Obraz zawierający tekst, Czcionka, logo, Grafika

Opis wygenerowany automatycznie

inż. Marta Glanowska

Nr albumu: 244815

PRACA DYPLOMOWA MAGISTERSKA

na kierunku Sztuczna Inteligencja i Uczenie Maszynowe

**Klasyfikacja typu osobowości**

**na podstawie postów internetowych**

Instytut Informatyki Stosowanej

Promotor: dr inż. Robert Susik

Łódź, 2024

**Streszczenie**

dodać

Słowa kluczowe:

**Abstract**

dodać

Keywords:

# Spis treści

[Spis treści 5](#_Toc160312659)

[1 Wprowadzenie 8](#_Toc160312660)

[2 Cel i zakres pracy 10](#_Toc160312661)

[2.1 Cel i zakres pracy 10](#_Toc160312662)

[2.2 Model osobowości MBTI 10](#_Toc160312663)

[2.3 Zbiór danych 12](#_Toc160312664)

[3 Przegląd literatury 14](#_Toc160312665)

[4 Wstęp teoretyczny 16](#_Toc160312666)

[4.1 Przetwarzanie języka naturalnego 16](#_Toc160312669)

[4.2 Przygotowanie danych tekstowych 16](#_Toc160312670)

[4.3 Ekstrakcja dodatkowych cech 17](#_Toc160312671)

[4.4 Reprezentacja numeryczna tekstu 17](#_Toc160312672)

[4.4.1 Metody uwzględniające kolejność wyrazów 18](#_Toc160312673)

[4.4.2 Metody statystyczne 19](#_Toc160312674)

[4.5 Klasyczne modele uczenia maszynowego 19](#_Toc160312675)

[4.5.1 Wielomianowa regresja logistyczna 19](#_Toc160312676)

[4.5.2 Liniowy klasyfikator SVM 20](#_Toc160312677)

[4.5.3 Wielomianowy naiwny klasyfikator bayesowski 21](#_Toc160312678)

[4.5.4 Drzewo decyzyjne 22](#_Toc160312679)

[4.5.5 Ekstremalne wzmocnienie gradientu 22](#_Toc160312680)

[4.6 Rekurencyjne sieci neuronowe 22](#_Toc160312681)

[4.7 Mechanizm Attention 23](#_Toc160312682)

[4.8 Trening i ewaluacja modeli 24](#_Toc160312683)

[4.8.1 Walidacja krzyżowa 24](#_Toc160312684)

[4.8.2 Sprawdzian prosty 24](#_Toc160312685)

[4.8.3 Funkcja kosztu – entropia krzyżowa 24](#_Toc160312686)

[4.8.4 Miara dokładności 24](#_Toc160312687)

[5 Materiały i metody 26](#_Toc160312688)

[5.1 Wstępne przetwarzanie tekstu 26](#_Toc160312690)

[5.2 Dodatkowe charakterystyki numeryczne 28](#_Toc160312691)

[5.3 Podział danych 29](#_Toc160312692)

[5.4 Klasyfikacja za pomocą tradycyjnych modeli uczenia maszynowego 30](#_Toc160312693)

[5.4.1 Przygotowanie reprezentacji wektorowej 30](#_Toc160312694)

[5.4.2 Dobór modeli 31](#_Toc160312695)

[5.5 Modelowanie z zastosowaniem głębokich sieci neuronowych 34](#_Toc160312696)

[5.5.1 Kodowanie danych 34](#_Toc160312697)

[5.5.2 Sieć rekurencyjna z warstwą GRU 35](#_Toc160312698)

[6 Wyniki 37](#_Toc160312699)

[6.1 Wieloklasowość 37](#_Toc160312704)

[7 Podsumowanie 38](#_Toc160312705)

[8 Roboczy generator schematów 39](#_Toc160312710)

[Spis tabel 40](#_Toc160312711)

[Spis rysunków 41](#_Toc160312712)

[Wykaz symboli i oznaczeń 42](#_Toc160312713)

[Wykaz używanych skrótów 43](#_Toc160312714)

[Bibliografia 44](#_Toc160312719)

# Wprowadzenie

Kwestionariusz Myers-Briggs powstał w połowie ubiegłego wieku i jest jednym z najbardziej rozpowszechnionych modeli osobowości w biznesie, szczególnie w środowisku rekruterów, a także w szeroko pojętym doradztwie i coachingu. Przez lata był agresywnie reklamowany, aż na dobre osadził się w codziennych praktykach biznesowych, a także przyjął się w kulturze masowej jako sposób definicji i opisu natury poszczególnych jednostek. Mimo wątpliwości psychologów co do jego skuteczności, model ten wciąż znajduje szerokie zastosowanie we współczesnym świecie, gdyż bardzo często badani wspomnianym testem deklarują, że odnajdują się w opisie osobowości, do którego zostali zaklasyfikowali, tym samym czują się rozumiani. Tagi definiujące użytkownika z użyciem typów MBTI pojawiły się nawet w niektórych społecznościach platformy Reddit[[1]](#footnote-1), a także zostały wprowadzone do popularnych aplikacji randkowych, w celu poszukiwania jeszcze większej kompatybilności między potencjalnymi partnerami. Jedna z najbardziej popularnych stron internetowych[[2]](#footnote-2) oferujących darmowy kwestionariusz Myers-Briggs jest łatwo dostępna, atrakcyjna wizualnie, a ponadto oferuje test w wielu językach, co daje wspomnianemu modelowi znaczną przewagę nad innymi testami osobowości.

W tej pracy postawiono pytanie badawcze, czy przy użyciu modelu sztucznej inteligencji można ułatwić proces definiowania typu osobowości, oszczędzając dzięki temu czas poświęcany na wykonanie standardowego kwestionariusza. Zmiana źródła danych z odpowiedzi zaznaczanych w teście na treść postów internetowych ma dodatkową zaletę – ograniczenie lustrzanego efektu, kiedy opis danego typu zawiera informacje pochodzące z zaznaczonych odpowiedzi, przez co wydaje się być jeszcze wierniejszy. Jest to istotny aspekt, tym bardziej, że wyniki testu potrafią wahać się w zależności od okoliczności i samopoczucia danej osoby w chwili wykonywania kwestionariusza.

Ta praca ma szansę przysłużyć się do pełnego wykorzystania potencjału modelu MBTI. Skuteczność eksperymentu oznaczała zwiększenie autentyczności wyników wraz ze spotęgowaniem ich stałości, przy większej tolerancji na czynniki zewnętrzne wpływające na daną jednostkę.

Stworzony model sztucznej inteligencji może zostać wprowadzony do systemów rekrutacyjnych, by przeprowadzać wstępną ocenę osobowości kandydata na podstawie próbek tekstu, przykładowo pozyskanych z jego profesjonalnego profilu zawodowego. Takie rozpoznanie może zostać potwierdzone przez kwestionariusz na dalszym etapie rekrutacji, oszczędzając jednak czas obu stronom we wstępnym jej stadium. Należy jednak pamiętać, że kwestionariusze osobowości, chociaż mogą być pomocne w scharakteryzowaniu danego kandydata, nie powinny być czynnikiem kluczowym dla podejmowania decyzji o zatrudnieniu.

Model może być także wsparciem dla początkujących przedsiębiorców lub pracowników zastanawiających się nad swoimi naturalnymi rolami w grupie, a także rozpoznającymi swoje mocne i słabe strony. System może również zaistnieć jako wtyczka do popularnych portali społecznościowych, na których ludzie przez lata generują dziesiątki, a nawet setki postów. Takie narzędzie samorozwojowe za sprawą jednego kliknięcia byłoby w stanie określić typ osobowości danego użytkownika, a tym samym przybliżyć mu informacje na temat jego codziennych motywacji, czy sposobu komunikacji, a więc przysłużyć się do wykorzystania jego potencjału, a ponadto poprawy jego relacji z bliskimi ludźmi.

# Cel i zakres pracy

## Cel i zakres pracy

Celem pracy jest przegląd literatury, opracowanie modelu klasyfikacji, przeprowadzenie badań i analiza wyników. Zakres pracy obejmuje eksplorację możliwości zastosowania metod sztucznej inteligencji w klasyfikacji typu osobowości na podstawie postów internetowych oraz ewaluacja uzyskanych wyników.

[Na jakich efektach się skupiałam, co chciałam osiągnąć, a co nie było istotne? Na czym pracowałam (ogólnie) – zbiór danych, modele ogólnikowo.]

