawania intencji użytkownika w celu uzyskiwania jak najbardziej trafnych wyników wyszukiwania. Rozpoznawanie intencji klienta metodą uczenia maszynowego zostało zastosowane również w pracy [80]. W tym celu wykorzystano Knowledge Graph wspierający wnioskowanie o rozmytych wymaganiach użytkownika. Wyżej wymienione metody rozpoznawania intencji bazują na analizie treści wypowiedzi oraz historii zachowania rozmówcy. Jak wykazano powyżej, nie zawsze jest to wystarczające i często boty nie rozpoznają poprawnie rzeczywistych intencji klienta, gdyż nie uwzględniają emocji.

Zgodnie z wiedzą autora pracy dotyczącą metod rozpoznawania intencji na infoliniach CC, nie istnieją rozwiązania biorące pod uwagę konkretne typy emocji dla języka polskiego, jakie towarzyszą klientowi podczas rozmowy z wirtualnym konsultantem (botem). Z semantycznego punktu widzenia obecność określonej emocji jest w wielu przypadkach niezmiernie istotna we właściwym rozpoznawaniu faktycznych intencji [81]. Przykładowo, często identyczne frazy wypowiadane z emocją Radość mają zupełnie inne znaczenie niż wypowiadane z emocją Złość, co też oznacza, że zupełnie inne są właściwe intencje rozmówcy.

Autor podkreśla, że w kontekście problematyki CC nie są mu znane badania, które uwzględniałyby wykorzystanie emocji do wykrywania rzeczywistych intencji z wypowiedzi klienta. Dlatego też istotnym zagadnieniem jest prowadzenie badań nad zastosowania mechanizmów rozpoznawania emocji dla potrzeb identyfikacji rzeczywistych (ukrytych) intencji rozmówcy w Contact Center dla języka polskiego. Cecha intencji, związana z wyrażaniem postaw emocjonalnych, wydaje się być szczególnie istotna z punktu widzenia interakcji komunikacyjnych. Uwzględnienie roli emocji, a w szczególności postaw emocjonalnych, które stanowią trzon intencji w przetwarzaniu fraz literalnie bądź nieliteralnie wyrażanych w systemie przetwarzania nowych danych przez bota, może znacznie podwyższyć poziom rozumienia wypowiedzi klienta przez automat, co będzie skutkować zwiększeniem skuteczności wirtualnych konsultantów.

3 Cel i teza pracy

W obecnych czasach jednym z kluczowych aspektów badań nad systemami Contact Center jest ich automatyzacja. W zakresie komunikacji z klientem dotychczasowa obsługa poprzez IVR jest zastępowana przez rozwiązania wirtualnych konsultantów bazujących na sztucznej inteligencji. Rozwiązania te stale się rozwijają i stają się coraz bardziej doskonałe. Jednym z głównych zadań wirtualnego konsultanta jest poprawne rozpoznawanie intencji klienta. Poprawność rozpoznawania intencji powinna osiągać jak najwyższy poziom, gdyż wraz ze wzrostem skuteczności rozpoznania intencji, wzrasta skuteczność obsługi klienta. Kluczowym komponentem wpływającym na skuteczność voicebotów i chatbotów są platformy konwersacyjne oparte o modele NLU. W przypadku voicebotów dochodzi jeszcze jeden istotny element jakim są systemy automatycznego rozpoznawania mowy ASR, które są odpowiedzialne za wykonywanie transkrypcji rozmów głosowych w czasie rzeczywistym. Obecne rozwiązania systemów ASR dla wielu mniej popularnych języków nie gwarantują w pełni zadowalającej jakości transkrypcji rozmów głosowych prowadzonych na infoliniach. Spowodowane jest to specyfiką generowanego tam sygnału dźwiękowego, którego parametry jakościowe znacznie odbiegają od nagrań studyjnych. W rozdziale 2 w kontekście systemów ASR zwrócono szczególną uwagę na zagadnienia związane ze słabą transkrypcją dla niskiej jakości rozmów z CC, co bezpośrednio przekłada się na obniżoną skuteczność rozpoznawania intencji na platformach NLU. Problemami związanymi z nieprawidłowym rozpoznawaniem i zapisem podobnie brzmiących wyrazów, a także słów zaczerpniętych z języków obcych. Stwierdzono, że brakuje mechanizmów, które pozwalałyby na ignorowanie dźwięków nieartykułowanych. Ponadto istnieje problem z normalizacją liczb, kwot, dat, godzin, interpunkcją czy wielkością liter. Zagadnienia związane z możliwościami poprawy jakości transkrypcji wykonywanych dla dedykowanych systemów CC o skomplikowanym korpusie językowym nie zostały dotychczas opisane w światowej literaturze. Z kolei platformy NLU bardzo dobrze sobie radzą z rozpoznawaniem intencji w wypowiedziach klientów, ale tylko wtedy, gdy rzeczywista intencja jest zgodna z semantyką wypowiedzi. Przeprowadzona analiza lingwistyczna rzeczywistych rozmów z infolinii Contact Center dla kanału głosowego i tekstowego wykazała, że istotny wpływ na rozpoznawanie rzeczywistych intencji człowieka mają, poza treścią wypowiedzi, również emocje w niej zawarte. Można zauważyć, że brak jest istotnych badań i publikacji, które poruszałyby wątek emocji w konwersacjach Contact Center w kontekście funkcjonowania wirtualnych konsultantów. W konsekwencji nie znaleziono dedykowanych rozwiązań, które znalazłyby zastosowanie w obszarze komunikacji voicebotów i chatbotów z klientem uwzględniając zarówno treść wypowiedzi jak i emocje w niej zawarte.

Problematyka transkrypcji automatycznej dla niskiej jakości rozmów telefonicznych Contact Center oraz wpływ emocji na rozpoznawanie rzeczywistych intencji klientów stanowiły dla autora motywację do podjęcia badań. W efekcie analizy istniejącego stanu wiedzy z zakresu proponowanej tematyki badawczej sformułowana została następująca teza:

Możliwa jest poprawa skuteczności wirtualnych konsultantów w Contact Center poprzez:

1. *minimalizację niepoprawnie rozpoznawanych wypowiedzi klientów z zastosowaniem preprocessingu i postprocessingu danych systemu ASR,*
2. *rozpoznawanie ukrytych intencji klientów na podstawie emocji zawartych w ich wypowiedziach.*

W celu udowodnienia postawionej tezy określone zostały następujące cele pracy:

- Opracowanie nowej metody poprawy jakości transkrypcji automatycznej dla rozmów Contact Center.
- Opracowanie nowej metody rozpoznawania ukrytych intencji klientów na podstawie emocji zawartych w ich wypowiedziach.
- Budowa kompleksowego systemu zwiększającego skuteczność voicebotów i chatbotów w Contact Center z zastosowaniem nowo opracowanych metod (poprawy transkrypcji i rozpoznawania ukrytych intencji).
- Ocena skuteczności opracowanych metod poprzez wykonanie eksperymentów dla rzeczywistych rozmów CC i porównanie uzyskanych wyników ze skutecznością istniejących rozwiązań.

Zakres prac badawczych służących do realizacji wyżej postawionych celów oraz do udowodnienia tezy obejmował:

1. Krytyczną analizę aktualnego stanu wiedzy z zakresu systemów automatycznego rozpoznawania mowy.
2. Krytyczną analizę aktualnego stanu wiedzy dotyczącą platform NLU w zakresie mechanizmów rozpoznawania intencji.
3. Opracowanie metody poprawy jakości transkrypcji automatycznych oraz implementacja jej w formie programu komputerowego.
4. Opracowanie i implementacja metody rozpoznawania ukrytych intencji na podstawie emocji oraz integracja z modułem rozpoznawania emocji [82].
5. Integrację oprogramowania z pięcioma systemami ASR (Microsoft, Google, VoiceLab, Techmo, Nuance) oraz z dwoma platformami NLU (Google Dialogflow oraz Rasa).
6. Implementację oprogramowania przygotowującego dane uczące i testowe oraz opracowanie mechanizmu przekazywania rozpoznanych emocji jako encji.
7. Opracowanie zasad budowania mechanizmu reguł wnioskowania na podstawie przekazanych emocji. Implementacja tego mechanizmu na dwóch wybranych platformach NLU (Google Dialogflow oraz Rasa).
8. Przygotowanie środowiska testowego złożonego z oprogramowania narzędziowego oraz danych testowych.
9. Przeprowadzenie szczegółowych badań i testów z wykorzystaniem rzeczywistych rozmów Contact Center.

4 Formalny model wirtualnego konsultanta

W pracy w kontekście wirtualnych konsultantów będą stosowane następujące definicje.

Definicja 4.1

Fraza P_i jest to pojedyncza wypowiedź klienta.

Fraza jest to sekwencja słów wypowiedziana lub napisana jednorazowo przez rozmówcę. W przypadku rozmowy głosowej zwykle jest ona zakończona ciszą stanowiącą element kończący wypowiedź. Taka wypowiedź jako strumień audio podlega transkrypcji automatycznej, w efekcie której powstaje tekst. W przypadku rozmowy czat fraza to wiadomość tekstowa wysłana przez klienta przez jednokrotne wciśnięcie przycisku „Wyślij".

Konwersacja to sekwencja naprzemiennych wypowiedzi rozmówców. Jednym rozmówcą jest zawsze klient, natomiast drugim jest wirtualny konsultant (bot).

Definicja 4.2

Emocja E_f jest atrybutem frazy.

Emocja używana jest zawsze w kontekście pojedynczej wypowiedzi i ma na celu określenie czy dana wypowiedź ma charakter neutralny czy zawiera jakąś emocję spośród zdefiniowanego zbioru. W pracy zakłada się, że emocje w wypowiedzi są znane, metoda automatycznego rozpoznawania emocji jest przedstawiona w pracy [82].

S_E oznacza zbiór używanych emocji:

$$S_E = \{E_1, E_2, \dots, E_n\} \qquad (4.1)$$

Definicja 4.3

Intencja INT_j określa klasę frazy P_i.

Intencja określa zamiar klienta wynikający z wypowiedzi. Przykład fraz sklasyfikowanych w dwie intencje:

- INT_1*: Informacje o umowie:*
  - P_1*: Kiedy kończy się umowa?*
  - P_2*: Jakie mam usługi na umowie?*
- INT_2*: Przedłużenie umowy:*
  - P_3*: Chcę przedłużyć umowę.*
  - P_4*: Czy mogę przedłużyć umowę?*

Definicja 4.4

Akcja A_k określa odpowiedź bota wynikającą z intencji INT_j.

Akcja bota A_k to akcja wykonywana przez bota na podstawie rozpoznanej intencji i dodatkowych warunków zdefiniowanych na platformie NLU. W najprostszym przypadku akcja bota udziela odpowiedzi tekstowej, która jest przekazywana do klienta – w chatbocie jest wyświetlana w postaci tekstu, a w voicebocie odtwarzana w postaci dźwięku z użyciem mechanizmu zamiany tekstu na mowę (*Text To Speech*).

Przykład: Jeżeli rozpoznana zostanie intencja *Przedłużenie umowy* i będzie warunek *IF (klientZadłużonyVar = true)* to odpowiedzi będą następujące:

- Dla klienta zadłużonego: *Przygotujemy nową umowę po spłacie zadłużenia.*
- Dla klienta nie zadłużonego: *Przygotujemy nową umowę niezwłocznie.*

Scenariusz rozmowy to opis przebiegu rozmowy na platformie NLU, w najprostszym przypadku jest to określenie, że jeżeli wystąpi dana intencja, to ma zostać wybrana określona akcja bota.

Zbiór intencji S_{INT} określa wszystkie możliwe intencje rozmówcy:

$$S_{INT} = \{INT_1, INT_2, \dots, INT_n\} \quad (4.2)$$

Zbiór akcji S_A określa wszystkie możliwe akcje bota dla poszczególnych intencji i fraz wejściowych:

$$S_A = \{A_1, A_2, \dots, A_n\} \quad (4.3)$$

Definicja 4.5

Funkcja F_{NLU} określa intencję frazy P:

$$F_{NLU}: P \to S_{INT} \quad (4.4)$$

gdzie P jest zbiorem wszystkich możliwych fraz $P_i \in P$.

Definicja 4.6

Funkcja F_A określa akcję bota dla intencji INT_j:

$$F_A: S_{INT} \to S_A \quad (4.5)$$

gdzie $INT_j \in S_{INT}$.

Definicja 4.7

Bot jest złożeniem funkcji $F_{NLU} \circ F_A$.

Projektowanie bota polega na określeniu zbioru intencji, zbioru akcji oraz funkcji F_{NLU} oraz F_A.

Funkcja F_{NLU} jest wyznaczana poprzez stosowanie uczenia maszynowego. Dla potrzeb uczenia maszynowego należy zdefiniować zbiór uczący i testowy.

Zbiór S_{PU} jest zbiorem fraz uczących wykorzystywanym do uczenia platformy NLU:

$$S_{PU} = \{P_1^U, P_2^U, \dots, P_n^U\} \quad (4.6)$$

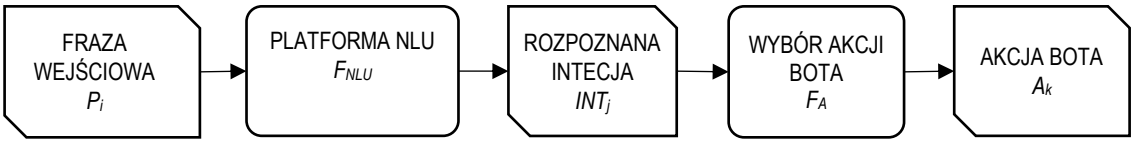
Zbiór S_{PT} jest zbiorem fraz testowych na bazie którego wykonywane jest testowanie platformy NLU:

$$S_{PT} = \{P_1^T, P_2^T, \dots, P_n^T\} \quad (4.7)$$

Procedura uczenia bota składa się z następujących kroków:

1. Opracowanie zbiorów S_{INT} i S_A.
2. Uczenie platformy NLU:
   - Opracowanie zbiorów S_{PU} i S_{PT}.
   - Trenowanie platformy: wykorzystując zbiór S_{PU} należy nauczyć system rozpoznawania intencji ze zbioru S_{INT}.
   - Weryfikacja poprawności rozpoznanych intencji.
3. Przyporządkowanie do każdej intencji odpowiedniej akcji ze zbioru S_A.
4. Walidacja bota: testowanie i ocena działania bota za pomocą zbioru S_{PT}. W przypadku niezadowalających wyników powrót do punktu 2.

Dotychczas stosowany proces określania odpowiedzi przez boty pokazano na rysunku 4.1. Aby wirtualny konsultant mógł rozpocząć konwersację konieczne jest dostarczenie tekstu na wejście platformy NLU w postaci frazy tekstowej. Na podstawie dostarczonej wypowiedzi platforma NLU rozpoznaje właściwą intencję klienta. Po prawidłowym rozpoznaniu intencji, wybierana jest akcja bota, której wykonanie udziela odpowiedzi klientowi [83].



Rysunek 4.1. Proces działania bota

Do oceny skuteczności bota zostanie wykorzystany zbiór fraz testowych S_{PT}.

Załóżmy, że w wyniku testowania bota rozpoznano poprawnie intencje dla *trueINT* fraz, natomiast niepoprawnie dla *falseINT* fraz, gdzie:

$$\overline{S_{PT}} = trueINT + falseINT \qquad (4.8)$$

Do oceny wyników rozpoznawania intencji, zastosowano popularną miarę jaką jest dokładność (ang. accuracy) [84] zdefiniowaną jako:

$$accINT = \frac{trueINT}{trueINT + falseINT} \qquad (4.9)$$

Załóżmy, że w wyniku testowania bota wybrano poprawne akcje dla *trueA* fraz, natomiast niepoprawnie dla *falseA* fraz, gdzie:

$$\overline{S_{PT}} = trueA + falseA \qquad (4.10)$$

Do oceny wyników wyboru akcji bota, zastosowano miarę dokładność zdefiniowaną jako:

$$accA = \frac{trueA}{trueA + falseA} \qquad (4.11)$$

W pracy końcową miarą skuteczności bota będzie *accA*.

Załóżmy, że w wypowiedziach rozpoznano poprawnie emocje dla *trueE* fraz, natomiast niepoprawnie dla *falseE* fraz, gdzie:

$$\overline{S_{PT}} = trueE + falseE \qquad (4.12)$$

Do oceny wyników rozpoznawania emocji, zastosowano miarę dokładność zdefiniowaną jako:

$$accE = \frac{trueE}{trueE + falseE} \qquad (4.13)$$

5 Poprawa jakości transkrypcji w celu zwiększenia skuteczności rozpoznawania intencji

W niniejszym rozdziale przedstawione są wyniki badań, które miały na celu zbadanie wpływu jakości transkrypcji automatycznej na rozpoznawanie intencji przez boty. Celem tych badań było opracowanie metody poprawy jakości transkrypcji dla rozmów CC o niskiej jakości. Zaprezentowano opracowane moduły preprocessingu i postprocessingu oraz wyniki ich skuteczności w zakresie metryk WER oraz LER. Na końcu przedstawiono efekty jakie uzyskano po zastosowaniu tych metod podczas testów rozpoznawania intencji.

5.1 Metodologia badań

Celem badań jest poprawa skuteczności funkcji F_{NLU} poprzez poprawę jakości transkrypcji automatycznej. Słaba jakość transkrypcji automatycznej może spowodować obniżenie *accINT*. Załóżmy, że część fraz w zbiorze testowym ma nieprawidłową transkrypcję w wyniku czego zostanie dla nich niepoprawnie rozpoznana intencja:

n – liczba wszystkich fraz testowych

n_1 – liczba fraz testowych z niepoprawną transkrypcją

Założono, że celem prac będzie uzyskanie jakości transkrypcji, w której zostanie zminimalizowana:

n_2 – liczba fraz testowych z niepoprawną transkrypcją (gdzie $n_1 > n_2$)

Dokładność $accINT_1$ dla n_1:

$$accINT_1 = \frac{n - n_1}{n} \tag{5.1}$$

Dokładność $accINT_2$ dla n_2:

$$accINT_2 = \frac{n - n_2}{n} \tag{5.2}$$

Można zauważyć, że $accINT_1 < accINT_2$, ponieważ:

$$n_1 > n_2 \Longrightarrow \frac{n - n_1}{n} < \frac{n - n_2}{n} \tag{5.3}$$

Warunkiem wyżej wymienionych zależności, jest założenie, że niepoprawne transkrypcje są odwzorowywane w niepoprawne intencje.

Zatem jeżeli poprawi się jakość transkrypcji automatycznych dla niepoprawnie odwzorowywanych fraz to poprawi się dokładność odwzorowywania fraz w intencję.

Jak zostanie wykazane w podrozdziale 7.6 spowoduje to również poprawę dokładności *accA* odwzorowywania fraz w intencje i następnie intencji w akcje bota.

Na podstawie przeglądu aktualnego stanu wiedzy w zakresie systemów automatycznego rozpoznawania mowy zdefiniowano założenie, że rozpoznawanie mowy będzie wykorzystywać istniejące systemy ASR – zakłada się, że istniejące metody są wystarczająco skuteczne dla wysokiej jakości rozmów głosowych, a głównym problemem badawczym będzie niwelacja wpływu słabej jakości rozmów telefonicznych CC na jakość rozpoznawania mowy.

Na podstawie tego założenia, określono kierunki badań:

1. Pierwszym kierunkiem badań była próba poprawy sygnału wejściowego do systemu ASR.
2. Drugim kierunkiem badań była próba poprawy transkrypcji wyjściowej z systemu ASR, tzn. minimalizacja błędnie rozpoznanych wyrazów.

Przyjęcie takiego podejścia pozwoliło na skorzystanie z aktualnych osiągnięć nauki i jednocześnie zwiększyło szanse uzyskania oczekiwanych rezultatów.

Prace badawcze dotyczące wpływu jakości transkrypcji na rozpoznawanie intencji podzielono na dwie zasadnicze części. Część pierwszą stanowiły badania modułów preprocessingu, natomiast część drugą badania modułów postprocessingu. W zakresie preprocessingu zbadano kolejno: wpływ rozdzielenia kanałów dla poszczególnych rozmówców (klient/agent); możliwości douczania badanych systemów ASR oraz możliwości korekcji parametrów sygnału dźwiękowego (redukcja szumów, normalizacja poziomów głośności kanałów agenta oraz klienta). W zakresie postprocessingu badano z kolei: możliwości korekty tekstu wraz z normalizacją wybranych elementów, problemy związane z podobnie brzmiącymi słowami, wyrazami pochodzącymi spoza korpusu języka polskiego oraz mechanizm lematyzacji.

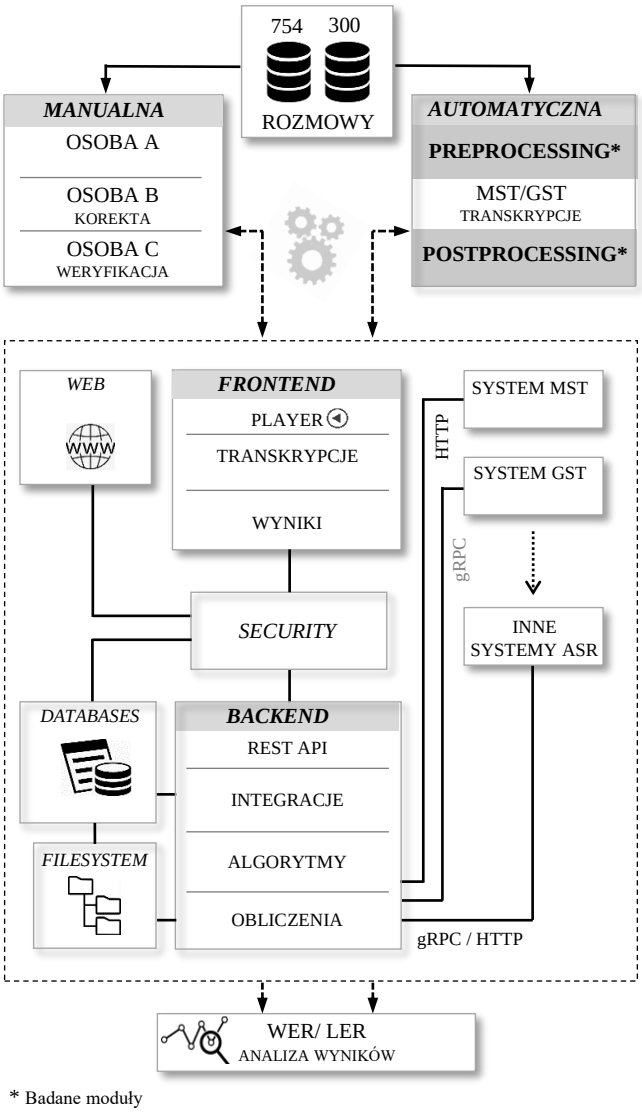
Celem badań wykonanych dla wyżej wymienionych modułów preprocessingu oraz postprocessingu było określenie możliwości wykorzystania ich jako elementów składowych nowej metody poprawy jakości transkrypcji dedykowanej dla systemów CC obsługujących język polski. Podstawowym celem prowadzonych prac było podniesienie jakości transkrypcji automatycznych dla rozmów prowadzonych na infoliniach CC. Wykonane prace pozwoliły docelowo określić moduły współpracujące z systemami ASR, które najlepiej wpływają na podniesienie jakości otrzymywanych transkrypcji.

Metodyka badawcza polegała na opracowaniu oraz integracji z wybranymi systemami ASR poszczególnych modułów preprocessingu i postprocessingu, a następnie ocenie ich skuteczności poprzez porównanie uzyskanych z ich zastosowaniem transkrypcji z transkrypcjami bez stosowania proponowanych modułów. Dla potrzeb oceny jakości transkrypcji opracowano środowisko przedstawione na rysunku 5.1.

W początkowej fazie prac, utworzono bazę danych nagrań dźwiękowych obejmującą 1054 rzeczywiste rozmowy telefoniczne (ang. real phone calls, dalej w pracy określane jako RLPHDB), pomiędzy klientem oraz agentem na infolinii CC. Z bazy tej wylosowano 300 nagrań walidacyjnych (dalej w pracy określanych jako RLPHDB-VLD), służących do końcowej weryfikacji opracowanej metody poprawy jakości transkrypcji. Pozostałe 754 nagrania oznaczono jako badawcze (dalej w pracy określane jako RLPHDB-RSR) i wykorzystano podczas prac nad opracowaniem i optymalizacją metody transkrypcji przeznaczonej dla systemów CC obsługujących język polski. Rozmowy prowadzone były podczas rzeczywistych kampanii realizowanych przez komercyjny duży system CC. W celach testowych wybrane zostały trzy podstawowe tematy rozmów:

1. faktury i płatności,
2. informacje techniczne,
3. umowy oraz aneksy.

Na potrzeby badań założono, że czas trwania każdej pełnej rozmowy powinien być nie krótszy niż 3 minuty oraz nie dłuższy niż 20 minut, co pozwoliło na stworzenie zbioru obejmującego łącznie 100 godzin 53 minuty i 41 sekund nagrań. Wszystkie próbki nagrań zostały zanonimizowane (usunięto wszelkie dane wrażliwe z punktu widzenia ustawy RODO – Rozporządzenie o Ochronie Danych Osobowych).



Rysunek 5.1. Schemat środowiska testowego dla potrzeby oceny jakości transkrypcji

Przygotowanie bazy danych nagrań wymagało stworzenia docelowego środowiska badawczego automatyzującego ich przetwarzanie oraz analizy, jak również gwarantującego powtarzalność otrzymywanych wyników. W środowisku tym:

- wprowadzono algorytmy wspierające procesy wykonywania manualnych transkrypcji referencyjnych (transkrypcji wykonanych przez człowieka),
- zintegrowano wybrane w podrozdziale 2.3 systemy ASR,
- opracowano algorytmy umożliwiające porównywanie przygotowanych transkrypcji referencyjnych z transkrypcjami wykonanymi automatycznie,
- wprowadzono algorytmy obliczające jakość transkrypcji oraz szereg innych funkcji, które powstawały wraz z rozwojem tego środowiska badawczego.

Przygotowane narzędzia w postaci programów komputerowych pozwoliły na efektywną pracę z bazą nagrań, zapewniły możliwość łatwego prowadzenia wielokrotnych eksperymentów porównawczych wykonywanych na dużych zbiorach danych oraz zapewniły łatwą możliwość zmiany parametrów testowych.

Opracowane środowisko badawcze wykorzystano zarówno w procesach transkrypcji manualnych, jak i procesach transkrypcji automatycznych. W tym celu szeroko wykorzystano przygotowane narzędzie odtwarzacza nagrań z możliwością:

- prezentacji widma akustycznego z podziałem na kanały klienta oraz agenta,
- regulacji szybkości odtwarzania,
- łatwej nawigacji pomiędzy nagraniami,
- automatycznego wstawiania znacznika czasu wypowiedzi rozmówcy.

Często również wykorzystywano funkcję wyciszania wybranego kanału oraz możliwości powiększania widma. Uzyskane manualnie transkrypcje rozmów stanowiły dane referencyjne przeznaczone do porównania z danymi wygenerowanymi automatycznie przez badane systemy ASR. Należy podkreślić, że poprawne przeprowadzenie transkrypcji manualnych jest procesem niezmiernie trudnym, kosztownym i czasochłonnym, a jednocześnie kluczowym z punktu widzenia wiarygodności wyników eksperymentów. Dlatego też, proces transkrypcji manualnych obejmował aż trzy kroki:

- W kroku pierwszym wykonana została pierwotna transkrypcja przez pierwszą osobę (osoba A).
- Krok drugi obejmował korektę wykonanej transkrypcji manualnej. Korekty tej dokonywała druga osoba oznaczona na rysunku 5.1 jako osoba B.
- Natomiast w kroku trzecim, kolejna osoba (osoba C) weryfikowała transkrypcję skorygowaną przez osobę B.

W pracach związanych z transkrypcją manualną oraz jej korektą i weryfikacją uczestniczył lingwista specjalizujący się w zagadnieniach języka polskiego. Wykonane transkrypcje zapisywane były w relacyjnej bazie danych w celu dalszego ich przetwarzania. Przeprowadzony w ten sposób proces transkrypcji manualnych zagwarantował ich bardzo wysoką jakość niezbędną do dalszych badań.

Dysponując transkrypcjami referencyjnymi, w kolejnym kroku wykonywane były badania wpływu modułów preprocessingu oraz postprocessingu na jakość transkrypcji automatycznych systemów MST oraz GST. Bezpośrednia integracja wybranych do testowania systemów ASR z opracowanym środowiskiem badawczym wpłynęła na znaczne uproszczenie zadań przetwarzania i porównywania danych otrzymywanych w wyniku transkrypcji automatycznych oraz ich archiwizacji. Samo środowisko zapewnia możliwości docelowej integracji również innych systemów ASR. Z punktu widzenia użytkownika końcowego w opisywanym środowisku główną rolę odgrywają funkcje zaimplementowane w bloku FRONTEND. Blok ten odpowiada za użyteczność oraz ergonomię pracy. Istotnym jego elementem jest zintegrowany odtwarzacz nagrań zapewniający wiele ważnych, wymienionych wcześniej funkcji, które wspierają procesy transkrypcji. Dzięki wbudowanym narzędziom porównawczym transkrypcji manualnych oraz automatycznych i możliwości bieżącej prezentacji otrzymywanych wyników blok ten wspiera również procesy postprocessingu. Realizacja wszystkich funkcji środowiska możliwa jest dzięki wykorzystaniu elementów

BACKEND. W bloku tym, przy wykorzystaniu interfejsu REST API udostępniane są serwisy dla bloku FRONTEND. Blok BACKEND odpowiada za zapisywanie danych w DATABASES oraz FILESYSTEM, zarządzanie danymi oraz ich spójność. BACKEND zapewnia ponadto integrację z systemami ASR oraz modułami postprocessingu. Z wykorzystaniem API (ang. application programming interface) bazującego na protokole gRPC (ang. remote procedure calls opracowany w Google w 2015 roku) zapewniono integrację z systemem GST. Natomiast, przy pomocy API bazującego na protokole HTTP (REST API) zintegrowany został system MST. Środowisko posiada wkomponowane bazy danych, w których stworzono odpowiednie struktury pozwalające na efektywną współpracę bloku DATABASES z innymi elementami systemu. W DATABASES dla każdego z plików dźwiękowych przechowywana jest jego manualna transkrypcja referencyjna oraz wykonane automatyczne transkrypcje dla MST i GST. Każdy plik dźwiękowy przetworzony przez moduł preprocessingu jest przechowywany jako kopia pliku oryginalnego z dedykowanymi mu transkrypcjami automatycznymi. Dla każdej transkrypcji automatycznej przechowywane są obliczone metryki WER oraz LER – zarówno przed jak i po zastosowaniu modułów postprocessingu. Każde nagranie klasyfikowane jest ze względu na temat rozmowy, czas jej trwania, częstotliwość próbkowania i rozdzielczość bitową. Zbiór taki przechowywany jest w bloku FILESYSTEM. Bezpośredni dostęp do zbioru możliwy jest tylko dla uprawnionych i uwierzytelnionych użytkowników z poziomu środowiska badawczego. Blok ten wykorzystywany jest również w prezentacji użytkownikom końcowym transkrypcji automatycznych, które zapisywane są w bloku DATABASES. Zgromadzone w systemie rzeczywiste nagrania rozmów telefonicznych, mimo że zostały zanonimizowane i nie zawierają bezpośrednio danych osobowych, adresowych i kontaktowych, to nadal powinny być chronione z należytą starannością. Dlatego też, konieczne było zastosowanie mechanizmów bezpieczeństwa pozwalających na identyfikację oraz zarządzanie dostępem do zasobów, co zaimplementowano w bloku SECURITY. Środowisko badawcze obsługiwane jest z poziomu dowolnie wybranej przeglądarki internetowej.

Jakość transkrypcji automatycznych, zgodnie z zaleceniami opisywanymi w literaturze [85] określa się poprzez metryki WER oraz LER. Z punktu widzenia CC, metryka WER jest najbardziej przydatna do mierzenia skuteczności systemów ASR [86]. Opiera się ona na metryce edycyjnej Levenshteina dla słów [87]. Wyznaczenie parametrów WER jest możliwe poprzez porównanie ze sobą transkrypcji nagrań pochodzących z danego systemu ASR z transkrypcjami referencyjnymi wykonanymi manualnie przez człowieka. WER jest zdefiniowane następująco [88]:

$$WER = (sw + dw + iw)/nw \qquad (5.4)$$

gdzie: *nw* – liczba słów w transkrypcji referencyjnej, *sw* – liczba zamienionych słów, *dw* – liczba usuniętych słów i *iw* – liczba wstawionych słów. Należy zauważyć, że liczbę poprawnie rozpoznanych słów *cw* określa formuła:

$$cw = nw - dw - sw \qquad (5.5)$$

Ważnych informacji dostarcza również metryka LER, która obliczana jest dla określonych liter jako stosunek między ich nieprawidłowymi transkrypcjami, a całkowitą liczbą jej wystąpień.

Metrykę tę określa równanie [89]:

$$LER = (s + d + i)/n \quad (5.6)$$

gdzie: n – całkowita liczba liter, s – liczba zamiany liter, d – liczba usunięcia liter, i – liczba wstawień liter.

W niniejszej pracy prezentowane są eksperymenty porównawcze oraz analizy dla obu wyżej opisanych metryk. Otrzymane wyniki eksperymentów opisano szczegółowo w podrozdziale 5.3.

5.2 Metoda poprawy jakości transkrypcji dedykowana dla Contact Center

W niniejszym podrozdziale opisane zostały moduły preprocessingu danych wejściowych oraz postprocessingu danych wynikowych systemu ASR, które mogą mieć wpływ na jakość transkrypcji mowy wykonywanej dla próbek pozyskanych z systemów CC obsługujących język polski.

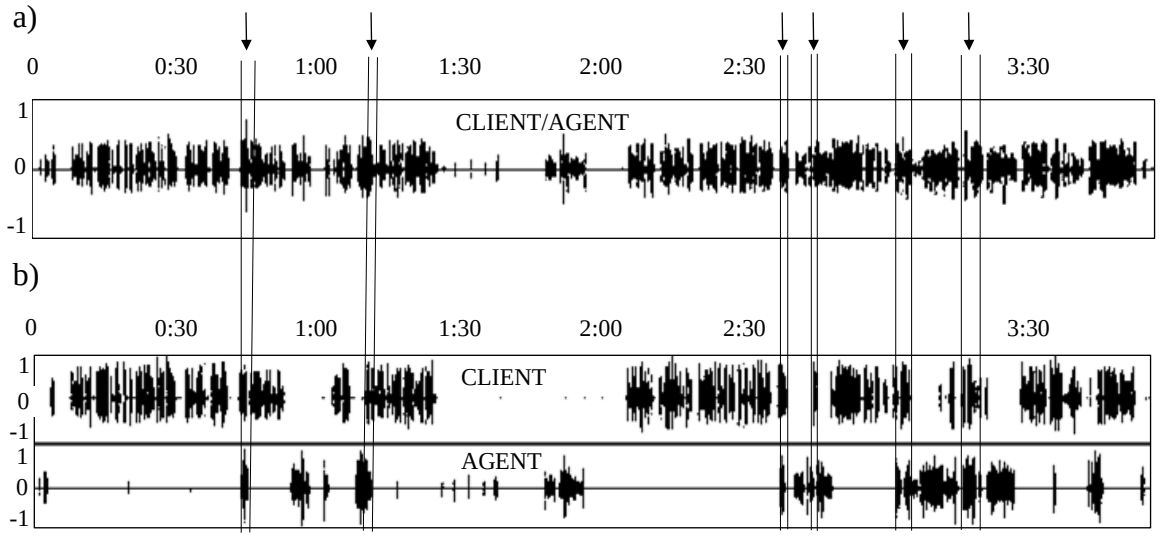
5.2.1 Moduły preprocessingu

Preprocessing danych wejściowych ma na celu poprawę jakości transkrypcji poprzez dostosowanie metody przetwarzania sygnału mowy oraz transkryptora do specyfiki systemu CC. W prezentowanym w tej pracy podejściu zaproponowano zastosowanie następujących trzech modułów:

1. rozdzielenie kanałów klienta oraz agenta,
2. douczanie systemów ASR,
3. korekcję sygnału dźwiękowego.

5.2.1.1 Rozdzielenie kanałów klienta oraz agenta

Wypowiedzi rozmówców często nakładają się na siebie uniemożliwiając tym samym poprawną ich transkrypcję. Dlatego też, w kroku pierwszym postanowiono sprawdzić wpływ rozdzielenia kanałów agenta oraz klienta na dwa niezależne od siebie strumienie danych, poddając je niezależnym transkrypcjom. Podstawowym założeniem jest, aby rozmowy zapisywane były w formacie stereo. Systemy ASR po przetworzeniu dźwięku zwracają tekst, ale również czas rozpoczęcia każdej wypowiedzi. Obie dane należy zapisywać jako pary, gdyż ma to znaczenie dla zachowania właściwej kolejności tekstów w całej transkrypcji. Na rysunku 5.2 przedstawiono część widma jednej z rzeczywistych rozmów badanych w czasie prowadzonych badań. Rysunek 5.2a) pokazuje sygnał, w którym kanały klienta oraz agenta są na siebie nałożone, natomiast rysunek 5.2b) przedstawia ich wyodrębnione widma. Należy zwrócić uwagę, że w prezentowanym dość krótkim fragmencie rozmowy występuje wiele miejsc, w których klient oraz agent mówią jednocześnie. Miejsca te oznaczone zostały strzałkami. Uwzględniając specyfikę rozmów prowadzonych w CC, opisywany przypadek powtarza się wielokrotnie niemal w każdej rozmowie. W odniesieniu do badanych systemów ASR, rozwiązanie GST ma wbudowane mechanizmy separujące oba kanały i radzi sobie z tym problemem względnie dobrze. Rozwiązanie MST nie posiada natomiast wkomponowanych wewnętrznie opcji pozwalających na zoptymalizowaną separację wypowiedzi rozmówców traktując dostarczany na potrzeby transkrypcji plik stereo, jak zwykły plik mono. Dlatego też, aby uzyskać odpowiednie efekty należy strumienie danych klienta oraz agenta zapisać w osobnych plikach, a następnie poddawać je osobnym równoległym transkrypcjom. Szczegółowe wyniki eksperymentów wpływu rozdzielenia kanałów rozmówców na jakość transkrypcji zaprezentowane są na rysunku 5.8 w kolumnie oznaczonej jako SEP.



Rysunek 5.2. Fragment widma rozmowy na infolinii CC: a) kanały rozmówców połączone, b) kanały rozmówców odseparowane

5.2.1.2 Douczanie systemów ASR

Kolejnym badanym mechanizmem preprocessingu było wykorzystanie możliwości douczania udostępnianych w różnej formie przez oba z rozpatrywanych systemów ASR. Rozwiązanie MST umożliwia tworzenie własnego niestandardowego modelu danego języka, na podstawie dobranych próbek nagrań oraz ich transkrypcji referencyjnych. Z kolei, mechanizm uczący w rozwiązaniu GST opiera się na wykorzystaniu wspólnego zestawu przygotowywanych fraz wzorcowych, które najczęściej powtarzają się w wypowiedziach agentów oraz klientów. Oba rozwiązania zostały przetestowane.

MST posiada zestaw narzędzi *Custom Speech* [90], które umożliwiają douczanie dla wielu różnych języków, w tym również dla języka polskiego. W tym celu wykorzystywane mogą być zarówno zintegrowane podstawowe modele językowe, jak i tworzone przez użytkowników własne modele niestandardowe. Przy czym, każdy ze wspieranych języków posiada co najmniej jeden lub więcej modeli podstawowych. Modele niestandardowe tworzone są poprzez douczanie wybranego modelu podstawowego w oparciu o dostarczane próbki nagrań dźwiękowych oraz ich transkrypcje referencyjne zawierające słownictwo specyficzne z punktu widzenia dalszych zastosowań. Douczanie modelu podstawowego systemu MST zrealizowano tworząc własne modele, które w procesie uczenia wykorzystywały różne ilości danych. W tabeli 5.1 zestawiono wykaz przygotowanych modeli wraz z wynikami pokazującymi skuteczność douczania.

Douczanie realizowano z wykorzystaniem 4 modeli, dla każdego zwiększając sukcesywnie liczbę danych uczących od 20 do 158 nagrań wraz z ich transkrypcjami referencyjnymi. Nagrania wybierano spośród próbek zgromadzonych w bazie RLPHDB-RSR. Weryfikację procesów douczania realizowano natomiast na zbiorze 30 testowych nagrań stereo z odseparowanymi kanałami klienta i agenta z bazy RLPHDB-VLD. Dla każdego z modeli wykonywano automatyczne transkrypcje nagrań testujących oraz obliczano metrykę WER. Jak pokazano w tabeli 5.1, zwiększenie ilości danych uczących do 158 pozwoliło na poprawę jakości transkrypcji o 1%. Dalsze douczanie nie poprawiało rezultatów. Po dokładnej analizie

tych transkrypcji zauważono, że słowa, które wielokrotnie powtarzały się w danych uczących, były częściej prawidłowo rozpoznawane nawet gdy sygnał był słabej jakości, ale z kolei pojawiały się błędnie rozpoznane słowa, które w domyślnym modelu języka polskiego rozpoznawane były prawidłowo. Można z tego wyciągnąć wnioski, że domyślny model języka polskiego dostarczany przez MST jest wyuczony z zachowaniem odpowiednich proporcji i douczanie go wcale nie musi przynieść zadowalających efektów.

Tabela 5.1. Parametry uczące dla systemu MST

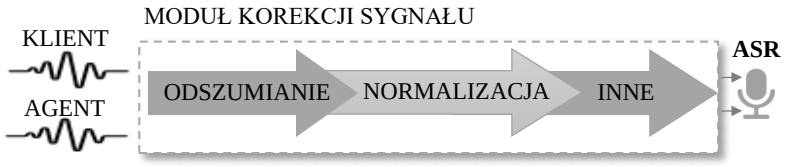
Nazwa	Lista próbek nagrań	Czas trwania rozmów	Poprawa WER
	[szt.]	[hh:mm:ss]	[%]
MODEL 1	20	05:29:52	**0,03**
MODEL 2	50	13:24:40	**0,18**
MODEL 3	100	27:09:23	**0,98**
MODEL 4	158	41:47:16	**1,00**

W przypadku douczania systemu GST wykorzystano z kolei funkcję adaptacji mowy [91] (ang. speech context). Na jej potrzeby przygotowano dwa zbiory po 100 najczęściej powtarzających się fraz, które najczęściej powodowały problemy w transkrypcjach. Jeden dla fraz występujących w kanale klienta, natomiast drugi dla fraz występujących w kanale agenta. Opracowane frazy były różnej długości, od 11 do 96 znaków. Opracowane zbiory wykorzystano w procesie uczenia systemu GST. Po analizie uzyskanych transkrypcji okazało się, że o ile wprowadzenie niektórych słów lub fraz (np. „mąż") poprawiło rozpoznawanie niektórych wypowiedzi (np. lepiej był rozpoznawany wyraz „mąż") to w innych wypowiedziach rozpoznawanie pogorszyło się (np. zamiast wyrazu „masz" rozpoznawany był wielokrotnie „mąż"). Poprawiło się natomiast rozpoznawanie takich słów jak „IMEI", które wcześniej było częściej mylone z „email". Natomiast najbardziej zaskakujący był efekt umieszczenia we frazach uczących słowa „przerwało", otóż powodowało ono, że dalsza część wypowiedzi w transkrypcji była pomijana – wyglądało to na słowo sterujące, ale nie znaleziono żadnej wzmianki na ten temat w dokumentacji systemu GST [91]. W funkcji adaptacji mowy bardzo trudno było dobrać zestaw wspólnych fraz uczących dla wszystkich nagrań, w efekcie wskaźnik WER nie ulegał poprawie przy zastosowaniu tej funkcji GST. Wydaje się, że funkcja adaptacji mowy mogłaby być skuteczna przy bardzo zawężonym i powtarzalnym scenariuszu rozmowy np. doładowanie Internetu jednym ze zdefiniowanych dodatkowych pakietów (2 GB, 5 GB, 10 GB).

Dla tak przygotowanych systemów wykonano transkrypcje automatyczne dla zbioru wybranych 30 nagrań testowych. Szczegółowe wyniki eksperymentów przedstawiono w punkcie 5.3.1 w kolumnach oznaczonych jako AI/ML.

5.2.1.3 Korekcja sygnału dźwiękowego

Bardzo ważnym czynnikiem wpływającym na transkrypcje automatyczne systemów ASR jest jakość sygnału dźwiękowego. Jakość tą można poprawić poprzez redukcję szumów (np. szum pojazdów, wiatru), trzasków (np. stuknięcia w biurko, wciskanie klawiszy na klawiaturze) lub innych niepożądanych dźwięków dobiegających z otoczenia (np. rozmowy innych agentów, rozmowy domowników). Wskazana może być również normalizacja poziomów głośności dla obu odseparowanych kanałów. Można też wykorzystywać inne szeroko dostępne filtry. Mechanizmy obróbki dźwięku są dzisiaj bardzo dobrze rozpoznane i implementowane wewnętrznie w wielu systemach ASR, w tym również w systemach MST oraz GST. Jednakże, w przypadku integracji systemów ASR, które nie optymalizują omawianych parametrów może pojawiać się potrzeba ich korekcji przed wykonaniem transkrypcji. W związku z powyższym jako kolejny element preprocessingu proponowany jest moduł korekcji sygnału dźwiękowego realizujący wyżej wspomniane mechanizmy. Na rysunku 5.3 przedstawiono poglądowy schemat modułu korekcji sygnału dźwiękowego.



Rysunek 5.3. Schemat modułu korekcji sygnału

Testy proponowanego modułu w odniesieniu do systemów MST ora GST nie wykazują potrzeby jego zastosowania, co potwierdzają wyniki eksperymentów prezentowane na rysunku 5.9 w kolumnach oznaczonych jako DEN oraz NORM. Pomimo, że dla człowieka poprawa dźwięku jest istotnie zauważalna to wyniki dla systemów ASR pogarszają się, z uwagi na to, że systemy te często mają dużo lepiej rozwinięte mechanizmy analizy dźwięku i odseparowywania z nich zakłóceń. Dlatego też, w głównej mierze należy zadbać o możliwie jak najlepszą rejestrację nagrania źródłowego, w tym uwzględniającą stereo i wyodrębnienie kanału klienta i agenta.

Proponowany moduł może okazać się jednak potrzebny w przypadku potencjalnej implementacji innych niż systemy MST oraz GST, w których sygnały dźwiękowe nie byłyby dostatecznie dobrze zoptymalizowane przez składowe wewnętrzne tych systemów. Aspekt ten wymaga przeprowadzenia dalszych badań, których prowadzenie nie było celowe z punktu widzenia niniejszej pracy.

5.2.2 Moduły postprocessingu

Moduły postprocessingu mają na celu zwiększyć jakość wyniku końcowego poprzez poprawę tekstu wynikowego systemu ASR. W proponowanym rozwiązaniu zastosowano następujące moduły:

1. korekty tekstu,
2. podobnie brzmiących oraz obcych słów,
3. lematyzacji.

5.2.2.1 Moduł korekty tekstu

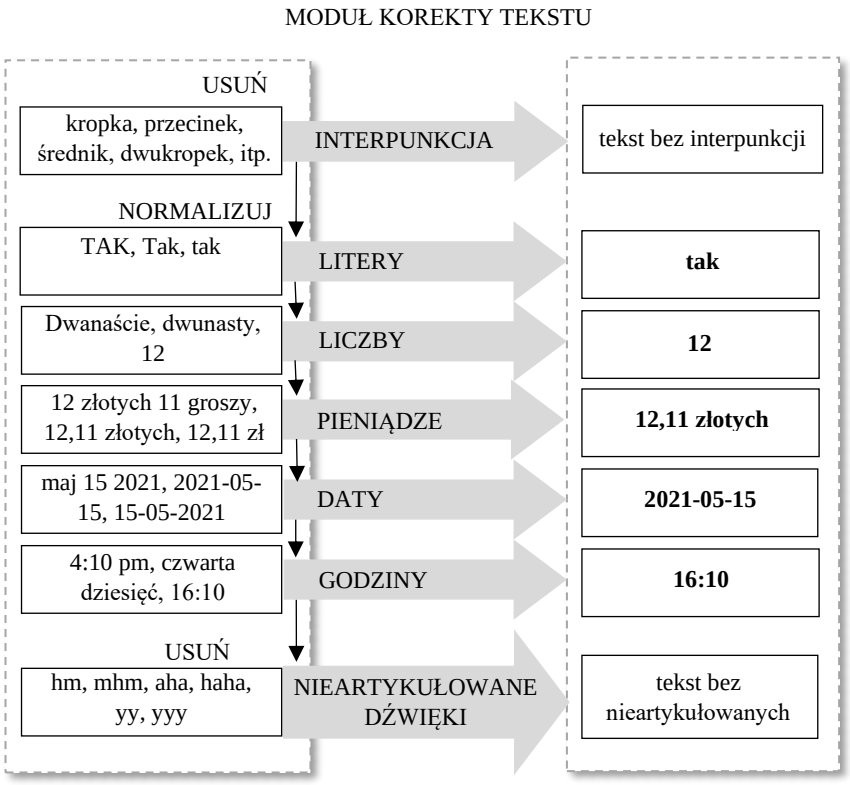
Przeprowadzona szczegółowa analiza porównawcza transkrypcji otrzymanych z uwzględnieniem mechanizmów preprocessingu z transkrypcjami manualnymi wykazała, że

duży wpływ na jakość transkrypcji automatycznych mają dźwięki nieartykułowane pojawiające się bardzo często w rozmowach CC. Z punktu widzenia semantycznego dźwięki te nie mają znaczenia dla rozumienia prowadzonych w CC konwersacji. Wpływają jednak negatywnie na jakość transkrypcji automatycznych, gdyż ich zapis zwykle był niepoprawny. Dlatego też na etapie postprocessingu postanowiono je wyeliminować. W tabeli 5.2 zestawiono wykaz tego typu dźwięków, które najczęściej występowały w bazie RLPHDB rozmów na linii klient – agent.

Tabela 5.2. Najczęstsze nieartykułowane dźwięki człowieka w bazie RLPHDB

Lp.	Dźwięki nieartykułowane w polskiej wymowie	Liczba wystąpień
1	“hm”, “hmm”, “hmmm”, “mhm”	762
2	“aha”, “haha”	458
3	“ee”, “eee”, “eeee”	252
4	“yy”, “yyy”, “yyyy”	110
5	“em”, “emm”, “emmm”	49
6	“mm”, “mmm”	19
7	inne nieartykułowane dźwięki	93

W ramach opisywanego w niniejszym punkcie modułu znormalizowano również sposób zapisu takich elementów jak: liczby, kwoty, daty oraz godziny. Ponadto, zniwelowano wpływ na jakość transkrypcji zapisu małych oraz wielkich liter i znaków interpunkcyjnych. Algorytm działania modułu korekty tekstu przedstawiono na rysunku 5.4.

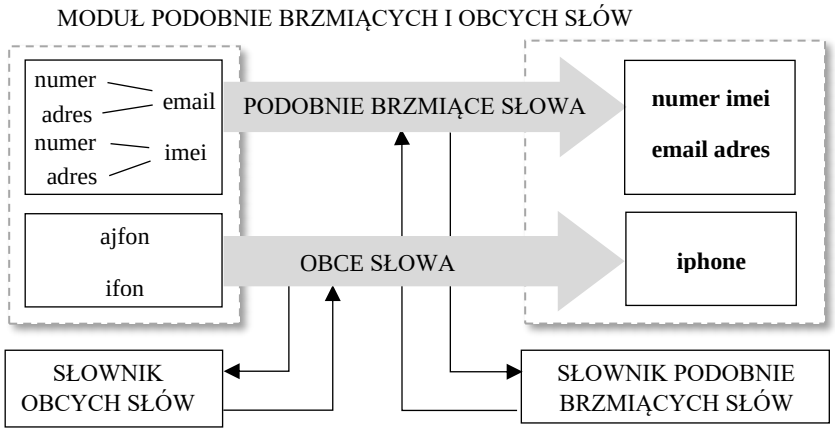


Rysunek 5.4. Schemat modułu korekty tekstu

Wyniki eksperymentów tego modułu postprocessingu przedstawione są w punkcie 5.3.2 oznaczano je na wykresach jako COR.

5.2.2.2 Moduł podobnie brzmiących i obcych słów

Z punktu widzenia jakości badanych transkrypcji poważny problem stanowią również podobnie brzmiące słowa. Jako przykład dla języka polskiego wymienić można zwrot „numer IMEI", który zwykle w czasie transkrypcji zamieniany był na zwrot „numer email". Oczywiste jest, że znaczeniowo w tym przypadku poprawny zwrot to „numer IMEI", gdyż przy poczcie elektronicznej mówimy o jej adresie, a nie o numerze. Wobec powyższego dla wybranych do badania tematów rozmów można przyjąć, że błędnie zapisany zwrot „numer email" należy zamienić na frazę „numer IMEI". Dla tematów rozmów określonych w podrozdziale 5.1 zbudowano więc bazę danych składającą się ze słów/wyrażeń o podobnym brzmieniu, które często były mylone przez badane systemy ASR. Zastosowanie utworzonego zbioru danych wpłynęło pozytywnie na jakość otrzymywanych transkrypcji. Algorytm działania modułu podobnie brzmiących i obcych słów przedstawiono na rysunku 5.5.



Rysunek 5.5. Moduł podobnie brzmiących i obcych słów

Innym elementem stwarzającym systemom ASR problemy są słowa pochodzące spoza korpusu języka polskiego (np. nazwy firm, produktów, marek, czy też słowa spolszczone, które zostały zapożyczone z innych języków – głównie angielskiego). Wiele z tych słów zapisywanych jest przez systemy ASR w sposób fonetyczny, np.: zamiast „iphone" system zwraca zapis fonetyczny w języku polskim jako „ajfon" lub „ifon".

Opisywane elementy można skorygować z wykorzystaniem badanych algorytmów postprocessingu. Dlatego stworzony został dla nich słownik w postaci bazy danych, zawierający słowa problematyczne z punktu widzenia rozpatrywanych w pracy tematów rozmów. Baza ta służy do automatycznej korekty w czasie rzeczywistym słów rozpoznawanych błędnie. Może ona być stale rozbudowywana w zależności od tematyki prowadzonych rozmów. Wyniki eksperymentów skuteczności dla tego modułu przedstawione są w punkcie 5.3.2 oznaczono je symbolem NCS.

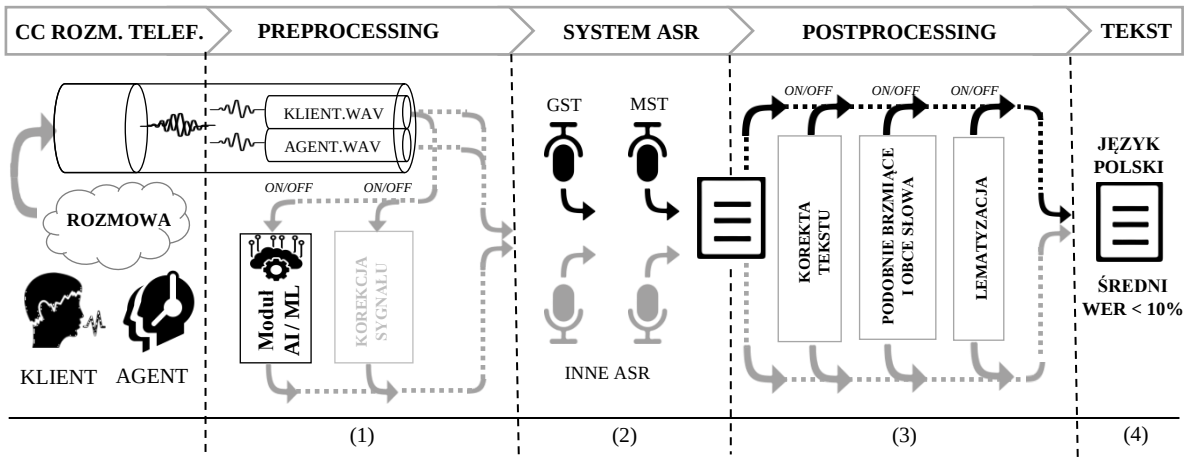
5.2.2.3 Moduł lematyzacji

W związku ze skomplikowaną naturą języka polskiego, związaną z licznymi złożonymi odmianami poszczególnych wyrazów, postanowiono opracować, zaimplementować oraz zbadać wpływ algorytmu lematyzacji na jakość badanych transkrypcji. Jak wiadomo, algorytm ten odpowiada za sprowadzenie danego słowa do jego formy podstawowej [92], [93], która przykładowo może znaleźć zastosowanie w algorytmach kategoryzacji danych tekstowych lub algorytmach służących do grupowania wypowiedzi tożsamych wykorzystywanych w budowie

wirtualnych konsultantów. Do celów lematyzacji wykorzystano ogólnodostępną bibliotekę *morfologic-stemming* [94]. Wyniki eksperymentów modułu lematyzacji oznaczone zostały symbolem LEM.

5.2.3 Metoda poprawy jakości transkrypcji

Na rysunku 5.6 zaprezentowano proponowaną metodę poprawy jakości transkrypcji automatycznej (w dalszej części pracy określaną jako metoda *IAT* ang. improvement automatic transcription) dedykowaną na potrzeby zastosowań w rozwiązaniach systemów CC [95] obsługujących język polski. W początkowej fazie, sygnał rejestrowany podczas rozmowy prowadzonej pomiędzy klientem oraz agentem poddawany jest zadaniom preprocessingu (1). Zadania te wykonywane są przed przesłaniem danych do dalszej analizy przez zintegrowane inteligentne systemy ASR (2). Wybrane systemy ASR tworzą transkrypcję prowadzonej rozmowy, która dalej trafia do bloku wykonującego algorytmy postprocessingu (3). Blok ten odpowiada za optymalizację wykonanej przez systemy ASR transkrypcji automatycznej. Wynikiem jego działania jest docelowa transkrypcja rozmowy (4).



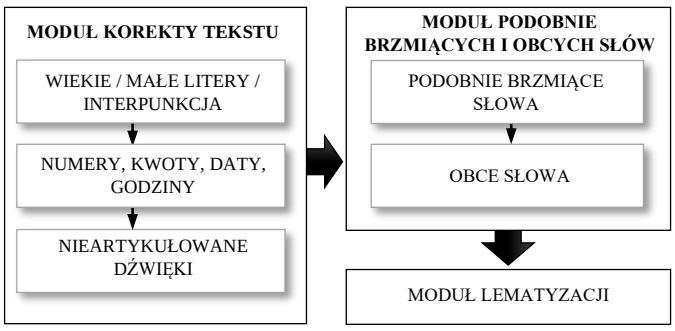
Rysunek 5.6. Metoda poprawy jakości transkrypcji dedykowana na potrzeby systemów CC w języku polskim

Pierwszą czynnością jest bezpośrednia rejestracja nagrań w formacie stereo. Nagrania te zawierają dwa osobno wydzielone kanały dla każdego z rozmówców (klienta oraz agenta). Strumień dźwiękowy z poszczególnych kanałów zapisywany jest w formacie WAV (ang. waveform audio format). Ten element w proponowanej metodzie zawsze powinien być wykonywany jako pierwszy. W dalszej kolejności w zależności od potrzeb sygnał może podlegać separacji na dwa niezależne kanały dla poszczególnych rozmówców oraz trafiać do kolejnych modułów realizujących zadania preprocessingu. Opisywane moduły preprocessingu mogą być stosowane w zależności od potrzeb jako potencjalne dodatkowe elementy poprawiające jakość transkrypcji wykorzystywanego rozwiązania ASR. W przypadku wyłączenia wszystkich opisywanych modułów preprocessingu, sygnał stereo może być przekazywany bezpośrednio do wybranego systemu ASR.

W fazie drugiej, po wykonaniu wybranych zadań preprocessingu, otrzymany sygnał dźwiękowy kierowany jest do jednego z wielu systemów ASR. Jak wyjaśniono wcześniej, na potrzeby realizacji postawionych w pracy celów wyselekcjonowane zostały dwa rozwiązania: MST oraz GST. Nie ogranicza to jednak możliwości wykorzystywania proponowanej metody dla innych systemów ASR, jak również w innych zastosowaniach niż transkrypcja na potrzeby Contact Center. Wynikiem pracy każdego z systemów ASR jest transkrypcja nagrania

dźwiękowego zapisana w formie tekstowej, która dalej, w fazie trzeciej podlega przetwarzaniu przez moduły postprocessingu.

W opracowanej metodzie *IAT* przygotowano i zaimplementowano zestawy algorytmów mających na celu zwiększenie jakości otrzymanych transkrypcji. Parametry tych algorytmów dobrano na podstawie analiz porównawczych wykonanych dla transkrypcji referencyjnych i otrzymanych transkrypcji automatycznych. Badania porównawcze obu transkrypcji pozwoliły na identyfikację najsłabszych cech badanych systemów ASR z punktu widzenia poprawności transkrypcji rozmów prowadzonych w CC w języku polskim. Pozwoliło to na określenie algorytmów postprocessingu, które w zdecydowanym stopniu wpływają na poprawę jakości transkrypcji. Ustalona została również najlepsza kolejności w jakiej algorytmy postprocessingu powinny być wykonywane. Na rysunku 5.7 zaprezentowano diagram opisujący kolejność pracy poszczególnych modułów. Jako pierwsze wywoływane są algorytmy zaimplementowane w module korekty tekstu, gdzie na początku niwelowany jest wpływ znaczenia wielkich i małych liter oraz znaków interpunkcyjnych na jakość transkrypcji wykonywanych przez zintegrowane systemy ASR. W kolejnym kroku pracy algorytmów tego modułu normalizowane są zapisy liczb, kwot, dat, godzin. Po wykonaniu tych operacji uruchamiane są algorytmy usuwające znaki powstałe w wyniku zapisu dźwięków nieartykułowanych nieistotnych z punktu widzenia znaczenia zdania, a często występujące w konwersacjach. Drugi wykonywany moduł zawiera mechanizmy optymalizujące poprawność transkrypcji dla podobnie brzmiących słów z punktu widzenia ich wymowy, które jednak mają zupełnie inne znaczenie semantyczne. Ta czynność wykonywana jest w opisywanym module jako pierwsza, po czym system normalizuje zapis wyrazów obcojęzycznych. Ostatnim z wykonywanych modułów postprocessingu jest lematyzacja. Algorytm ten wzbogaca blok postprocessingu, a jego zastosowanie powinno być przydatne z punktu widzenia skomplikowanej struktury języka polskiego. Wykorzystanie lematyzacji nie zawsze musi być zasadne i może zależeć od potrzeb dla jakich wykonywane są transkrypcje.



Rysunek 5.7. Diagram kolejność wywoływania modułów postprocessingu

5.3 Wyniki eksperymentów opracowanej metody poprawy jakości transkrypcji

Na potrzeby przeprowadzenia końcowej weryfikacji opracowanej metody poprawy jakości transkrypcji *IAT* wykorzystano bazę RLPHDB-VLD. Wszystkie nagrania zostały poddane transkrypcji manualnej wykonanej przez człowieka oraz transkrypcji automatycznej wykonywanej przez system.

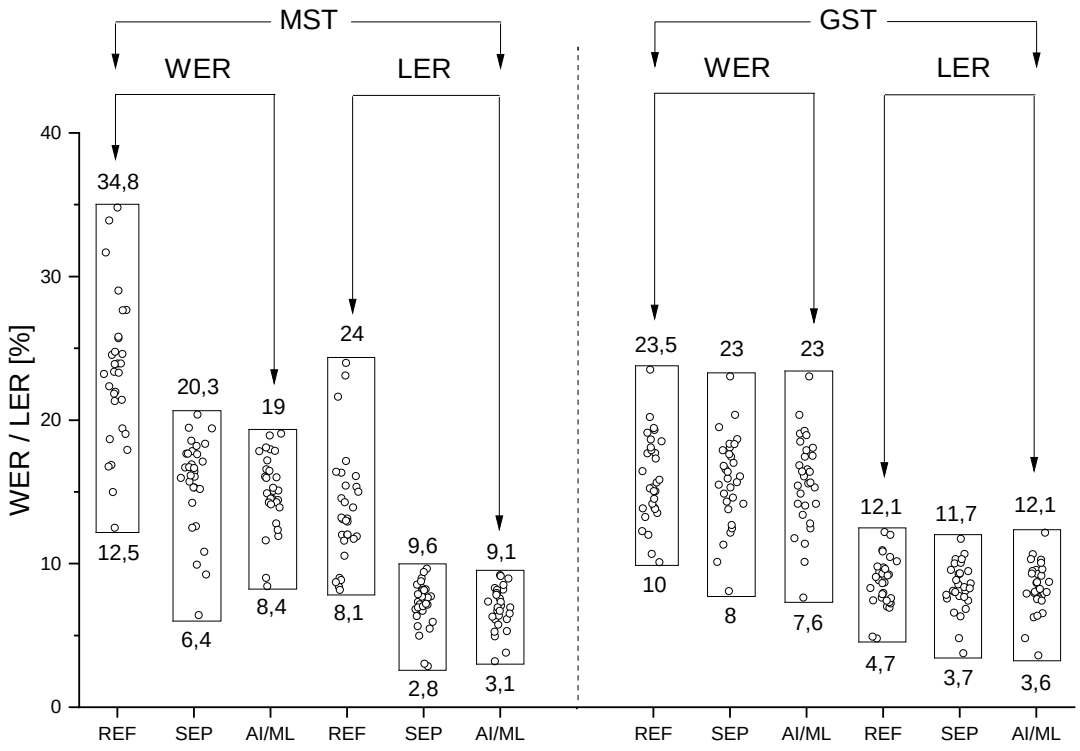
W kolejnych podrozdziałach opisano wyniki eksperymentów otrzymane dla wymienionych wcześniej modułów preprocessingu oraz postprocessingu.

5.3.1 Wyniki eksperymentów dla modułów preprocessingu

Rysunek 5.8 oraz rysunek 5.9 przedstawiają wpływ zastosowanych modułów preprocessingu na jakość transkrypcji wyrażoną metrykami WER oraz LER. Na potrzeby oceny skuteczności badanych modułów, z bazy RLPHDB-VLD utworzono dwa zbiory nagrań zawierające po trzydzieści nagrań rozmów przeprowadzonych podczas rzeczywistych kampanii CC. Eksperymenty wpływu wydzielenia kanałów rozmówców oraz możliwości douczania systemów ASR na jakość transkrypcji wykonano przy pomocy pierwszego zbioru nagrań. Natomiast, na potrzeby badań modułu korekcji jakości sygnału wykorzystano drugi zbiór. Zbiór ten zawierał nagrania, w których jakość dźwięku słyszanego przez człowieka była najgorsza. Wpływ na to miały występujące w nagraniach wyraźne szumy, trzaski, odgłosy samochodu lub zwierząt, różne poziomy głośności kanałów klienta i agenta, jak również wiele innych niepożądanych dźwięków zakłócających przebieg prowadzonych konwersacji. Podczas badań tego modułu preprocessingu wszystkie wymienione zakłócenia zostały wyeliminowane.

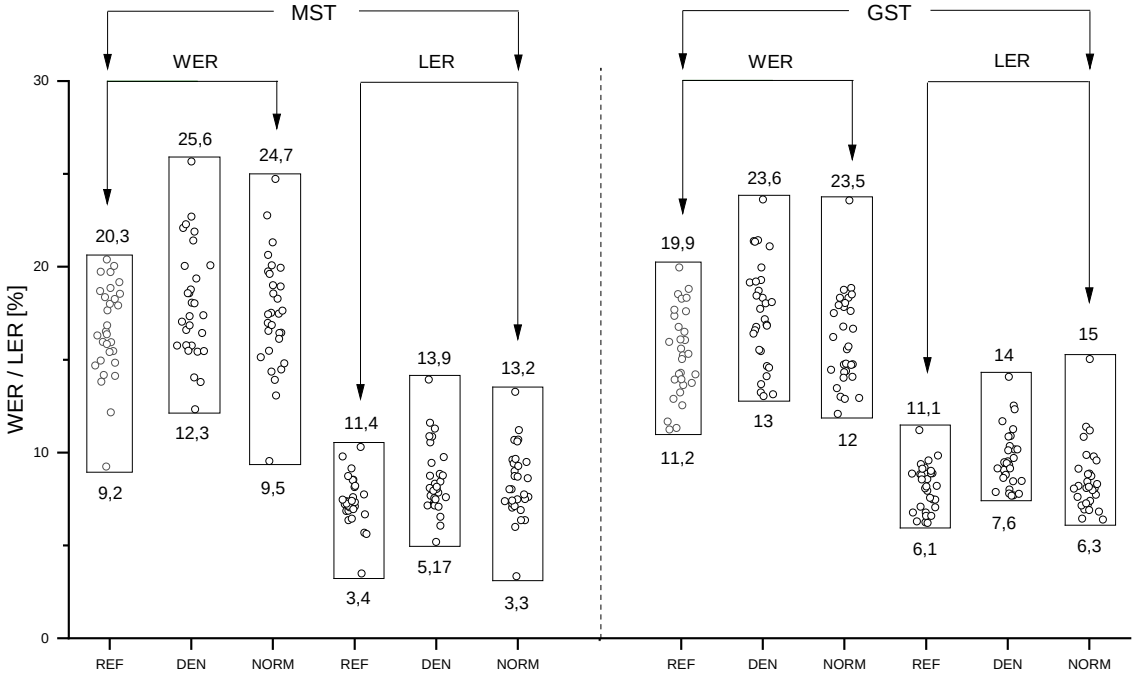
Przygotowany zbiór nagrań poddano transkrypcji automatycznej w systemach MST oraz GST. Z eksperymentów przeprowadzonych dla rozwiązania MST wynika, że transkrypcja sygnału podstawowego jest niskiej jakości, na co wskazuje uzyskana wartość metryki WER (12,5% - 34,8%). Metryka ta ulega zauważalnej poprawie przy podaniu do transkryptora odseparowanych kanałów agenta oraz klienta (6,4% - 20,3%). Analizując otrzymane wyniki, widać, że MST nie dysponuje funkcją separacji kanałów, dlatego też wykorzystując ten system na potrzeby CC należy zaadoptować opisywany moduł preprocessingu jako rozwiązanie bazowe. Analogiczne eksperymenty wykonano przy implementacji drugiego z wybranych do testowania w pracy rozwiązań – systemu GST. Z przeprowadzonych eksperymentów tego modułu preprocessingu wynika, że wartości metryki WER (10% - 23,5%) dla transkrypcji automatycznych wykonanych z sygnału podstawowego była już stosunkowej dobrej jakości. Metryka ta uległa nieznacznej poprawie, gdy kanały agenta oraz klienta były osobno podane do transkryptora (8% - 23%). Poprawa ta jest nieznaczna ze względu na funkcjonujący, dobrze zoptymalizowany mechanizm separacji ścieżek rozmówców w GST. Dlatego też, wykonując transkrypcje z wykorzystaniem tego systemu można podawać tylko jeden strumień danych stereo. Odpowiednie wartości średnich arytmetycznych dla prezentowanych eksperymentów zestawiono w tabeli 5.3.

Kolejnym krokiem badań modułów preprocessingu była ocena skuteczności procesów douczania (opisanych w punkcie 5.2.1.2) wykonywana na odseparowanych kanałach agenta i klienta zarówno dla MST i GST. Moduł separacji kanałów we wszystkich kolejnych eksperymentach został włączony jako bazowy.