## Model osobowości MBTI

Powstanie testu MBTI ma swoje korzenie w teorii osobowości szwajcarskiego psychiatry i psychologa Carla Gustava Junga[[3]](#footnote-3). Jego koncepcja została rozwinięta w pierwszej połowie XX wieku przez amerykańską pisarkę i badaczkę Katharine Cook Briggs oraz jej córkę Isabel Briggs Myers, a dwadzieścia lat ich obserwacji doprowadziły ostatecznie do powstania kwestionariusza MBTI. Rzetelność i ważność modelu zostały potwierdzone i opisane w raporcie The Myers-Briggs Company[[4]](#footnote-4).

Klasyfikacja MBTI opiera się na określeniu wartości dla czterech par kategorii[[5]](#footnote-5), opisanych w *Tabeli 1*.

Tabela 1 – Opis par kategorii modelu MBTI (Źródło: opracowanie własne)

|  |  |  |
| --- | --- | --- |
|  | Para kategorii | |
| Poziom 1 – Źródło czerpania energii i komunikacja. | **Ekstrawersja (E)**  Koncentracja na świecie zewnętrznym. Komunikacja poprzez rozmowę. Towarzyskość. | **Introwersja (I)**  Koncentracja na świecie wewnętrznym. Komunikacja poprzez pismo. Powściągliwość i nieufność. |
| Poziom 2 – Sposób zbierania informacji i operowania nimi. | **Poznanie (S)**  Zorientowanie na fakty. Rzeczowość. Zapamiętywanie szczegółów. Praktyka i doświadczenie. | **Intuicja (N)**  Zorientowanie na możliwości. Kreatywność. Zapamiętywanie ogólnego obrazu. Analiza teoretyczna i inspiracja. |
| Poziom 3 – Motywacja podejmowania decyzji. | **Myślenie (T)**  Rozum ponad serce.  Rachunek „za i przeciw”.  Sprawiedliwość i logika.  Skupienie na osiągnięciu celu. | **Odczuwanie (F)**  Serce ponad rozum.  Empatia i harmonia w grupie.  Skupienie na byciu docenianym przez innych. |
| Poziom 4 – Styl życia i pracy. | **Osądzanie (J)**  Systematyczność. Planowanie. Sztywne ramy. | **Obserwacja (P)**  Spontaniczność. Swoboda.  Otwartość na modyfikacje. |

Istnieje 16 typów osobowości modelu MBTI. Czteroelementowa nazwa każdego typu jest prostym złożeniem liter z każdej z czterech par, która bardziej odpowiada naturze danej jednostki. W ten sposób, osoba o usposobieniu introwertyka (I), zorientowana na fakty i rzeczowość (S), kierująca się empatią i uczuciami (F), charakteryzująca się swobodnym stylem życia i elastycznością (P) najprawdopodobniej zostanie zaklasyfikowana do typu ISFP.

Czasami stosuje się koncentrację 16 typów w cztery grupy, aby sprawniej zorientować się w charakterystyce ich natury i pełnionych naturalnie funkcjach, są to: dążący do celu i pomysłowi Analitycy (INTJ, INTP, ENTP, ENTJ), kreatywni i zorientowani na ludzi Dyplomaci (INFJ, INFP, ENFJ, ENFP), praktyczni i skuteczni Strażnicy (ISTJ, ISFJ, ESTJ, ESFJ) oraz otwarci i spontaniczni Odkrywcy (ISTP, ISFP, ESTP, ESFP).

Wadą systemu MBTI niewątpliwie jest brak wartości pośrednich dla poszczególnych kategorii. Na przykład, w tej klasyfikacji badany nie może zostać uznany za ambiwertyka, czyli osobę posiadającą cechy z pogranicza ekstrawersji i introwersji. Może być to problematyczne w przypadku osób „uniwersalnych”, u których przychylność do jednego z biegunów w parze kategorii nie jest ewidentna. Tacy ludzie nierzadko są postrzegani jako elastyczni, potrafiący dostosować się do sytuacji, korzystający z różnych swoich atrybutów zależnie od potrzeb i okoliczności. Wykonanie wspomnianego testu może okazać się dla nich frustrującym doświadczeniem, ponieważ oscylując pośrodku dwóch biegunów kategorii, wynik klasyfikacji jest narażony na dużą zmienność wraz z kolejnymi wykonaniami testu.

## Zbiór danych

*(MBTI) Myers-Briggs Personality Type Dataset* to zbiór danych dostępny publicznie na platformie Kaggle[[6]](#footnote-6). Zawiera on 8675 wierszy, przy czym każdy z nich odpowiada ostatnim pięćdziesięciu postom udostępnionym przez danego użytkownika na forum PersonalityCafe[[7]](#footnote-7) wraz z charakteryzującym go typem osobowości w konwencji modelu MBTI. Zbiór odznacza się różnorodnością w kontekście opisywanych przez niego osób, a także samego stylu pisania. Zebranie danych z istniejącego forum, a nie w sztucznych warunkach, przyczyniło się do powstania próbek, które można określić jako autentyczne i odwzorowujące realne życie.

*Rys. 1* przedstawia dystrybucję próbek między klasami. Każdy z wierszy zawiera maksymalnie pięćdziesiąt postów internetowych. Zbiór nie jest zbalansowany – zebrano ponad czterdziestokrotnie więcej informacji na temat osób o typie INFP niż osób scharakteryzowanych jako ESTJ czy ESFJ. Cztery najwyższe słupki wykresu opisują liczbę postów należących do osób określonych jako introwertyczne. Ponadto, około 85% danych odnosi się do użytkowników forum o typie osobowości bazującym na intuicji (składowa N), a nie zmysłach (składowa S).

Rys. 1 – Dystrybucja próbek między klasami (Źródło: opracowanie własne)



# Przegląd literatury

Klasyfikacja typów osobowości modelu MBTI na podstawie próbek tekstu przy użyciu uczenia maszynowego jest zagadnieniem chętnie podejmowanym przez badaczy na przestrzeni ostatnich lat.

W 2017 roku Rayne Hernandez i Ian Scott Knight opisali problem badawczy w swoim artykule[[8]](#footnote-8) i odnotowali, że istniejące ówcześnie rozwiązania są nieliczne. Ich praca jako jedna z pierwszych bazowała na zbiorze *(MBTI) Myers-Briggs Personality Type Dataset.* Posty internetowe zostały wstępnie przetworzone, poszczególne słowa zakodowane w przestrzeni 50-wymiarowej za pomocą algorytmu GloVe, a następnie wymodelowane za pomocą rekurencyjnych sieci neuronowych z komórkami LSTM, osobno dla każdej pary kategorii. Wspomniany artykuł dał podwaliny kolejnym pracom i okazał się jednym z najczęściej cytowanych w późniejszych latach.

We wspomnianym roku, próbę modelowania dokładnie tego samego zbioru danych podjęli również Brandon Cui i Calvin Qi[[9]](#footnote-9). Ich działania były szerzej zakrojone – zajęli się zarówno problemem klasyfikacji szesnastoklasowej jak i tworzeniem czterech odrębnych klasyfikatorów binarnych odpowiedzialnych za kolejne pary kategorii. Autorzy postanowili skupić się na drugim z wymienionych podejść, deklarując niskie wartości dokładności klasyfikatorów wieloklasowych, wynoszących poniżej 25%. W tej pracy dane tekstowe zostały wyrażone numerycznie w głównej mierze za pomocą tradycyjnej techniki worka słów (ang. *Bag of Words*).

Bianca Antonio wraz z zespołem wykonali w 2018 roku wnikliwy projekt[[10]](#footnote-10), w którym udało im się wyczerpująco opisać sam zbiór danych, przetestować kilka metod próbkowania w celu zniwelowania braku zbalansowania na przestrzeni szesnastu klas, przeprowadzić szczegółową ekstrakcję cech, włączając zliczanie wystąpień poszczególnych elementów postów, analizę sentymentu i otagowanie części mowy. Podobnie jak w poprzednich pracach, autorzy nie zdecydowali się jednak na modelowanie wieloklasowe. Po treningu modeli regresji logistycznej, maszyn wektorów nośnych (ang. *Support Vector Machines*), ekstremalnego wzmocnienia gradientu (ang. *Extreme Gradient Boosting*) oraz lasów losowych, te ostatnie osiągnęły najwyższy iloczyn dokładności dla czterech modeli binarnych, jednak wynik ten był niewiele lepszy od pracy Brandona Cui i Calvina Qi9.