Rysunek 5.8. Wpływ rozdzielenia kanałów klienta i agenta oraz procesów douczania na jakość transkrypcji systemów MST oraz GST

Legenda: REF - Wszystkie moduły przetwarzania wstępnego są wyłączone, SEP - moduł separacji kanałów jest włączony, AI/ML - moduł uczenia się jest włączony



Rysunek 5.9. Wpływ redukcji szumów oraz normalizacji poziomów głośności kanałów klienta oraz agenta na jakość transkrypcji systemów MST oraz GST

Legenda: REF - Wszystkie moduły przetwarzania wstępnego są wyłączone, DEN - Moduł redukcji szumu jest włączony, NORM - Moduł normalizacji jest włączony

Analiza wpływu mechanizmów douczania na jakość transkrypcji wykonywanych z wykorzystaniem systemu MST wskazała na dalsze możliwości jej optymalizacji. Otrzymane wyniki dla metryki WER kształtują się na poziomie od 8,4% do 19%, co jest poprawą tylko o 1% w stosunku wartości bazowych (z separacją kanałów). Douczanie rozwiązania GST nie przyniosło natomiast rezultatów na oczekiwanym poziomie, a średnie dla obu badanych metryk kształtowały się bardzo podobnie do wyników referencyjnych. Biorąc powyższe pod uwagę wnioskować można, że w przypadku obu systemów modele języka polskiego dostarczane przez dostawców systemów ASR są dobrze przygotowane. Chociaż jak wykazano, zaadoptowany w rozwiązaniu MST proces uczenia może jeszcze poprawić jakość transkrypcji automatycznych.

Ostatnim z badanych modułów preprocessingu był moduł korekcji sygnału dźwiękowego. Wyniki eksperymentów obrazujące wpływ tego modułu na poziom metryk WER oraz LER dla systemów GST oraz MST pokazano na rysunku 5.9. Korekcja w zakresie redukcji szumu oraz normalizacji sygnału była wykonywana niezależnie, należy więc ją zawsze porównywać z sygnałem podstawowym. Pomimo, że w analizowanym zbiorze nagrań poprawa jakości dźwięku jest istotnie zauważalna dla człowieka to jakość transkrypcji nie poprawia się, a dla niektórych nagrań ulega pogorszeniu. Wynika to z tego, że podczas redukcji szumów i normalizacji sygnału są usuwane również dźwięki istotne z punktu widzenia rozpoznania słów przez systemy ASR. Zakładać można, że zaimplementowane systemy ASR posiadają już wbudowane oraz dobrze rozwinięte mechanizmy korekcji sygnałów dźwiękowych, które są zoptymalizowane pod kątem wykonywanych transkrypcji. Dlatego ingerencja w sygnał wejściowy wpływa niekorzystnie na ich działanie. Z punktu widzenia rozpatrywanych systemów najważniejsze jest to, aby zadbać o rejestrację stereofonicznego nagrania źródłowego w możliwie najlepszej jakości.

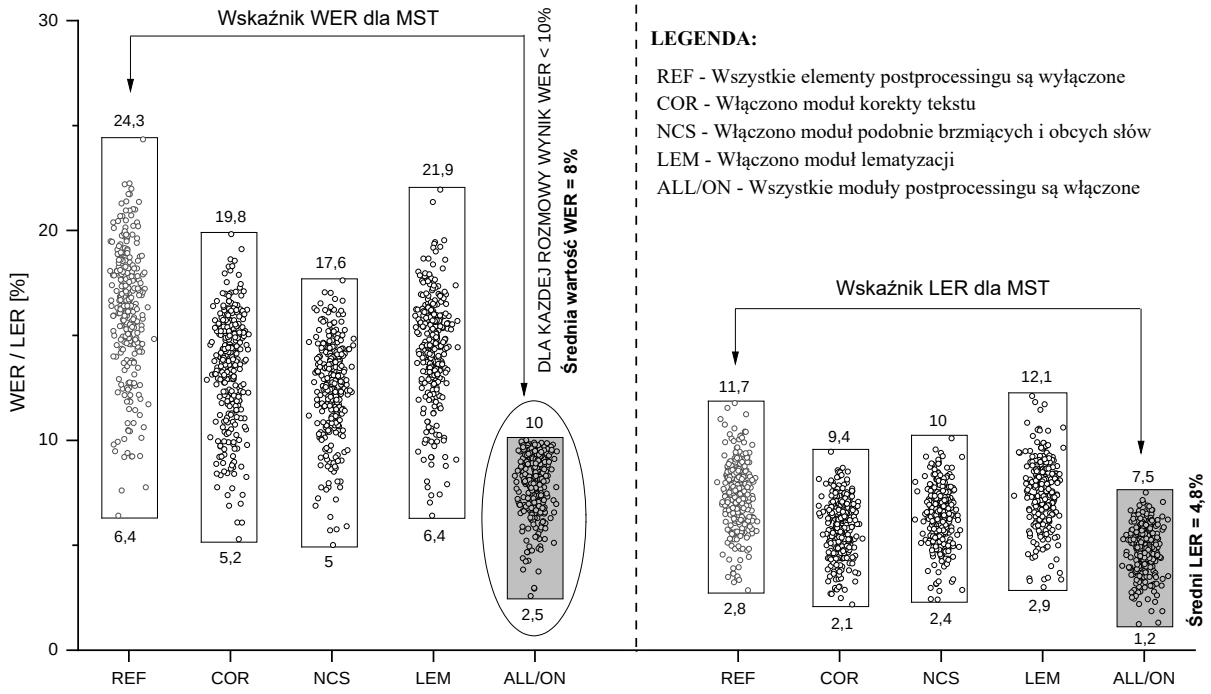
W tabeli 5.3 zestawiono uśrednione wyniki eksperymentów metryk WER i LER dla omawianych modułów preprocessingu. Analizując prezentowane dane widać, że z perspektywy możliwości zastosowań badanych rozwiązań na potrzeby CC obsługującego klientów w języku polskim nieco lepsze wyniki osiągnięto wykorzystując system MST.

Tabela 5.3. Średnie wartości WER i LER dla modułów preprocessingu

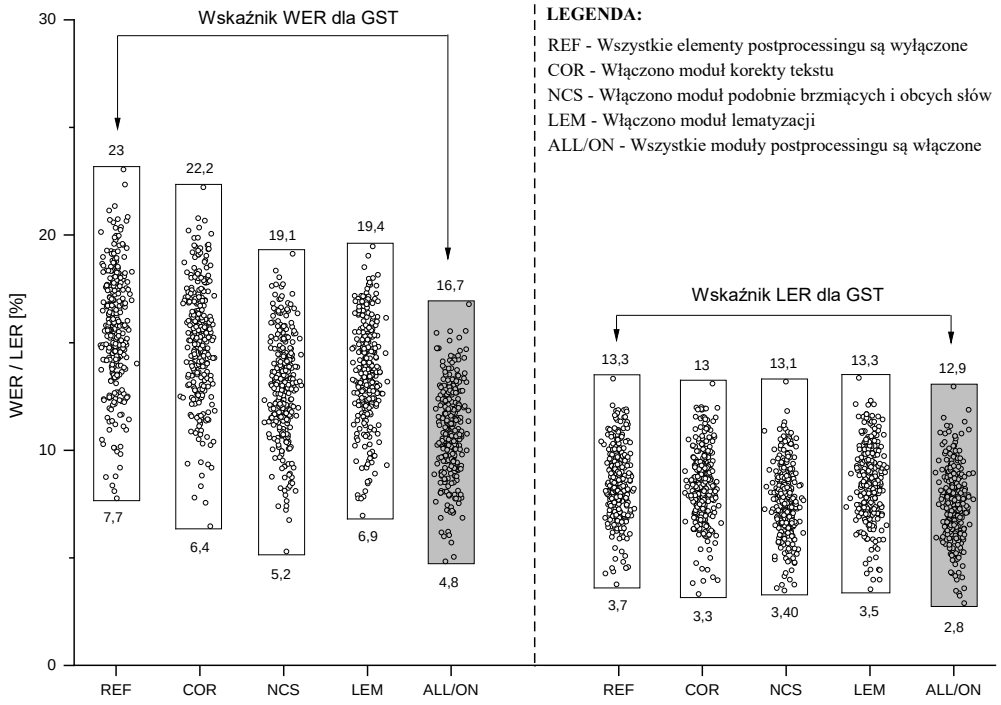
Wskaźnik	**ASR**	**Zbiór 1**			**Zbiór 2**		
		REF	**SEP**	**AI/ML**	**REF**	**DEN**	**NORM**
		[%]	**[%]**	**[%]**	**[%]**	**[%]**	**[%]**
WER	MST	23,2	15,7	14,7	15,8	15,4	17,4
	GST	15,9	15,7	15,7	15,9	16,8	15,7
LER	MST	13,8	7,1	6,8	7,4	7,0	7,4
	GST	8,3	8,3	8,3	9,0	9,1	8,0

5.3.2 Wyniki eksperymentów dla modułów postprocessingu

Eksperymenty poszczególnych modułów postprocessingu wykonano dla utworzonego zbioru 300 losowo wybranych próbek nagrań. Uwzględniono trzy tematy rozmów: faktury i płatności; informacje techniczne; umowy i aneksy. Dla każdego z tematów wybrano po 100 nagrań. Zarówno dla MST jak i GST zastosowano moduł preprocessingu rozdzielenia kanałów agenta oraz klienta na dwa niezależne od siebie strumienie danych.



Rysunek 5.10. Wpływ postprocessingu na jakość transkrypcji systemu MST



Rysunek 5.11. Wpływ postprocessingu na jakość transkrypcji systemu GST

Rysunek 5.10 oraz rysunek 5.11 przedstawiają wyniki eksperymentów wpływu zastosowanych modułów postprocessingu na jakość transkrypcji badanych systemów ASR wyrażoną metrykami WER oraz LER. Na rysunku 5.10 przedstawiono wyniki otrzymane podczas testów rozwiązania MST, natomiast na rysunku 5.11 zestawiono dane z testów rozwiązania GST.

Najpierw zbadano moduł korekty tekstu. Okazał się on bardziej skuteczny dla MST niż dla GST, gdyż błędy występujące w MST były częściej rozpoznawane i poprawiane przez ten moduł niż dla GST. Implementacja modułu podobnie brzmiących i obcych słów pozwoliła na znaczącą poprawę jakości transkrypcji względem wartości referencyjnych zarówno dla MST jak i GST. Okazał się on też najskuteczniejszy spośród wszystkich zastosowanych modułów postprocessingu. Wynika to z tego, że systemy ASR często popełniają błędy podczas rozpoznawania podobnie brzmiących słów. Zastosowanie dla obu transkrypcji (referencyjnej oraz automatycznej) modułu lematyzacji w przypadku rozwiązania opartego o system GST okazało się skuteczniejsze w porównaniu do systemu MST. Wynika to z tego, że GST słabiej sobie radzi z rozpoznawaniem końcówek wyrazów niż MST. Z badań wynika, że algorytm lematyzacji może być wykorzystywany w przypadku języka polskiego, natomiast dla innych języków sens jego implementacji wymaga dodatkowych badań korpusu danego języka.

W tabeli 5.4 zaprezentowano średnie wartości WER i LER. Przy wykorzystaniu rozwiązania GST uzyskano wynik na poziomie 10,8%, natomiast przy wykorzystaniu rozwiązania MST uzyskano wynik na poziomie 8%. Są to wyniki bardzo dobre, pozwalające na wykorzystywanie wykonywanych transkrypcji na potrzeby branży CC. Zgodnie z artykułem [96] za dobrą jakość i gotowość do użycia transkrypcji uznaje się, gdy jej WER mieści się w zakresie 5%-10%. Przyjmuje się, że akceptowalny poziom WER to 20%, a poziom 30% i więcej wskazuje na zbyt niską jakość transkrypcji.

Tabela 5.4. Średnie wartości WER i LER dla modułów postprocessingu

Wskaźnik	ASR	REF	COR	NCS	LEM	ALL/ON
		[%]	[%]	[%]	[%]	[%]
WER	MST	16,2	13,4	12,5	14,3	**8**
	GST	15,6	15,1	12,9	13,6	**10,8**
LER	MST	7,3	5,9	6,4	7,5	4,8
	GST	8,5	8,3	7,7	8,4	7,5

Opracowana metoda *IAT* zapewnia satysfakcjonujący poziom transkrypcji automatycznej dla rozmów prowadzonych w kanałach głosowych CC pomiędzy klientami i agentami. Warto podkreślić, że poziom ten jest odpowiedni do jej dalszego wykorzystywania w rozwiązaniach inteligentnych metod rozpoznawania intencji klienta. Wysoka jakość transkrypcji pozwala na skuteczniejsze funkcjonowanie voicebotów w środowiskach Contact Center, gdzie jakość rozmów głosowych jest często na niskim poziomie z uwagi na zakłócenia dźwięku m.in.: szumy z tła, strumień audio przesyłany technologią VoIP, czy słaba dykcja klientów. Z uwagi na to, że automatyzacja rozmów stanowi duże wyzwanie w branży Contact Center, to bardzo istotne jest, aby jakość wdrażanych rozwiązań była jak najwyższa i nie spowodowała zniechęcenia klientów.

5.4 Wyniki eksperymentów wpływu metody poprawy jakości transkrypcji na rozpoznawanie intencji

W systemach CC wyróżniamy dwa zasadnicze kanały komunikacji, jakimi są kanał głosowy oraz kanał tekstowy. Ponieważ metody rozpoznawania intencji klienta na większości platform NLU funkcjonują przy wykorzystaniu jedynie danych tekstowych, dlatego też rozmowy

głosowe muszą podlegać wcześniejszemu procesowi transkrypcji. W zależności od zastosowanego systemu ASR, jakość tej transkrypcji jest zwykle bardzo różna. Należało więc zweryfikować, jak wpływa jakość transkrypcji rozmów prowadzonych w CC na proces rozpoznawania intencji. W tym celu wykonano eksperymenty porównawcze, w których wykorzystano transkrypcje automatyczne wygenerowane przez MST oraz transkrypcje wykonane z wykorzystaniem nowej metody *IAT* opisanej w podrozdziale 5.2.

W oparciu o zgromadzone w bazie RLPHDB rozmowy wyselekcjonowano zbiór S_{INT} piętnastu najczęściej występujących intencji, które podzielono na trzy grupy (tabela 5.5). Z transkrypcji manualnych z bazy RLPHDB-RSR wybrano 375 fraz uczących (po 25 fraz dla każdej intencji). Z transkrypcji automatycznych z bazy RLPHDB-VLD wybrano 225 fraz testowych (po 15 fraz dla każdej intencji).

Tabela 5.5. Grupy wyselekcjonowanych intencji dla kanału głosowego

Grupa	**Wyselekcjonowane intencje INT_j dla kanału głosowego**	
	ID	**Nazwa**
1	V1	Potwierdzenie
	V2	Zaprzeczenie
	V3	Autoryzacja
	V4	Wypowiedzenie umowy
	V5	Zadowolenie z obsługi
2	V1	Stan usługi
	V2	Informacje o umowie
	V3	Aktywacja usług internetowych
	V4	Naprawa urządzenia
	V5	Przedłużenie umowy
3	V1	Informacje o fakturze
	V2	Problem z połączeniem internetowym
	V3	Połącz z agentem
	V4	Opis uszkodzeń
	V5	Odzyskiwanie hasła

Legenda: V1-V5 – oznaczenie intencji w danej grupie dla kanału głosowego.

Z uwagi na to, że baza RLPHDB pochodzi z rozmów o tematyce telekomunikacyjnej, zbudowano dla niej wyspecjalizowane słowniki, które są wykorzystywane przy metodzie *IAT*. Takie słowniki należy dostosowywać do określonej kampanii Contact Center. Na przykład w transkrypcji automatycznej wytworzonej przez MST pojawia się często tekst „psl" jako transkrypcja wypowiedzianego słowa „pesel" co stanowi błąd. Świadomość, że kampania dotyczy obsługi klienta telekomunikacyjnego, pozwala przyporządkować w słowniku słowo „psl" do słowa bazowego „pesel" (zamiana „psl" na „pesel"). Zupełnie inaczej jest w przypadku, gdyby była to kampania Contact Center badającą sympatie polityczne, wówczas nie można byłoby zamieniać słowa „psl" z uwagi na to, że jest to skrót jednej z partii politycznych, czyli pojawienie się go w rozmowie jest prawdopodobne.

Budując słowniki należy brać pod uwagę związki frazeologiczne (np. „numer imei", „adres email"). Jeżeli transkryptor automatyczny zwraca frazę „numer email" to należy zamienić ją na „numer imei". Wynika to z tego, że gdyby chodziło o słowo „email", to klient powiedziałby „adres email", a nie „numer email".

Prowadząc badania porównawcze transkrypcji manualnej z automatyczną wyodrębniono dużo przypadków takich zamian słów. Przy tworzeniu nowych kampanii, część zamian słów z poprzednich kampanii można wykorzystać. Natomiast musi być ona zweryfikowana pod

kątem danej dziedziny, specyfiki prowadzonych rozmów i rozbudowana o nowo występujące zamiany słów.

W całej bazie eksperymentalnej takich słów jak „pesel" czy „numer imei" było znacznie więcej, część z nich zaprezentowano w tabeli 5.6.

Tabela 5.6. Przykładowe słowniki podobnie brzmiących i obcych słów

Lp.	Słowo bazowe	Słowa zamieniane na słowo bazowe
1	mail	meil, mailem, meilowy, mailowym, meilowym, mejlowego, mejlowym, mailowego, mejl, emaila, meila, mailowy, mailo, e-mail, email, mailowe, mailową, mailowo, maila, mejlowy, mejla
2	sms	sms owo, smsie, sms em, sms owe, smsa, smsy, esemesa, smsm, smsowym, smsami, sms ami, sso, sama son, smsem, smska, smsowe, sms ow, smsowo
3	youtube	yuo tubę, yuo tuba, jutub, jutubie, outubie, jutuba, jutube, jutuby, yuo tube, ju tubę
4	gigabyte	globali, gigabajt, gigabajty, gigacon, giga, gigabajtowa, gigabajta, gq, gigabajtu, giganoto, gb, giga bajtów, gigabajtów
5	iphone	ajfona, iphonów, ajfonów, iphona, ajphone, ajfon
6	pesel	ep sl, zalesie, posel, peselu, psl, pesa, absl, peszek
7	doładować	doładowywać, doładowuje, do ładować, doładowałem, doładowywałem, ładować, doładowuję, doładowałam
8	uzupełnić	uzupełnie, zupełnie, ul wypełnienia, uzupełniać, uzupełnienie, uzupełnię
9	żeby	że by, ażeby, że bym, żebym, żeby śmy, żebyśmy
10	samsung 10	samsung s10, samsung dziesiątkę
11	amazon prime video	ama ten kraj wideo, amazon kraj video, ama son pry video
12	prime video	kraj vidio, pry video, pra ymy video
13	google play	google plej, gugle plej
14	youtube music	jutube muzik, yuo tuba muzik
15	numer email	numer imei
16	adres imei	adres email

Warto podkreślić, że testowanie na platformie NLU odbywało się w ten sposób, że frazy uczące i frazy testowe były zawsze przygotowywane w sposób jednakowy – przez system ASR lub metodę *IAT*. Na przykład, gdy zamierzano przeprowadzić weryfikację fraz testowych dla metody *IAT* to tę samą metodę stosowano dla fraz uczących. Wynika to z tego, że użycie różnych metod transkrypcji (systemu ASR i metody *IAT*) dla fraz uczących i testowych generowałoby dodatkowe błędy w rozpoznawaniu intencji, wynikające z różnicy pomiędzy tymi transkrypcjami, co wpłynęłoby niekorzystnie na wiarygodność wyników eksperymentów.

Natomiast w metodzie *IAT* poza mechanizmem separacji kanałów zastosowano opracowane moduły postprocessingu: moduł korekty tekstu oraz moduł podobnie brzmiących i obcych słów. Przy czym z modułu korekty tekstu został wyłączony komponent dźwięków nieartykułowanych, gdyż ma on negatywny wpływ na rozumienie intencji z wypowiedzi klientów nacechowanych emocjami. Niepożądane byłoby usuwanie fraz mających wydźwięk emocjonalny np. „aha", „ojej", „yyy", „eee", „mhm". Odrzucono również zastosowanie modułu lematyzacji ze względu na to, że w procesie lematyzacji jest gubiony czas gramatyczny czasownika (np. „będę", „byłem", „jestem" podczas lematyzacji są zamieniane na „być"). Brak czasu gramatycznego w niektórych przypadkach zaburza sens zrozumienia danej wypowiedzi.

Tabela 5.7. Frazy uczące dla intencji dla kanału głosowego

Lp.	Gr.	Intencja INT_j	Frazy uczące P_i	
			Transkrypcja manualna	**Metoda *IAT***
1	1	V1 Potwierdzenie	*Tak, tak. Oczywiście, oczywiście*	*tak tak oczywiście oczywiście*
2			*Dobra, akceptuję jego warunki*	*dobra akceptuję jego warunki*
3		V2 Zaprzeczenie	*Nie, nie zgadzam się*	*nie nie zgadzam się*
4			*Nie, nie dziękuję*	*nie nie dziękuję*
5		V3 Autoryzacja	*mogę pesel bo ja nie pamiętam hasła*	*mogę pesel bo ja nie pamiętam hasła*
6			*będę miał po prostu ten nr IMEI*	*będę miał po prostu ten nr imei*
7		V4 Wypowiedzenie umowy	*Mam takie pytanie, chciałabym zrezygnować. Chciałabym wypowiedzieć umowę.*	*mieć takie pytanie chciałabym rezygnować chciałabym wypowiedzieć umowa*
8			*aha no to w takim razie to zrezygnuję z tej usługi*	*aha no to w takim razie to zrezygnuję z tej usługi*
9		V5 Zadowolenie z obsługi	*Dobra, no to świetnie. Bardzo dziękuję w takim razie tyle jutro się zarejestruję na tym, żeby było już spokojnie mi to aktywowało. Nie bardzo panu dziękuję za Wszystko, no.*	*dobra no to świetnie bardzo dziękuję w takim razie tyle jutro się zarejestruję na tym żeby było już spokojnie mnie to aktywowało nie bardzo panu dziękuję za wszystko no*
10			*bardzo super Pani mnie obsłużyła*	*bardzo super pani mnie obsłużyła*
11	2	V1 Stan usługi	*Może tam jakaś usługa jest włączona*	*może tam jakaś usługa jest włączona*
12			*Ile jest tego pakietu tego internetu?*	*ile jest tego pakietu tego internet*
13		V2 Informacje o umowie	*to nie zobaczę do kiedy mam umowę?*	*to nie zobaczę do kiedy mieć umowa*
14			*Ja mam takie pytanko bo ja bym chciała się dowiedzieć, ile nie zostały jeszcze do końca umowy tego numeru. Co mam?*	*ja mieć takie pytanko bo ja bym chciała się dowiedzieć ile nie zostały jeszcze do końca umowa tego nr co mieć*
15		V3 Aktywacja usług internetowych	*mi chodzi o to, że chcę dokupić pakiet internetowy.*	*mnie chodzi o to że chcę dokupić pakiet internetowy*
16			*Ja chciałem kupić pakiet internetu za 10 zł 5 giga*	*ja chciałem kupić pakiet internet za 10 złotych 5 gigabyte*
17		V4 Naprawa urządzenia	*coś tam z tyłu klapka taka pękła*	*coś tam z tyłu klapka taka pękła*
18			*rozbiła mi się szybka w telefonie i chciałam zgłosić*	*rozbiła mnie się szybka w telefonie i chciałam zgłosić*
19		V5 Przedłużenie umowy	*a jeżeli ona się kończy w kwietniu maju to ja mam możliwość przedłużenia*	*a jeżeli ona się kończyć w kwietniu maju to ja mieć możliwość przedłużenia*
20			*tak to chce zawrzeć nową, bo nie została. Hmm nie została zatwierdzona taka umowa*	*tak to chcę zawrzeć nową bo nie została hmm nie została zatwierdzona taka umowa*
21	3	V1 Informacje o fakturze	*jeszcze by mi pani sprawdziła dlaczego ta faktura jest tak wysoka, ostatnia 260 zł?*	*jeszcze by mnie pani sprawdziła dlaczego ta faktura jest tak wysoka ostatnia 260 złotych*
22			*czyli obydwie faktury muszę zapłacić do 9 grudnia?*	*czyli obydwie faktury muszę zapłacić do 9 grudnia*
23		V2 Problem z połączeniem internetowym	*No ale nie ma internetu, by wychodzi cały czas brak dostępu do sieci.*	*no ale nie ma internet by wychodzi cały czas brak dostępu do sieci*
24			*jest aktywny tak cały czas , no bo on mówi że nie ma Internetu, więc myślę*	*jest aktywny tak cały czas no bo on mówi że nie ma internet więc myślę*
25		V3 Połącz z agentem	*Nie, ja już nie mam ochoty rozmawiać, po prostu jestem wkurwiony no.*	*nie ja już nie mieć ochoty rozmawiać po prostu jestem wkurwiony no*
26			*Jak pan nie może rozwiązać tej sprawy proszę mnie z kierownikiem połączyć*	*jak pan nie może rozwiązania tej sprawa poproszę mnie z kierownikiem połączyć*
27		V4 Opis uszkodzeń	*jest rozbity wyświetlacz i te gniazdo od słuchawek się też uszkodziło*	*jest rozbity wyświetlacz i te gniazdo od słuchawek się też uszkodziło*
28			*Pęknięty ekran i rozlany i z tyłu coś tam jest.*	*pęknięty ekran i rozlany i z tyłu coś tam jest*
29		V5 Odzyskiwanie hasła	*Nie pamiętam hasła do konta, musimy w jakiś inny sposób zweryfikować.*	*nie pamiętam hasła do konta musimy w jakichś inny sposób zweryfikować*
30			*Hasła nie znam.*	*hasła nie znam*

W tabeli 5.7 dla każdej intencji zaprezentowano po dwie frazy uczące. Z uwagi na to, że frazy uczące pochodzą z transkrypcji manualnej, to liczba korekt wykonanych przez metodę *IAT* nie była duża. Głównie sprowadziły się one do normalizacji tekstu do małych liter, eliminacji znaków interpunkcyjnych, sprowadzenia części słów do słów bazowych. W większości wierszy widać normalizację, natomiast w wierszach 12, 13, 14, 15, 16 widać też zamianę słów na słowa bazowe: „internetu" na „internet", „umowę" na „umowa", „mam" na „mieć", „mi" na „mnie", „zł" na „złotych", „giga" na „gigabyte".

W tabeli 5.8 dla każdej intencji zaprezentowano po dwie frazy testowe. Ze względu, że frazy testowe pochodzą z transkrypcji automatycznej to ilość poprawek tekstu w metodzie *IAT* w stosunku do systemu ASR jest znacznie większa niż w przypadku tabeli 5.7. Analogiczna sytuacja będzie miała miejsce w przypadku zastosowania metody *IAT* w połączeniach głosowych w czasie rzeczywistym. Poza normalizacją tekstu do małych liter oraz eliminacją znaków interpunkcyjnych, można zaobserwować wiele korekt słów i fraz. Ważniejsze korekty wyróżniono szarym tłem w tabeli 5.8. Sprowadzenie do słów bazowych „dobrze" na „dobra" (lp. 1, 10), „mi" na „mnie" (lp. 9, 10), "GB" na „gigabyte" (lp. 16) „wykupić" na „zakupić" (lp. 15), „mam" na „mieć" (lp. 13), „jakiś" na „jakichś" (lp. 15), „proszę" na „poproszę" (lp. 13). Poprawa błędów „Ten sam tag" na „ten sam tak" (lp. 2), „odwas" na „od was" (lp. 8). Zamiana wyrazów wynikających z charakterystyki kampanii „psl" na „pesel" (lp. 5). Poprawa związków frazeologicznych „numer email" na „numer imei" (lp. 6). „krutynia klapka" na „tylna klapka" (lp. 17), „z piłą szkiełko" na „zbiło szkiełko" (lp. 18), normalizacja liczb „jednego" na „1" (lp. 14). Warto podkreślić, że zamiana „mam" na „mieć" nie wynikała z procesu lematyzacji. Słowo to zostało uwzględnione w słownikach zamiany słów. Było to podyktowane tym, że występowanie słowa „mam" często zaburzało rozpoznawanie intencji, a sprowadzenie go do słowa „mieć" rozwiązywało ten problem.

Tabela 5.8. Frazy testowe dla intencji dla kanału głosowego

Lp.	Gr.	Intencja	Frazy testowe	
			System ASR	**Metoda *IAT***
1	1	V1 Potwierdzenie	*dobrze poproszę tak oczywiście tak*	*dobra poproszę tak oczywiście tak*
2			*Ten sam tag.*	*ten sam tak*
3		V2 Zaprzeczenie	*No nie, nie zapoznałem się*	*no nie nie zapoznałem się*
4			*Nie absolutnie nie tak oczywiście nie*	*nie absolutnie nie tak oczywiście nie*
5		V3 Autoryzacja	*Nie podam psl.*	*nie podam pesel*
6			*powiem numer email*	*powiem numer imei*
7		V4 Wypowiedzenie umowy	*Nie, myślę, że po prostu rezygnuję tak kompletnie.*	*nie myślę że po prostu rezygnuję tak kompletnie*
8			*Nie ja u was złożyłam wypowiedzenie i odwas chce zabrać numer przenieść do innego operatora.*	*nie ja u was złożyłam wypowiedzenie i od was chcę zabrać numer przenieść do innego operatora*
9		V5 Zadowolenie z obsługi	*dziękuję bardzo dużo mi Pan pomógł, dziękuję bardzo.*	*dziękuję bardzo dużo mnie pan pomógł dziękuję bardzo*
10			*Pomógł bardzo, naprawdę teraz mi Pani pomogła no bo ten Pan mi nic nie powiedział, tak? A ja też nie wiedziałam, że to jest, że to będzie parę dni trwało ale dobrze, Pani mi pomogła teraz.*	*pomógł bardzo naprawdę teraz mnie pani pomogła no bo ten pan mnie nic nie powiedział tak a ja też nie wiedziałam że to jest że to będzie parę dni trwało ale dobra pani mnie pomogła teraz*
11	2	V1 Stan usługi	*to znaczy jak w ogóle z tego nie korzystałam tego netfliksa kompletnie ani razu nie włączam tego pakietu*	*to znaczy jak w ogóle z tego nie korzystałam tego netflix kompletnie ani razu nie włączyłam tego pakietu*
12			*Czy ja mam jakiś pakiet włączony czy czy czy po prostu nie wiem?*	*czy ja mieć jakichś pakiet włączony czy czy czy po prostu nie wiem*

13		V2 Informacje o umowie	*Witam dzień dobry jest Sony Proszę pana chciałem się zapytać do kiedy mam umowę nie tylko że nie mam hasła abonenckiego od razu po z góry bo nie ma mnie w domu nie mam dostępu do tego hasła tak na mnie tak*	*witam dzień dobry jest sony poproszę pana chciałem się zapytać do kiedy mieć umowa nie tylko że nie mieć hasła abonenckiego od razu po z góry bo nie ma mnie w domu nie mieć dostępu do tego hasła tak na mnie tak*
14			*Dzień dobry. Chciałam się dowiedzieć odnośnie jednego numeru na moim koncie.*	*dzień dobry chciałam się dowiedzieć odnośnie 1 numer na moim koncie*
15		V3 Aktywacja usług internetowych	*No jakiś taki pakiet, jaki Jaki mogę sobie wykupić pakiet internetu?*	*no jakichś taki pakiet jaką jaką mogę sobie zakupić pakiet internet*
16			*Dobry wieczór, moje nazwisko, ja dzwonię do pani w takiej sprawie, bo ja mam intenet 35 GB tylko, a dzieci korzystają na lekcje online*	*dobry wieczór moje nazwisko ja dzwonię do pani w takiej sprawa bo ja mieć internet 35 gigabyte tylko a dzieci korzystają na lekcje online*
17		V4 Naprawa urządzenia	*Porysowana krutynia klapka.*	*porysowana tylny klapka*
18			*Około 2 tygodni temu oddałem telefon do naprawy mi się z piłą szkiełko w ramach po prostu takiego ubezpieczenia, które opłacam.*	*około 2 tygodni temu oddałem telefon do naprawy mnie się zbiło szkiełko w ramach po prostu takiego ubezpieczenia który opłacam*
19		V5 Przedłużenie umowy	*Ale to już teraz nie będę mogła zawrzeć umowy na 2,3 czy 3 lata tylko muszę na nieokreślony musi być?*	*ale to już teraz nie będę mogła zawrzeć umowa na 2 3 czy 3 lata tylko muszę na nieokreślony musi być*
20			*Bo, przedłużam umowę i tak po prostu jednorazowa taka nadpłata, że tak to nazwę.*	*bo przedłużam umowa i tak po prostu jednorazowa taka nadpłata że tak to nazwę*
21	3	V1 Informacje o fakturze	*Dobrze, do dobrze, to jest 196 zł za dwie faktury.*	*dobra do dobra to jest 196 złotych za 2 faktury*
22			*Dzień dobry, proszę Pana dzwonię odnośnie faktury z p..., bo jeden mam w r.., drugi mam w p..., i jedna mi faktura w ogóle nie dotarła do mnie, a ona przychodzi mi pocztą.*	*dzień dobry poproszę pana dzwonię odnośnie faktury z p... bo 1 mieć w r... 2 mieć w p... i 1 mnie faktura w ogóle nie dotarła do mnie a ona przychodzi mnie pocztą*
23		V2 Problem z połączeniem internetowym	*Dzień dobry ja mam tylko sprawę nie mam internetu od dwóch dni i dzwonię podpytać co się wydarzyło*	*dzień dobry ja mieć tylko sprawę nie mieć internet od 2 dni i dzwonię podpytać co się wydarzyło*
24			*No dobra no ale dlaczego skoro niby on został aktywowany ja nie mam internetu*	*no dobra no ale dlaczego skoro niby on został aktywowany ja nie mieć internet*
25		V3 Połącz z agentem	*Będę kurna czekał pół dnia, natychmiastową proszę odpowiedź. Niech ktoś do mnie zadzwoni, nie wiem kurna, napisze mejla, bo zaraz mnie kurna szlag trafi.*	*będę kurna czekał pół dnia natychmiastową poproszę odpowiedź niech ktoś do mnie zadzwoni nie wiem kurna napisze mail bo zaraz mnie kurna szlag trafi*
26			*kurna, jak ja czekam od piątku i to samo mi mówicie*	*kurna jak ja czekam od piątku i to samo mnie mówicie*
27		V4 Opis uszkodzeń	*No to znaczy tak, jest pobity ekran, chociaż on jest działający, ale jest pobity i co nie działa, wejście do ładowania.*	*no to znaczy tak jest pobity ekran chociaż on jest działający ale jest pobity i co nie działa wejście doładowania*
28			*No właśnie, to jest głównie pęknięcie tego wyświetlacza.*	*no właśnie to jest głównie pęknięcie tego wyświetlacza*
29		V5 Odzyskiwanie hasła	*No ale okej, ja nawet nie mam tego hasła swojego, więc jak to teraz mogę uzyskać tanie?*	*no ale ok ja nawet nie mieć tego hasła swojego więc jak to teraz mogę uzyskać tanie*
30			*Jakie hasło? Ja nie nam żadnego hasła.*	*jakie hasło ja nie nam żadnego hasła*

Poszczególne grupy wyselekcjonowanych intencji zaprezentowane w tabeli 5.8 oraz tabeli 5.9 zostały wykorzystane również w eksperymentach opisanych w kolejnych rozdziałach tej pracy.

W tabeli 5.9 zestawiono wyniki wykonanych eksperymentów obrazujące dokładność rozpoznawania intencji klientów przy zastosowaniu wymienionych sposobów transkrypcji automatycznej.