Podejście zaproponowane w 2020 roku przez inną grupę badaczy[[11]](#footnote-11) zawierało porównanie skuteczności maszyn wektorów nośnych, modeli regresji logistycznej, algorytmów k najbliższych sąsiadów (ang. *K-Nearest Neighbors*) oraz lasów losowych bazując na danych zakodowanych za pomocą techniki *Word2Vec ,* oraz dodatkowych cechach wybranych po analizie wskaźników korelacji Pearsona. W tym przypadku, las losowy osiągnął 100% dokładności, jednak również był trenowany odrębnie dla każdej pary kategorii.

Interesujące wnioski wysnuli Amirhosseini i Kazemian[[12]](#footnote-12), udowadniając, że metoda ekstremalnego wzmocnienia gradientu może być skuteczniejsza niż rekurencyjna sieć neuronowa w rozwiązaniu zadanego problemu. Istotnie, cząstkowe dokładności ich modeli były wyższe niż tych w artykule z 2017 roku8.

Jedna z nowszych prac[[13]](#footnote-13) podejmująca wspomniany problem badawczy odznacza się wyższą dokładnością klasyfikacji próbek tekstowych niż wymienione powyżej. W 2022 roku zespołowi badaczy w Tajlandii udało się osiągnąć ponad 40% dokładności zarówno dla naiwnych klasyfikatorów bayesowskich (ang. *Naive Bayes*) jak i maszyn wektorów nośnych. Przygotowanie danych do tych klasyfikatorów obejmowało zastosowanie techniki worka słów oraz ważenia częstością termów (ang. *Term Frequency-Inverse Document Frequency*). Ponadto, zaproponowana przez autorów dwukierunkowa sieć rekurencyjna z użyciem komórek LSTM osiągnęła niemal 50% skuteczności dla danych zakodowanych za pomocą tokenizera biblioteki *Keras*.

# Wstęp teoretyczny



## Przetwarzanie języka naturalnego

Przetwarzanie języka naturalnego (ang. *Natural Language Processing*) to dziedzina zajmująca się zastosowaniem technik analizy danych oraz modeli sztucznej inteligencji w celu generowania i rozumienia, a także rozwiązywania innych podobnych zadań opartych na tekście pisanym i mowie ludzkiej. Ta prężnie rozwijająca się w dzisiejszych czasach gałąź informatyki jest używana między innymi w systemach umożliwiających odczytanie tekstu ludzkim głosem, lub odwrotnie – narzędziach umożliwiającym głosowe wprowadzanie tekstu. Ma swoje odzwierciedlenie również w pracy wirtualnych doradców w postaci chatów na witrynach internetowych lub asystentów podczas rozmów telefonicznych z punktami usługowymi. Ponadto, jest ceniona jako fundamentalna składowa internetowych tłumaczy języków obcych, narzędzi do streszczania lub klasyfikacji dokumentów. Innym jej zastosowaniem może być automatyczna analiza sentymentu opinii konsumentów na forum.

## Przygotowanie danych tekstowych

Tekst jest wymagającym źródłem danych. Zrozumienie kontekstu słów wymaga przeanalizowania nie tylko ich kolejności i sąsiedztwa innych wyrazów w zdaniach, ale także sytuacji, w jakiej wypowiedź się pojawiła. Należy pamiętać, że w języku naturalnym występowanie żartów sytuacyjnych, ironii czy związków frazeologicznych jest zjawiskiem codziennym, co oprócz ubarwienia języka skutkuje też niemożnością jednoznacznego zdefiniowania znaczenia poszczególnych słów.

Celem wstępnego przetwarzania tekstu jest sprowadzenie go do zwartej formy zawierającej maksymalną ilość kluczowych informacji, przy czym w zależności od rozwiązywanego problemu zestaw najbardziej akcentowanych cech może być inny.

Jedną z podstawowych metod obróbki tekstu jest tokenizacja (ang. *Tokenization*). Polega ona na rozbiciu ciągłego tekstu na sekwencję tokenów, czyli mniejszych części. Często stosuje się podział na słowa, znaki, lub zdania. Równocześnie odrzuca się wybrane znaki interpunkcyjne.

W przygotowaniu tekstu istotne może być również zmniejszenie wszystkich liter – taki zabieg ma dwie zalety. Po pierwsze, automatycznie uszczupla rozmiar słownika. Ponadto, wskazuje modelowi zgodność dwóch wystąpień tego samego słowa, spośród których jedno pojawiło się na początku zdania, a więc było pisane wielką literą. Z drugiej strony, może to również generować problemy, na przykład w przypadku wyrazu „Jagoda”, gdzie wielkość litery wskazuje na fakt, że mowa jest o pewnej osobie, a nie o owocu.

Eliminacja słów, które same w sobie nic nie znaczą (ang. *Stopwords*) to kolejna popularna technika wstępnego przetwarzania tekstu. Istnieją zaimki, spójniki, partykuły i inne wyrazy, które występują w języku pisanym nieraz częściej niż pozostałe słowa, równocześnie nie niosąc ze sobą znaczącej informacji. Świadczy o tym fakt, że są one pomijane przez wyszukiwarki. Takie wyrazy usuwa się zazwyczaj przygotowując tekst do klasyfikacji, z kolei nie jest to zalecane w pracy z generatywnymi modelami uczenia maszynowego, ponieważ tworzone przez nie wypowiedzi nie będą brzmieć naturalnie. Listę usuwanych słów należy jednak zweryfikować, gdyż niektóre standardowo na niej umieszczane wyrazy mogą okazać się znaczące w kontekście specyficznych zadań – na przykład badania częstości występowania zaimków „ja” i „moje” w wykrywaniu osobowości narcystycznych.

Innym prostym pomysłem jest praca nad poszczególnymi elementami tekstu, na przykład usuwanie linków czy emotikon, oflagowywanie części mowy, wydobywanie ze słów ich rdzeni (ang. *Stemming*) lub lematów (ang. *Lemmatization*).

Wybór technik różni się od specyfiki problemu, nie istnieje jeden uniwersalny pomysł na przygotowanie danych tekstowych.

## Ekstrakcja dodatkowych cech

Oprócz modelowania samej wstępnie przetworzonej treści danego tekstu czy jego wycinków, można dodatkowo zdefiniować cechy opisujące numerycznie poszczególne próbki. Zależnie od problemu, inne charakterystyki mogą być znaczące. Fragment tekstu można zdefiniować zliczając występujące w nim poszczególne części mowy, wyznaczając jego długość bezwzględną lub mierzoną ilością zdań, przeprowadzając analizę sentymentu i przypisując wartość binarną, czy oznaczając emocje wiązane z poszczególnymi zwrotami.

## Reprezentacja numeryczna tekstu

Wyrażenie tekstu liczbowo jest jedną z kluczowych operacji przetwarzania wstępnego ze względu na to, że modele uczenia maszynowego są trenowane właśnie na reprezentacji numerycznej. Większość algorytmów opisanych poniżej jest dodatkowo specyficznymi ekstraktorami cech, ponieważ pozyskują one informacje na temat współwystępowania słów i ich częstości lub o kontekście.

### Metody uwzględniające kolejność wyrazów

Najbardziej podstawową techniką kodowania tekstu jest stworzenie słownika istniejących w danym zbiorze danych wyrazów, a następnie zamiana słów w próbkach według jego indeksów. W zależności od potrzeb, wielkość słownika można ograniczyć. Uzyskany ciąg liczb całkowitych reprezentujących poszczególne słowa nazywany jest ciągiem tokenów.

Mechanizm mapowania wyrazów zakodowanych tokenami na wielowymiarową reprezentację wektorową (ang. *Embedding*) jest natomiast pomysłem na rozszerzenie powyższej techniki o pewne informacje semantyczne. Ta technika ma dwie wyraźne zalety. Po pierwsze, można zauważyć, że w celu stworzenia binarnej reprezentacji próbek, zastosowanie kodowania one-hot przeprowadzonego na tokenach, mając na uwadze, że dla wielu tekstowych baz danych ich jest ich dziesiątki lub setki tysięcy, nie będzie optymalnym działaniem. Z pomocą przychodzi wówczas umiejętność przetworzenia informacji zawartej w tokenach w taki sposób, aby każde słowo zostało wyrażone jako punkt w przestrzeni o wcześniej jasno ustalonej, o wiele niższej liczbie wymiarów. Zmniejsza to ilość mocy obliczeniowej potrzebnej do samego wytrenowania modelu. Druga zaleta takiego podejścia, to możliwość uchwycenia w reprezentacji relacji zachodzących między wyrazami. Dzięki zrozumieniu kontekstu, słowa o podobnym znaczeniu semantycznym znajdują się w przestrzeni blisko siebie[[14]](#footnote-14).

Do tworzenia reprezentacji numerycznej słów można również zastosować pretrenowany model typu BERT[[15]](#footnote-15) będący transformerem zawierającym w swojej architekturze jedynie koder, z pominięciem dekodera. W tym przypadku w reprezentacji wektorowej poszczególnych słów kodowane są nie tylko podstawowe informacje o ich tokenach (ang. *Token Embeddings*), ale również oznaczenia fragmentów tekstu, z których pochodzą, w przypadku gdy zawierająca je próbka jest złączeniem kilku (ang. *Segment Embeddings*), a także ich pozycje w zawierających je próbkach (ang. *Position Embeddings*).

Powyższe metody są skuteczne w połączeniu z zastosowaniem rekurencyjnych sieci neuronowych, które następnie analizują otrzymane reprezentacje, a mając możliwość wychwycenia informacji o kolejności elementów sekwencji, badają strukturę zdań.

### Metody statystyczne

Inną często stosowaną metodą zamiany tekstu na formę liczbową jest technika worka słów (ang. *Bag of Words*) budująca uproszczoną reprezentację próbki na podstawie częstości występowania w niej poszczególnych wyrazów, jednak w odróżnieniu do poprzednio opisanych technik, nie uwzględnia ona ich kolejności. Z tego powodu nie jest odpowiednia, aby przygotować dane wejściowe do sieci rekurencyjnych i łączy się ją z tradycyjnymi modelami uczenia maszynowego.

Rozwinięciem idei worka słów jest koncepcja ważenia częstością termów, bardziej znana jako TF-IDF. To metoda statystyczna, w której kluczową rolę grają dwa czynniki – częstość występowania danego słowa w konkretnej próbce, a także jego ważność, czyli frekwencja występowania na przestrzeni wszystkich próbek. Algorytm nadaje wagi słowom – największą przypisuje rzadkim wyrazom, które w danej próbce wystąpiły często. Dopuszcza również traktowanie par, trójek lub większych zestawień słów, tak zwanych n-gramów, jako całości, co oznacza, że dla każdego n-gramu liczona jest wówczas pojedyncza reprezentacja. [wzór?]

## Klasyczne modele uczenia maszynowego

### Wielomianowa regresja logistyczna

Klasyczna regresja logistyczna to procedura dostosowana do problemów klasyfikacyjnych natury binarnej. [Wyjaśnić podstawy?]

[Regresja logistyczna przypisuje punktowi jego etykietę klasy w zależności od tego, czy leży on powyżej, czy poniżej linii separującej. Ponadto używa funkcji logit do przekształcenia odległości od punktu do linii na prawdopodobieństwo, że dana próbka należy do zidentyfikowanej klasy. Rozważania optymalizacyjne w regresji logistycznej polegają na minimalizowaniu sumy prawdopodobieństw błędnej klasyfikacji we wszystkich punktach.]

W przypadku wieloklasowości, wielomianowa odmiana tego modelu przewiduje prawdopodobieństwo dla każdej z etykiet, a dodatkowo posługuje się entropią krzyżową do obliczenia kosztu. [Entropia krzyżowa i model wieloklasowy + wzory]

[Parametr C + regularyzacja L2]

### Liniowy klasyfikator SVM

Celem liniowego klasyfikatora SVM (z ang. *Support Vector Machine*) nazywanego również maszyną wektorów nośnych jest zbudowanie maksymalnie dużego marginesu separującego w przestrzeni punkty należące do dwóch odrębnych klas.

Pierwszym krokiem budowania takiego modelu jest zdefiniowanie hiperpłaszczyzny za pomocą prostego równania liniowego , gdzie symbolizuje wektor współczynników, a oznacza wektor zmiennych niezależnych. Wokół niej należy następnie wyznaczyć dwie hiperpłaszczyzny równoległe, w tej samej odległości. Każda z nich będzie wyznaczała granicę występowania próbek konkretnej klasy, dzięki czemu powstanie wspomniany margines. Zakładając, że etykiety próbek pochodzą ze zbioru , margines zostaje opisany równaniami: oraz . W jego przestrzeni nie mogą występować żadne punkty, dlatego dla próbek należących do klasy pierwszej zachodzi nierówność , a dla pozostałych prawdą jest, że .[[16]](#footnote-16)

Próbki należące do zbioru i spełniające nierówność nazywane są wektorami nośnymi (ang. *support vectors*), ponieważ ich lokalizacja wyznacza położenie marginesu separującego. Celem algorytmu jest znalezienie takich współczynników hiperpłaszczyzny, aby margines był maksymalny. Wówczas skuteczność klasyfikacji jest większa, model rzadziej się myli.

Rys. 2 – Wyznaczanie hiperpłaszczyzny liniowego klasyfikatora SVM oraz maksymalizacja marginesu separującego (Źródło: <https://en.wikipedia.org/wiki/Support_vector_machine>)

Obraz zawierający tekst, zrzut ekranu, linia, diagram

Opis wygenerowany automatycznie

W przypadku klasyfikacji wieloklasowej, najczęściej stosowaną strategią dla tego modelu jest jeden-przeciw-pozostałym (ang. *one-versus-rest*), czyli konstrukcja tylu klasyfikatorów ile jest klas, nie uwzględniając jednej, a następnie analiza uzyskanych wyników dla jednoznacznej predykcji.

[Parametr C + Regularyzacja L2 + Squared hinge loss?]

Maszyna wektorów nośnych jest klasyfikatorem podobnym do regresji logistycznej, celem obu algorytmów jest optymalna separacja próbek, choć każdy z nich optymalizuje inne kryteria, aby to osiągnąć.

### Wielomianowy naiwny klasyfikator bayesowski

Naiwne klasyfikatory bayesowskie to rodzina klasyfikatorów probabilistycznych. Oznacza to, że przewidują one rozkład prawdopodobieństwa przynależności próbki do zestawu klas, zamiast proponować predykcję do pojedynczej klasy. Ze względu na założenie, że predyktory są parami niezależne, co rzadko jest możliwe w przypadku pracy na realnych danych, nazywa się je naiwnymi. Oparte są na twierdzeniu Bayesa *(1).*

(1)

W powyższym wzorze oznacza zmienną celu, a to zależny wektor cech. Zapis określa natomiast prawdopodobieństwo warunkowe, czyli w tym wypadku prawdopodobieństwo wystąpienia pod warunkiem zajścia .

Do zadań klasyfikacji tekstu, odpowiednim wariantem z tej rodziny modeli jest wielomianowy naiwny klasyfikator bayesowski. Stara się on parametryzować wielomianowy rozkład danych dla każdej klasy za pomocą wektorów . Oznaczenie symbolizuje liczbę cech, czyli w wypadku klasyfikacji tekstu – wielkość słownictwa. określa prawdopodobieństwo wystąpienia cechy w próbce należącej do klasy . Parametry są szacowane metodą liczenia częstości względnych *(2).*

(2)

W przedstawionym wzorze, wyraża ilość wystąpień cechy w próbkach klasy w zbiorze treningowym, natomiast jest całkowitą liczbą wystąpień wszystkich cech w próbkach klasy zbioru treningowego. Parametr wygładzania addytywnego został oznaczony literą .

### Drzewo decyzyjne

[Do analizy częstości]

### Ekstremalne wzmocnienie gradientu

[Gradient]

[Wzmocnienie gradientowe GBM]

[Co to znaczy „ekstremalne”, jakie są zalety?]

## Rekurencyjne sieci neuronowe

Co to są dane sekwencyjne

Gated Recurrent Unit - Cho et al. 2014.,

Batch + BatchNormalization

Dense + ReLU

Dropout

Softmax

optymalizacja Adam

## Mechanizm Attention

W 2017 roku, w szeroko później cytowanym artykule pod tytułem „*Attention Is All You Need*”[[17]](#footnote-18), w dosłownym tłumaczeniu „Attention to wszystko, czego potrzebujesz” autorzy przedstawili propozycję nowej przełomowej architektury sieci neuronowej typu transformer, której skuteczność oparta była przede wszystkim na istniejącym wówczas lecz niezbyt spopularyzowanym mechanizmie Attention.