Dla danych z tabeli 5.9, które zacieniowano oraz oznaczono symbolami (a-c), na rysunku 5.12 zestawiono odpowiednie macierze konfuzji. Obrazują one szczegółowe rozkłady otrzymanych wyników pokazujące poprawność rozpoznania poszczególnych intencji dla fraz testowych. Wiersz pierwszy macierzy we wszystkich kolumnach prezentuje dane uzyskane przy wykorzystaniu systemu ASR, natomiast wiersz drugi macierzy zawiera rozkłady dla danych otrzymanych w wyniku zastosowania metody *IAT*.

Na podstawie wykonanych eksperymentów, dla zdefiniowanych poszczególnych grup (1, 2, 3) intencji (V1-V5), obserwuje się, że zastosowana metoda poprawy jakości transkrypcji *IAT* dedykowanej dla branży CC zwiększa wyraźnie dokładność *accINT* rozpoznawania intencji. W zależności od badanej próby, dla platformy Dialogflow poprawa następuje w zakresie od 1,15% do 27,94% natomiast dla platformy Rasa od 2,11% do 13,58%.

Tabela 5.9. Wpływ jakości transkrypcji na dokładność rozpoznawania intencji dla kanału głosowego

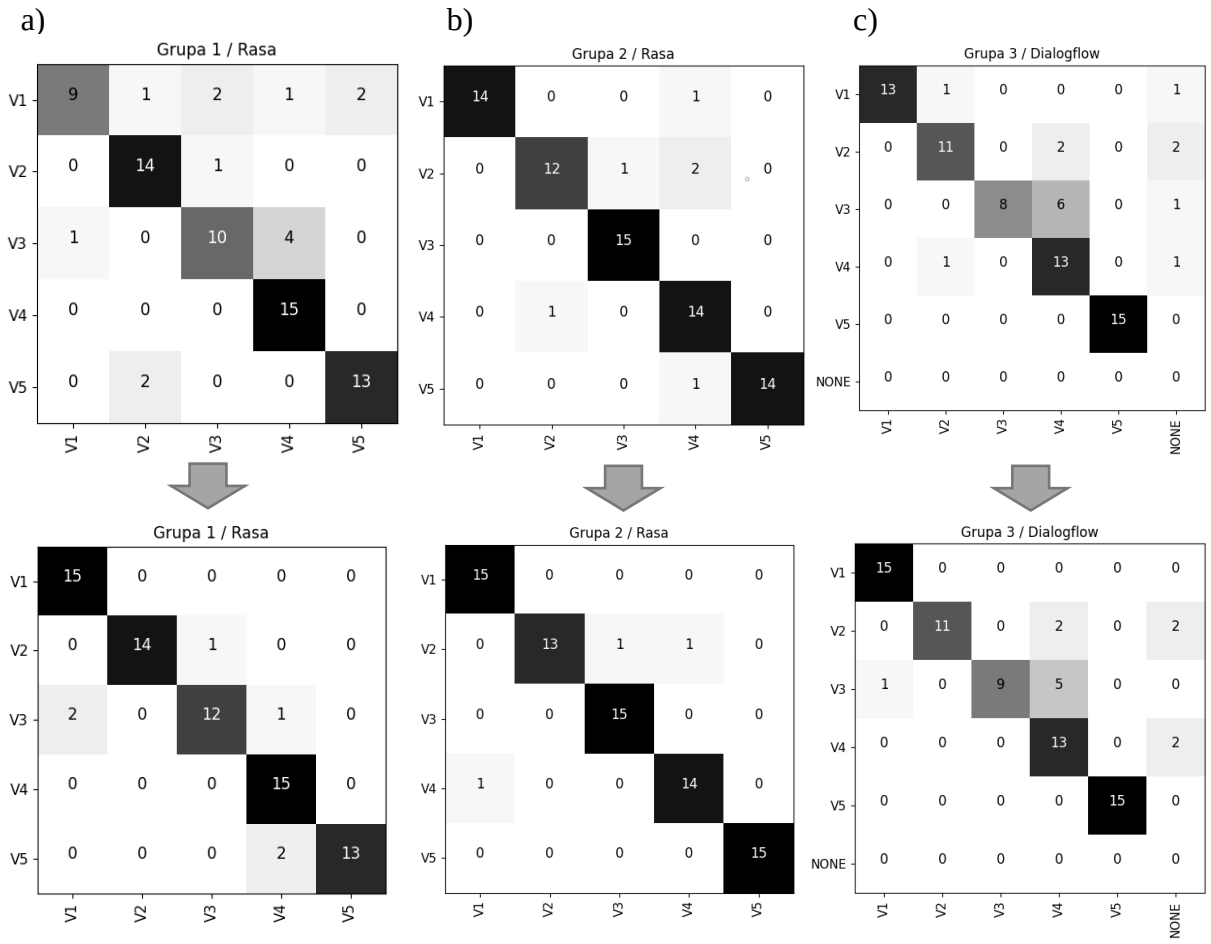
Grupa	**Platforma NLU**	**Rozpoznawanie intencji F_{NLU}**		
		System ASR	**Metoda *IAT***	**Poprawa**
		accINT	***accINT***	**[%]**
1	Dialogflow	0,68 (51/75)	0,87 (65/75)	**27,94**
	Rasa	0,81[(a)] (61/75)	0,92[(a)] (69/75)	**13,58**
2	Dialogflow	0,87 (65/75)	0,88 (66/75)	**1,15**
	Rasa	0,92[(b)] (69/75)	0,96[(b)] (72/75)	**4,35**
3	Dialogflow	0,80[(c)] (60/75)	0,84[(c)] (63/75)	**5,00**
	Rasa	0,95 (71/75)	0,97 (73/75)	**2,11**

(a), (b), (c) – dane zaprezentowane na rysunku 5.12

Wykorzystywany model NLU wytrenowany był frazami uczącymi pochodzącymi z transkrypcji manualnych, które były w pełni poprawne. Błędy występowały we frazach testowych utworzonych na podstawie transkrypcji automatycznych. Po zastosowaniu metody *IAT* duża liczba błędów transkrypcji automatycznych została skorygowana.

Na rysunku 5.12 zaprezentowano rozkład danych w postaci macierzy konfuzji dla wybranych wartości z tabeli 5.9. W kolumnie a) można zauważyć, że największy rozrzut błędów miała intencja V1 przy zastosowaniu systemu ASR, a po zastosowaniu metody *IAT* nastąpiła bardzo znacząca poprawa. Nieznaczna poprawa nastąpiła dla intencji V3. Natomiast dla intencji V2, V4, V5 poprawa transkrypcji nie miała znaczenia. W kolumnie b) poziom rozpoznawania intencji był bardzo wysoki już przy systemie ASR, natomiast zastosowanie metody *IAT* jeszcze ten poziom podniosło dla intencji V1, V2, V5. W kolumnie c) spory rozrzut błędów można zauważyć dla intencji V1, V2, V3. Natomiast po zastosowaniu metody *IAT* poprawa zaistniała dla intencji V1 i nieznaczna dla V3, natomiast dla V2 pozostała na niezmienionym poziomie. Można wnioskować z tego, że dla V2 problemy z rozpoznawaniem intencji wynikają z innych powodów niż poprawność transkrypcji (np. niewystarczająco dobrze dobrane frazy uczące).

Na rysunku 5.12 w kolumnie c) można również zaobserwować, że dla platformy Dialogflow dość często zdarza się, że żadna intencja nie zostaje rozpoznana (co oznaczono jako NONE), natomiast platforma Rasa zawsze próbuje dopasować jakąś intencję co widać w kolumnach a) i b).



Rysunek 5.12. Macierze konfuzji dla wybranych danych z tabeli 5.9

Legenda: **V1 – V5** – Intencje w poszczególnych grupach (1, 2 lub 3) zgodnie z tabelą 5.9; **NONE** – nierozpoznana żadna intencja; Osie rzędnych (y) opisują wartości **PRAWDA**; Osie odciętych (x) opisują wartości **PREDYKCJA**

W tabeli 5.10 zaprezentowano dla „Grupa 1 / Rasa" szczegółowe dane w rozbiciu na poszczególne intencje z zastosowaniem systemu ASR oraz metody *IAT*. W wierszu zawierającym intencję „V3 Autoryzacja" można dostrzec, że prawidłowość rozpoznawania poprawiła się w sposób znaczący przy zastosowaniu metody *IAT*. Dlatego w tabeli 5.11 zaprezentowano frazy uczące, a w tabeli 5.12 frazy testowe dla intencji „V3 Autoryzacja".

Tabela 5.10. Wyniki accINT dla fraz testowych z „Grupa 1 / Rasa" dla kanału głosowego

Zbiór fraz testowych dla Intencji	**Rozpoznawanie intencji *accINT***		
	System ASR	**Metoda *IAT***	**Poprawa**
	accINT	***accINT***	**[%]**
V1 Potwierdzenie	0,67 (10/15)	0,80 (12/15)	**19,40**
V2 Zaprzeczenie	0,87 (13/15)	0,87 (13/15)	**0,00**
V3 Autoryzacja	0,60 (9/15)	1,00 (15/15)	**66,67**
V4 Wypowiedzenie umowy	1,00 (15/15)	1,00 (15/15)	**0,00**
V5 Zadowolenie z obsługi	0,93 (14/15)	0,93 (14/15)	**0.00**
Cały zbiór	**0,81 (61/75)**	**0,92 (69/75)**	**13,58**

Tabela 5.11. Frazy uczące dla „V3 Autoryzacja" z „Grupa 1 / Rasa" dla kanału głosowego

Lp.	**Frazy uczące P_i**	
	Transkrypcja manualna	**Metoda *IAT***
1	Czego imei?	czego imei
2	Nr imei właśnie szukam.	nr imei właśnie szukam
3	te i IMEI IMEI 1 dobrze to z kodem kreskowym tak na pudełku	te i imei imei 1 dobra to z kodem kreskowym tak na pudełku
4	Jest. IMEI 1 IMEI 2, tak, to?	jest imei 1 imei 2 tak to
5	IMEI 1 czy IMEI 2?	imei 1 czy imei 2
6	tylko niech mi pani teraz powie, to będzie IMEI jeden czy IMEI dwa?	tylko niech mnie pani teraz powie to będzie imei 1 czy imei 2
7	mogę podać jeszcze raz ten numer IMEI, bo to jest na pudełku	mogę podać jeszcze raz ten nr imei bo to jest na pudełku
8	na umowie e mam poprawny numer imei	na umowie e mieć poprawny nr imei
9	będę miał po prostu ten nr IMEI	będę miał po prostu ten nr imei
10	mogę pesel bo ja nie pamiętam hasła	mogę pesel bo ja nie pamiętam hasła
11	pesel zawsze podawałem	pesel zawsze podawałem
12	mogę numer pesel podać.	mogę nr pesel podać
13	Nie mam hasła, PESEL jest.	nie mieć hasła pesel jest
14	po prostu nie pamiętam, wcisnęłam PESEL	po prostu nie pamiętam wcisnęłam pesel
15	ja jakoś załatwię ten numer IMEI i się z wami skontaktuję	ja jakichś załatwię ten nr imei i się z wami skontaktuję
16	będę miała numer IMEI tego mojego telefonu	będę miał nr imei tego mojego telefonu
17	Jest tak IMEI	jest tak imei
18	ten numer imei to jest z tyłu na telefonie	ten nr imei to jest z tyłu na telefonie
19	Myślałam, że IMEI jest na obudowie	myślałam że imei jest na obudowie
20	ten numer imei, on będzie jakby sprawdzany	ten nr imei on będzie jakby sprawdzany
21	podam swój numer PESEL	podam swój nr pesel
22	podam pani pesel	podam pani pesel
23	No to nie można po peselu? Parę razy dzwoniłam i po peselu zawsze mi ktoś sprawdzał.	no to nie można po pesel parę razy dzwoniłam i po pesel zawsze mnie ktoś sprawdzał
24	PESEL mogę podać	pesel mogę podać
25	mogę podać PESEL?	mogę podać pesel

Frazy uczące były selekcjonowane z transkrypcji manualnej pochodzącej z nagrań rozmów telefonicznych. Metoda *IAT* zmodyfikowała frazy uczące w zakresie zamiany tekstu do małych liter oraz eliminacji znaków interpunkcyjnych.

Przedstawiony w tabeli 5.12 zbiór fraz testowych był tematycznie zawężony do autoryzacji w obsłudze klienta telekomunikacyjnego (przypisany do intencji „V3 Autoryzacja”), stąd w wypowiedziach klientów powtarzały się wyrażenia „pesel” oraz „numer imei”. Natomiast w transkrypcji automatycznej często te frazy były błędnie zapisywane jako „psl”, „posel” czy „numer email”, co można zaobserwować w większości wierszy. Z uwagi na te błędy platforma NLU nie poradziła sobie z prawidłowym rozpoznawaniem intencji „V3 Autoryzacja” w wierszach 1, 3, 5, 8, 9, 10. W tym przypadku dla systemu ASR poprawnie rozpoznano jedynie 9 z 15 intencji, natomiast po zastosowaniu metody *IAT* uzyskano rozpoznano 15 z 15 intencji.

We frazach testowych dla innych intencji takiej powtarzalności błędów nie zaobserwowano, ale zazwyczaj pojawiało się kilka przypadków, co zaprezentowano w tabeli 5.13 na przykładzie intencji „V1 Potwierdzenie”. W wierszu 2 i 7 metoda *IAT* poprawiła błędne słowa „tag” oraz „taak” na prawidłowe słowo „tak”, w wyniku czego intencja została rozpoznana prawidłowo.

Tabela 5.12. Frazy testowe „V3 Autoryzacja” z „Grupa 1 / Rasa” dla kanału głosowego

Lp.	**Frazy testowe P_i**		**Rozpoznawanie intencji F_{NLU}**	
	System ASR	**Metoda *IAT***	**System ASR**	**Metoda *IAT***
1	Gdy psl.	gdy pesel	Nie	Tak
2	I psl bym poprosiła.	i pesel bym poprosiła	Tak	Tak
3	Ale psl pani widzi początek psl u pani widzi?	ale pesel pani widzi początek pesel u pani widzi	Nie	Tak
4	Psl to jest.	pesel to jest	Tak	Tak
5	Psl już zresztą też n.	pesel już zresztą też n	Nie	Tak
6	Dobrze i psl oprawami?	dobra i pesel oprawami	Tak	Tak
7	E psl.	e pesel	Tak	Tak
8	Hmm psl jest.	hmm pesel jest	Nie	Tak
9	Nie, nie psl mogę podać pani?	nie nie pesel mogę podać pani	Nie	Tak
10	Nie podam psl.	nie podam pesel	Nie	Tak
11	Nigdzie nie mogę znaleźć numer email?	nigdzie nie mogę znaleźć numer imei	Tak	Tak
12	powiem numer email	powiem numer imei	Tak	Tak
13	ja mogę podać Jeszcze raz ten numer email bo to jest na pudełku prawda	ja mogę podać jeszcze raz ten numer imei bo to jest na pudełku prawda	Tak	Tak
14	Zaraz sobie numer email pudełka ee urządzenia podaj 15000.	zaraz sobie numer imei pudełka ee urządzenia podaj 15000	Tak	Tak
15	O nie mamy to ja zawsze proszę podawałam i adres zameldowania, bo naprawdę ten hasło to rzadko to się w ogóle go używa i wyleciał z głowy najczęściej się posel używa.	o nie mieć to ja zawsze poproszę podawałam i adres zameldowania bo naprawdę ten hasło to rzadko to się w ogóle go używa i wyleciał z głowy najczęściej się pesel używa	Tak	Tak
		Podsumowanie	**9/15**	**15/15**

Tabela 5.13. Frazy testowe „V1 Potwierdzenie” z „Grupa 1 / Rasa” dla kanału głosowego

Lp.	Frazy testowe P_i		Rozpoznawanie intencji F_{NLU}	
	System ASR	Metoda *IAT*	System ASR	Metoda *IAT*
1	No właśnie.	no właśnie	Tak	Tak
2	Ten sam tag.	ten sam tak	Nie	Tak
3	No dobrze, akceptuję no potwierdzam.	no dobra akceptuję no potwierdzam	Tak	Tak
4	E tak to będzie pierwszy naprawa.	e tak to będzie 1 naprawa	Nie	Nie
5	Tak oczywiście będzie ten sam z którego dzwonię	tak oczywiście będzie ten sam z którego dzwonię	Nie	Nie
6	Tak tak, tak właśnie	tak tak tak właśnie	Tak	Tak
7	taak	Tak	Nie	Tak
8	dobrze poproszę tak oczywiście tak	dobra poproszę tak oczywiście tak	Tak	Tak
9	No zwykle do tej pory akceptowałem, czyli teraz też akceptuję.	no zwykle do tej pory akceptowałem czyli teraz też akceptuję	Nie	Nie
10	Tak, ale jeszcze się zastanowię	tak ale jeszcze się zastanowię	Tak	Tak
11	Hmm dobra dobra dobra dobra super.	hmm dobra dobra dobra dobra super	Tak	Tak
12	No rozumiem, rozumiem, dobrze, super	no rozumiem rozumiem dobra super	Tak	Tak
13	No a co mam zrobić, no akceptuję.	no a co mieć zrobić no akceptuję	Tak	Tak
14	Tak, tak, tak, tak oczywiście będą będą spłacane.	tak tak tak tak oczywiście będą będą spłacane	Tak	Tak
15	Aha no to dobra, no to dobra	aha no to dobra no to dobra	Tak	Tak
		Podsumowanie	**10/15**	**12/15**

Powyższa dyskusja potwierdziła celowość implementacji w voicebotach CC modułów poprawy jakości transkrypcji automatycznej. Dzięki zwiększonej jakości transkrypcji poprawiają się wyniki rozpoznawania intencji na platformach NLU. Dokładna analiza tabeli 5.9 pokazuje, że użycie metody *IAT* dla każdego zbioru zwiększyło dokładność rozpoznawania intencji dla Dialogflow od 1,15% do 27,94%, a dla platformy Rasa od 2,11% do 13,58%. Natomiast patrząc na bardziej szczegółowe dane w ramach zbioru testowego „Grupa 1 / Rasa” (tabela 5.10) można dostrzec, że poprawa dla „V3 Autoryzacja” osiągnęła aż 66,67%. Oczywiście należy uznać to za skrajny przypadek, gdyż tutaj system ASR notorycznie popełniał błędy w słowach „pesel” i „numer imei”. Jednak takie przypadki często stanowią problem w systemach ASR, gdyż trudno je zmusić do poprawnego rozpoznawania takich specyficznych słów. Dlatego też zastosowanie metody *IAT* w takich sytuacjach może być najskuteczniejszym rozwiązaniem w porównaniu do podejścia douczania systemu ASR (MST) lub mechanizmu podpowiedzi (GST).

Opracowana metoda poprawy jakości transkrypcji automatycznej *IAT* dla CC pomimo swojej dużej skuteczności nie rozwiązuje wszystkich aspektów wpływających na udzielanie im jak najbardziej trafnych odpowiedzi przez voicebota CC. Dlatego też jako kolejny cel badań wyznaczono analizę wpływu emocji na rozpoznawanie intencji, aby osiągnąć maksymalną skuteczność funkcjonowania wirtualnego konsultanta.

6 Wpływ emocji na rozpoznawanie intencji

Niniejszy rozdział przedstawia przeprowadzone badania, w zakresie wpływu emocji na rozpoznawanie rzeczywistych intencji klientów, w celu zwiększenia skuteczność funkcjonowania wirtualnego konsultanta. Zaprezentowano analizę lingwistyczną i przykłady konwersacji, w których kluczowym elementem do zrozumienia intencji klienta jest rozpoznawanie jego emocji. Wskazano mechanizmy (encję określającą emocję, reguły wnioskowania), które zostały wykorzystane przy tworzeniu nowej metody rozpoznawania ukrytych intencji. Scharakteryzowano sposób tworzenia zbiorów uczących i testujących. Zaproponowano koncepcję nowego procesu funkcjonowania bota. Na końcu przedstawiono wyniki eksperymentów z zastosowaniem nowej metody rozpoznawania ukrytych intencji uwzględniającej emocje, zarówno dla kanału głosowego jak i kanału czat.

6.1 Metodologia badań

Dane eksperymentalne stanowiły rzeczywiste rozmowy Contact Center pomiędzy agentem i klientem. W badaniach wykorzystano bazę RLPHDB opisaną w podrozdziale 5.1 obejmującą 1054 rzeczywiste rozmowy telefoniczne i nowo utworzoną bazę danych zawierającą 545 rzeczywiste rozmowy czat, oznaczoną jako RLCHDB (ang. real chat conversations). W wyniku współpracy z psychologami i lingwistami specjalizującymi się w obszarze CC określono pięć głównych typów emocji, które w problematyce rozmów na linii klient-agent mają największe znaczenie. Są to emocje: Radość, Złość, Smutek, Strach oraz Neutralność [97].

W tabeli 6.1 przedstawiono przykłady, w których rozumienie wypowiedzi klienta wymaga posiadania pewnej wiedzy o świecie, konsytuacji oraz umiejętności odczytywania innych pozawerbalnych znaków komunikacji. W tabeli 6.1 oznaczono również typy emocji, jakie zostały zidentyfikowane przy współpracy z psychologami w prezentowanych dialogach CC.

Tabela 6.1. Wybrane intencje informacyjne skorelowane z emocjami

Lp.	K/A	Fragment konwersacji	Znaczenie dosłowne (bez tła emocjonalnego)	Rozpoznana emocja	Znaczenie implikowane z tłem emocjonalnym
PRZYKŁADY RZECZYWISTYCH KONWERSACJI Z KANAŁU GŁOSOWEGO					
1	K	*Przepraszam, ale to jest nieporozumienie żebym płacił za coś czego nie otrzymałem*	Jestem wdzięczny za to, że mogę korzystać z usług firmy	Złość	Mam dość współpracy z tą firmą.
	A	*Rozumiem*			
	K	***Dziękuję za taką firmę***			
		Analiza konwersacji: Stwierdzenie, zdanie oznajmujące.			
2	A	*Czy zapoznał się pan z regulaminem wybranej usługi serwisowej i akceptuje Pan jego warunki?*	Zapoznałem się z regulaminem usługi serwisowej.	Strach	Nie zapoznałem się z regulaminem, natomiast żeby moja sprawa była rozpatrywana potwierdziłem ten fakt.
	K	*No dobrze,* ***akceptuję, potwierdzam***			
		Analiza konwersacji: Zdanie oznajmujące, użycie partykuły *no* wzmacnia fakt formalnego zatwierdzenia regulaminu usługi, ale użycie jej przed czasownikiem *potwierdzam,* wskazuje na to, że klient nie ma pewności, czy zna treść regulaminu.			
3	A	*Czy kontakt z infolinią pomógł w rozwiązaniu sprawy. Po rozmowie otrzyma Pani SMS*	Dziękuję, na pewno odpowiem	Radość	Dziękuję na pewno odpowiem, ocena będzie pozytywna.
	K	***Dziękuję, na pewno odpowiem***			
		Analiza konwersacji: Zdanie oznajmujące, użycie wyrażenia przyimkowego *na pewno* wskazuje, że klient, którego sprawa została rozwiązana zgodnie z oczekiwaniami, pochwali firmę za jej skuteczność.			
4	A	*Czy chce Pan zrezygnować z numeru z którego Pan dzwoni?*	Mam zamiar zrezygnować z wszystkich numerów.	Smutek	Jestem rozczarowany i jak nie rozwiążecie sprawy to zrezygnuję ze wszystkich numerów.
	K	***Będę rezygnował ze wszystkich numerów.***			

<table>
<tr><td></td><td></td><td colspan="4">Analiza konwersacji: Zdanie oznajmujące. Użycie zaimka upowszechniającego w wyrażeniu „wszystkich numerów” wskazuje na duże rozczarowanie klienta, służy do uogólnienia i podkreślenia negatywnych emocji, które klient żywi w stosunku do firmy. Generalizacja tego rodzaju stanowi jedną z barier komunikacyjnych polegających na przesadnym przypisaniu firmie nawet tych błędów, których nie popełniła.</td></tr>
<tr><td colspan="6">PRZYKŁADY RZECZYWISTYCH KONWERSACJI Z KANAŁU TEKSTOWEGO</td></tr>
<tr><td rowspan="4">1</td><td>A</td><td>W czym mogę pomóc?</td><td rowspan="3">O, ponownie się spotykamy.</td><td rowspan="3">Smutek</td><td rowspan="3">Nie jestem zadowolony, że znowu będzie mnie Pan obsługiwał</td></tr>
<tr><td>K</td><td>O znowu Pan! :(</td></tr>
<tr><td>A</td><td>Postaram się Panu pomóc jak najlepiej.</td></tr>
<tr><td></td><td colspan="4">Wykrzyknienie, z użyciem emotikona ze smutną miną, wskazuje na rozczarowanie klienta, że będzie obsługiwany przez osobę, z której usług nie był zadowolony.</td></tr>
<tr><td rowspan="4">2</td><td>K</td><td>Poproszę o aktywowanie subskrypcji</td><td rowspan="3">Dziękuję za pomoc, subskrypcja została aktywowana.</td><td rowspan="3">Radość</td><td rowspan="3">Pomógł mi Pan. O to mi chodziło.</td></tr>
<tr><td>A</td><td>Subskrypcja została aktywowana. Czy mogę jeszcze jakoś pomóc?</td></tr>
<tr><td>K</td><td>Póki co dzięki za pomoc. Pozdrawiam :)</td></tr>
<tr><td></td><td colspan="4">Analiza konwersacji: Użycie w wypowiedzi wyrazów potocznych, takich jak dzięki i póki co (w znaczeniu-jak do tej pory), podkreślają zakończenie danej czynności.</td></tr>
<tr><td rowspan="4">3</td><td>K</td><td>To kto mi jest w stanie odpowiedzieć?</td><td rowspan="3">Jestem bardzo usatysfakcjonowany</td><td rowspan="3">Złość</td><td rowspan="3">Jestem niezadowolony, wręcz wściekły, że nikt mi nie udzielił odpowiedzi na moje pytanie.</td></tr>
<tr><td>A</td><td>Niestety nikt</td></tr>
<tr><td>K</td><td>Jestem bardzo usatysfakcjonowany!!! :(Do widzenia!!!</td></tr>
<tr><td></td><td colspan="4">Analiza konwersacji: Zdanie wykrzyknikowe, niezadowolenie klienta wzmacniają wykrzykniki na końcu frazy oraz emotikon ze smutną minką.</td></tr>
<tr><td rowspan="4">4</td><td>A</td><td>Zgłoszenie zostało wysłane, numer Pana zgłoszenia to 123-456</td><td rowspan="3">Przez dłuższy czas będę bez numeru telefonu</td><td rowspan="3">Strach</td><td rowspan="3">Zbyt długo nie będę miał telefonu i obawiam się tego.</td></tr>
<tr><td>K</td><td>Co mi z numeru zgłoszenia, tydzień będą naprawiać i przez ten okres będę bez telefonu :(</td></tr>
<tr><td>A</td><td>Rozpoczęcie świadczenia usług może nastąpić w terminie 24 godzin od Aktywacji.</td></tr>
<tr><td></td><td colspan="4">Analiza konwersacji: Zdanie oznajmujące, którego treść została wzmocniona emotikonem ze smutną minką, wzmacnia obawy klienta, że nie będzie mógł korzystać z telefonu i przez to będzie narażony na jakieś niedogodności.</td></tr>
</table>
Legenda: K – Klient, A – Agent

Jak wynika z analizy przykładów zamieszczonych w tabeli 6.1, interpretacja słów lub wyrażeń w powiązaniu z konkretnym typem emocji może zmieniać zupełnie kierunek rozmowy. Jest to związane z rozumieniem znaczeń przekazywanych przez klienta werbalnie, ale również i emocjonalnie. Te same frazy wypowiadane z określoną emocją mają inne znaczenie niż te same frazy pozbawione tła emocjonalnego [98]. Zadaniem wirtualnych konsultantów powinna być zatem możliwość prowadzenia rozmów w taki sposób, aby możliwe było uzależnienie dialogów zarówno od rozpoznawanych intencji jak również rozpoznawanych emocji, jeśli takie wystąpią. Wówczas wirtualny konsultant będzie mógł traktować frazę z emocją negatywną inaczej niż tę samą frazę z emocją pozytywną lub neutralną. Analiza lingwistyczna wykazała, że w wielu przypadkach funkcja F_{NLU} może zwracać fałszywe intencje. Prawdziwe (ukryte intencje) można zidentyfikować poprzez uwzględnienie emocji z wypowiedzi klienta. Są możliwe dwa rozwiązania:

- Modyfikacja funkcji F_{NLU}
- Modyfikacja funkcji F_A

Pierwsze podejście wymagałoby opracowanie nowej platformy NLU. Ponieważ założeniem pracy było wykorzystanie standardowych platform NLU z tego względu wybrano drugie podejście i wprowadzono nową funkcję F_{AE}.

Definicja 6.1

Funkcja F_{AE} określa akcję bota dla intencji INT_j i emocji wypowiedzi E_f:

$$F_{AE}: (S_{INT}, S_E) \rightarrow S_A \quad (6.1)$$

gdzie $INT_j \in S_{INT}$ *i* $E_f \in S_E$.

Definicja 6.2

Jeżeli dla pewnej frazy P_i zachodzi:

$$INT_j = F_{NLU}(P_i) \quad (6.2)$$

i istnieje E_f takie, że:

$$F_{AE}(INT_j, E_f) \neq F_A(INT_j) \quad (6.3)$$

wtedy INT_j nazywamy fałszywą intencją w kontekście frazy P_i.

Funkcja F_{AE} definiuje prawidłową akcję bota A_k w przypadku, gdy z frazą P_i związana jest emocja E_f:

$$A_k = F_{AE}(INT_j, E_f) \quad (6.4)$$

Natomiast nieuwzględnienie emocji powoduje nieprawidłową akcję bota A_y:

$$A_y = F_A(INT_j) \quad (6.5)$$

W celu spowodowania wyboru akcji bota należy zidentyfikować ukrytą intencję zawartą w wypowiedzi klienta.

Definicja 6.3

Niech INT_j będzie fałszywą intencją w kontekście frazy P_i taką, że

$$F_{AE}(INT_j, E_f) = A_k \quad (6.6)$$

Jeśli istnieje INT_x takie, że:

$$F_A(INT_x) = A_k \quad (6.7)$$

wtedy INT_x nazywamy ukrytą intencją.

Ukryta intencja wynika z intencji rozpoznanej na podstawie tekstu wypowiedzi, która jest korygowana na podstawie rozpoznanej emocji w tej wypowiedzi. Gdy wystąpi emocja inna niż

Neutralna to akcja bota może zostać zmieniona. Na przykład: Jeżeli rozpoznana zostanie intencja *Zadowolenie z obsługi* to w przypadku:

- Gdy rozpoznana emocja będzie Neutralna zostanie wybrana akcja: *Klient zadowolony, podziękuj.*
- Gdy rozpoznana emocja będzie Złość zostanie wybrana akcja: *Klient niezadowolony, zapytaj, dlaczego.*

Przy założeniu, że dla części fraz w zbiorze testowym rozpoznawane są fałszywe intencje, w wyniku czego zostanie dla nich wybrana niepoprawna akcja.

m – liczba wszystkich fraz testowych

m_1 – liczba fraz testowych z fałszywi intencjami (wybrane niepoprawne akcje)

Założono, że celem prac będzie uzyskanie funkcji F_{AE}, która będzie uwzględniać emocje (wykrywać ukryte intencje).

m_2 – liczba fraz testowych z ukrytymi intencjami (wybrane poprawne akcje dla części fraz z fałszywymi intencjami, gdzie $0 < m_2 <= m_1$).

Dokładność $accA_1$ dla m_1:

$$accA_1 = \frac{m - m_1}{m} \qquad (6.8)$$

Dokładność $accA_2$ dla m_2:

$$accA_2 = \frac{m - (m_1 - m_2)}{m} \qquad (6.9)$$

Można zauważyć, że $accA_1 < accA_2$, ponieważ:

$$m_2 > 0 \Longrightarrow \frac{m - m_1}{m} < \frac{m - (m_1 - m_2)}{m} \qquad (6.10)$$

Warunkiem wyżej wymienionych zależności, jest założenie, że fałszywe intencje są odwzorowywane w niepoprawne akcje.

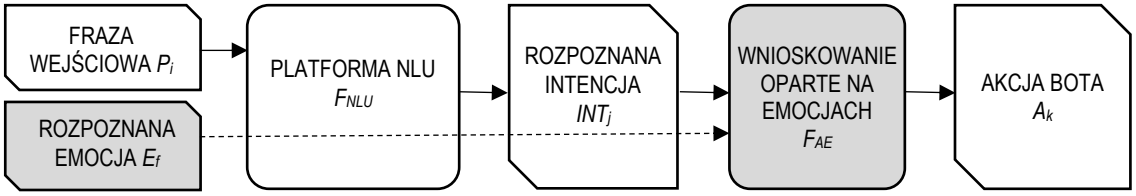
Zatem jeżeli zwiększy się liczba rozpoznawanych ukrytych intencji to poprawi się dokładność odwzorowania fraz w akcje bota *accA.*

6.2 Metoda rozpoznawania ukrytych intencji

Jak pokazano w podrozdziale 6.1, specyfika rozmów w CC wskazuje, że jest wiele przypadków, gdy odpowiedzi na rozpoznane automatycznie intencje klienta powinny być różne w zależności od pojawiających się w poszczególnych konwersacjach emocji. Dlatego też, w prezentowanym podejściu wykorzystana została metoda rozpoznawania emocji klienta w kanałach dźwiękowych oraz kanałach tekstowych CC [82]. W proponowanej metodzie (rysunek 6.1) na wejście platformy NLU dostarczana jest fraza testowa (lub wejściowa) oraz wartość rozpoznanej emocji. Fraza testowa podlega standardowemu przetwarzaniu przez wybraną platformę NLU w wyniku czego określona zostaje rozpoznana intencja. Następnie rozpoznana intencja jest korygowana w zależności od rozpoznanych emocji w wypowiedzi klienta, w ten sposób scenariusz rozmowy przełączany jest zgodnie ze zdefiniowanymi regułami wnioskowania na podstawie przekazanych emocji. Na wyjście zwracana jest odpowiedź do

klienta generowana przez akcję bota wynikająca z predykcji intencji klienta z uwzględnieniem jego stanu emocjonalnego.

Na rysunku 6.1 szary kolor tła wskazuje na nowe elementy zaproponowane w pracy. W bloku WNIOSKOWANIE OPARTE NA EMOCJACH jest użyta funkcja F_{AE}, określona wzorem (6.1), której argumentami jest zarówno ROZPOZNANA INTENCJA INT_j jak i ROZPOZNANA EMOCJA E_f, a wynikiem AKCJA BOTA A_k.



Rysunek 6.1. Proces działania bota uwzględniający emocje

W tabeli 6.2 przedstawiono wybrane fragmenty wypowiedzi klientów, wyselekcjonowane bezpośrednio z rzeczywistych rozmów CC, dla których w zależności od stanów emocjonalnych klienta wnioskowane akcje wykonywane przez bota powinny być zupełnie inne. Zobrazowane przykłady dotyczą zarówno kanału głosowego jak również kanału tekstowego.