Jak wskazuje nazwa, celem algorytmu jest znalezienie i podkreślenie cech próbki, na których należy się skupić podczas modelowania, poprzez przypisanie im wyższych wag. Spośród reprezentacji poszczególnych słów w próbce (ang. *Keys*), mechanizm wychwytuje te wysoko skorelowane z wybranym zapytaniem (ang. *Query*) kalkulując iloczyn skalarny tablic. Następnie aktywuje funkcję *Softmax*, dzięki której poszczególne prawdopodobieństwa sumują się do 1. Dzięki porównaniu z wagami opisującymi cechy znaczące dla rozwiązywanego zadania (ang. Value) skutkuje wzmocnieniem wag dla istotnych komponentów tekstu i osłabieniem wag dla mniej ważnych. [Softmax? Wzory?]

Autorzy wspomnieli również o drugim wariancie mechanizmu. Dopuszcza on równoległe działanie kilku modułów obliczeniowych, tak zwanych głów (ang. *Heads*). Takie podejście umożliwia równoczesne analizowanie kilku kontekstów, dzięki czemu możliwe jest jednoczesne wychwycenie kilku znaczeń danego słowa. Schematy obu architektur przedstawiono na *Rys. 3.*

Rys. 3 - Wizualizacja architektur mechanizmu Attention (Źródło: Ashish Vaswani i in., Attention Is All You Need”, 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 2023)

Obraz zawierający tekst, zrzut ekranu, diagram, linia

Opis wygenerowany automatycznie

## Trening i ewaluacja modeli

### Walidacja krzyżowa

Zastosowanie walidacji krzyżowej ma na celu ewaluację wybranej architektury modelu uczenia nadzorowanego poprzez przeprowadzenie kilkukrotnego szkolenia i testu, za każdym razem na podstawie innych zestawów próbek. Najpopularniejszym wariantem tej metody jest walidacja -krotna, która polega na rozbiciu zbioru danych na podzbiorów, a następnie przeprowadzeniu treningów wybranej architektury, za każdym razem traktując jako zbiór testowy inną spośród stworzonych grup. Ogólną jakość modelu można ocenić uśredniając wyniki takich klasyfikacji. Jako kryterium oceny skuteczności modelu zazwyczaj przyjmuje się miarę dokładności.

### Sprawdzian prosty

Jedno z najczęściej stosowanych podejść do trenowania modeli uczenia maszynowego polega na podziale danych na podzbiory treningowy i testowy, gdzie pojemność zbioru testowego wynosi zazwyczaj 30% wszystkich próbek lub mniej. Wykonywany jest wówczas tylko jeden trening modelu na zbiorze treningowym, natomiast ewaluacja odbywa się poprzez inferencję zbioru testowego, porównanie otrzymanych predykcji z realnymi etykietami i kalkulację pożądanych metryk.

### Funkcja kosztu – entropia krzyżowa

[Opisać]

[+ softmax loss?]

### Miara dokładności

Dokładność (ang. *Accuracy*) jest jedną z prostych metryk oceny skuteczności modelu klasyfikacyjnego. Jest to stosunek ilości poprawnie zaklasyfikowanych próbek do ilości wszystkich próbek. Należy pamiętać, że ta miara może nie być wystarczająca w przypadku zbioru o nierównomiernym rozkładzie danych na przestrzeni klas, ponieważ mimo otrzymania wysokiej wartości pomiaru, istnieje ryzyko, że model nie rozpoznaje prawidłowo żadnej próbki należącej do klasy rzadkiej.

# Materiały i metody



## Wstępne przetwarzanie tekstu

Rozważając problem klasyfikacji próbek tekstu w postaci postów internetowych do kilkunastu kategorii należy najpierw zdefiniować, jakie cechy wpisów mogą okazać się najbardziej istotne dla ich rozróżnienia między sobą i te właśnie atrybuty należy następnie wyeksponować. Jest to istotne, ponieważ ostatecznym celem projektu jest znalezienie indywidualnych schematów obecnych w postach osób będących w obrębie tego samego typu osobowości. Z perspektywy treningu modelu, ważne jest natomiast ograniczenie ilości mniej ważnych słów w bazie danych, ponieważ dzięki temu uczenie trwa krócej przy zachowaniu tej samej ilości informacji.

Początkowo, zmapowano typy osobowości wyrażone literami do wartości liczbowych z zakresu od 0 do 15 i dołączono dodatkową kolumnę z tymi wartościami do ramki danych powstałej po wczytaniu zbioru *(MBTI) Myers-Briggs Personality Type Dataset.* Poszczególne posty danego użytkownika, oddzielone wstępnie symbolami „|||” zostały złączone ze sobą i były później traktowane jako pojedyncza próbka.

Wszystkie litery zostały zmniejszone. W celu tokenizacji przetestowano dwa rozwiązania języka Python: standardowy *word\_tokenize[[18]](#footnote-19)* oraz *TweetTokenizer[[19]](#footnote-20)* , oba pochodzące z zestawu narzędzi języka naturalnego *NLTK*. W *Tabeli 2* przedstawiono porównanie ich działania dla przykładowych fragmentów tekstu. *TweetTokenizer* okazał się być lepiej przystosowany do pracy z postami internetowymi, ze względu na większą czułość na obecność emotikon w tekście, co w praktyce oznacza, że rzadziej rozdziela te charakterystyczne zestawienia znaków. Częściej niż *word\_tokenize* odseparowuje on natomiast złączenia losowych symboli z wyrazami („’Now”), co może mieć znaczenie w przypadku błędów w postach pisanych naprędce na forum. Można też zauważyć, że traktuje on linki jako spójną całość, co jest jego niewątpliwą zaletą. *TweetTokenizer* nie rozdziela również angielskich form skróconych („I’m”, ‘it’s’), co jest cechą neutralną – należy zastanowić się, jaka informacja na temat próbek może być decydująca, a jej wydobycie może przełożyć się na większą skuteczność późniejszego modelowania – separacja takich złączeń, pozostawienie ich w pierwotnej formie, a może nawet rozwinięcie wspomnianych skrótów. W każdym z tych podejść można odnaleźć inne charakterystyki – można, na przykład, analizować częstość użycia słów wskazujących na siebie przez poszczególne osoby, wnioskować o temperamencie lub stylu życia ludzi na postawie ich skłonności do używania krótszych form, albo skupić się na badaniu samej treści wypowiedzi po ujednoliceniu języka.

Tabela 2 – Porównanie tokenizacji tekstu przez word\_tokenizer i TweetTokenizer (Źródło: opracowanie własne)

|  |
| --- |
| **Fragment próbki 1** |
| ‘Now I'm interested. But too lazy to go research it, because it's time-consuming :( |
| Tokeny zwrócone przez *word\_tokenizer* po zmniejszeniu liter |
| "'now", 'i', "'m", 'interested', '.', 'but', 'too', 'lazy', 'to', 'go', 'research', 'it', ',', 'because', 'it', "'s", 'time-consuming', ':', '(' |
| Tokeny zwrócone przez *TweetTokenizer* po zmniejszeniu liter |
| "'", 'now', "i'm", 'interested', '.', 'but', 'too', 'lazy', 'to', 'go', 'research', 'it', ',', 'because', "it's", 'time-consuming', ':(' |
| **Fragment próbki 2** |
| http://www.youtube.com/watch?v=u8ejam5DP3E On repeat for most of today. |
| Tokeny zwrócone przez *word\_tokenizer* po zmniejszeniu liter |
| '//www.youtube.com/watch', '?', 'v=u8ejam5dp3e', 'on', 'repeat', 'for', 'most', 'of', 'today', '.' |
| Tokeny zwrócone przez *TweetTokenizer* po zmniejszeniu liter |
| 'http://www.youtube.com/watch?v=u8ejam5dp3e', 'on', 'repeat', 'for', 'most', 'of', 'today', '.' |

W przypadku opisywanej pracy angielskie formy skrócone zostały rozdzielone względem apostrofu, uznane za stosunkowo mało istotne dla badanego problemu klasyfikacji i zupełnie usunięte z próbek wraz z pozostałymi słowami, które same w sobie nic nie znaczą (ang. *Stopwords*). Cały proces wstępnego przetworzenia próbki został zwizualizowany w *Tabeli 3* na przykładowym fragmencie posta.