Tabela 6.2. Przykład budowy scenariusza rozmowy w zależności od emocji

Kanał	**Fraza klienta** P_i	**Intencja** INT_j	**Emocja** E_f		**Akcja bota**	
					F_A	F_{AE}
Głos	*Super firma*	Zadowolenie z obsługi	Złość, Smutek, Strach	→	Klient zadowolony, podziękuj	Klient niezadowolony, zapytaj, dlaczego
			Radość	→	Klient zadowolony, podziękuj	Klient zadowolony, podziękuj i wyślij ankietę
			Neutralność	→	Klient zadowolony, podziękuj	Klient zadowolony, podziękuj
Głos	*No dobrze, akceptuję, potwierdzam*	Potwierdzenie	Smutek, Strach	→	Przyjmij odmowę	Potwierdza, ale upewnij się
			Radość, Złość, Neutralność	→	Przyjmij odmowę	Przyjmij potwierdzenie
Głos	*No, nie dostałam*	Zaprzeczenie	Złość, Smutek, Strach	→	Przyjmij odmowę	Zaprzecza, ale upewnij się
			Radość, Neutralność	→	Przyjmij odmowę	Przyjmij odmowę
Głos	*Będę rezygnował ze wszystkich numerów*	Wypowiedzenie umowy	Smutek, Strach	→	Przyjmij rezygnację	Upewnij się czy jest zamiar rezygnacji
			Złość	→	Przyjmij rezygnację	Rezygnuje, zaproponuj lepszą ofertę
			Neutralność	→	Przyjmij rezygnację	Przyjmij rezygnację
Tekst	*Nie mogę zresetować hasła*	Odzyskiwanie hasła	Złość, Smutek	→	Reset hasła przez email	Reset hasła przez agenta
			Radość, Strach, Neutralność	→	Reset hasła przez email	Reset hasła przez email
Tekst	*Reinstalacja aplikacji nie dała rezultatów*	Problemy z aplikacją	Złość	→	Samodzielna reinstalacja aplikacji	Reset aplikacji przez agenta
			Smutek, Strach, Radość, Neutralność	→	Samodzielna reinstalacja aplikacji	Samodzielna reinstalacja aplikacji
Tekst	*Proszę o połączenie z agentem*	Połącz z agentem	Złość	→	Spróbuj rozwiązać problem automatycznie	Połącz z agentem natychmiast
			Smutek, Strach	→	Spróbuj rozwiązać problem automatycznie	Wstaw do kolejki oczekującej do agenta
			Radość, Neutralność	→	Spróbuj rozwiązać problem automatycznie	Spróbuj rozwiązać problem automatycznie

Legenda: Zaznaczone pola zawierają błędną wartość akcji bota.

W zakresie opracowywanej metody rozpoznawania ukrytych intencji F_{AE} ważnym elementem wymagającym rozstrzygnięcia było określenie sposobu przekazywania informacji o rozpoznawanych emocjach klientów. Najprostszym z możliwych rozwiązań w tym zakresie jest uzupełnianie frazy uczącej stosownym tekstem opisującym dany typ emocji. Rozwiązanie to nie jest jednak skuteczne ze względu na niekorzystny wpływ na przeprowadzane w dalszych krokach procesy uczenia. W konsekwencji takie rozwiązanie dość często skutkuje błędnym rozpoznawaniem intencji. Inną możliwością jest wykorzystanie wejściowych parametrów sesji. Jednak przekazywanie parametrów na każdej platformie jest realizowane w inny sposób, co stanowi stosunkowo duże ograniczenie mając na uwadze uniwersalność proponowanej metody. Ostatecznie wybrane zostało rozwiązanie najbardziej uniwersalne, wykorzystujące mechanizm encji [99], [100]. Jest to mechanizm będący standardem dla różnych platform NLU [101].

Encja definiuje typ konkretnych informacji, które można wyodrębnić z danych wejściowych. Odpowiadają one wielu różnym popularnym typom danych, np.: encje dopasowujące daty, adresy e-mail czy godziny. Istnieje ponadto możliwość tworzenia własnych encji, w celu dopasowywania danych niestandardowych.

6.2.1 Reprezentowanie emocji za pomocą encji

Encje na platformach NLU w głównym zamierzeniu służą do zbierania danych od klienta np. marki telefonu, koloru telefonu itd. Fraza testowa przekazywana do platformy NLU stanowi podstawę do rozpoznania intencji, a encje, które się znajdują w frazie testowej są wyodrębniane w celu doprecyzowania informacji od klienta. Przykład definiowania encji we frazie uczącej i wyodrębnienia wartości encji z frazy testowej przedstawiono w tabeli 6.3. W przytoczonym przykładzie platforma NLU wyodrębnia z transkrypcji wypowiedzi klienta (z frazy testowej) informację dotyczącą marki i koloru telefonu tj. czarny Samsung.

Tabela 6.3. Przykład definiowania encji we frazie uczącej i wyodrębniania wartości encji z frazy testowej

Fraza ucząca P_i		
Zamawiam telefon [Xiaomi](markaTelefonuEntity) koloru [czarnego](kolorTelefonuEntity)		
Proces uczenia	Intencja	*Zamówienie telefonu*
	Encja 1	*markaTelefonuEntity*
	Encja 2	*kolorTelefonuEntity*
Fraza testowa P_i		
Chcę zamówić telefon Samsung najlepiej koloru czarnego.		
Proces rozpoznawania	Intencja	*Zamówienie telefonu* - intencja rozpoznana na podstawie: *Chcę zamówić telefon najlepiej koloru*
	Encja 1	*markaTelefonuEntity=Samsung*
	Encja 2	*kolorTelefonuEntity=czarny*

Z uwagi na to, że wszystkie platformy NLU posiadają funkcjonalność encji jak i możliwość wyodrębniania z nich wartości uznano, że zastosowanie ich do przekazywania rozpoznanych emocji będzie stanowiło najbardziej adekwatne podejście.

Podejście, które zastosowano polega na tym, że w pierwszej kolejność następuje proces uczenia poprzez tworzenie fraz uczących przypisanych do każdej z intencji ze zdefiniowaną encją *emoEnt* na końcu frazy. Przy czym jako wartość encji wstawiane są wszystkie możliwe wartości: *emo_happy* dla Radość, *emo_sad* dla Smutek, *emo_angry* dla Złość, *emo_fear* dla Strach. Dodatkowo część fraz uczących jest tworzona bez *emoEnt* na końcu frazy co odpowiada

stanowi Neutralność. W tabeli 6.4 zaprezentowano przykładowe frazy uczące z uwzględnieniem obsługi emocji.

Tabela 6.4. Frazy uczące dla intencji „Wypowiedzenie umowy" z uwzględnieniem emocji

Emocja	Frazy uczące P_i z encją *emoEnt* określającą E_f
Radość	*Ja myślę że rezygnuję tak kompletnie [**emo_happy**](emoEnt)*
	*Chce zrezygnować z tego [**emo_happy**](emoEnt)*
Smutek	*To w takim razie rezygnuję [**emo_sad**](emoEnt)*
	*Czy mogłabym zrezygnować z tego abonamentu [**emo_sad**](emoEnt)*
Złość	*chcielibyśmy właśnie zrezygnować [**emo_angry**](emoEnt)*
	*mogę w takim układzie zrezygnować [**emo_angry**](emoEnt)*
Strach	*aha no to w takim razie to zrezygnuję z tej usługi [**emo_fear**](emoEnt)*
	*ja bym zrezygnował z tego [**emo_fear**](emoEnt)*
Neutralność	*jako składam pani dyspozycję hmm rezygnacji*
	to w takim razie to zrezygnuję z tej usługi

Warto zwrócić uwagę na fakt, że przy frazach uczących wcale emocje nie muszą w rzeczywistości wystąpić w wypowiedziach klientów tzn. mogą zostać przypisane do dowolnej frazy. Chodzi o to, aby przekazać do platformy NLU informację, że taka encja z taką wartością może wystąpić i żeby była uwzględniona w procesie uczenia.

W przypadku fraz testowych, jeżeli w wypowiedzi klienta wystąpi jedna z emocji to wówczas jest ona dodawana na końcu transkrypcji tej wypowiedzi przed dostarczeniem jej do platformy NLU co zaprezentowano w tabeli 6.5.

Tabela 6.5. Frazy testowe dla intencji „Wypowiedzenie umowy" z uwzględnieniem emocji

Emocja	Frazy testowe P_i z encją *emoEnt* określającą E_f
Radość	*chciałbym zrezygnować od razu z tego ubezpieczenia.* ***emo_happy***
	Ja chciałem zrezygnować z podpisanej umowy internetowej ***emo_happy***
Smutek	*chciałbym w tej chwili zrezygnować* ***emo_sad***
	bardzo proszę w tej chwili ee rozwiązać tą umowę ***emo_sad***
Złość	*Rezygnuję, tak* ***emo_angry***
	chyba będę chyba będę rezygnował ze wszystkich numerów ***emo_angry***
Strach	*Nie, myślę, że po prostu rezygnuję tak kompletnie.* ***emo_fear***
	Wie pani co ja zrezygnuję z tej usługi ***emo_fear***
Neutralność	*żeby żeby po prostu z tego rezygnować*
	Nie ja u was złożyłam wypowiedzenie i od was chce zabrać numer przenieść do innego operatora.

W metodzie rozpoznawania ukrytych intencji, encje przekazywane są więc jako określone rodzaje emocji. Kluczowe jest to, że zastosowanie mechanizmów encji pozwala przekazywać we frazach dowolne wartości i nie ma to wpływu na proces uczenia oraz testowania, gdyż w tych procesach jest uwzględniany tylko tekst bez zdefiniowanych lub wyodrębnionych encji.

6.2.2 Procedura uczenia bota

Zgodnie z zaproponowanym podejściem, przygotowanie bota dla danego CC będzie składać się z następujących kroków:

1. Opracowanie zbiorów S_{INT} i S_A: zbiory te są określane na podstawie analizy spraw będących przedmiotem rozmów w danej kampanii CC.
2. Uczenie platformy NLU:
   - Opracowanie zbiorów S_{PU} i S_{PT}*:* zbiory fraz powinny zostać opracowane na podstawie rzeczywistych rozmów agentów i klientów przeprowadzonych w danym CC. Frazy uczące S_{PU} powinny występować zarówno z dodanymi encjami określającymi emocje jak i bez nich.
   - Trenowanie platformy: wykorzystując zbiór S_{PU} należy nauczyć system rozpoznawania intencji ze zbioru S_{INT}.

- Weryfikacja poprawności rozpoznanych intencji.

3. Opracowanie reguł wnioskowania opartych o rozpoznane emocje (zbiór S_E): zbiór reguł będzie tworzył mechanizm wnioskowania wybierający odpowiednią akcję bota ze zbioru S_A.
4. Walidacja bota: testowanie i ocena działania bota za pomocą zbioru S_{PT}. W przypadku niezadowalających wyników powrót do punktu 2.

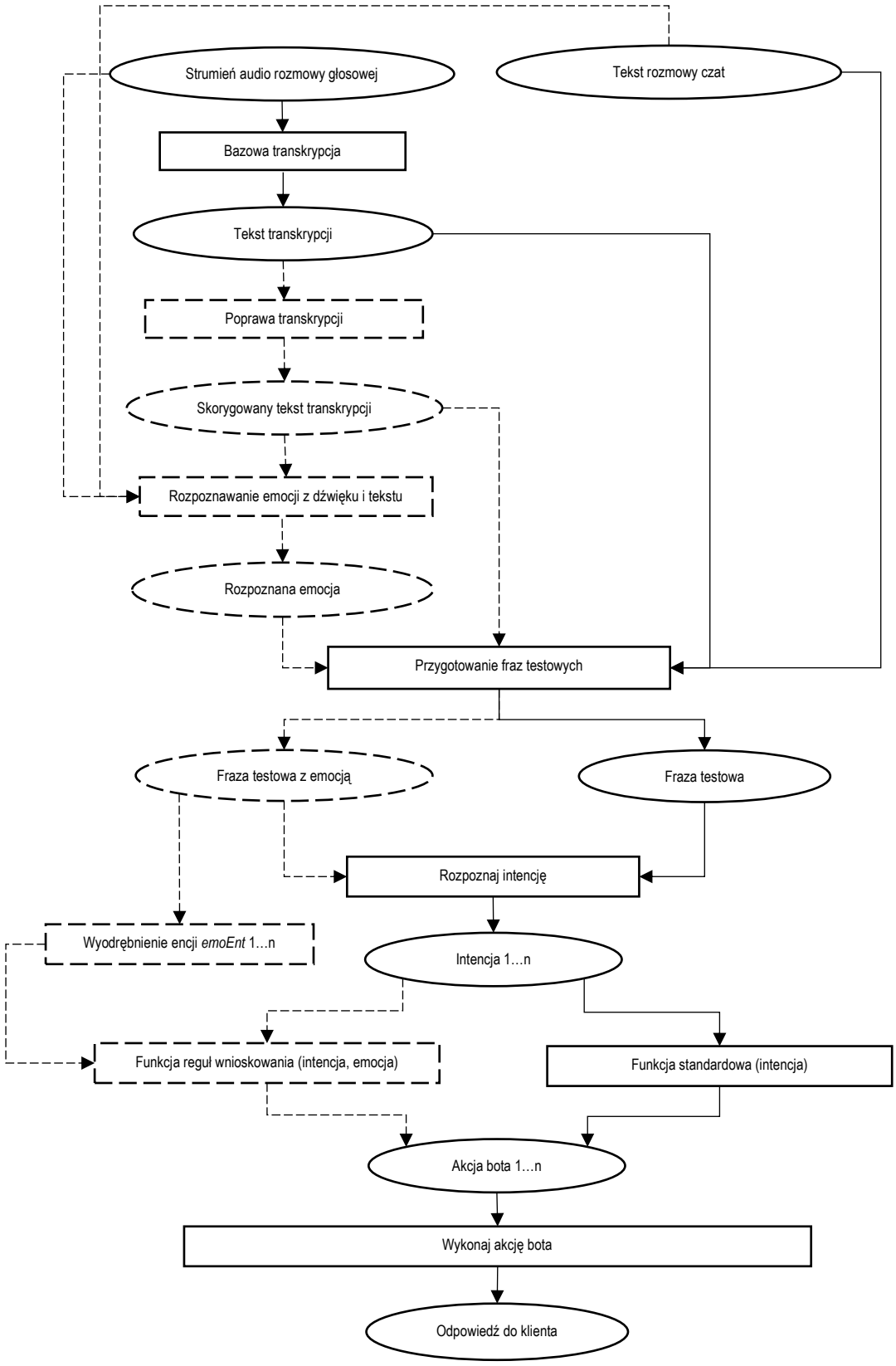
6.2.3 Proces działania bota rozpoznającego ukryte intencje z metodą poprawy jakości transkrypcji

W oparciu o nowe metody *IAT* i F_{AE} poprawiające skuteczność działania bota opracowano zmodyfikowany proces działania bota (rysunek 6.2). Linią ciągłą oznaczono dotychczasowe standardowe metody, natomiast linią przerywaną oznaczono nowe elementy związane z zastosowaniem metody rozpoznawaniem ukrytych intencji oraz metody poprawy jakości transkrypcji opisanej w podrozdziale 5.2.

Proces działania bota musi być wcześniej poprzedzony fazą uczenia opisaną w punkcie 6.2.2. W przypadku standardowego procesu frazy uczące zawierają tylko tekst wypowiedzi, natomiast w proponowanym rozwiązaniu frazy uczące są rozbudowane o encję *emoEnt* opisaną w punkcie 6.2.1.

Zmodyfikowany proces działania bota, rozbudowany o nowe metody poprawiające skuteczność działania bota, ma bardziej złożony przebieg (rysunek 6.2):

- Strumień audio rozmowy jest procesowany standardową metodą transkrypcji automatycznej.
- Tekst transkrypcji jest przetwarzany przez metodę poprawy jakości transkrypcji *IAT*.
- Następuje rozpoznawanie emocji:
  - Dla rozmowy głosowej na podstawie strumienia audio.
  - Dla rozmowy czat na podstawie tekstu.
- Frazy testowe są tworzone na podstawie skorygowanego tekstu transkrypcji lub na podstawie tekstu rozmowy czat. Dodatkowo do frazy testowej dodawana jest rozpoznana emocja w postaci encji *emoEnt*.
- Fraza testowa wraz z dodaną encją *emoEnt* jest przekazywana na wejście do platformy NLU.
- Platforma NLU w pierwszej kolejności rozpoznaje intencję funkcją F_{NLU} na podstawie tekstu frazy testowej P_i (nie uwzględnia encji *emoEnt*).
- Następuje przekierowanie do funkcji reguł wnioskowania F_{AE} (6.1), która na podstawie intencji INT_j oraz encji *emoEnt* określa docelową akcję bota A_k. Warto podkreślić, że w tym miejscu dla standardowego procesu jest wykonywana funkcja standardowa F_A (4.5) przyjmująca jako argument tylko intencję INT_j.
- Następuje wybór akcji bota A_k, która generuje odpowiedź klientowi.



Rysunek 6.2. Zmodyfikowany proces działania bota, rozbudowany o metody poprawiające skuteczność jego działania

6.3 Wyniki eksperymentów oceniających skuteczność metody rozpoznawania ukrytych intencji

Skuteczność zaproponowanej metody rozpoznawania ukrytych intencji klientów F_{AE} została wykazana poprzez wykonanie eksperymentów porównawczych, w których porównano wyniki wyboru akcji bota dla standardowej metody F_A (standardowa platforma NLU) z nową metodą F_{AE} (platforma NLU rozbudowana o nową metodę).

W oparciu o zgromadzone w bazie RLCHDB rozmowy czat wyselekcjonowano zbiór S_{INT} piętnastu najczęściej występujących intencji, które podzielono na trzy grupy (tabela 6.6). Z rozmów czat wybrano 375 fraz uczących (po 25 fraz dla każdej intencji) oraz 225 fraz testowych (po 15 fraz dla każdej intencji). Frazy uczące i testowe pochodziły z różnych zbiorów wypowiedzi klientów z rozmów czat.

Dla rozmów głosowych skorzystano z wcześniej przygotowanych fraz uczących i testowych dla piętnastu intencji z bazy RLPHDB opisanych w podrozdziale 5.4 (tabela 5.5).

Tabela 6.6. Grupy wyselekcjonowanych intencji dla kanału tekstowego

Grupa	**Wyselekcjonowane intencje INT_j dla kanału tekstowego**	
	ID	**Nazwa**
1	T1	Potwierdzenie
	T2	Zaprzeczenie
	T3	Autoryzacja
	T4	Wypowiedzenie umowy
	T5	Zadowolenie z obsługi
2	T1	Odzyskiwanie hasła
	T2	Aktywacja subskrypcji
	T3	Problemy z aplikacją
	T4	Aktywacja usług internetowych
	T5	Status sprawy
3	T1	Stan usługi
	T2	Połącz z agentem
	T3	Aktywacja SIM
	T4	Problem z połączeniem internetowym
	T5	Termin rozwiązania

Legenda: T1-T5 – oznaczenie intencji w danej grupie dla kanału tekstowego.

W proponowanej metodzie emocje występujące w wypowiedziach są wykorzystywane do rozpoznawania ukrytych intencji klienta. Na tej podstawie mechanizm wnioskowania dokonuje wyboru poprawnej akcji bota jaka ma być wykonana. Mechanizm ten korzysta z reguł wnioskowania warunkowanych emocjami, które tworzy specjalista CC na platformie NLU jako nierozłączna część budowanego scenariusza rozmowy. Zarówno scenariusz rozmowy jak i reguły wnioskowania tworzone są w graficznym lub tekstowym środowisku projektowym udostępnianym przez każdą platformę NLU. Kluczową zasadą jest to, aby dla danej intencji wyodrębnić przypadki, w których znaczenie fraz testowych zostaje zmienione poprzez pojawiające się w nich emocje. Jeżeli dana emocja zmienia znaczenie intencji rozpoznanej z frazy testowej na tyle, że powinna być wybrana inna akcja bota, to należy zdefiniować regułę uwzględniającą tę emocję, która wskaże inną akcję bota, niż pierwotna akcja bota (dla stanu Neutralność). W tabeli 6.7 pokazano zdefiniowane reguły wnioskowania związane z intencjami wyszczególnionymi w podrozdziale 5.4.

Zbiór przesłanek jest określony przez zbiór możliwych emocji $S_{E.}$ Zbiór wniosków jest określony przez zbiór akcji bota S_A dla każdej grupy w ramach kanału głosowego i tekstowego. Jako wartości zbiorów używane są identyfikatory akcji bota w celu skrócenia zapisu:

$$\begin{aligned} S_{AVG1} &= \{VG1A1, \dots, VG1A11\} & S_{ATG1} &= \{TG1A1, \dots, TG1A11\} \\ S_{AVG2} &= \{VG2A1, \dots, VG2A7\} & S_{ATG2} &= \{TG2A1, \dots, TG2A7\} \\ S_{AVG3} &= \{VG3A1, \dots, VG3A7\} & S_{ATG3} &= \{TG3A1, \dots, TG3A6\} \end{aligned} \quad (6.11)$$

Jak można zauważyć w niektórych przypadkach emocje nie mają znaczenia (np. V3). Oznacza to, że niezależnie od emocji w wypowiedzi klienta nie ma ukrytych intencji. W innych przypadkach pojawienie się wybranych emocji może oznaczać, że faktyczna intencja zależy od emocji, a chęć zmiany decyzji może być uznana za intencję ukrytą (np. V4). Interesującym przypadkiem jest intencja V5. W tym przypadku emocje mogą oznaczać zupełnie inną ukrytą intencję niż wykryta na podstawie analizy samego tekstu wypowiedzi. Nieuwzględnienie takich przypadków prowadzi do całkowicie błędnego działania bota.

Tabela 6.7. Akcje bota wynikające z reguł wnioskowania

Grupa Kanał	**ID**	**Intencja INT_j**	**Reguła wnioskowania dla intencji ⇒ ID akcji bota**	**Nazwa akcji bota A_k**
Grupa 1 Głos	V1	Potwierdzenie	`emo_sad ∨ emo_fear ⇒ VG1A6`	Potwierdza, ale upewnij się
			`~(emo_sad ∨ emo_fear) ⇒ VG1A5`	Przyjmij potwierdzenie
	V2	Zaprzeczenie	`emo_angry ∨ emo_sad ∨ emo_fear ⇒ VG1A11`	Zaprzecza, ale upewnij się
			`~(emo_angry ∨ emo_sad ∨ emo_fear) ⇒ VG1A10`	Przyjmij odmowę
	V3	Autoryzacja	`VG1A1`	Zautoryzuj
	V4	Wypowiedzenie umowy	`emo_sad ∨ emo_fear ⇒ VG1A9`	Upewnij się czy jest zamiar rezygnacji
			`emo_angry ⇒ VG1A8`	Rezygnuje, zaproponuj lepszą ofertę
			`~(emo_angry ∨ emo_sad ∨ emo_fear) ⇒ VG1A7`	Rezygnuje
	V5	Zadowolenie z obsługi	`emo_angry ∨ emo_sad ∨ emo_fear ⇒ VG1A4`	Klient niezadowolony, zapytaj, dlaczego
			`emo_happy ⇒ VG1A2`	Klient zadowolony, podziękuj i wyślij ankietę
			`~(emo_angry ∨ emo_sad ∨ emo_fear ∨ emo_happy) ⇒ VG1A3`	Klient zadowolony, podziękuj
Grupa 2 Głos	V1	Status usługi	`VG2A3`	Podaj status usługi
	V2	Informacje o umowie	`emo_angry ∨ emo_fear ⇒ VG2A6`	Podaj szczegółowe informacje o umowie
			`~(emo_angry ∨ emo_fear) ⇒ VG2A5`	Podaj informacje o umowie
	V3	Aktywacja usług internetowych	`⇒ VG2A4`	Aktywuj usługę Internetu
	V4	Naprawa urządzenia	`emo_angry ∨ emo_fear ⇒ VG2A2`	Zweryfikuj zasadność zgłoszenia i przyjmij urządzenie do naprawy
			`~(emo_angry ∨ emo_fear) ⇒ VG2A1`	Przyjmij urządzenie do naprawy
	V5	Przedłużenie umowy	`VG2A7`	Przedłuż umowę
Grupa 3 Głos	V1	Informacje o fakturze	`VG3A6`	Podaj informacje o fakturze
	V2	Problem z połączeniem internetowym	`VG3A7`	Przyjmij zgłoszenie problemu z Internetem
	V3	Połącz z agentem	`emo_angry ⇒ VG3A3`	Połącz z agentem natychmiast
			`emo_sad ∨ emo_fear ⇒ VG3A2`	Zbierze informację od klienta i następnie połącz z agentem
			`~(emo_angry ∨ emo_sad ∨ emo_fear) ⇒ VG3A4`	Umieść klienta w kolejce oczekiwania na agenta
	V4	Opis uszkodzeń	`VG3A5`	Zarejestruj opis uszkodzeń

	V5	Odzyskiwanie hasła	VG3A1	Resetuj hasło klienta
Grupa 1 Tekst	T1	Potwierdzenie	emo_sad ∨ emo_fear ⇒ TG1A6	Potwierdza, ale upewnij się
			~(emo_sad ∨ emo_fear) ⇒ TG1A5	Przyjmij potwierdzenie
	T2	Zaprzeczenie	emo_angry ∨ emo_sad ∨ emo_fear ⇒ TG1A11	Zaprzecza, ale upewnij się
			~(emo_angry ∨ emo_sad ∨ emo_fear) ⇒ TG1A10	Przyjmij odmowę
	T 3	Autoryzacja	TG1A1	Zautoryzuj
	T4	Wypowiedzenie umowy	emo_sad ∨ emo_fear ⇒ TG1A9	Upewnij się czy jest zamiar rezygnacji
			emo_angry ⇒ TG1A8	Rezygnuje, zaproponuj lepszą ofertę
			~(emo_sad ∨ emo_fear ∨ emo_angry) ⇒ TG1A7	Rezygnuje
	T5	Zadowolenie z obsługi	emo_angry ∨ emo_sad ∨ emo_fear ⇒ TG1A4	Klient niezadowolony, zapytaj, dlaczego
			emo_happy ⇒ TG1A2	Klient zadowolony, podziękuj i wyślij ankietę
			~(emo_angry ∨ emo_sad ∨ emo_fear ∨ emo_happy) ⇒ TG1A3	Klient zadowolony, podziękuj
Grupa 2 Tekst	T1	Odzyskiwanie hasła	emo_sad ∨ emo_angry ⇒ TG2A5	Reset hasła przez agenta
			~(emo_sad ∨ emo_angry) ⇒ TG2A6	Reset hasła przez email
	T2	Aktywacja subskrypcji	TG2A7	Włącz subskrypcję
	T3	Problemy z aplikacją	emo_sad ∨ emo_angry ⇒ TG2A2	Reset aplikacji przez agenta
			~(emo_sad ∨ emo_angry) ⇒ TG2A1	Samodzielna reinstalacja aplikacji
	T4	Aktywacja usług internetowych	TG2A4	Włącz usługę Internetu
	T5	Status sprawy	TG2A3	Podaj status sprawy
Grupa 3 Tekst	T1	Status usługi	TG3A7	Podaj status usługi
	T2	Połącz z agentem	emo_angry ⇒ TG3A2	Połącz z agentem natychmiast
			emo_sad ∨ emo_fear ⇒ TG3A1	Zbierz informację od klienta i następnie połącz z agentem
			~(emo_angry ∨ emo_sad ∨ emo_fear) ⇒ TG3A3	Umieść klienta w kolejce oczekiwania na agenta
	T3	Aktywacja SIM	TG3A4	Aktywuj kartę SIM
	T4	Problem z połączeniem internetowym	TG3A5	Przyjmij zgłoszenie problemu z Internetem
	T5	Termin rozwiązania	TG3A6	Podaj, ile potrwa rozwiązanie sprawy

Legend: V1-V5 - intencje w danej grupie w kanale głosowym; T1-T5 - intencje w danej grupie w kanale tekstowym; VG - poszczególne akcje wykonywane przez bota w kanale głosowym; TG - poszczególne akcje wykonywane przez bota w kanale tekstowym

Wyniki eksperymentów obrazujących wpływ rozpoznanych emocji klienta na przebieg rozmowy na infolinii CC przedstawiono w tabeli 6.8 i tabeli 6.9. Aby w eksperymentach uwzględnić tylko wpływ proponowanego przez autora mechanizmu wnioskowania, dla kanałów głosowych wykorzystany został standardowy system transkrypcji MST. W ten sposób różnice w uzyskanych wynikach są jedynie efektem uwzględnienia emocji.

Analizując wyniki zawarte w tabeli 6.8 i tabeli 6.9 obserwuje się wyraźny wpływ emocji klienta na poprawność prowadzenia rozmowy przez wirtualnego konsultanta. Teoretycznie, eksperymenty z wykorzystaniem jednakowych fraz testowych dla przypadków, gdy emocje były dołączone oraz gdy emocje nie występowały powinny dać ten sam wynik w zakresie rozpoznawania intencji. W rzeczywistości jednak, wyniki rozpoznawania intencji czasem nieznacznie się różnią. Wynika to z faktu, że przy każdym inaczej sparametryzowanym (bez encji *emoEnt*, z encją *emoEnt*) teście wykorzystywane są inne modele NLU. Platformy NLU nie są deterministyczne i dlatego nie gwarantują uzyskania tych samych wyników dla różnych modeli NLU nawet wytrenowanych tymi samymi frazami uczącymi. W eksperymentach pierwszy model NLU powstawał na podstawie fraz uczących bez encji *emoEnt*, a drugi model

NLU na podstawie fraz uczących z encjami *emoEnt*. Należy jednak zwrócić uwagę, że różnice w zakresie rozpoznawania intencji są na tyle małe, że można je pominąć w dalszej ocenie wyników dotyczących analizy akcji bota.

Tabela 6.8. Wpływ emocji na poprawność wyboru akcji bota dla kanału głosowego

Grupa	**Platforma NLU**	**Rozpozn. emocji**	**Rozpoznawanie intencji**		**Wybór akcji bota**		
			F*$_{NLU}$ bez *emoEnt	***F*$_{NLU}$ z *emoEnt***	***F*$_{A}$ bez *emoEnt***	***F*$_{AE}$ z *emoEnt***	**Poprawa**
		accE	***accINT***	***accINT***	***accA***	***accA***	**[%]**
1	Dialogflow	0,80 (60/75)	0,68 (51/75)	0,68 (51/75)	0,52 (39/75)	0,57 (43/75)	**9,62**
	Rasa		0,81 (61/75)	0,81 (61/75)	0,61[(a)] (46/75)	0,69[(a)] (52/75)	**13,11**
2	Dialogflow	0,79 (59/75)	0,87 (65/75)	0,84 (63/75)	0,79 (59/75)	0,81 (61/75)	**2,53**
	Rasa		0,92 (69/75)	0,95 (71/75)	0,81[(b)] (61/75)	0,91[(b)] (68/75)	**12,35**
3	Dialogflow	0,65 (49/75)	0,80 (60/75)	0,81 (61/75)	0,71[(c)] (53/75)	0,76[(c)] (57/75)	**7,04**
	Rasa		0,95 (71/75)	0,95 (71/75)	0,83 (62/75)	0,88 (66/75)	**6,02**

(a), (b), (c) – dane zaprezentowane na rysunku 6.3

Tabela 6.9. Wpływ emocji na poprawność wyboru akcji bota dla kanału tekstowego

Grupa	**Platforma NLU**	**Rozpozn. emocji**	**Rozpoznawanie intencji**		**Wybór akcji bota**		
			F*$_{NLU}$ bez *emoEnt	***F*$_{NLU}$ z *emoEnt***	***F*$_{A}$ bez *emoEnt***	***F*$_{AE}$ z *emoEnt***	**Poprawa**
		accE	***accINT***	***accINT***	***accA***	***accA***	**[%]**
1	Dialogflow	0,84 (63/75)	0,84 (63/75)	0,84 (63/75)	0,60[(a)] (45/75)	0,77[(a)] (58/75)	**28,33**
	Rasa		0,93 (70/75)	0,93 (70/75)	0,71 (53/75)	0,88 (66/75)	**23,94**
2	Dialogflow	0,92 (69/75)	0,76 (57/75)	0,77 (58/75)	0,69 (52/75)	0,77 (58/75)	**11,59**
	Rasa		0,91 (68/75)	0,91 (68/75)	0,84[(b)] (63/75)	0,91[(b)] (68/75)	**8,33**
3	Dialogflow	0,89 (67/75)	0,71 (53/75)	0,75 (56/75)	0,71[(c)] (53/75)	0,75[(c)] (56/75)	**5,63**
	Rasa		0,76 (57/75)	0,75 (56/75)	0,72 (54/75)	0,75 (56/75)	**4,17**

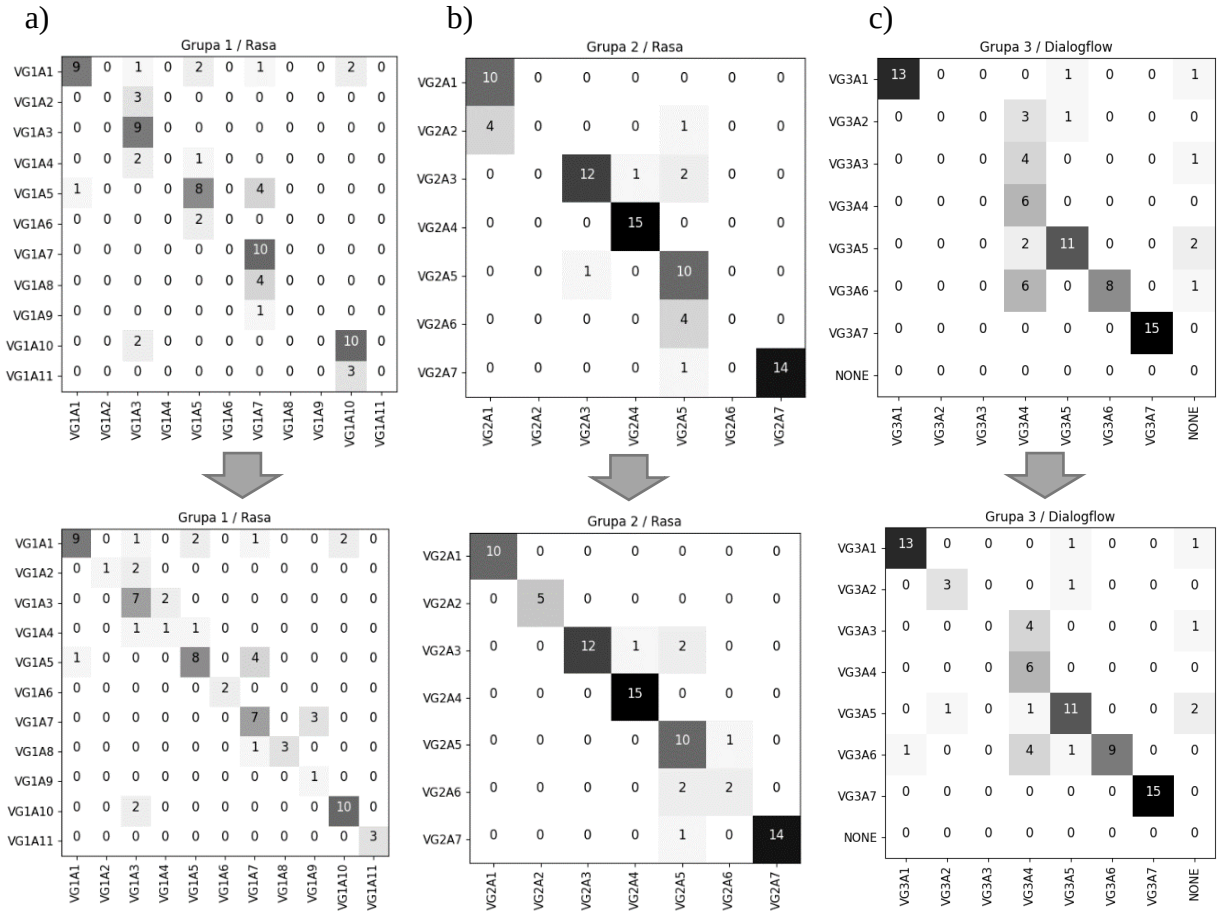
(a), (b), (c) – dane zaprezentowane na rysunku 6.3

W zależności od badanej platformy oraz wyselekcjonowanego zbioru intencji poprawność wyboru akcji bota dla platformy Dialogflow wzrasta odpowiednio: w kanale głosowym od 2,53% do 9,62%, a w kanale tekstowym od 5,63% do 28,33%. Natomiast dla platformy Rasa w kanale głosowym od 6,02% do 13,11%, a w kanale tekstowym od 4,17% do 23,94%. Otrzymane wyniki jednoznacznie potwierdzają istotne znaczenie metody F_{AE} w zwiększeniu skuteczności działania bota.

W punkcie 6.3.1 i 6.3.2 zaprezentowano szczegółowe wyniki dla przykładowych fraz testowych dla jednej wybranej intencji dla kanału głosowego i jednej dla kanału testowego. Skorzystano w nich z reguł wnioskowania zawartych w tabeli 6.7.

6.3.1 Szczegółowe wyniki rozpoznawania ukrytych intencji dla kanału głosowego

Dla danych z tabeli 6.8, które zacieniowano oraz oznaczono symbolami (a-c), na rysunku 6.3 zestawiono odpowiednie macierze konfuzji. Wiersz pierwszy we wszystkich kolumnach prezentuje dane uzyskane przy badaniu rozwiązania z zastosowaniem standardowej metody F_A, natomiast wiersz drugi zawiera rozkłady dla metody rozpoznawania ukrytych intencji F_{AE} uwzględniającej reguły wnioskowania na podstawie rozpoznanych emocji.



Rysunek 6.3. Macierze konfuzji dla wybranych danych głosowych z tabeli 6.8

Legenda: **VG1-3A1–11** – Akcje bota w poszczególnych grupach (1, 2 lub 3) dla kanału głosowego; **NONE** – niewybrana żadna akcja bota; Osie rzędnych (y) opisują wartości **PRAWDA**; Osie odciętych (x) opisują wartości **PREDYKCJA**

W tabeli 6.10 zaprezentowano dla „Grupa 1 / Rasa" szczegółowe dane w rozbiciu na poszczególne intencje z zastosowaniem metody F_A oraz F_{AE}. W wierszu zawierającym intencję „V1 Potwierdzenie" można dostrzec, że prawidłowość rozpoznawania akcji bota poprawiła się w sposób znaczący przy zastosowaniu metody F_{AE}. Dlatego w tabeli 6.11 zaprezentowano frazy testowe dla intencji „V1 Potwierdzenie" wraz z wynikami akcji bota dla obu metod rozpoznawania intencji.

Tabela 6.10. Wyniki accINT, accA w podziale na intencje dla zbioru „Grupa 1 / Rasa”

Zbiór fraz testowych dla intencji	**Akcje bota A_k z S_{AVG1}**	**Rozpozn. emocji**	**Rozpoznawanie intencji**		**Wybór akcji bota**		
			F_{NLU} bez *emoEnt*	**F_{NLU} z *emoEnt***	**F_A bez *emoEnt***	**F_{AE} z *emoEnt***	**Poprawa**
		accE	***accINT***	***accINT***	***accA***	***accA***	[%]
V1 Potwierdzenie	VG1A5 VG1A6	0,80 (12/15)	0,67 (10/15)	0,67 (10/15)	0,53 (8/15)	0,67 (10/15)	**26,42**
V2 Zaprzeczenie	VG1A10 VG1A11	1,00 (15/15)	0,87 (13/15)	0,87 (13/15)	0,67 (10/15)	0,87 (13/15)	**29,85**
V3 Autoryzacja	VG1A1	0,80 (12/15)	0,60 (9/15)	0,60 (9/15)	0,60 (9/15)	0,60 (9/15)	**0,00**
V4 Wypowiedzenie umowy	VG1A7 VG1A8 VG1A9	0,73 (11/15)	1,00 (15/15)	1,00 (15/15)	0,67 (10/15)	0,73 (11/15)	**8,96**
V5 Zadowolenie z obsługi	VG1A2 VG1A3 VG1A4	0,67 (10/15)	0,93 (14/15)	0,93 (14/15)	0,60 (9/15)	0,60 (9/15)	**0,00**
Cały zbiór		**0,80 (60/75)**	**0,81 (61/75)**	**0,81 (61/75)**	**0,61 (46/75)**	**0,69 (52/75)**	**13,11**

Reguły wnioskowania dla intencji „V1 Potwierdzenie” (tabela 6.7), należy interpretować jako instrukcję otrzymaną od menedżera biura obsługi klienta. Oczekuje on, że klientom, którzy w wypowiedzi zawierającej „V1 Potwierdzenie” wyrażą emocję Smutek lub Strach zostanie zadane pytanie w celu upewnienia się odpowiedzi klienta, a w przypadku innych emocji „V1 Potwierdzenie” zostanie przyjęte.

W przypadku metody F_A wszędzie, gdzie rzeczywiście w wypowiedzi wystąpi Smutek lub Strach, będzie wybrana nieprawidłowa akcja bota, bo dla tej metody do frazy testowej nie jest dodana encja *entEmo*. Natomiast w przypadku metody F_{AE}, jeżeli zostanie wykryta prawidłowa emocja to zostanie udzielona prawidłowa odpowiedź. Oczywiście w przypadku tych konkretnych reguł wnioskowania na wynik nie wpłynie błędna zamiana Smutku ze Strachem, gdyż dla nich odpowiedź jest taka sama (VG1A6). To samo dotyczy pomyłki w ramach zbioru w zakresie emocji Złość, Radość, Neutralność, odpowiedź dla nich również się nie różni (VG1A5).

W tabeli 6.11 przedstawiono wyniki dla fraz testowych dla intencji „V1 Potwierdzenie”. Wynikiem prawidłowym dla tej intencji w zależności od rozpoznanych emocji może być akcja bota VG1A5 lub VG1A6.

W tabeli 6.11 w wierszach 1 i 3 widać, że obydwie wypowiedzi zawierały Smutek. Ze względu, że metoda rozpoznawania emocji w tym przypadku prawidłowo wykryła emocję Smutek to metoda F_{AE} wybrała prawidłową akcję bota VG1A6. Te dwa wiersze zaważyły na wyniku całej intencji V1, dla której metoda F_{AE} jest o dwa przypadki wyboru akcji bota lepsza niż standardowa metoda F_A.

Dla fraz testowych przypisanych do intencji „V1 Potwierdzenie” można zauważyć, że dokładność *accE* rozpoznawania emocji nie jest bezbłędna, gdyż wyniosła 0,80. W wierszu 4 widać, że źle została wykryta emocja Złość zamiast Neutralność, ale z uwagi, że i tak intencja została źle wykryta to wynik akcji bota był nieprawidłowy, zarówno dla metody F_A jak i F_{AE}. Gdyby jednak intencja została dobrze rozpoznana, a nie udało się wykryć prawidłowo emocji Smutek lub Strach zaważyłoby to negatywnie na wynikach tego eksperymentu.

Tabela 6.11. Wyniki dla fraz testowych dla intencji „V1 Potwierdzenie”

Lp.	Fraza testowa P_i	Prawdziwa emocja E_f	Rozpoznawanie/wybór bez *emoEnt*			Rozpoznawanie/wybór z *emoEnt*		
			Brak rozp. emocji	Intencji F_{NLU}	Akcji bota F_A	Emocji E_f	Intencji F_{NLU}	Akcji bota F_{AE}
1	*No właśnie.*	Smutek	Neutralność	Tak	Nie VG1A5	Smutek	Tak	Tak VG1A6
2	*Ten sam tag.*	Neutralność	Neutralność	Nie	Nie INNA	Neutralność	Nie	Nie INNA
3	*No dobrze, akceptuję no potwierdzam.*	Smutek	Neutralność	Tak	Nie VG1A5	Smutek	Tak	Tak VG1A6
4	*E tak to będzie pierwszy naprawa.*	Neutralność	Neutralność	Nie	Nie INNA	Złość	Nie	Nie INNA
5	*Tak oczywiście będzie ten sam z którego dzwonię*	Neutralność	Neutralność	Nie	Nie INNA	Neutralność	Nie	Nie INNA
6	*Tak tak, tak właśnie*	Neutralność	Neutralność	Tak	Tak VG1A5	Neutralność	Tak	Tak VG1A5
7	*Taak*	Neutralność	Neutralność	Nie	Nie INNA	Neutralność	Nie	Nie INNA
8	*dobrze poproszę tak oczywiście tak*	Neutralność	Neutralność	Tak	Tak VG1A5	Neutralność	Tak	Tak VG1A5
9	*No zwykle do tej pory akceptowałem, czyli teraz też akceptuję.*	Neutralność	Neutralność	Nie	Nie INNA	Neutralność	Nie	Nie INNA
10	*Tak, ale jeszcze się zastanowię*	Złość	Neutralność	Tak	Tak VG1A5	Złość	Tak	Tak VG1A5
11	*Hmm dobra dobra dobra dobra super.*	Neutralność	Neutralność	Tak	Tak VG1A5	Neutralność	Tak	Tak VG1A5
12	*No rozumiem, rozumiem, dobrze, super*	Neutralność	Neutralność	Tak	Tak VG1A5	Neutralność	Tak	Tak VG1A5
13	*No a co mam zrobić, no akceptuję.*	Złość	Neutralność	Tak	Tak VG1A5	Radość	Tak	Tak VG1A5
14	*Tak, tak, tak, tak oczywiście będą będą spłacane.*	Neutralność	Neutralność	Tak	Tak VG1A5	Neutralność	Tak	Tak VG1A5
15	*Aha no to dobra, no to dobra*	Neutralność	Neutralność	Tak	Tak VG1A5	Radość	Tak	Tak VG1A5
Podsumowanie			Błędy istotne dla akcji bota: 2	10/15	8/15	Błędy istotne dla akcji bota: 0	10/15	10/15

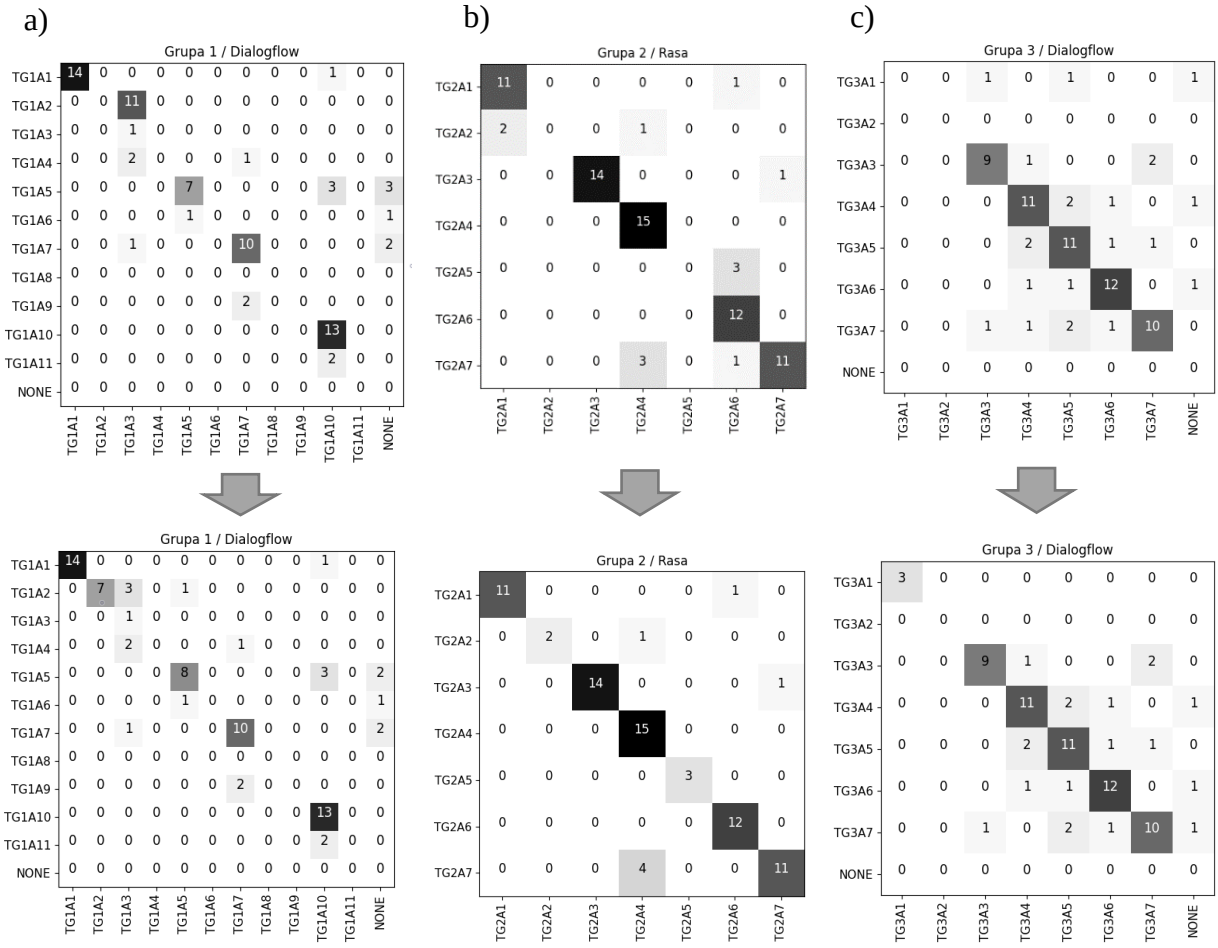
W wierszu 10 została prawidłowo rozpoznana emocja Złość, natomiast nie jest ona warunkowana w regułach wnioskowania, więc nie wpłynęło to na ostateczny wynik. Tym samym obie metody F_A i F_{AE} w tym przypadku rozpoznały prawidłowo akcję bota. W wierszu 13 została wykryta emocja Radość zamiast Złość, natomiast nie wpłynęło to na wynik, gdyż zarówno Złość i Radość nie są ujęte w regułach wnioskowania i prawidłowe rozpoznawanie intencji wystarczyło do prawidłowego rozpoznania akcji bota.

W tabeli 6.10 wybór akcji bota dla „V1 Potwierdzenie”, „V2 Zaprzeczenie” i „V4 Wypowiedzenie umowy” zwiększyło swoją dokładność *accA* dla metody F_{AE} odpowiednio o 26,42%, 29,85% i 8,96%. Natomiast dla „V3 Autoryzacja” i „V5 Zadowolenie z obsługi” pozostało na tym samym poziomie jak dla metody F_A. W działaniu operacyjnym bota dla metody F_{AE} zdarzały się przypadki, gdy nie następowało polepszenie rozpoznania akcji bota z uwagi na małe znaczenie emocji dla danej intencji lub brak emocji we frazach testowych. Zjawisko to tylko potwierdza, że metoda rozpoznawania ukrytych intencji F_{AE} w sytuacjach,

gdy emocje nie mają znaczenia nie pogarsza wyników. W sumarycznym bilansie dla testowanych zbiorów rozmów głosowych zawsze metoda F_{AE} uzyskiwała znacząco lepsze wyniki od standardowej metody F_A, dlatego należy przyjąć, że jest to rozwiązanie, które poprawia skuteczność działania voicebotów i potwierdza słuszność postawionej tezy w pracy.

6.3.2 Szczegółowe wyniki rozpoznawania ukrytych intencji dla kanału tekstowego

Na rysunku 6.4 zaprezentowano dane dotyczące kanału tekstowego w postaci macierzy konfuzji. Zróżnicowanie liczby możliwych akcji wykonywanych przez bota w zależności od nacechowania emocjonalnego poszczególnych zbiorów jest naturalne. Dlatego też, dla grupy 1 określono jedenaście różnych możliwości odpowiedzi akcji bota (rysunek 6.4a), a dla grupy 2 i 3 jest ich siedem (rysunek 6.4b, rysunek 6.4c).



Rysunek 6.4. Macierze konfuzji dla wybranych danych tekstowych z tabeli 6.9

Legenda: **TG1-3A1–11** – Akcje bota w poszczególnych grupach (1, 2 lub 3) dla kanału tekstowego; **NONE** – niewybrana żadna akcja bota; Osie rzędnych (y) opisują wartości **PRAWDA**; Osie odciętych (x) opisują wartości **PREDYKCJA**

Dla grupy 1, stworzony scenariusz rozmowy jest w dużym stopniu uwarunkowany rozpoznanymi emocjami, a dzięki dużej dokładności *accE* 0,84 poprawnie rozpoznanych emocji wyniki w zakresie akcji bota również uległy znacznej poprawie. Dla grupy 2 emocje rozpoznane zostały poprawnie z dokładnością *accE* 0,92, dzięki czemu wyniki w zakresie akcji bota również ulegają znacznej poprawie. Najmniejszą poprawę w zakresie akcji bota zaobserwowano dla grupy 3.

W tabeli 6.12 dla zbioru „Grupa 2 / Rasa” zaprezentowano szczegółowe dane w rozbiciu na poszczególne intencje z zastosowaniem standardowej metody F_A oraz metody rozpoznawania ukrytych intencji F_{AE}. W wierszach zawierających intencje „T1 Odzyskiwanie hasła” oraz „T3 Problemy z aplikacją” można dostrzec, że prawidłowość wyboru akcji bota poprawiła się w sposób wyraźny przy zastosowaniu metody F_{AE}. Dlatego w tabeli 6.13 zaprezentowano frazy testowe dla intencji „T1 Odzyskiwanie hasła” wraz z wynikami akcji bota dla obu metod F_A oraz F_{AE}.

Tabela 6.12. Wyniki accE, accINT, accA w podziale na intencje dla zbioru „Grupa 2 / Rasa”

Zbiór fraz testowych dla Intencji	**Akcje bota A_k z S_{ATG1}**	**Rozpozn. emocji**	**Rozpoznawanie intencji**		**Wybór akcji bota**		
			F_{NLU} bez *emoEnt*	**F_{NLU} z *emoEnt***	**F_A bez *emoEnt***	**F_{AE} z *emoEnt***	**Poprawa**
		accE	***accINT***	***accINT***	***accA***	***accA***	[%]
T1 Odzyskiwanie hasła	TG2A5 TG2A6	1,00 (15/15)	1,00 (15/15)	1,00 (15/15)	0,80 (12/15)	1,00 (15/15)	**25**
T2 Aktywacja subskrypcji	TG2A7	0,80 (12/15)	0,73 (11/15)	0,73 (11/15)	0,73 (11/15)	0,73 (11/15)	**0,00**
T3 Problemy z aplikacją	TG2A1 TG2A2	1,00 (15/15)	0,87 (13/15)	0,87 (13/15)	0,73 (11/15)	0,87 (13/15)	**19,18**
T4 Aktywacja usług internetowych	TG2A4	0,87 (13/15)	1,00 (15/15)	1,00 (15/15)	1,00 (15/15)	1,00 (15/15)	**0,00**
T5 Status sprawy	TG2A3	0,93 (14/15)	0,93 (14/15)	0,93 (14/15)	0,93 (14/15)	0,93 (14/15)	**0,00**
Cały zbiór		**92% (69/75)**	**90.67% (68/75)**	**90.67% (68/75)**	**84% (63/75)**	**90.67% (68/75)**	**7,94**

Dla intencji „T1 Odzyskiwanie hasła” (tabela 6.7) w przypadku emocji Złość, Smutek celem jest wykrycie klientów, którym należy niezwłocznie zresetować hasło z pomocą agenta (co skutkuje akcją bota „TG2A2 Reset hasła przez agenta”), aby nie pogłębiać ich negatywnych emocji, gdyż mogłyby one wpłynąć na negatywny wizerunek firmy w ich odbiorze. Natomiast w przypadku innych emocji klienta dla tej intencji następuje wykonanie akcji „TG2A1 Reset hasła przez email” co powoduje z kolei oszczędność czasu agenta i nie wpłynie negatywnie na odbiór firmy przez klienta, a jednocześnie zmusi go do większej samodzielności. W przypadku metody F_A wszędzie, gdzie rzeczywiście w wypowiedzi wystąpi Złość, Smutek, akcja bota będzie wybrana nieprawidłowo (TG2A1 zamiast TG2A2). Natomiast w przypadku metody F_{AE}, jeżeli zostanie wykryta prawidłowa emocja i intencja to wybrana akcja bota będzie prawidłowa: TG2A2 dla emocji Złość oraz Smutek, a dla pozostałych emocji TG2A1.

W tabeli 6.13 przedstawiono wyniki dla fraz testowych dla intencji „T1 Odzyskiwanie hasła”. Wynikiem prawidłowym dla tej intencji w zależności od rozpoznanych emocji może być akcja bota TG2A1, TG2A2. W tabeli 6.13 w wierszach 1, 8, 11 w rzeczywistych wypowiedziach tekstowych wystąpiły słowa i/lub emotikony, które wskazują na emocję Smutek. Wykrył to prawidłowo mechanizm rozpoznawania emocji, dlatego też metoda F_{AE} rozpoznała prawidłową akcję bota TG2A2. Standardowa metoda F_A ze względu, że nie ma wbudowanego mechanizmu warunkowania emocjami w każdym z tych przypadków wybrała nieprawidłową akcję bota TG2A1 (w regułach wnioskowania jest ona przypisana do stanu Neutralność).

Tabela 6.13. Wyniki dla fraz testowych dla intencji „T1 Odzyskiwanie hasła”

Lp.	Fraza P_i	Prawdz. emocja E_f	Rozpoznawanie/wybór bez *emoEnt*			Rozpoznawanie/wybór z *emoEnt*		
			Brak rozp. emocji	Intencji F_{NLU}	Akcji bota F_A	Emocji E_f	Intencji F_{NLU}	Akcji bota F_{AE}
1	*Dzień dobry, nie mogę zresetować hasła :(*	Smutek	Neutral.	Tak	Nie TG2A1	Smutek	Tak	Tak TG2A2
2	*Pisałam z mojej komórki ale nie zalogowana. Nie pamiętam hasła a próba zmiany nie powiodła się. Nie otrzymałam linka na maile*	Neutral.	Neutral.	Tak	Tak TG2A1	Neutral.	Tak	Tak TG2A1
3	*Dzień dobry Czy chciałem zapytać czy jest możliwość sprawdzenia emaila na jaki zostało stworzone konto. Nie mogę się zalogować przy podaniu swojego emaila a także przy resecie hasła nie otrzymuje maila*	Neutral.	Neutral.	Tak	Tak TG2A1	Neutral.	Tak	Tak TG2A1
4	*Nie może zmienić mi Pan hasła, lub wysłać sms-em?*	Neutral.	Neutral.	Tak	Tak TG2A1	Neutral.	Tak	Tak TG2A1
5	*Zapomniałem hasła i chciałem zmienić hasło*	Neutral.	Neutral.	Tak	Tak TG2A1	Neutral.	Tak	Tak TG2A1
6	*I Wyślij hasło, ale daje błąd.*	Neutral.	Neutral.	Tak	Tak TG2A1	Neutral.	Tak	Tak TG2A1
7	*jak zmienić hasło i login do logowania skoro nr tel nie działa i mail jest nie aktualny?*	Neutral.	Neutral.	Tak	Tak TG2A1	Neutral.	Tak	Tak TG2A1
8	*Dzień dobry, nie mogę zresetować hasła. Nie dostaję e-maila z linkiem :(*	Smutek	Neutral.	Tak	Nie TG2A1	Smutek	Tak	Tak TG2A2
9	*I gdyby się jeszcze dało to e-mail z resetem hasła na ten nowy bym poprosił*	Neutral.	Neutral.	Tak	Tak TG2A1	Neutral.	Tak	Tak TG2A1
10	*a co ma telefon dorzeczy w pisuję hasło na stronie do resetu hasła z linku który przychodzi e-meiem i mam informację o błędzie*	Neutral.	Neutral.	Tak	Tak TG2A1	Neutral.	Tak	Tak TG2A1
11	*Odzyskiwanie hasła nie działa. Nie dostaje wiadomości mail :(*	Smutek	Neutral.	Tak	Nie TG2A1	Smutek	Tak	Tak TG2A2
12	*Rozumiem, nie mogę zrestrtować hasła do konta może mi Pan pomóć z tym?*	Neutral.	Neutral.	Tak	Tak TG2A1	Neutral.	Tak	Tak TG2A1
13	*Nawet jak chcę zmienić hasło, to*	Neutral.	Neutral.	Tak	Tak TG2A1	Neutral.	Tak	Tak TG2A1

	wyskakuje, że nie da się zmienić hasła							
14	*zresetowałem hasło i nadal pisze ze dane nie prawidlowe w aplikacji*	Neutral.	Neutral.	Tak	Tak TG2A1	Neutral.	Tak	Tak TG2A1
15	*Mogę prosić numer do biura obsługi aby zresetowali mi hasło*	Neutral.	Neutral.	Tak	Tak TG2A1	Neutral.	Tak	Tak TG2A1
Podsumowanie			Błędy istotne dla akcji bota: 3	15/15	12/15	Błędy istotne dla akcji bota: 0	15/15	15/15

Dla intencji „T1 Odzyskiwanie hasła” (tabela 6.12) uzyskano poprawę wyboru akcji na poziomie 25%. Warunkowanie emocjami występowało jeszcze dla intencji „T3 Problemy z aplikacją” i tu poprawa również była znacząca 19,18%. Natomiast dla pozostałych zbiorów T2, T4, T5 emocje nie były uwzględnione w regułach wnioskowania, dlatego wyniki metody F_{AE} są takie same jak dla F_A co oznaczono jako poprawa 0%.

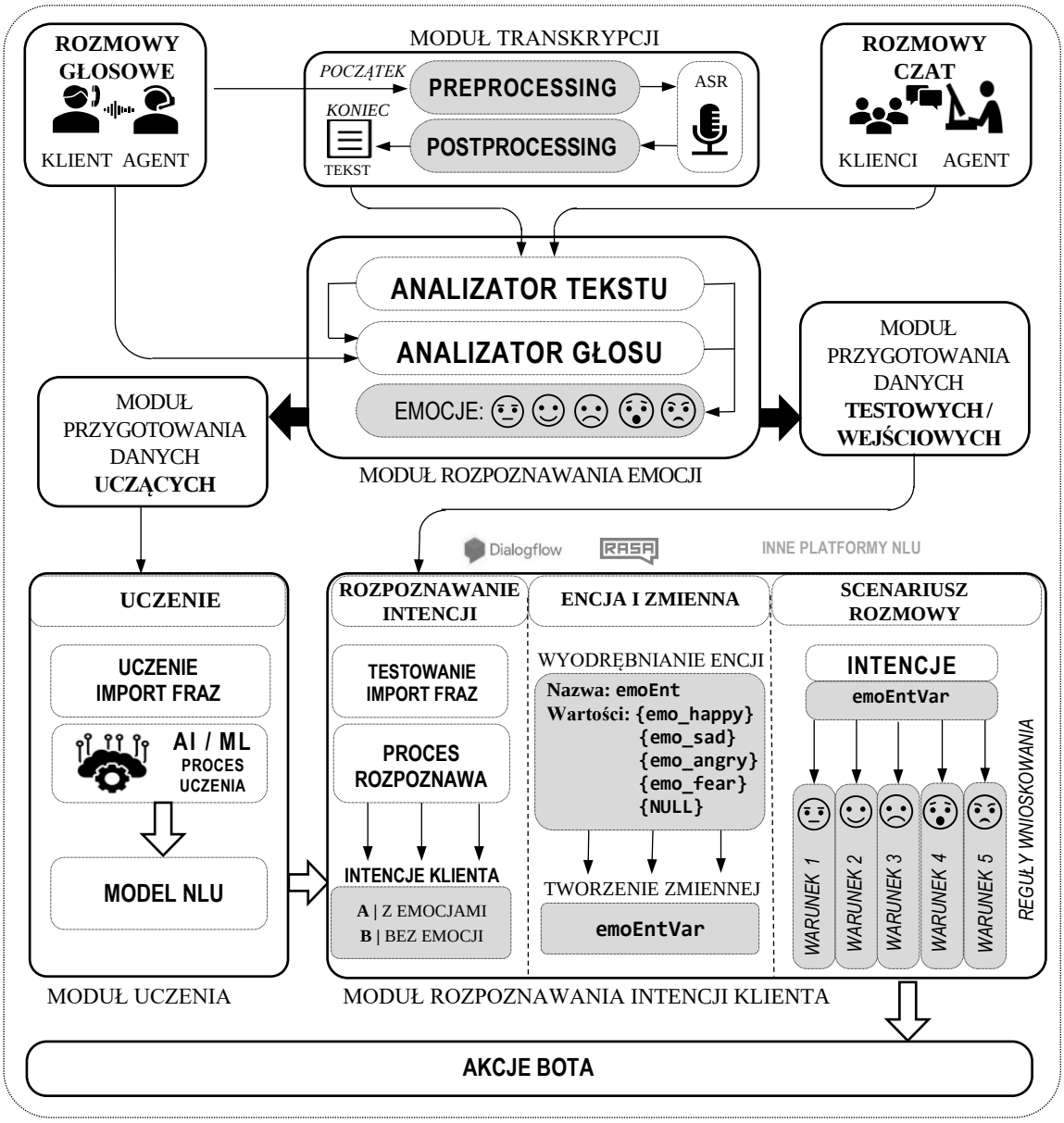
W działaniu operacyjnym bota dla metody F_{AE} mogą zdarzyć się przypadki, gdy nastąpi też pogorszenie dokładności *accA* wyboru akcji bota z uwagi na niską dokładność *accE* rozpoznawania emocji, ale wówczas należy usprawnić moduł rozpoznawania emocji lub zastąpić go bardziej efektywnym. W sumarycznym bilansie dla testowanych zbiorów rozmów czat zawsze metoda rozpoznawania ukrytych intencji F_{AE} uzyskiwała znacząco lepsze wyniki od standardowej metody F_A, dlatego należy przyjąć, że jest to rozwiązanie, które poprawia skuteczność działania chatbotów i potwierdza słuszność postawionej tezy w pracy.

6.3.3 Podsumowanie wyników dla kanału głosowego i tekstowego

Przypadek intencji „V1 Potwierdzenie” i „T1 Odzyskiwanie hasła” pokazuje jak istotny jest wybór akcji bota w oparciu o reguły wnioskowania bazujące na emocjach, szczególnie wtedy, gdy zrozumienie ukrytych intencji klienta wynikających z jego Złości, Smutku, Strachu czy Radości ma kluczowe znaczenie dla adekwatności udzielanych odpowiedzi. Dla człowieka zrozumienie rzeczywistej intencji rozmówcy jest procesem naturalnym, bo w ramach swojego rozwoju uczy się rozpoznawać komunikaty pozawerbalne, a także ich „synonimy” w postaci znaków tekstowych jakimi są emotikony. Reakcja człowieka na prawidłowo zrozumianą intencję rozmówcy jest również czynnością naturalną. Natomiast w przypadku botów konieczne jest ich uczenie nie tylko rozumienia tekstu wypowiedzi, ale również rzeczywistych i niekiedy ukrytych intencji często wynikających z emocji. Kluczowym elementem jest zatem możliwość udzielania adekwatnej odpowiedzi klientowi bazując na wszystkich czynnikach, które mogą na nią wpłynąć tzn. treści wypowiedzi, wyrażanych emocjach w głosie, jak również emocjach wyrażanych w tekście (np. w postaci emotikon). Wyniki przeprowadzonych testów potwierdzają wysoką skuteczność opracowanej metody rozpoznawania ukrytych intencji F_{AE} zarówno w rozmowach głosowych jak i tekstowych, szczególnie wtedy, gdy odpowiedź kierowana do klienta ściśle zależy od emocji zawartej w jego wypowiedzi. Warto też podkreślić, że wybrany sposób przekazywania emocji poprzez encje oraz opracowane zasady dotyczące budowania reguł wnioskowania są rozwiązaniem uniwersalnym dla platform NLU.

7 System zwiększający skuteczność wirtualnego konsultanta w Contact Center

Proponowana metoda rozpoznawania ukrytych intencji F_{AE} została zintegrowana z metodą poprawy jakości transkrypcji *IAT* tworząc kompletne środowisko do uczenia i uruchamiania chatbotów i voicebotów (rysunek 7.1). Opracowywana metoda F_{AE} uwzględnia rozpoznawane emocje w kanałach głosowych oraz kanałach tekstowych. Na rysunku 7.1 szary kolor tła wskazuje na nowe elementy zaproponowane w pracy.



Rysunek 7.1. System zwiększający skuteczność wirtualnego konsultanta w Contact Center

Środowisko uczenia i uruchamiania systemu wirtualnego konsultanta dedykowane dla CC zawiera następujące składowe:

- moduł przygotowania danych uczących oraz danych testowych/wejściowych,
- moduł transkrypcji rozmów pochodzących z kanałów dźwiękowych CC,
- moduł rozpoznawania emocji w wypowiedziach klientów CC,
- moduł uczenia modelu NLU,

- moduł rozpoznawania ukrytych intencji uwzględniający rozpoznane emocje.

W fazie uczenia definiowane są intencje, w ramach których grupowane są frazy uczące wraz z przypisanymi encjami *emoEnt*. Następnie projektowany jest scenariusz rozmowy z regułami wnioskowania. Na podstawie tych danych następuje proces uczenia, którego efektem działania jest wytworzenie modelu NLU.

W fazie działania metody F_{AE} następuje uruchomienie wszystkich modułów. Na podstawie rozpoznania emocji, a następnie intencji oraz wyodrębnienia wartości encji *emotEnt* do zmiennej *emoEntVar* następuje uruchomienie reguł wnioskowana, które decydują o wyborze akcji bota.

DANE UCZĄCE (zbiór S_{PU}) przygotowywane są w postaci fraz uczących przez specjalistów CC analizujących dane w postaci rozmów chat (dla kanału tekstowego) lub w postaci transkrypcji rozmów telefonicznych (dla kanału głosowego). Teksty rozmów uzupełniane są o wartości encji określające rozpoznane emocje rozmówcy. Frazy uczące zawierają zatem docelowo tekst wybranej wypowiedzi, który ewentualnie uzupełniony może być o wartość rozpoznanej emocji. DANE UCZĄCE przekazywane są do modułu uczenia platformy NLU.

7.1 Moduł przygotowania danych uczących i testujących

Moduł przygotowania danych testujących oraz wejściowych, przeznaczony jest do testowania rozwiązania w środowisku produkcyjnym w czasie bezpośrednich rozmów prowadzonych na infoliniach CC. DANE TESTOWE (zbiór S_{PT}) stosowane są podczas weryfikacji pracy systemu w warunkach laboratoryjnych, natomiast DANE WEJŚCIOWE są danymi wykorzystywanymi już podczas bezpośredniej eksploatacji opracowanego rozwiązania. Moduł ten analizuje prowadzone rozmowy wraz z określanymi dla nich emocjami i wybrane frazy przekazuje do platformy NLU. FRAZY UCZĄCE podobnie jak FRAZY TESTOWE zawierają tekst wypowiedzi oraz ewentualnie dodaną encję określającą emocję w postaci *emotEntValue*.

7.2 Moduł transkrypcji rozmów głosowych

Stosowana metoda poprawy jakości transkrypcji rozmów głosowych *IAT* [95] dedykowana jest bezpośrednio dla systemów CC. Metoda ta pozwala na zwiększenie skuteczności transkrypcji automatycznej rozmów poprzez stosowanie wybranych algorytmów preprocessingu nagrania rozmowy oraz postprocessingu wykonanej pierwotnie transkrypcji automatycznej (danych systemu ASR). W zakresie preprocessingu funkcjonują trzy moduły: moduł separacji kanałów, moduł uczący systemów ASR, moduł korekcji sygnału audio. Z kolei w zakresie postprocessingu w zależności od potrzeb wykorzystywać można moduły: moduł korekty tekstu, moduł podobnie brzmiących i obcych słów, moduł lematyzacji. Wykazano, że opracowane metody preprocessingu i postprocessingu umożliwiają uzyskanie jakości transkrypcji na poziomie 90-92%, czyli wyższej w porównaniu z aktualnie znanymi rozwiązaniami, które zostały przedstawione w podrozdziale 2.3.