Tabela 3 – Proces wstępnej obróbki tekstu (Źródło: opracowanie własne)

|  |
| --- |
| **Fragment próbki** |
| 'You're fired. That's another silly misconception. |
| Zmniejszenie liter i tokenizacja z użyciem *TweetTokenizer* |
| "'", "you're", 'fired', '.', "that's", 'another', 'silly', 'misconception', '.' |
| Rozdzielenie form skróconych względem apostrofu |
| '', '', 'you', 're', 'fired', '.', 'that', 's', 'another', 'silly', 'misconception', '.' |
| Usunięcie słów bez znaczenia |
| '', '', 'fired', '.', 'another', 'silly', 'misconception', '.' |

## Dodatkowe charakterystyki numeryczne

Za pomocą operacji na wyrażeniach regularnych (ang. *RegEx*), w próbkach zostały zidentyfikowane i oflagowane linki (‘[LINK]’), emotikony min nie zawierające liczb oraz emotikony w kształcie serc, ze względu na ich częste użytkowanie w popkulturze jako symbole pozytywnych emocji (‘[EMOJI]’). Wśród znaków interpunkcyjnych wyróżniono wielokropek, znak zapytania, znak wykrzyknienia oraz kropkę ("[ELLIPSIS]", "[QUESTION\_MARK]", "[EXCLAMATION\_MARK]", "[PERIOD]"), jako te niosące potencjalnie największą informację. Inne symbole zostały usunięte ze zbioru. Oznaczono również samodzielne liczby (‘[NUMBER]’) oraz usunięto hashtagi, czyli słowa poprzedzone znakiem *#*, funkcjonujące jako słowa kluczowe podczas wyszukiwania i grupowania treści [[20]](#footnote-21). Na koniec, listy tokenów zostały oczyszczone z pustych elementów powstałych w trakcie procesu. W *Tabeli 4* przedstawiono przykłady fragmentów próbek przed i po wykonaniu całego opisanego procesu.

Tabela 4 – Przykłady oznaczania charakterystycznych elementów w próbkach (Źródło: opracowanie własne)

|  |  |
| --- | --- |
| Przed | '', '', 'http://www.youtube.com/watch?v=qsxhcwe3krw', 'http://41.media.tumblr.com/tumblr\_lfouy03pma1qa1rooo1\_500.jpg', 'enfp', 'intj', 'moments' |
| Po | '[LINK]', '[LINK]', 'enfp', 'intj', 'moments' |
| Przed | '', '', ',', 'draw', 'nails', '(', 'haha', ')', '.', 'done', 'professionals', 'nails', '.', 'yes', ',', 'gel', '.', 'mean', 'posted', 'done', 'nails', '?', 'awesome', '!' |
| Po | 'draw', 'nails', 'haha', '[PERIOD]', 'done', 'professionals', 'nails', '[PERIOD]', 'yes', 'gel', '[PERIOD]', 'mean', 'posted', 'done', 'nails', '[QUESTION\_MARK]', 'awesome', '[EXCLAMATION\_MARK]' |
| Przed | '', '', 'always', 'thought', 'tony', 'stark', 'entj', '...', '(' |
| Po | 'always', 'thought', 'tony', 'stark', 'entj', '[ELLIPSIS]' |
| Przed | '', '', 'got', '593', '.', 'read', 'enneagram', '953', ',', 'though', '.' |
| Po | 'got', '[NUMBER]', '[PERIOD]', 'read', 'enneagram', '[NUMBER]', 'though', '[PERIOD]' |
| Przed | '', '', 'thanks', 'jaydubs', 'miss', 'bingley', 'littledreamer', 'advices', 'important', '!', ':d' |
| Po | 'thanks', 'jaydubs', 'miss', 'bingley', 'littledreamer', 'advices', 'important', '[EXCLAMATION\_MARK]', '[EMOJI]' |

W celu orientacyjnego sprawdzenia, czy zbiór jest opisywalny garstką stworzonych dodatkowych cech, przekazano próbki wyrażone zawieranymi przez nie licznościami linków, emotikon, liczb oraz poszczególnych znaków interpunkcyjnych do binarnego drzewa decyzyjnego[[21]](#footnote-22), na którym następnie zastosowano walidację krzyżową. Wraz ze wzrostem głębokości drzewa, średnia dokładność malała. Z tego powodu określono maksymalną dopuszczalną głębokość drzewa decyzyjnego wartością 5. Uzyskany uśredniony wynik walidacji krzyżowej dla 10 podzbiorów wyniósł 22.84%, i jest on znacznie wyższy od poziomu losowego, który wynosi 6.25%. To potwierdza, że wyodrębnione cechy zawierają istotne informacje na temat poszczególnych typów osobowości, choć nie definiują ich jednoznacznie.

[Tabela średnich liczności lub wykres]

## Podział danych

Trening nadzorowanych modeli uczenia maszynowego należy przeprowadzić na pewnym wycinku danych, podczas gdy pozostała część próbek jest wówczas używana do ewaluacji jego skuteczności, a przy tym oceny optymalności wytrenowanych parametrów.

W wypadku tego projektu, zbiór danych został rozbity na podgrupę treningową i testową w stosunku 80:20. Podział nastąpił w sposób stratyfikowany (ang. *stratified split*) na podstawie zmiennej celu, co oznacza zachowanie oryginalnych proporcji pomiędzy klasami decyzyjnymi w każdej z podgrup. W tym celu zastosowano funkcję *train\_test\_split* biblioteki   
*scikit-learn[[22]](#footnote-23)*. Podzbiór treningowy wyniósł ponad 6900 próbek, natomiast testowy zawierał powyżej 1700 wierszy.

Posty należące do jednego użytkownika zostały uprzednio scalone w jeden wiersz, dlatego potencjalny problem znalezienia się wpisów tej samej osoby zarówno w zbiorze treningowym jak i testowym nie wystąpił. Taka sytuacja poddałaby w wątpliwość wyniki późniejszej ewaluacji i umiejętności przystosowania się modelu do zupełnie nowych danych.

## Klasyfikacja za pomocą tradycyjnych modeli uczenia maszynowego

### Przygotowanie reprezentacji wektorowej

Proces konwersji danych tekstowych do postaci wektorowej w przypadku tradycyjnych modeli klasyfikacyjnych obejmował zastosowanie konceptu ważenia częstością termów, czyli rozszerzonej metody worka słów.

W tym celu najpierw zastosowano technikę worka słów z modułu *sklearn.feature\_extraction.text*, czyli narzędzie *CountVectorizer[[23]](#footnote-24)*. Parametr *max\_features* definiujący rozważaną wielkość słownika ustawiono na wartość 5000. Wynik tej modyfikacji stanowił wejście do funkcji *TfidfTransformer[[24]](#footnote-25)* będącej cyfrową reprezentacją modyfikacji   
TF-IDF. Oba obiekty zostały dopasowane do danych treningowych metodą *fit\_transform*. Reprezentacja danych testowych została natomiast zbudowana przy użyciu metody *transform* na bazie informacji ze zbioru treningowego.

Podczas budowania worka słów najbardziej naturalnym podejściem jest przyjmowanie jako samodzielny element słownika każdego pojedynczego słowa. Istnieje jednak również możliwość traktowania jako jeden człon tak zwanych bigramów, czyli par wyrazów. Jest to uzasadnione podejście, ponieważ w języku naturalnym istnieją słowa, które często występują w parze. Dlatego w tej pracy przeprowadzono trzy różne próby treningu tradycyjnych architektur. Pierwsza z nich bazowała na worku słów zbudowanym z indywidualnych wyrazów, w drugiej słownik zawierał zarówno jedno- jak i dwuczłonowe elementy, trzecia próba opierała się jedynie na bigramach. Należy dodać, że eksperymenty dotyczące znajdowania najlepszych parametrów modeli były wykonywane na słowniku zawierającym jedynie pojedyncze wyrazy.