7.3 Moduł rozpoznawania emocji

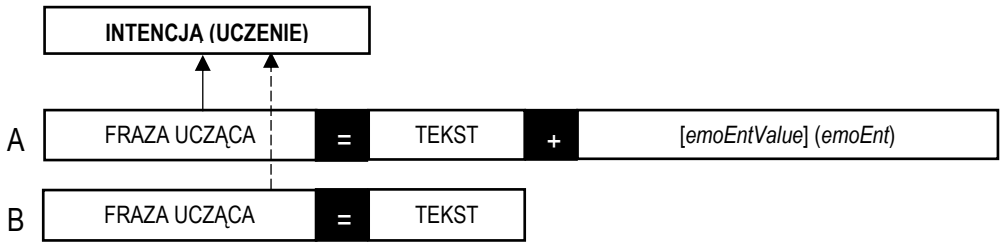
Metoda rozpoznawania emocji [82] wykorzystuje mechanizmy sztucznej inteligencji, uczenia maszynowego oraz metody przetwarzania języka naturalnego do rozpoznawania emocji w wypowiedziach rozmówców. W skład narzędzia rozpoznawania emocji wchodzą analizator głosu oraz analizator tekstu. Za ich pomocą wyodrębniane są cechy zawierające informacje niezbędne do określenia właściwego stanu emocjonalnego klienta [102], [103]. Cechy te są danymi wejściowymi klasyfikatora, odpowiedzialnego za określenie typu emocji.

7.4 Moduł uczenia modelu NLU

Moduł UCZENIA odpowiada za proces importu fraz uczących oraz proces uczenia z wykorzystaniem zaimportowanych fraz, które powinny zawierać również wartości ze zbioru S_E. Kluczowym wynikiem działania procesu uczenia jest utworzenie modelu NLU, który jest podstawowym elementem wykorzystywanym przez platformy NLU w rozpoznawaniu intencji z fraz testowych oraz wyodrębnianiu z nich wartości encji. W proponowanym rozwiązaniu model ten jest istotny w pracy modułu ROZPOZNAWANIE INTENCJI oraz w procesie WYODRĘBNIANIE ENCJI. W celu aktywacji mechanizmów encji, każda z określonych intencji jest uczona, co pokazano na rysunku 7.2. Uczenie następuje:

- frazami zawierającymi encje *emoEnt* określającą poszczególne emocje (A),
- frazami nie zawierającymi encji *emoEnt* (B).

Frazy B zapewniają możliwość określania neutralnego stanu emocjonalnego.

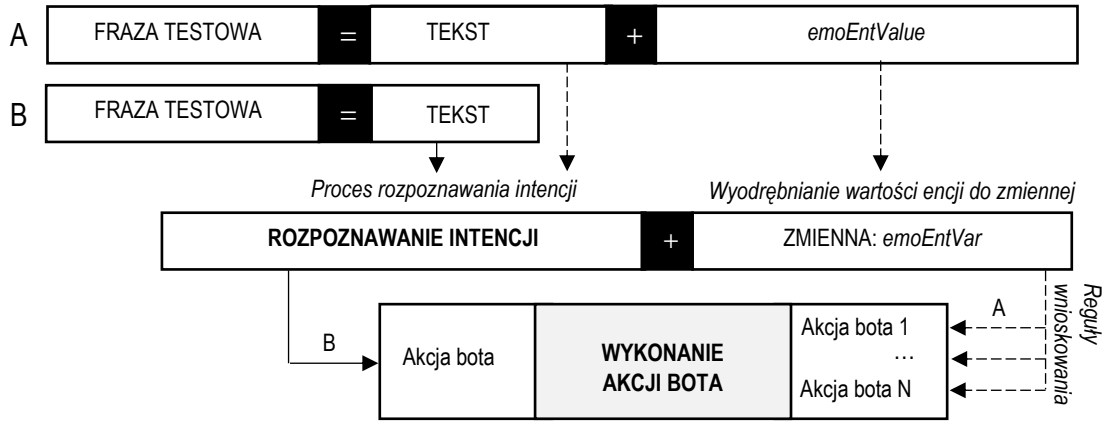


Rysunek 7.2. Schemat procesu uczenia wykorzystany w metodzie rozpoznawania ukrytych intencji

7.5 Moduł rozpoznawania ukrytych intencji

Najistotniejszym komponentem funkcjonalnym jest opracowany moduł rozpoznawania intencji klienta, z którym zintegrowano platformy NLU: Dialogflow oraz Rasa. Z punktu widzenia działania proponowanej metody F_{AE} nie ma ograniczania do dwóch powyższych platform, tym samym wykorzystywane mogą być również inne platformy NLU. Moduł zaprezentowany na rysunku 7.1 obejmuje następujące bloki: ROZPOZNAWANIE INTENCJI, ENCJA I ZMIENNA oraz SCENARIUSZ ROZMOWY. Dla modelu wytrenowanego według opisanych założeń, jako dane testujące do bloku ROZPOZNAWANIE INTENCJI trafiać mogą frazy testowe zawierające określoną emocję lub też wskazujące na stan neutralny rozmówcy. W przypadku (A), gdy emocja zostanie rozpoznana, wówczas na końcu frazy testowej dołączana jest informacja o występującej emocji w postaci *emoEntValue*. Później, informacja ta jest wyodrębniana i zapamiętywana w zmiennej o nazwie *emoEntVar*, co dzieje się w bloku ENCJA I ZMIENNA. W momencie, gdy informacja o konkretnej emocji sprowadzona zostaje do postaci zmiennej, wówczas możliwe jest uruchomienie reguł wnioskowania określających akcje bota. W kolejnym kroku blok SCENARIUSZ ROZMOWY realizuje docelowy scenariusz rozmowy wykorzystując nowe podejście. Bazuje ono nie tylko na rozpoznanej intencji, ale również pozwala na uzależnienie akcji bota od rozpoznanej emocji klienta zapisanej w zmiennej *emoEntVar*. O wyborze odpowiedniej akcji bota decydują opracowane reguły wnioskowania zbudowane na podstawie emocji, które mogą się pojawić w wypowiedziach klientów i są istotne z punktu widzenia wyboru akcji bota. Dzięki temu SCENARIUSZ ROZMOWY uwzględnia zarówno tekst wypowiedzi jak również emocje w niej zawarte. Wszystko to skutkuje udzieleniem odpowiedzi adekwatnej do rzeczywistej sytuacji, z której wynika faktyczna intencja klienta (A). W przypadku, gdy emocje nie mają znaczenia dla danej

intencji, boty nie zawierają reguł wnioskowania, tylko wprost wyznaczają akcję bota (B). Na rysunku 7.3 przedstawiono schemat procesu wyboru akcji bota w zależności od otrzymanych parametrów i zdefiniowanych reguł wnioskowania dla danej intencji (A) lub wyznaczenia akcji bota wprost (B).



Rysunek 7.3. Schemat procesu wyboru akcji bota wykorzystany w metodzie rozpoznawania ukrytych intencji

Implementacja sposobu prowadzenia rozmowy w zależności od wartości *emoEnt* jest różna dla obu badanych platform. W przypadku platformy Dialogflow wykorzystywane są parametry osadzane w poszczególnych intencjach. Parametry te są automatycznie uzupełniane wartością encji, a w celu określenia dalszego przebiegu rozmowy stosowany jest język warunków dostępnych na tej platformie. Na podstawie warunków (reguł wnioskowania) określane są kolejne etapy rozmowy. W przypadku platformy Rasa wykorzystywane są sloty, które pełnią rolę zmiennych przechowujących wartości encji. W oparciu o wypełnione sloty, w zależności od ich wartości, wybierana jest określona akcja bota, czyli odpowiedź jakiej udziela wirtualny konsultant klientowi.

7.6 Wyniki eksperymentów wpływu metody poprawy jakości transkrypcji na wybór akcji bota

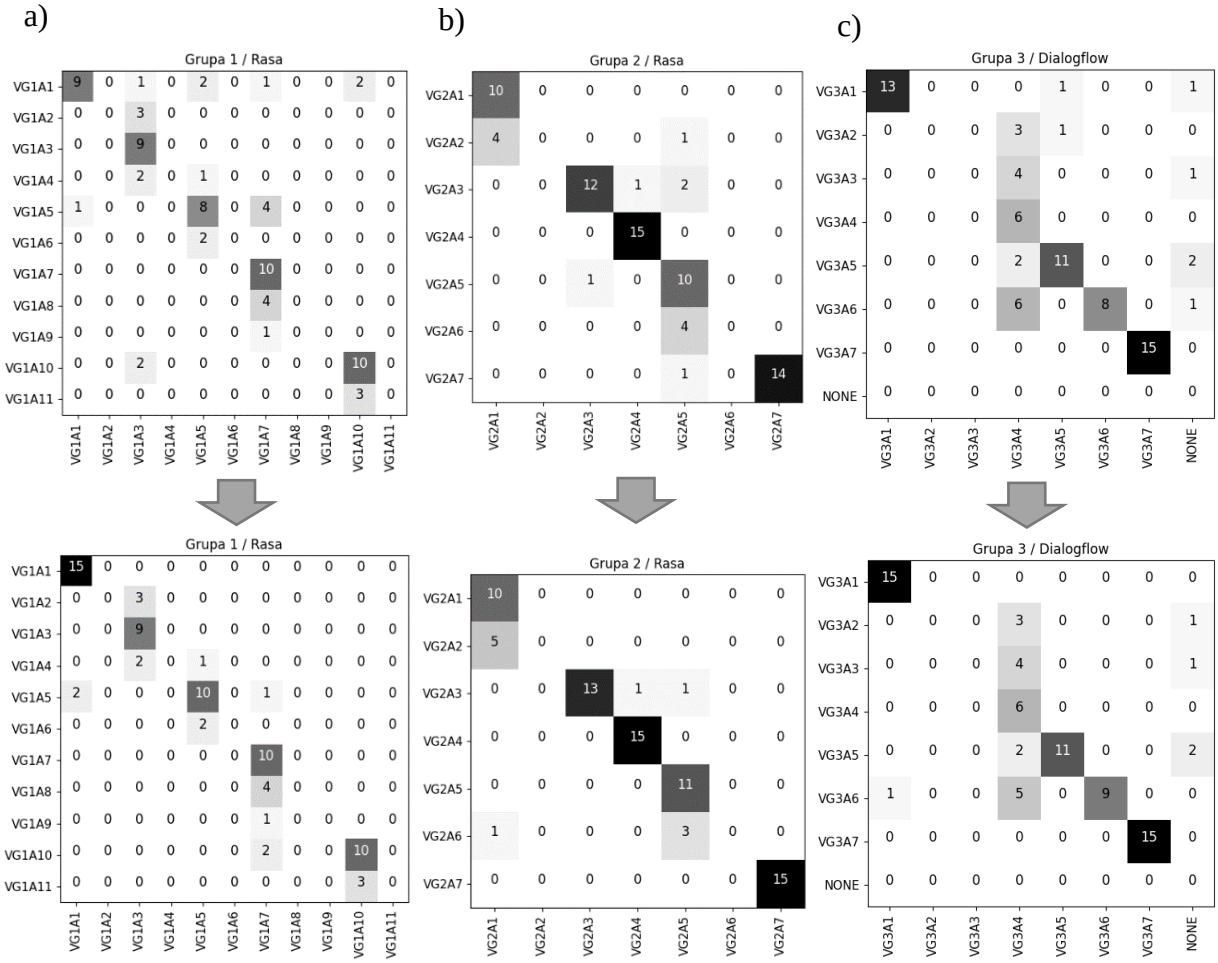
W tabeli 7.1 zestawiono wyniki eksperymentów obrazujące dokładność *accA* wyboru akcji bota przy zastosowaniu metody poprawy jakości transkrypcji *IAT*. W zakresie rozpoznawania intencji są to te same dane co w podrozdziale 5.4, natomiast zostały one uzupełnione o wyniki wyboru akcji bota, aby stanowiły dane porównawcze dla wyników zaprezentowanych w podrozdziale 7.7, które zostały uzyskane na podstawie zintegrowania metody poprawy jakości transkrypcji *IAT* oraz metody rozpoznawania ukrytych intencji F_{AE} w jeden system zwiększający skuteczność wirtualnego konsultanta w Contact Center.

Tabela 7.1. Wpływ jakości transkrypcji na poprawność wyboru akcji bota dla kanału głosowego

Grupa	**Platforma NLU**	**Rozpoznawanie intencji F_{NLU}**		**Wybór akcji bota F_A**		
		System ASR	**Metoda *IAT***	**System ASR**	**Metoda *IAT***	**Poprawa**
		accINT	***accINT***	***accA***	***accA***	**[%]**
1	Dialogflow	0,68 (51/75)	0,87 (65/75)	0,52 (39/75)	0,71 (53/75)	**36,54**
	Rasa	0,81 (61/75)	0,92 (69/75)	0,61[(a)] (46/75)	0,72[(a)] (54/75)	**18,03**
2	Dialogflow	0,87 (65/75)	0,88 (66/75)	0,79 (59/75)	0,80 (60/75)	**1,27**
	Rasa	0,92 (69/75)	0,96 (72/75)	0,81[(b)] (61/75)	0,85[(b)] (64/75)	**4,94**
3	Dialogflow	0,80 (60/75)	0,84 (63/75)	0,71[(c)] (53/75)	0,75[(c)] (56/75)	**5,63**
	Rasa	0,95 (71/75)	0,97 (73/75)	0,83 (62/75)	0,87 (65/75)	**4,82**

(a), (b), (c) – dane zaprezentowane na rysunku 7.4

W tabeli 7.1 można zaobserwować wyraźną poprawę wyników w przypadku zastosowania metody *IAT* zawierającej elementy preprocessingu oraz postprocessingu. W zależności od badanej próby, dla platformy Dialogflow poprawa w zakresie akcji bota następuje w zakresie od 1,27% do 36,54%, natomiast dla platformy Rasa od 4,82% do 18,03%. Jest to także widoczne w prezentowanych na rysunku 7.4 wybranych macierzach konfuzji, szczególnie dla danych z grupy 1 (rysunek 7.4a). W grupie 2 i 3 liczba błędów w transkrypcjach automatycznych była mała, stąd wyniki nie uległy aż tak znacznej poprawie (rysunek 7.4b i rysunek 7.4c).



Rysunek 7.4. Macierze konfuzji dla wybranych danych z tabeli 7.1

Legenda: **VG1-3A1–11** – Akcje bota w poszczególnych grupach (1, 2 lub 3) dla kanału głosowego; **NONE** – niewybrana żadna akcja bota; Osie rzędnych (y) opisują wartości **PRAWDA**; Osie odciętych (x) opisują wartości **PREDYKCJA**

W tabeli 7.2 przedstawiono wyniki w podziale na poszczególne intencje dla „Grupa 1 / Rasa".

Tabela 7.2. Wyniki accINT, accA w podziale na intencje dla zbioru „Grupa 1 / Rasa"

Zbiór fraz testowych dla intencji	**Akcje bota A_k**	**Intencje F_{NLU}**		**Akcje bota F_A**		
		System ASR	**Metoda *IAT***	**System ASR**	**Metoda *IAT***	**Poprawa**
		accINT	***accINT***	***accA***	***accA***	**[%]**
V1 Potwierdzenie	VG1A5 VG1A6	0,67 (10/15)	0,80 (12/15)	0,53 (8/15)	0,67 (10/15)	**26,42**
V2 Zaprzeczenie	VG1A10 VG1A11	0,87 (13/15)	0,87 (13/15)	0,67 (10/15)	0,67 (10/15)	**0,00**
V3 Autoryzacja	VG1A1	0,60 (9/15)	1,00 (15/15)	0,60 (9/15)	1,00 (15/15)	**66,67**
V4 Wypowiedzenie umowy	VG1A7 VG1A8 VG1A9	1,00 (15/15)	1,00 (15/15)	0,67 (10/15)	0,67 (10/15)	**0,00**
V5 Zadowolenie z obsługi	VG1A2 VG1A3 VG1A4	0,93 (14/15)	0,93 (14/15)	0,60 (9/15)	0,60 (9/15)	**0,00**
Cały zbiór		**0,81 (61/75)**	**0,92 (69/75)**	**0,61 (46/75)**	**0,72 (54/75)**	**18,03**

W tabeli 7.3 przedstawiono wyniki dla fraz testowych dla intencji „V1 Potwierdzenie”. Wynikiem prawidłowym dla tej intencji może być tylko akcja VG1A5 z uwagi na to, że została zastosowana tylko metoda *IAT*, a w zakresie wyboru akcji bota standardowa metoda F_A, która nie uwzględnia emocji.

Tabela 7.3. Wyniki dla fraz testowych dla intencji „V1 Potwierdzenie”

Lp.	Prawdz. emocja E_f	Brak rozp. emocji	System ASR			Metoda *IAT*		
			Tekst ASR	Intencje F_{NLU}	Akcje bota F_A	Tekst po poprawie *IAT*	Intencje F_{NLU}	Akcje bota F_A
1	Smutek	Neutralność	*No właśnie.*	Tak	Nie VG1A5	*no właśnie*	Tak	Nie VG1A5
2	Neutralność	Neutralność	*Ten sam tag.*	Nie	Nie INNA	*ten sam tak*	Tak	Tak VG1A5
3	Smutek	Neutralność	*No dobrze, akceptuję no potwierdzam.*	Tak	Nie VG1A5	*no dobra akceptuję no potwierdzam*	Tak	Nie VG1A5
4	Neutralność	Neutralność	*E tak to będzie pierwszy naprawa.*	Nie	Nie INNA	*e tak to będzie 1 naprawa*	Nie	Nie INNA
5	Neutralność	Neutralność	*Tak oczywiście będzie ten sam z którego dzwonię*	Nie	Nie INNA	*tak oczywiście będzie ten sam z którego dzwonię*	Nie	Nie INNA
6	Neutralność	Neutralność	*Tak tak, tak właśnie*	Tak	Tak VG1A5	*tak tak tak właśnie*	Tak	Tak VG1A5
7	Neutralność	Neutralność	*Taak*	Nie	Nie INNA	*Tak*	Tak	Tak VG1A5
8	Neutralność	Neutralność	*Dobrze poproszę tak oczywiście tak*	Tak	Tak VG1A5	*dobra poproszę tak oczywiście tak*	Tak	Tak VG1A5
9	Neutralność	Neutralność	*No zwykle do tej pory akceptowałem, czyli teraz też akceptuję.*	Nie	Nie INNA	*do tej pory akceptowałem czyli teraz też akceptuję*	Nie	Nie INNA
10	Złość	Neutralność	*Tak, ale jeszcze się zastanowię*	Tak	Tak VG1A5	*tak ale jeszcze się zastanowię*	Tak	Tak VG1A5
11	Neutralność	Neutralność	*Hmm dobra dobra dobra super.*	Tak	Tak VG1A5	*hmm dobra dobra dobra super*	Tak	Tak VG1A5
12	Neutralność	Neutralność	*No rozumiem, rozumiem, dobrze, super*	Tak	Tak VG1A5	*no rozumiem rozumiem dobra super*	Tak	Tak VG1A5
13	Złość	Neutralność	*No a co mam zrobić, no akceptuję.*	Tak	Tak VG1A5	*no a co mieć zrobić no akceptuję*	Tak	Tak VG1A5
14	Neutralność	Neutralność	*Tak, tak, tak, tak oczywiście będą będą spłacane.*	Tak	Tak VG1A5	*tak tak tak tak oczywiście będą będą spłacane*	Tak	Tak VG1A5
15	Neutralność	Neutralność	*Aha no to dobra, no to dobra*	Tak	Tak VG1A5	*aha no to dobra no to dobra*	Tak	Tak VG1A5
Podsumowanie		Błędy istotne dla akcji bota: 2	-	10/15	8/15	-	12/15	10/15

Ze względu, że nie ma warunkowania emocjami to nie są one również dołączane do frazy testowej jako encja, zmienna *emoEntVar* ma więc zawsze wartość *NULL*. Można przyjąć, że jest to równoznaczne ze stanem Neutralność, który w przypadku metody rozpoznawania ukrytych intencji F_{AE} również nie powoduje dodania emocji jako encji *emoEnt* do frazy testowej. W efekcie, jeżeli w wypowiedzi pojawi się emocja Smutek lub Strach to pomimo

prawidłowego rozpoznania intencji, akcja bota dla bazowej metody będzie nieprawidłowa VG1A5, gdyż dla tych emocji powinna to być VG1A6 zgodnie z regułami wnioskowania (tabela 6.7).

W wierszach 2 i 4 system ASR z uwagi na błędy popełnione w transkrypcji „tag" i „taak" dostarczył na wejście frazy testowe, dla których platforma NLU błędnie rozpoznała intencje. Metoda *IAT* poprawiła te błędy na „tak" i spowodowała prawidłowe rozpoznawanie intencji. Końcowy wynik rozpoznawania akcji bota w przypadku tych wierszy również okazał się poprawny, z uwagi na to, że rzeczywista emocja w tych wypowiedziach była Neutralna, czyli spoza zbioru emocji (Smutek, Strach) na które były zdefiniowane warunki (tabela 6.7).

W wierszach 1 i 3 można zauważyć, że pomimo prawidłowego rozpoznania intencji wynik jest nieprawidłowy. Wynika to z tego, że fraza testowa powstała z wypowiedzi, która zawierała prawdziwą emocję Smutek, a ze względu, że w tym badaniu nie były dołączone rozpoznane emocje w wypowiedzi w postaci encji *emoEnt* to odpowiedź nie była zgodna z regułą wnioskowania (czyli na przykład z oczekiwaniami menedżera biura obsługi klienta).

Jak już wspomniano w podrozdziale 5.4 przy omawianiu wyników rozpoznawania intencji, tak samo w przypadku wyników akcji bota widać, iż zasadne jest użycie metody *IAT*, gdyż zwiększa ona poziom rozpoznawania intencji, który bezpośrednio przekłada się na wybór prawidłowych akcji bota. Dokładna analiza tabeli 7.1 pokazuje, że użycie metody *IAT* dla każdego zbioru zwiększyło wskaźniki rozpoznawania akcji bota. Natomiast patrząc na bardziej szczegółowe dane w ramach jednego zbioru testowego „Grupa 1 / Rasa" (tabela 7.2) można dostrzec, że zdarzają się przypadki, gdy metoda *IAT* nie polepsza wyników (V2, V4, V5). Gdy jednak spojrzymy na wynik wszystkich akcji bota w danym zbiorze, to wynik dla metody *IAT* jest zawsze lepszy od systemu ASR. W analizowanym zbiorze zwiększenie dokładności widać dla intencji V1 i V3, gdzie wzrost akcji bota nastąpił odpowiednio o 26,42% oraz 66,67%. Należy też podkreślić, że nawet jeżeli w wynikach dla metody poprawy jakości transkrypcji *IAT* pojawiają się błędy to istnieje potencjał rozbudowania jej słowników o zauważone problematyczne przypadki celem zwiększenia jej skuteczności. Należy jednak podkreślić, że już obecny poziom rozpoznawania akcji bota dla metody *IAT* w sposób znaczy przewyższa system ASR.

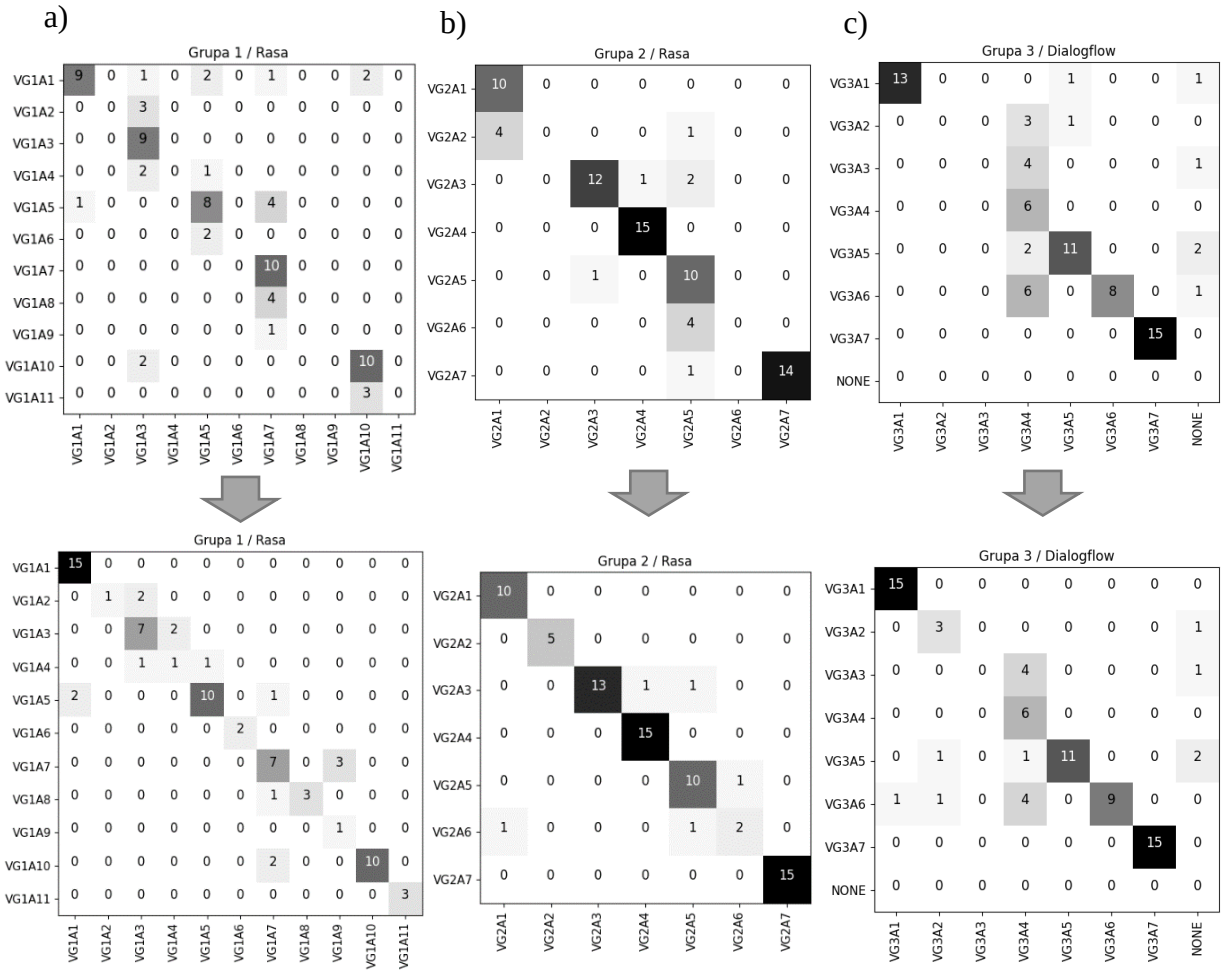
7.7 Wyniki eksperymentów wpływu zintegrowanych metod na wybór akcji bota

W niniejszej sekcji zaprezentowano wyniki eksperymentów proponowanego w pracy rozwiązania, w którym zintegrowano zarówno metodę poprawy jakości transkrypcji *IAT* jak i metodę rozpoznawania ukrytych intencji F_{AE}. Wyniki eksperymentów uzyskane podczas końcowych testów opracowanego w pracy rozwiązania zawiera tabela 7.4.

Tabela 7.4. Wpływ jakości transkrypcji oraz występujących emocji na poprawność wyboru akcji bota dla kanału głosowego

Grupa	**Platforma NLU**	**Rozpoznawanie emocji**	**Wybór akcji bota**		
			System ASR, F_A bez *emoEnt*	**Metoda *IAT*, F_{AE} z *emoEnt***	**Poprawa**
		accE	***accA***	***accA***	**[%]**
1	Dialogflow	0,80	0,52	0,79	**51,92**
	Rasa		0,61(a)	0,80(a)	**31,15**
2	Dialogflow	0,79	0,79	0,84	**6,33**
	Rasa		0,81(b)	0,93(b)	**14,81**
3	Dialogflow	0,65	0,71(c)	0,79(c)	**11,27**
	Rasa		0,83	0,91	**9,64**

(a), (b), (c) – dane zaprezentowane na rysunku 7.5



Rysunek 7.5. Macierze konfuzji dla wybranych danych z tabeli 7.4

Legenda: **VG1-3A1–11** – Akcje bota w poszczególnych grupach (1, 2 lub 3) dla kanału głosowego; **NONE** – niewybrana żadna akcja bota; Osie rzędnych (y) opisują wartości **PRAWDA**; Osie odciętych (x) opisują wartości **PREDYKCJA**

Kolumna „Poprawa” zawiera dane określające wartości procentowe pokazujące poprawę wyboru akcji bota po zastosowaniu metody *IAT* oraz F_{AE} w porównaniu ze standardowymi metodami (systemem ASR i metodą F_A). Dane w tabeli 7.4 potwierdzają, że poprawność wyboru akcji bota bardzo wyraźnie wzrasta w przypadku zastosowania proponowanego systemu wirtualnego konsultanta w CC zaprezentowanego na rysunku 7.1.

Wyniki eksperymentów, potwierdzają macierze konfuzji, które pokazano na rysunku 7.5. Macierze odnoszą się do wybranych przypadków w kanale głosowym, które zostały zaciemnione w tabeli 7.4. Na rysunku 7.5 wiersz pierwszy we wszystkich kolumnach prezentuje dane uzyskane dla standardowych metod, natomiast wiersz drugi zawiera dane dla metody *IAT* oraz F_{AE}.

W przypadku metody F_{AE} prawidłowy wybór akcji bota będzie zależał od tego czy w danej frazie testowej zostały prawidłowo wykryte emocje. Z kolei na rozpoznawanie intencji pozytywny wpływ ma metoda poprawy transkrypcji *IAT*, szczególnie wtedy, gdy we frazach testowych pojawiają się błędy w transkrypcji (literówki, zamiany podobnie brzmiących słów itd.). Szczegółowe wyniki dla „Grupa 1 / Rasa” dla kanału głosowego przedstawiono w tabeli 7.5.

Tabela 7.5. Wyniki accE, accINT, accA dla zbioru „Grupa 1 / Rasa”

Zbiór fraz testowych dla intencji	**Akcje bota A_k**	**Rozpozn. emocji**	**Intencje F_{NLU}**		**Akcje bota**		
			System ASR bez *emoEnt*	**Metoda *IAT* z *emoEnt***	**System ASR, F_A bez *emoEnt***	**Metoda *IAT*, F_{AE} z *emoEnt***	**Poprawa**
		accE	***accINT***	***accINT***	***accA***	***accA***	**[%]**
V1 Potwierdzenie	VG1A5 VG1A6	0,80 (12/15)	0,67 (10/15)	0,80 (12/15)	0,53 (8/15)	0,80 (12/15)	**50,94**
V2 Zaprzeczenie	VG1A10 VG1A11	1,00 (15/15)	0,87 (13/15)	0,87 (13/15)	0,67 (10/15)	0,87 (13/15)	**29,85**
V3 Autoryzacja	VG1A1	0,80 (12/15)	0,60 (9/15)	1,00 (15/15)	0,60 (9/15)	1,00 (15/15)	**66,67**
V4 Wypowiedzenie umowy	VG1A7 VG1A8 VG1A9	0,73 (11/15)	1,00 (15/15)	1,00 (15/15)	0,67 (10/15)	0,73 (11/15)	**8,96**
V5 Zadowolenie z obsługi	VG1A2 VG1A3 VG1A4	0,67 (10/15)	0,93 (14/15)	0,93 (14/15)	0,60 (9/15)	0,60 (11/15)	**0**
Cały zbiór		**0,80 (60/75)**	**0,81 (61/75)**	**0,92 (69/75)**	**0,61 (46/75)**	**0,80 (60/75)**	**31,15**

W tabeli 7.6 przedstawiono wyniki dla fraz testowych dla intencji „V1 Potwierdzenie”, wynikiem prawidłowym dla tej intencji w zależności od rozpoznanych emocji może być tylko akcja bota VG1A5 i VG1A6. Skuteczność działania metody F_{AE}, która bierze pod uwagę rozpoznane emocje widać w wierszu 1 i 3. Gdzie zostały prawidłowo rozpoznane emocje (*emoEntVar=emo_sad*) i zadziałały reguły wnioskowania dla tej intencji, w efekcie czego w każdym przypadku prawidłowo została rozpoznana akcja bota VG1A6. Dla metody F_A pomimo prawidłowego rozpoznania intencji, rozpoznana akcja bota okazała się nieprawidłowa VG1A5, gdyż reguły wnioskowania na wejściu otrzymały stan Neutralność (*emoEntVar=NULL*). W wierszu 2 i 7 pozytywny wpływ na rozpoznawanie intencji miała z kolei metoda poprawy jakości transkrypcji *IAT*, a dzięki temu, że rozpoznana emocja również była prawidłowa Neutralność (*emoEntVar=NULL*) to końcowa akcja bota VG1A5 została prawidłowo wybrana.

Tabela 7.6. Wyniki dla fraz testowych dla intencji „V1 Potwierdzenie”

Lp.	Prawdz. emocja E_f	System ASR bez *emoEnt*				Metoda IAT z *emoEnt*			
		Tekst z ASR	**Brak rozp. emocji**	**Intencje F_{NLU}**	**Akcje bota F_A**	**Tekst po poprawie *IAT***	**Emocje**	**Intencje F_{NLU}**	**Akcje bota F_{AE}**
1	Smutek	*No właśnie.*	Neutral.	Tak	Nie VG1A5	*no właśnie*	Smutek	Tak	Tak VG1A6
2	Neutral.	*Ten sam tag.*	Neutral.	Nie	Nie INNA	*ten sam tak*	Neutral.	Tak	Tak VG1A5
3	Smutek	*No dobrze, akceptuję no potwierdzam.*	Neutral.	Tak	Nie VG1A5	*no dobra akceptuję no potwierdzam*	Smutek	Tak	Tak VG1A6
4	Neutral.	*E tak to będzie pierwszy naprawa.*	Neutral.	Nie	Nie INNA	*e tak to będzie 1 naprawa*	Złość	Nie	Nie INNA
5	Neutral.	*Tak oczywiście będzie ten sam z którego dzwonię*	Neutral.	Nie	Nie INNA	*tak oczywiście będzie ten sam z którego dzwonię*	Neutral.	Nie	Nie INNA
6	Neutral.	*Tak tak, tak właśnie*	Neutral.	Tak	Tak VG1A5	*tak tak tak właśnie*	Neutral.	Tak	Tak VG1A5
7	Neutral.	*taak*	Neutral.	Nie	Nie INNA	*tak*	Neutral.	Tak	Tak VG1A5
8	Neutral.	*dobrze poproszę tak oczywiście tak*	Neutral.	Tak	Tak VG1A5	*dobra poproszę tak oczywiście tak*	Neutral.	Tak	Tak VG1A5
9	Neutral.	*No zwykle do tej pory akceptowałem, czyli teraz też akceptuję.*	Neutral.	Nie	Nie INNA	*no zwykle do tej pory akceptowałem czyli teraz też akceptuję*	Neutral.	Nie	Nie INNA
10	Złość	*Tak, ale jeszcze się zastanowię*	Neutral.	Tak	Tak VG1A5	*tak ale jeszcze się zastanowię*	Złość	Tak	Tak VG1A5
11	Neutral.	*Hmm dobra dobra dobra super.*	Neutral.	Tak	Tak VG1A5	*hmm dobra dobra dobra super*	Neutral.	Tak	Tak VG1A5
12	Neutral.	*No rozumiem, rozumiem, dobrze, super*	Neutral.	Tak	Tak VG1A5	*no rozumiem rozumiem dobra super*	Neutral.	Tak	Tak VG1A5
13	Złość	*No a co mam zrobić, no akceptuję.*	Neutral.	Tak	Tak VG1A5	*no a co mieć zrobić no akceptuję*	Radość	Tak	Tak VG1A5
14	Neutral.	*Tak, tak, tak, tak oczywiście będą będą spłacane.*	Neutral.	Tak	Tak VG1A5	*tak tak tak tak oczywiście będą będą spłacane*	Neutral.	Tak	Tak VG1A5
15	Neutral.	*Aha no to dobra, no to dobra*	Neutral.	Tak	Tak VG1A5	*aha no to dobra no to dobra*	Radość	Tak	Tak VG1A5
Podsumowanie		Błędy istotne dla akcji bota: 2		10/15	8/15	Błędy istotne dla akcji bota: 1		12/15	12/15

W zaprezentowanych przypadkach (tabela 7.6) widać, że dla wypowiedzi nacechowanych emocjami, które mają znaczenie w zakresie wyboru akcji bota, o sukcesie decydują dwa elementy: prawidłowa transkrypcja, bo od niej zależy rozpoznawanie intencji, prawidłowe rozpoznawanie emocji, bo od niej zależy prawidłowe działanie reguł wnioskowania na podstawie emocji. Oczywiście rozpoznawanie emocji musi być wsparte mechanizmem

przekazywania ich jako encji *emoEnt* w ramach platformy NLU. W wierszu 10 i 13 widać, że prawidłowe rozpoznawanie emocji jest tylko istotne w przypadku, gdy dana emocja występuje w regułach wnioskowania. Po dokładniejszej analizie wiersza 13 widać, że prawidłowa emocja to Złość, a rozpoznana emocja to Radość – zarówno jedna jak i druga nie występuje w regułach wnioskowania. Stąd wniosek, że jeżeli emocje jak w tym przypadku Złość, Radość, Neutralność nie występują w regułach wnioskowania mogą być rozpoznawane zamiennie bez wpływu na końcową akcję bota. Jest to zasada dość logiczna biorąc pod uwagę fakt, że skoro dla reguł wnioskowania określone emocje nie mają znaczenia, to jest istotne tylko czy prawdziwa i rozpoznana emocja występuje w zbiorze będącym poza regułami wnioskowania (jak w tym przypadku w zbiorze: Złość, Radość, Neutralność). Ostatni wiersz warty omówienia to 4. Pokazuje on, że jeżeli intencja nie zostanie prawidłowo wykryta to końcowy wynik już z definicji jest błędny. Nie ma już dalej znaczenia poprawność wykrycia emocji, gdyż proces zostanie skierowany do reguł wnioskowania innej intencji, które zawierają inny zbiór akcji w tabeli 7.6 zapisanych jako akcja INNA. Oznacza to, że nie jest to ani akcja VG1A5, ani VG1A6, tylko to jest jedna z akcji pochodząca z innej intencji (VG1A1 – VG1A4, VG1A7 – VGA11).

Zastosowanie kompleksowego środowiska systemu zwiększającego skuteczność wirtualnego konsultanta w Contact Center dla badanych zbiorów przyniosło bardzo dobre rezultaty uzyskując „Poprawę” od 6,33% do 51,92% (tabela 7.4). Poprawa wyników jest szczególnie widoczna, gdy scenariusze rozmów („Grupa 1 / Dialogflow” i „Grupa 1 / Rasa”) mają rozbudowane reguły wnioskowania, wynikające z dużego znaczenia emocji w zakresie wyboru akcji bota.

8 Podsumowanie i wnioski

Celem pracy było opracowanie metod poprawy skuteczności wirtualnych konsultantów poprzez minimalizację niepoprawnie rozpoznawanych intencji z wypowiedzi klientów.

W pierwszym kroku zwrócono szczególną uwagę na niską jakość sygnałów dźwiękowych, która jest główną przyczyną problemów z wystarczającym poziomem transkrypcji automatycznych wykonywanych przez znane rozwiązania systemów ASR. Ponadto, zwrócono uwagę, że modele języka angielskiego w systemach ASR są dużo lepiej dopracowane i zoptymalizowane niż modele innych mniej popularnych języków komunikacji. Sprawia to, że docelowe zastosowanie popularnych obecnie systemów ASR na potrzeby rozwiązań implementowanych w CC obsługujących klientów w różnych językach nie spełnia zakładanych oczekiwań. Natomiast, odpowiednio wysoka jakość transkrypcji jest bardzo istotna z punktu widzenia możliwości dalszego jej wykorzystywania między innymi w voicebotach.

W drugim kroku przy współpracy ze specjalistą z dziedziny lingwistyki przeprowadzono szczegółową analizę lingwistyczną struktur wybranych aktów mowy w koincydencji z czynnikiem emocjonalnym klienta. Materiał do analizy stanowiły rzeczywiste rozmowy CC pomiędzy agentem i klientem. Wyciągnięto wnioski, że prawidłowe rozumienie wypowiedzi klienta wymaga rozumienia pozawerbalnych znaków komunikacji. Przy współpracy z psychologami z rozmów tych wyodrębniono różne typy emocji. Analizowano jaki wpływ mają emocje w wypowiedziach na ukryte intencje klientów i jak się to przekłada na udzielanie im właściwych odpowiedzi.

W efekcie prowadzonych badań w niniejszej pracy wykazano, że jakość transkrypcji automatycznych może być znacznie poprawiona przez odpowiednie zastosowanie opisanych mechanizmów preprocessingu oraz postprocessingu. Dla wylosowanej reprezentatywnej próbki danych 300 rozmów głosowych średni wynik dla metryki WER w przypadku implementacji rozwiązania MST osiągnął poziom 8%. Przeprowadzone eksperymenty przedstawione w podrozdziale 5.4 rozpoznawania intencji z zastosowaniem metody *IAT* osiągnęły poprawę od 1,15% do 27,94% (tabela 5.9) i potwierdziły postawioną tezę, że możliwa jest poprawa skuteczności wirtualnych konsultantów w Contact Center poprzez minimalizację niepoprawnie rozpoznawanych wypowiedzi klientów z zastosowaniem preprocessingu i postprocessingu danych systemu ASR. Uzyskany poziom jakości transkrypcji pozwolił na zbudowanie skutecznych metod rozpoznawania intencji klientów na platformach NLU, co było solidną podstawą do kolejnych prac w postaci uwzględniania emocji w procesie rozpoznawania intencji.

Efektem badań nad wpływem emocji na wybór prawidłowych akcji bota było zaproponowanie nowej metody F_{AE} rozpoznawania ukrytych intencji klienta kontaktującego się z infolinią CC. Główną nową cechą funkcjonalną opracowanej metody F_{AE} są możliwości uwzględniania stanów emocjonalnych klientów rozpoznawanych w trakcie prowadzonych konwersacji. Jak wykazano w niniejszej pracy pozwala to na odpowiednie oraz zdecydowanie bardziej skuteczne poprowadzenie rozmowy przez voicebota lub chatbota z klientem. Przeprowadzone eksperymenty przedstawione w podrozdziałach 6.3.1 i 6.3.2 wyboru akcji bota z zastosowaniem metody F_{AE} osiągnęły wyniki poprawy dla kanału głosowego od 2,53% do 13,11% (tabela 6.8), a dla kanału czat od 4,17% do 28,33% (tabela 6.9) i potwierdziły postawioną tezę, że możliwa jest poprawa skuteczności wirtualnych konsultantów w Contact Center poprzez rozpoznawanie ukrytych intencji klientów na podstawie emocji zawartych

w ich wypowiedziach. W eksperymentach uwzględniane były emocje: Radość, Złość, Smutek, Strach oraz stan Neutralny. Natomiast metoda F_{AE} potrafi działać z dowolnym zbiorem emocji.

W ramach pracy zostało stworzone środowisko systemu wirtualnego konsultanta dedykowane dla Contact Center, w którym opracowana metoda poprawy jakości transkrypcji *IAT* została wbudowana w mechanizmy chatbotów i voicebotów stosowanych w CC. Powstały moduły przygotowania danych uczących oraz testowych posiadające funkcjonalność dołączania emocji do fraz w postaci encji. Został wbudowany moduł rozpoznawania emocji. Na platformach NLU zostały zaimplementowane reguły wnioskowania akcji bota, których argumentem wejściowym jest zmienna, do której przekazywana jest wyodrębniona wartość z encji *emoEnt*. Zostały opracowane zasady dotyczące budowania zbiorów fraz uczących oraz testowych. Określono w jaki sposób najbardziej efektywnie budować reguły wnioskowania i implementować je na platformach NLU oraz jak uczyć modele NLU w oparciu o przygotowane dane. W podrozdziale 7.7 przeprowadzono eksperymenty połączonych metod *IAT* i F_{AE}, dla których poprawa wyboru akcji bota wyniosła od 6,33% do 51,92% (tabela 7.4), które potwierdziły słuszność celów pracy w zakresie opracowania systemu ze zintegrowanymi metodami *IAT* i F_{AE}, zwiększającymi skuteczność wirtualnych konsultantów w Contact Center.

Oczekuje się, że rozwiązanie zaproponowane w pracy przyczyni się do poprawy jakości obsługi klienta oraz optymalizacji wskaźników KPI [3]. Prezentowane w pracy wyniki eksperymentów jednoznacznie wykazały, że w bardzo wielu przypadkach obecność emocji w konwersacjach powoduje konieczność zbudowania odpowiedniego scenariusza rozmowy uwzględniającego reguły wnioskowania, co znacznie poprawia skuteczność udzielania poprawnych odpowiedzi przez boty. Konkluzje te odnoszą się zarówno do kanału głosowego jak również do kanału tekstowego. Bardzo istotny jest również fakt, że proponowane metody sprawdziły się dla obu wybranych platform NLU (jednej komercyjnej, a drugiej Open Source). Zastosowanie standardowych mechanizmów (m.in. fraz, encji, intencji, akcji bota) otwiera możliwości integracji opracowanych metod *IAT* i F_{AE} z innymi platformami NLU, co stanowi duży potencjał do powszechnego ich wykorzystywania.

W branży CC coraz większe znaczenie ma wysoki poziom obsługi klienta. W ostatnich latach do bezpośredniej obsługi coraz częściej wykorzystywane są rozwiązania inteligentnych voicebotów oraz chatbotów, co w dużej mierze eliminuje czynnik ludzki. Jednym z podstawowych zadań bota jest poprawne rozpoznawanie intencji klienta, które w wielu przypadkach zależą również od emocji towarzyszących rozmówcy. W wielu kampaniach prowadzonych przez CC prawidłowe określenie emocji może mieć wręcz kluczowe znaczenie dla wysokiej jakości obsługi. Zaliczyć możemy do nich między innymi: windykację, likwidację szkód ubezpieczeniowych, przeprowadzanie wszelkiego rodzaju ankiet satysfakcji. Tego typu kampanie są bardzo skuteczne, gdy obsługuje je człowiek, dla którego właściwe rozumienie treści wypowiadanych emocjonalnie jest procesem naturalnym. Dla voicebotów oraz chatbotów poprawna obsługa kampanii, w których rozmowy zwykle nacechowane są wieloma skrajnymi emocjami stanowi duże wyzwanie. Zaproponowana przez autora metoda F_{AE} ma na celu pomóc w rozwiązywaniu tych problemów.

W dalszych pracach nad rozwojem proponowanej tematyki przewidziano badania związane z budową oraz implementacją voicebotów oraz chatbotów przygotowanych docelowo dla różnych tematycznie kampanii CC. Przewiduje się również, że proponowane w pracy metody będą mogły być wykorzystywane w systemach CC przyszłości, w których jak wynika z przeglądu literatury [3] zintegrowane mogą zostać popularne obecnie technologie Internetu

rzeczy IoT (ang. Internet of Things) implementowane obecnie w wielu obszarach techniki [104]–[106]. Ponadto, przewiduje się, że znacznie większe znaczenie w obsłudze klienta zyskają technologie wideo [107], które mogłyby również wspomóc rozpoznawanie emocji.

W ramach komunikacji międzyludzkiej wgląd w osobliwości danej osoby służy do dostosowania sposobu, w jaki rozmawiamy z tą osobą: inaczej rozmawiamy z dzieckiem niż z dorosłym, inaczej z kobietą niż z mężczyzną, zmieniamy nasze strategie dyskursu, gdy zdamy sobie sprawę, że osoba, z którą rozmawiamy, jest pod presją czasu lub ma doświadczenie w danym temacie. Wszystkie tego typu udoskonalenia pozwalają na skuteczną komunikację i unikanie nieporozumień [71]. Uwzględnienie tych czynników w regułach wnioskowania metody F_{AE} rozpoznawania ukrytych intencji jest również elementem, który będzie warto zbadać w dalszych pracach badawczych i ocenić ich wpływ na zwiększenie skuteczności voicebotów i chatbotów.

W pracy przedstawiono badania realizowane w ramach projektu „EMOTICA AI – Inteligentny System Contact Center”, akronim: EMOTICA AI nr umowy: POIR.04.01.04-00-0079/19. Projekt współfinansowany przez Narodowe Centrum Badań i Rozwoju z Programu Operacyjnego Inteligentny Rozwój w ramach podziałania 4.1.4 – Projekty aplikacyjne.

Bibliografia

[1] R. Rijo, J. Varajao, and R. Gonçalves, “Contact center: Information systems design,” *Journal of Intelligent Manufacturing*, vol. 23, no. 3, pp. 497–515, Jun. 2012, doi: 10.1007/S10845-010-0389-0.

[2] M. Saberi, O. Khadeer Hussain, and E. Chang, “Past, present and future of contact centers: a literature review,” *Business Process Management Journal*, vol. 23, no. 3, pp. 574–597, 2017, doi: 10.1108/BPMJ-02-2015-0018.

[3] M. Plaza and L. Pawlik, “Influence of the Contact Center Systems Development on Key Performance Indicators,” *IEEE Access*, vol. 9, 2021, doi: 10.1109/ACCESS.2021.3066801.

[4] G. Suciu, A. Pasat, T. Uşurelu, and E. C. Popovici, “Social media cloud contact center using chatbots,” *Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering, LNICST*, vol. 283, pp. 437–442, 2019, doi: 10.1007/978-3-030-23976-3_39.

[5] D. A. Dillman *et al.*, “Response rate and measurement differences in mixed-mode surveys using mail, telephone, interactive voice response (IVR) and the Internet,” *Social Science Research*, vol. 38, no. 1, pp. 1–18, Mar. 2009, doi: 10.1016/J.SSRESEARCH.2008.03.007.

[6] C. B. Nordheim, A. Følstad, and C. A. Bjørkli, “An Initial Model of Trust in Chatbots for Customer Service—Findings from a Questionnaire Study,” *Interacting with Computers*, vol. 31, no. 3, May 2019, doi: 10.1093/iwc/iwz022.

[7] G. Daniel, J. Cabot, L. Deruelle, and M. Derras, “Xatkit: a Multimodal Low-Code Chatbot Development Framework,” *IEEE Access*, vol. 8, pp. 15332–15346, 2020, doi: 10.1109/ACCESS.2020.2966919.

[8] D. Yu and L. Deng, “Signals and Communication Technology Automatic Speech Recognition A Deep Learning Approach”, Accessed: Oct. 24, 2021. [Online]. Available: http://www.springer.com/series/4748

[9] S. Trott, M. Eppe, and J. Feldman, “Recognizing Intention from Natural Language: Clarification Dialog and Construction Grammar”, Accessed: Oct. 24, 2021. [Online]. Available: https://www.youtube.com/watch?v=

[10] N. Ilk, M. Brusco, and P. Goes, “Workforce management in omnichannel service centers with heterogeneous channel response urgencies,” *Decision Support Systems*, vol. 105, pp. 13–23, Jan. 2018, doi: 10.1016/J.DSS.2017.10.008.

[11] “Three Measurable Benefits of Video-Enabling Your Contact Center - Vidyo Blog.” https://blog.vidyo.com/customer-engagement/video-contact-center/ (accessed Oct. 24, 2021).

[12] R. Picek, D. Peras, and R. Mekovec, “Opportunities and challenges of applying omnichannel approach to contact center,” *2018 4th International Conference on Information Management, ICIM 2018*, pp. 231–235, Jun. 2018, doi: 10.1109/INFOMAN.2018.8392841.

[13] L. Kerkeni, Y. Serrestou, M. Mbarki, K. Raoof, M. Ali Mahjoub, and C. Cleder, "Automatic Speech Emotion Recognition Using Machine Learning," *Social Media and Machine Learning*, Feb. 2020, doi: 10.5772/INTECHOPEN.84856.

[14] J. T. H. Chung-How and D. R. Bull, "Robust H.263+ video for real-time internet applications," *IEEE International Conference on Image Processing*, vol. 3, 2000, doi: 10.1109/ICIP.2000.899496.

[15] A. Snyder, D. Sullivan, and B. Fernekees, "Best Practices Series THE FUTURE OF IVR IN CUSTOMER SUPPORT", Accessed: Oct. 24, 2021. [Online]. Available: www.genesys.com

[16] A. Khalid, A. Sarfaraz, S. Ahmed, and F. Malik, "Prevalence of Stress among Call Center Employees," *Pakistan Journal of Social & Clinical Psychology*, vol. 11, p. 58, Oct. 2013.

[17] Z. Liu, C. Long, X. Lu, Z. Hu, J. Zhang, and Y. Wang, "Which Channel to Ask My Question?: Personalized Customer Service Request Stream Routing Using Deep Reinforcement Learning," *IEEE Access*, vol. 7, 2019, doi: 10.1109/ACCESS.2019.2932047.

[18] L. Boonstra, "Introduction to Conversational AI," in *The Definitive Guide to Conversational AI with Dialogflow and Google Cloud*, Berkeley, CA: Apress, 2021. doi: 10.1007/978-1-4842-7014-1_1.

[19] M. Hrabí, "Call centres: going voice-first in the post-Covid world," *Biometric Technology Today*, vol. 2020, no. 8, pp. 10–12, 2020, doi: https://doi.org/10.1016/S0969-4765(20)30111-9.

[20] B. Ellway, "Design vs practice: How problematic call routing and rerouting restructures the call centre service system," *International Journal of Operations and Production Management*, vol. 36, no. 4, pp. 408–428, Apr. 2016, doi: 10.1108/IJOPM-11-2013-0487/FULL/XML.

[21] "Business Intelligence: 6th International Conference, CBI 2021, Beni Mellal ... - Google Books." https://books.google.pl/books?hl=en&lr=&id=6eAuEAAAQBAJ&oi=fnd&pg=PA415&dq=voicebot+call+center&ots=BjL58otnrs&sig=9el-4Utn6OEUYigdWaYUqL0JGwA&redir_esc=y#v=onepage&q&f=false (accessed Oct. 19, 2021).

[22] A. Pappu and A. Rudnicky, "Deploying speech interfaces to the masses," *International Conference on Intelligent User Interfaces, Proceedings IUI*, pp. 41–42, Mar. 2013, doi: 10.1145/2451176.2451189.

[23] tipi, "Implementing Voice Over IP Telephony in 2-1-1 Call Centers," 2003, Accessed: Oct. 24, 2021. [Online]. Available: http://www.utexas.edu/research/tipi

[24] P. Bahar, T. Bieschke, and H. Ney, "A Comparative Study on End-To-End Speech to Text Translation," *2019 IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2019 - Proceedings*, pp. 792–799, Dec. 2019, doi: 10.1109/ASRU46091.2019.9003774.

[25] R. Liu, B. Sisman, Y. Lin, and H. Li, “FastTalker: A neural text-to-speech architecture with shallow and group autoregression,” *Neural Networks*, vol. 141, pp. 306–314, Sep. 2021, doi: 10.1016/J.NEUNET.2021.04.016.

[26] D. Inupakutika *et al.*, “Integration of NLP and Speech-to-text Applications with Chatbots,” *Electronic Imaging*, vol. 2021, no. 3, pp. 35-1-35–6, Feb. 2021, doi: 10.2352/ISSN.2470-1173.2021.3.MOBMU-035.

[27] W. Xiong, L. Wu, F. Alleva, J. Droppo, X. Huang, and A. Stolcke, “The Microsoft 2017 Conversational Speech Recognition System,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5934–5938. doi: 10.1109/ICASSP.2018.8461870.

[28] F. Dernoncourt, T. Bui, and W. Chang, “A Framework for Speech Recognition Benchmarking”, Accessed: Oct. 19, 2021. [Online]. Available: https://www.microsoft.com/cognitive-services/en-us/speech-api,

[29] “Voicebot and Chatbot Design: Flexible conversational interfaces with Amazon ... - Rachel Batish - Google Books.” https://books.google.pl/books?hl=en&lr=&id=XSxxDwAAQBAJ&oi=fnd&pg=PP1&dq=%22Voicebot+and+Chatbot+Design%22&ots=Ra7lnAaX81&sig=Ki8PAPLX8a16e0lFLk3FZL6F1g0&redir_esc=y#v=onepage&q=%22Voicebot%20and%20Chatbot%20Design%22&f=false (accessed Oct. 19, 2021).

[30] “From Your Mouth to Your Screen, Transcribing Takes the Next Step - The New York Times.” https://www.nytimes.com/2019/10/02/technology/automatic-speech-transcription-ai.html (accessed Oct. 19, 2021).

[31] B. Smagowska, “Noise at Workplaces in the Call Center,” *Archives of Acoustics*, vol. 35, Oct. 2010, doi: 10.2478/v10168-010-0024-2.

[32] A. Narayanan *et al.*, “Toward Domain-Invariant Speech Recognition via Large Scale Training,” Oct. 2018, pp. 441–447. doi: 10.1109/SLT.2018.8639610.

[33] G. Suciu, A. Pasat, T. Uşurelu, and E.-C. Popovici, “Social Media Cloud Contact Center Using Chatbots,” 2019. doi: 10.1007/978-3-030-23976-3_39.

[34] “Speech to Text – Audio to Text Translation | Microsoft Azure.” https://azure.microsoft.com/en-us/services/cognitive-services/speech-to-text/ (accessed Oct. 19, 2021).

[35] “Speech-to-Text Accuracy Benchmark - June 2021.” https://www.voicegain.ai/post/speech-to-text-benchmark-june-2021 (accessed Oct. 19, 2021).

[36] “Contact Center AI | Intelligent Contact Center Automation | Nuance.” https://www.nuance.com/omni-channel-customer-engagement/contact-center-ai.html (accessed Oct. 19, 2021).

[37] J. Kodish-Wachs, E. Agassi, P. Kenny, and J. M. Overhage, “A systematic comparison of contemporary automatic speech recognition engines for conversational clinical

speech," *AMIA Annual Symposium Proceedings*, vol. 2018, p. 683, 2018, Accessed: Dec. 06, 2021. [Online]. Available: /pmc/articles/PMC6371385/

[38] "Speech-to-Text: Automatic Speech Recognition | Google Cloud." https://cloud.google.com/speech-to-text (accessed Oct. 19, 2021).

[39] "VoiceLab.ai ASR Automatic Speech Recognition." https://voicelab.ai/#technologies (accessed Oct. 19, 2021).

[40] "GitHub - techmo-pl/dictation-client: Techmo Dictation Automatic-Speech-Recognition (ASR) gRPC client." https://github.com/techmo-pl/dictation-client (accessed Oct. 19, 2021).

[41] "Primespeech ASR Server™ Telecom - Speech technologies ready to prime." https://primespeech.pl/oferta/primespeech-asr-server-telecom/ (accessed Oct. 19, 2021).

[42] "GitHub - mozilla/DeepSpeech: DeepSpeech is an open source embedded (offline, on-device) speech-to-text engine which can run in real time on devices ranging from a Raspberry Pi 4 to high power GPU servers." https://github.com/mozilla/DeepSpeech (accessed Oct. 19, 2021).

[43] "Amazon Transcribe – Speech to Text - AWS." https://aws.amazon.com/transcribe/ (accessed Oct. 19, 2021).

[44] "Wav2letter." https://ai.facebook.com/tools/wav2letter/ (accessed Oct. 19, 2021).

[45] "IBM Watson - przegląd | IBM Cloud." https://www.ibm.com/pl-pl/cloud/watson-speech-to-text (accessed Oct. 19, 2021).

[46] "Automatic Speech Recognition (ASR) Systems Compared | by Sciforce | Sciforce | Medium." https://medium.com/sciforce/automatic-speech-recognition-asr-systems-compared-6ad5e54fd65f (accessed Oct. 19, 2021).

[47] "Transcription Services Showdown: AWS vs. Google vs. IBM vs. Nuance." https://armedia.com/blog/transcription-services-aws-google-ibm-nuance/ (accessed Oct. 19, 2021).

[48] "Dragon speech recognition Nuance Dragon Software Developer Kits," 2016, Accessed: Oct. 19, 2021. [Online]. Available: www.nuance.com.

[49] "Offline ASR : Which solutions are available today? – Vivoka." https://vivoka.com/benchmark-comparison-embedded-asr/ (accessed Oct. 19, 2021).

[50] A. Hannun *et al.*, "Deep Speech: Scaling up end-to-end speech recognition," Dec. 2014, Accessed: Oct. 19, 2021. [Online]. Available: https://arxiv.org/abs/1412.5567v2

[51] Nate Drake, "Best cloud computing services of 2021: for Digital Transformation." https://www.techradar.com/best/best-cloud-computing-services (accessed Jan. 09, 2021).

[52] A. Singh, K. Ramasubramanian, and S. Shivam, "Introduction to Microsoft Bot, RASA, and Google Dialogflow," in *Building an Enterprise Chatbot*, Berkeley, CA: Apress, 2019. doi: 10.1007/978-1-4842-5034-1_7.

[53] “IBM Watson Assistant - Visual Chatbot Builder | IBM.” https://www.ibm.com/cloud/watson-assistant/visual-builder (accessed Oct. 19, 2021).

[54] “Interactions with the API | Dialogflow CX | Google Cloud.” https://cloud.google.com/dialogflow/cx/docs/quick/api (accessed Oct. 19, 2021).

[55] A. Abdellatif, K. Badran, D. Costa, and E. Shihab, “A Comparison of Natural Language Understanding Platforms for Chatbots in Software Engineering,” *IEEE Transactions on Software Engineering*, p. 1, 2021, doi: 10.1109/TSE.2021.3078384.

[56] V. J. J. Flores, O. Juan, J. Carlos, and J. Ubaldo, “Performance Comparison of Natural Language Understanding Engines in the Educational Domain,” *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 8, 2020, doi: 10.14569/ijacsa.2020.0110892.

[57] “Language Understanding (LUIS) Overview - Azure Cognitive Services | Microsoft Docs.” https://docs.microsoft.com/en-us/azure/cognitive-services/luis/what-is-luis (accessed Oct. 19, 2021).

[58] “Dialogflow | Google Cloud.” https://cloud.google.com/dialogflow (accessed Oct. 19, 2021).

[59] “Wit.ai.” https://wit.ai/ (accessed Oct. 19, 2021).

[60] “Intelligent Virtual Agent - IBM Watson Assistant | IBM.” https://www.ibm.com/cloud/watson-assistant (accessed Oct. 19, 2021).

[61] “VoiceLab.ai Cognitive Automation.” https://voicelab.ai/#products (accessed Oct. 19, 2021).

[62] “Open source conversational AI | Rasa.” https://rasa.com/ (accessed Oct. 19, 2021).

[63] “Stack Overflow - Where Developers Learn, Share, & Build Careers.” https://stackoverflow.com/ (accessed Oct. 19, 2021).

[64] Y. Sasaki and R. Fellow, “The truth of the F-measure,” 2007.

[65] “Natural Language API Basics | Google Cloud.” https://cloud.google.com/natural-language/docs/basics#sentiment_analysis (accessed Oct. 19, 2021).

[66] “Language reference | Dialogflow ES | Google Cloud.” https://cloud.google.com/dialogflow/es/docs/reference/language (accessed Oct. 19, 2021).

[67] Y. Ma, H. Drewes, and A. Butz, “Fake Moods: Can Users Trick an Emotion-Aware VoiceBot?,” May 2021. doi: 10.1145/3411763.3451744.

[68] A. Ghandeharioun, D. McDuff, M. Czerwinski, and K. Rowan, “EMMA: An Emotion-Aware Wellbeing Chatbot,” Sep. 2019. doi: 10.1109/ACII.2019.8925455.

[69] K. J. Oh, D. K. Lee, B. S. Ko, J. Hyeon, and H. J. Choi, “Empathy Bot: Conversational service for psychiatric counseling with chat assistant,” *Studies in Health Technology and Informatics*, vol. 245, p. 1235, 2017, doi: 10.3233/978-1-61499-830-3-1235.

[70] S. Ramakrishnan and I. M. M. el Emary, “Speech emotion recognition approaches in human computer interaction,” *Telecommunication Systems*, vol. 52, no. 3, Mar. 2013, doi: 10.1007/s11235-011-9624-z.

[71] T. S. Polzin and A. Waibel, “EMOTION-SENSITIVE HUMAN-COMPUTER INTERFACES”, Accessed: Oct. 19, 2021. [Online]. Available: http://www.iscaaspeech.org/archive

[72] “Thomas Polzin - Google Scholar.” https://scholar.google.com/citations?user=BpMwy64AAAAJ&hl=pl&oi=sra (accessed Oct. 19, 2021).

[73] “Alexander Waibel - Google Scholar.” https://scholar.google.com/citations?user=uHwTzpYAAAAJ&hl=pl&oi=ao (accessed Oct. 19, 2021).

[74] “Vokaturi emotion… | Understand the emotion in a speaker’s voice.” https://vokaturi.com/ (accessed Oct. 19, 2021).

[75] M. Mostafavi and M. D. Porter, “How emoji and word embedding helps to unveil emotional transitions during online messaging,” *15th Annual IEEE International Systems Conference, SysCon 2021 - Proceedings*, Apr. 2021, doi: 10.1109/SYSCON48628.2021.9447137.

[76] M. Karna, D. S. Juliet, and R. C. Joy, “Deep learning based Text Emotion Recognition for Chatbot applications,” *Proceedings of the 4th International Conference on Trends in Electronics and Informatics, ICOEI 2020*, pp. 988–993, Jun. 2020, doi: 10.1109/ICOEI48184.2020.9142879.

[77] C. Chen *et al.*, “AntProphet: an Intention Mining System behind Alipay’s Intelligent Customer Service Bot,” Aug. 2019. doi: 10.24963/ijcai.2019/935.

[78] J. Zhang, Y. Xie, F. Yu, D. Soukal, and W. Lee, “Intention and Origination: An Inside Look at Large-Scale Bot Queries,” 2013.

[79] V. Gupta, N. Garg, and T. Gupta, “Search Bot: Search Intention Based Filtering Using Decision Tree Based Technique,” Feb. 2012. doi: 10.1109/ISMS.2012.78.

[80] J. Tian, Z. Tu, Z. Wang, X. Xu, and M. Liu, “User Intention Recognition and Requirement Elicitation Method for Conversational AI Services,” Oct. 2020. doi: 10.1109/ICWS49710.2020.00042.

[81] R. R. Sehgal, S. Agarwal, and G. Raj, “Interactive Voice Response using Sentiment Analysis in Automatic Speech Recognition Systems,” in *2018 International Conference on Advances in Computing and Communication Engineering (ICACCE)*, 2018, pp. 213–218. doi: 10.1109/ICACCE.2018.8441741.

[82] “Artykuł o rozpoznawaniu emocji (do opublikowania)”.

[83] K. Ohta, R. Nishimura, and N. Kitaoka, “Response type selection for chat-like spoken dialog systems based on LSTM and multi-task learning,” *Speech Communication*, vol. 133, Oct. 2021, doi: 10.1016/j.specom.2021.07.003.

[84] “Macierz błędów, raport z klasyfikacji, dokładność, czułość i precyzja.” https://www.statystyczny.pl/macierz-bledow-raport-dokladnosc-czulosc-precyzja/ (accessed Oct. 23, 2021).

[85] N. Zeghidour, N. Usunier, G. Synnaeve, R. Collobert, and E. Dupoux, “End-to-End Speech Recognition from the Raw Waveform,” Oct. 2018, pp. 781–785. doi: 10.21437/Interspeech.2018-2414.

[86] Y.-Y. Wang, A. Acero, and C. Chelba, “Is word error rate a good indicator for spoken language understanding accuracy,” in *2003 IEEE Workshop on Automatic Speech Recognition and Understanding (IEEE Cat. No.03EX721)*, 2003, pp. 577–582. doi: 10.1109/ASRU.2003.1318504.

[87] A. Zgank and Z. Kacic, “Predicting the Acoustic Confusability between Words for a Speech Recognition System using Levenshtein Distance,” *Electronics and Electrical Engineering*, vol. 18, no. 8, Oct. 2012, doi: 10.5755/j01.eee.18.8.2628.

[88] “Word error rate - Wikipedia.” https://en.wikipedia.org/wiki/Word_error_rate (accessed Oct. 19, 2021).

[89] “2.3 Computing error rates - Text Digitisation.” https://sites.google.com/site/textdigitisation/qualitymeasures/computingerrorrates (accessed Oct. 19, 2021).

[90] “Custom Speech overview - Speech service - Azure Cognitive Services | Microsoft Docs.” https://docs.microsoft.com/en-US/azure/cognitive-services/speech-service/custom-speech-overview (accessed Oct. 19, 2021).

[91] “Improve transcription results with speech adaptation.” https://cloud.google.com/speech-to-text/docs/speech-adaptation (accessed Oct. 19, 2021).

[92] I. Ali, M. Asif, M. Shahbaz, A. Khalid, M. Rehman, and A. Guergachi, “Text Categorization Approach for Secure Design Pattern Selection Using Software Requirement Specification,” *IEEE Access*, vol. 6, 2018, doi: 10.1109/ACCESS.2018.2883077.

[93] M. Wolí, M. Miłkowski, M. Ogrodniczuk, A. Przepiórkowski, and Ł. Szałkiewicz, “PoliMorf: a (not so) new open morphological dictionary for Polish”, Accessed: Oct. 19, 2021. [Online]. Available: http://nkjp.pl

[94] “GitHub - morfologik/morfologik-stemming: Tools for finite state automata construction and dictionary-based morphological dictionaries. Includes Polish stemming dictionary.” https://github.com/morfologik/morfologik-stemming (accessed Oct. 19, 2021).

[95] M. Plaza, L. Pawlik, and S. Deniziak, “Call Transcription Methodology for Contact Center Systems,” *IEEE Access*, vol. 9, 2021, doi: 10.1109/ACCESS.2021.3102502.

[96] “Ocenianie i poprawianie dokładności usługi Custom Speech — usługa rozpoznawania mowy - Azure Cognitive Services | Microsoft Docs.” https://docs.microsoft.com/pl-

pl/azure/cognitive-services/speech-service/how-to-custom-speech-evaluate-data (accessed Dec. 12, 2021).

[97] “Artykuł o wyborze konkretnych emocji z psychologami i lingwistą (do opublikowania)”.

[98] J. M. Roche, B. Peters, and R. Dale, “‘Your Tone Says It All’: The processing and interpretation of affective language,” *Speech Communication*, vol. 66, pp. 47–64, 2015, doi: https://doi.org/10.1016/j.specom.2014.07.004.

[99] “Entities | Dialogflow CX | Google Cloud.” https://cloud.google.com/dialogflow/cx/docs/concept/entity (accessed Oct. 19, 2021).

[100] “Training Data Format.” https://rasa.com/docs/rasa/training-data-format#entities (accessed Oct. 20, 2021).

[101] “Conversational AI The next wave of customer and employee experiences.” https://www2.deloitte.com/content/dam/Deloitte/au/Documents/strategy/au-deloitte-conversational-ai.pdf (accessed Oct. 19, 2021).

[102] H. S. Cheang and M. D. Pell, “The sound of sarcasm,” *Speech Communication*, vol. 50, no. 5, May 2008, doi: 10.1016/j.specom.2007.11.003.

[103] M. B. Akçay and K. Oğuz, “Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers,” *Speech Communication*, vol. 116, Jan. 2020, doi: 10.1016/j.specom.2019.12.001.

[104] M. Płaza, R. Belka, and Z. Szcześniak, “TOWARDS A DIFFERENT WORLD – ON THE POTENTIAL OF THE INTERNET OF EVERYTHING,” *Informatyka Automatyka Pomiary w Gospodarce i Ochronie Środowiska*, vol. 9, pp. 8–11, Oct. 2019, doi: 10.5604/01.3001.0013.2539.

[105] P. Pięta, S. Deniziak, R. Belka, M. Płaza, and M. Płaza, “Multi-domain model for simulating smart IoT-based theme parks,” Oct. 2018, p. 199. doi: 10.1117/12.2501659.

[106] R. Belka *et al.*, “Integrated visitor support system for tourism industry based on IoT technologies,” Oct. 2018, p. 5. doi: 10.1117/12.2326403.

[107] M. Królikowski, M. Płaza, and Z. Szcześniak, “Chosen sources of signal interference in HD-TVI technology,” Aug. 2017. doi: 10.1117/12.2280534.

Spis tabel

Spis rysunków