Tabela 5 – Przykładowe rekordy słownika w zależności od numeru próby (Źródło: opracowanie własne)

|  |  |  |
| --- | --- | --- |
| Pierwsza próba | Druga próba | Trzecia próba |
| care  others  tattoos  would  never  40  fox  richard  encouraging  frequency | wings  sorta  pointing  gosh  label  would never  never get  something really  pretty sure  another one | also tend  new people  star wars  guess ellipsis  really mean  typing people  find one  come conclusion  making friends  people could |

### Dobór modeli

Pierwszym z zastosowanych algorytmów była wielomianowa regresja logistyczna zaimplementowana przy pomocy funkcji *LogisticRegression[[25]](#footnote-26)* z biblioteki *scikit-learn.* W fazie eksperymentów zauważono wzrost wartości miary dokładności wraz ze zwiększaniem parametru *C* (*Rys. 4*), dlatego dobrano dla niego wartość 1. Regularyzacja jest wbudowana w metodę i wykonywana domyślnie, w przypadku tego problemu była to regularyzacja L2. Podczas treningu jako funkcję kosztu przyjęto entropię krzyżową.

Rys. 4 – Wzrost dokładności regresji logistycznej na zbiorze testowym przy wzroście parametru C (Źródło: opracowanie własne)

Obraz zawierający tekst, linia, zrzut ekranu, Wykres

Opis wygenerowany automatycznie

Kolejne podejście również opierało się na próbie znalezienia optymalnych separatorów klas. W jego przebiegu zastosowano liniowy klasyfikator SVM przystosowany dla potrzeb wieloklasowego zbioru, czyli z użyciem strategii jeden-przeciw-pozostałym. Posłużyła do tego klasa *LinearSVC* pochodząca z biblioteki *scikit-learn[[26]](#footnote-27)*. Podobnie jak w przypadku regresji liniowej, przeprowadzono analizę parametru C, jednak na *Rys. 5* można zauważyć, że tendencja w tym wypadku nie była monotoniczna. Jako optymalną wartość wybrano 0.4. Ponadto, jako funkcję kosztu wybrano Squared Hinge, która dobrze sprawdza się w treningu maszyn wektorów nośnych w zadaniu optymalizacyjnym znajdywania maksymalnego marginesu. Podczas treningu stosowano regularyzację L2.

Rys. 5 – Dokładność klasyfikatora SVM na zbiorze testowym przy wzroście parametru C (Źródło: opracowanie własne)

Obraz zawierający tekst, diagram, Wykres, linia

Opis wygenerowany automatycznie

Następnym tradycyjnym modelem uczenia maszynowego wybranym do klasyfikacji szesnastu typów osobowości był wielomianowy naiwny klasyfikator bayesowski w implementacji *MultinomialNB*[[27]](#footnote-28). Wzrost parametru wygładzania addytywnego był jednoznaczny ze zmniejszaniem się dokładności modelu, dlatego ustawiono jego wartość na niewielką, równą 0.1. Zaobserwowane zjawisko przedstawiono na *Rys. 6*.

Rys. 6 – Spadek dokładności wielomianowego naiwnego klasyfikatora bayesowskiego na zbiorze testowym przy wzroście parametru alpha (Źródło: opracowanie własne)

Obraz zawierający tekst, linia, Wykres, diagram

Opis wygenerowany automatycznie

Ostatnim zaproponowanym algorytmem było ekstremalne wzmocnienie gradientu. W tym celu wytrenowano klasyfikator *XGBClassifier* wywodzący się z modułu *XGBoost[[28]](#footnote-29)*. Model automatycznie rozpoznał naturę klasyfikacyjną zadania i obliczał prawdopodobieństwo należenia próbek do więcej niż dwóch klas. Strojenie pary parametrów, czyli maksymalnej głębokości drzewa na zbiorze oraz lambdy związanej z regularyzacją L2 na zbiorze zakończyło się wyłonieniem najlepszej pary wartości – odpowiednio liczby 6 i 1.

## Modelowanie z zastosowaniem głębokich sieci neuronowych

### Kodowanie danych

Przygotowanie reprezentacji numerycznej tekstu dla modeli uczenia głębokiego rozpoczęto od stworzenia słownika zawierającego wyrazy występujące we wstępnie przygotowanych i oczyszczonych postach zbioru treningowego. Spośród ponad 108 000 słów wykrytych przez funkcję *Tokenizer[[29]](#footnote-30),*  100 000 z nich zostało uwzględnionych. Zastosowanie metody *fit\_on\_texts* zdeterminowało strategię budowy słownika opartej na częstości występowania wyrazów. To znaczy, że im dane słowo pojawiało się w próbkach częściej, tym znalazło się bliżej początku listy rekordów i tym niższy otrzymało indeks (*Rys. 7*). Po przygotowaniu opisanej struktury, przeprowadzono na jej podstawie proste mapowanie próbek do reprezentacji liczbowej (*Rys. 8*).

Rys. 7 - Pierwsze rekordy słownika z użyciem metody fit\_on\_texts klasy Tokenizer (Źródło: opracowanie własne)

Obraz zawierający tekst, paragon, Czcionka, biały

Opis wygenerowany automatycznie

Rys. 8 - Przykładowe fragmenty próbek i ich reprezentacji numerycznej (Źródło: opracowanie własne)

Obraz zawierający tekst, Czcionka, zrzut ekranu, biały

Opis wygenerowany automatycznie

Najdłuższy ze scalonych postów zawierał 1081 tokenów numerycznych. Przy medianie długości wynoszącej 761, nie uznano potrzeby przycinania próbek. Uzupełniono zerami od lewej strony krótsze listy tokenów przy użyciu funkcji *pad\_sequences[[30]](#footnote-31)*.

### Sieć rekurencyjna z warstwą GRU

Neuronowe sieci rekurencyjne są szczególnie ukierunkowane na przetwarzanie sekwencji. Posiadają one unikalne struktury umożliwiające zapamiętywanie cech poprzednich elementów w szeregu, a przez to, w przypadku tekstu, wychwytujące kontekst i relacje słów w zdaniach. Ze względu na to należy pamiętać, że zachowanie kolejności poszczególnych wyrazów w próbce tekstowej jest kluczowe dla skutecznej klasyfikacji.

Architektura pierwszego zaproponowanego modelu sieci rekurencyjnej składała się z warstwy *Embedding*[[31]](#footnote-32)osadzającej słowa w przestrzeni 64-wymiarowej, uwzględniając ich podobieństwa znaczeniowe. Ten segment można traktować jako dopełnienie dotychczasowych działań w kontekście tworzenia funkcjonalnej reprezentacji numerycznej próbek. Kolejnym komponentem modelu była pojedyncza warstwa *GRU*[[32]](#footnote-33) zawierająca 512 jednostek wraz z warstwą normalizującą *BatchNormalization*[[33]](#footnote-34). Po niej wystąpiła warstwa gęsta *Dense[[34]](#footnote-35)* złożona z 128 neuronów z funkcją aktywacyjną ReLU, następnie mechanizm *Dropout[[35]](#footnote-36)* odrzucający losowo30% danych podczas każdego kroku treningu. Ostatnią warstwę stanowiła wyjściowa warstwa gęsta o 16 neuronach z funkcją aktywacyjną Softmax. Schemat opisanej architektury został zobrazowany na *Rys. 9*.

Rys. 9 - Schemat architektury rekurencyjnej sieci neuronowej z warstwą GRU (Źródło: opracowanie włąsne)

Obraz zawierający tekst, zrzut ekranu, design

Opis wygenerowany automatycznie

Entropia krzyżowa w wariancie dla klasyfikacji wieloklasowej odgrywała rolę funkcji kosztu podczas treningu, a algorytm *Adam*[[36]](#footnote-37) o domyślnych parametrach został wybrany do optymalizacji. Jako metrykę oceny skuteczności modelu obrano implementację dokładności *CategoricalAccuracy[[37]](#footnote-38)*.

[Attention, BERT?]

[wymagania sprzętowe i programistyczne]]

# Wyniki



## Wieloklasowość

[Wyjaśnić, dlaczego moje podejście jest nietypowe – 16 klas, a nie binarne modele jak w wielu źródłach + wynikający z tego poziom losowy do porównania]

[Szczegóły treningów – epoki, jak poradziły sobie moje architektury, co było skuteczne, a co nie (dlaczego?), ewentualne porównanie do innych architektur. Czasy treningów.]

# Podsumowanie



dodać

[Cel pracy został osiągnięty. Opisać najlepsze osiągnięte rozwiązanie. Rozważania nt. efektów w kontekście osobowości. Możliwe rozwinięcia projektu. Zastosowania w codziennym życiu.]

Input

Embedding

GRU

Batch Normalization

Dense

Dropout

Dense

Output

# Roboczy generator schematów

# Spis tabel

Tabela 1 – Opis par kategorii modelu MBTI (Źródło: opracowanie własne) 10

Tabela 2 – Porównanie tokenizacji tekstu przez word\_tokenizer i TweetTokenizer (Źródło: opracowanie własne) 27

Tabela 3 – Proces wstępnej obróbki tekstu (Źródło: opracowanie własne) 27

Tabela 4 – Przykłady oznaczania charakterystycznych elementów w próbkach (Źródło: opracowanie własne) 28

Tabela 5 – Przykładowe rekordy słownika w zależności od numeru próby (Źródło: opracowanie własne) 31

# Spis rysunków

Rys. 1 – Dystrybucja próbek między klasami (Źródło: opracowanie własne) 13

Rys. 2 – Wyznaczanie hiperpłaszczyzny liniowego klasyfikatora SVM oraz maksymalizacja marginesu separującego (Źródło: https://en.wikipedia.org/wiki/Support\_vector\_machine) 21

Rys. 3 - Wizualizacja architektur mechanizmu Attention (Źródło: Ashish Vaswani i in., Attention Is All You Need”, 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 2023) 23

Rys. 4 – Wzrost dokładności regresji logistycznej na zbiorze testowym przy wzroście parametru C (Źródło: opracowanie własne) 32

Rys. 5 – Dokładność klasyfikatora SVM na zbiorze testowym przy wzroście parametru C (Źródło: opracowanie własne) 33

Rys. 6 – Spadek dokładności wielomianowego naiwnego klasyfikatora bayesowskiego na zbiorze testowym przy wzroście parametru alpha (Źródło: opracowanie własne) 33

Rys. 7 - Pierwsze rekordy słownika z użyciem metody fit\_on\_texts klasy Tokenizer (Źródło: opracowanie własne) 34

Rys. 8 - Przykładowe fragmenty próbek i ich reprezentacji numerycznej (Źródło: opracowanie własne) 35

Rys. 9 - Schemat architektury rekurencyjnej sieci neuronowej z warstwą GRU (Źródło: opracowanie włąsne) 36

# Wykaz symboli i oznaczeń

[O ile nie jest to inaczej wyszczególnione w tekście praca ta wykorzystuje poniższą notację:

* a (mała litera kursywą) oznacza skalar,
* **a** (mała pogrubiona litera) oznacza wektor]

# Wykaz używanych skrótów



* MBTI – Myers-Briggs Type Indicator®

# Bibliografia

1. <https://www.reddit.com/> [↑](#footnote-ref-1)
2. <https://www.16personalities.com/pl> [↑](#footnote-ref-2)
3. Carl Gustav Jung, *Typy psychologiczne,* Wydawnictwo KR, 2015 [↑](#footnote-ref-3)
4. The Myers-Briggs Company, *Reliability and validity of the MBTI® instrument,* <https://eu.themyersbriggs.com/-/media/Files/PDFs/Technical-information/MBTI_reliability_and_validity_info.pdf?la=en> [↑](#footnote-ref-4)
5. Charakterystyki w *Tabeli 1* są skrótem bardziej szczegółowych opisów przedstawionych w artykule <https://potencjalosobowosci.com/mbti/>, dostęp z dnia 01.02.2024 r. [↑](#footnote-ref-5)
6. <https://www.kaggle.com/datasets/datasnaek/mbti-type>, dostęp z dnia 17.10.2023 r. [↑](#footnote-ref-6)
7. <https://www.personalitycafe.com/>, dostęp z dnia 17.10.2023 r. [↑](#footnote-ref-7)
8. Hernandez R., Knight I.S., *Predicting Myers-Briggs Type Indicator with Text Classification*, 31st Conference on Neural Information Processing Systems, 2017 [↑](#footnote-ref-8)
9. Cui B., Qi C., *Survey Analysis of Machine Learning Methods for Natural Language Processing for MBTI Personality Type Prediction*, 2017 [↑](#footnote-ref-9)
10. Bianca Antonio i in., *Data Science Final Project: Myers-Briggs Prediction*, 2018, artykuł dostępny pod linkiem<https://medium.com/@bian0628/data-science-final-project-myers-briggs-prediction-ecfa203cef8>, dostęp z dnia 17.10.2023 r. [↑](#footnote-ref-10)
11. Abidin N.H.Z. i in., *Improving Intelligent Personality Prediction using Myers-Briggs Type Indicator and Random Forest Classifier*, International Journal of Advanced Computer Science and Applications(IJACSA), 11(11), 2020 [↑](#footnote-ref-11)
12. Amirhosseini M. H., Kazemian H., *Machine Learning Approach to Personality Type Prediction Based on the Myers–Briggs Type Indicator®*, *Multimodal Technologies and Interaction*, 2020 [↑](#footnote-ref-12)
13. Ontoum S., Chan J.H., *Personality Type Based on Myers-Briggs Type Indicator with Text Posting Style by using Traditional and Deep Learning*, 2022 [↑](#footnote-ref-13)
14. <https://medium.com/analytics-vidhya/understanding-embedding-layer-in-keras-bbe3ff1327ce>, dostęp z dnia 21.02.2024 r. [↑](#footnote-ref-14)
15. Jacob Devlin i in., *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*, 2019 [↑](#footnote-ref-15)
16. Steven S. Skiena, *The Data Science Design Manual*, Edition 1, Springer Cham, 2017, s. 366-369 [↑](#footnote-ref-16)
17. Ashish Vaswani i in., *Attention Is All You Need”,* 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 2023 [↑](#footnote-ref-18)
18. <https://www.nltk.org/api/nltk.tokenize.word_tokenize.html>, dostęp z dnia 05.11.2023 r. [↑](#footnote-ref-19)
19. <https://tedboy.github.io/nlps/generated/generated/nltk.tokenize.TweetTokenizer.html>, dostęp z dnia 05.11.2023 r. [↑](#footnote-ref-20)
20. <https://sjp.pl/hashtag>, dostęp z dnia 18.11.2023 r. [↑](#footnote-ref-21)
21. <https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html>, dostęp z dnia 18.11.2023 r. [↑](#footnote-ref-22)
22. <https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html>, dostęp z dnia 20.11.2023 r. [↑](#footnote-ref-23)
23. <https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html>, dostęp z dnia 10.12.2023 r. [↑](#footnote-ref-24)
24. <https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfTransformer.html>, dostęp z dnia 10.12.2023 r. [↑](#footnote-ref-25)
25. <https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html>, dostęp z dnia 20.12.2023 r. [↑](#footnote-ref-26)
26. <https://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC.html>, dostęp z dnia 20.12.2023 r. [↑](#footnote-ref-27)
27. <https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.MultinomialNB.html>, dostęp z dnia 21.12.2023 r. [↑](#footnote-ref-28)
28. <https://xgboost.readthedocs.io/en/stable/get_started.html>, dostęp z dnia 21.12.2023 r. [↑](#footnote-ref-29)
29. <https://www.tensorflow.org/api_docs/python/tf/keras/preprocessing/text/Tokenizer>, dostęp z dnia 27.12.2023 r. [↑](#footnote-ref-30)
30. <https://www.tensorflow.org/api_docs/python/tf/keras/utils/pad_sequences>, dostęp z dnia 27.12.2023 r. [↑](#footnote-ref-31)
31. <https://keras.io/api/layers/core_layers/embedding/>, dostęp z dnia 27.12.2023 r. [↑](#footnote-ref-32)
32. <https://keras.io/api/layers/recurrent_layers/gru/>, dostęp z dnia 27.12.2023 r. [↑](#footnote-ref-33)
33. <https://keras.io/api/layers/normalization_layers/batch_normalization/>, dostęp z dnia 27.12.2023 r. [↑](#footnote-ref-34)
34. <https://keras.io/api/layers/core_layers/dense/>, dostęp z dnia 27.12.2023 r. [↑](#footnote-ref-35)
35. <https://keras.io/api/layers/regularization_layers/dropout/>, dostęp z dnia 27.12.2023 r. [↑](#footnote-ref-36)
36. <https://keras.io/api/optimizers/adam/>, dostęp z dnia 27.12.2023 r. [↑](#footnote-ref-37)
37. <https://www.tensorflow.org/api_docs/python/tf/keras/metrics/CategoricalAccuracy>, dostęp z dnia 27.12.2023 r. [↑](#footnote-ref-38)