

Politechnika Warszawska

W Y D Z I A Ł E L E K T R O N I K I
I T E C H N I K I N F O R M A C Y J N Y C H



Instytut Radioelektroniki i Technik Multimedialnych

Praca dyplomowa

na Studium Podyplomowym Głębokie Sieci Neuronowe – Zastosowania w Mediach Cyfrowych

Klasyfikacja poezji polskiej przy użyciu
architektury Transformers w modelu herBERT

mgr Nikola Renata Janik

Numer albumu P-10535

Promotor
prof. dr hab. inż. Władysław Skarbek

WARSZAWA 2023

Streszczenie

Celem niniejszej pracy jest przeprowadzenie komparatystycznych badań literackich w ujęciu ilościowym, w odróżnieniu od podejść jakościowych stosowanych w literaturoznawstwie, skupiając się na próbach klasyfikacji wierszy wg stylu literackiego charakterystycznego dla każdego z analizowanych autorów. Wykorzystano metody uczenia maszynowego oraz nowoczesne modele językowe oparte o głębokie sieci neuronowe, model herBERT, w celu klasyfikacji zbioru czterystu utworów polskich poetów i poeteck z różnych epok literackich.

Zagadnienia teoretyczne z obszaru przetwarzania języka naturalnego przedstawiono w rozdziale 2. Zostały w nim wyjaśnione kolejne rozwiązania, jakie pojawiały się na drodze rozwoju tej dziedziny naukowej: tokenizacja oraz wektoryzacja TF-IDF, szukanie relacji między słowami przy pomocy word2vec, zastosowanie sieci rekurencyjnej oraz LSTM, co pozwoliło na osadzenie słów w kontekście, a także rewolucyjną architekturę Transformers, która umożliwiła maszynowe generowanie tekstów niemal nieodróżnialnych od tekstów tworzonych przez człowieka.

Wprowadzenie do podstawowych terminów literackich zawarto w rozdziale 3. Wyjaśniono w nim pojęcie stylu literackiego oraz stylizacji. Następnie przybliżono sylwetki poetów oraz krótką charakterystykę ich twórczość, która posłużyła jako materiał badawczy w niniejszym pracy dyplomowej.

Analizie ilościowej badanych wierszy, dla których przy pomocy modelu językowego herBERT przygotowano reprezentację wektorową, poświęcono rozdziały 4 i 5. W pierwszym z nich została zaprezentowana analiza skupień z zastosowaniem liniowych i nieliniowych algorytmów redukcji wymiarowości. Natomiast wyniki klasyfikacji tekstów z wykorzystaniem trzech typów klasyfikatorów: algorytmu najbliższego sąsiada, metody uczenia maszynowego XGBoost oraz głębokiej sieci neuronowej przedstawiono w rozdziale 5.

Podsumowanie otrzymanych wyników i analiz, wraz z nakreśleniem przyszłych perspektyw zastosowań neuronowych technik przetwarzania języka naturalnego w badaniach literackich umieszczone w ostatnim rozdziale.

Summary

The aim of this study is to conduct comparative literary research from a quantitative perspective, in contrast to the qualitative approaches used in literary studies, focusing on attempts to classify poems based on the literary style characteristic of each analyzed author. Machine learning methods and modern language models based on deep neural networks, the herBERT model, were employed to classify a set of four hundred works by Polish poets from different literary epochs.

The theoretical aspects of natural language processing in the field of computational linguistics are presented in Chapter 2. It explains the successive solutions that have emerged in the development of this scientific discipline: tokenization and TF-IDF vectorization, word-to-word relations using word2vec, the use of recurrent neural networks and LSTM to embed words in context, as well as the revolutionary architecture of Transformers, which enables the machine generation of texts almost indistinguishable from those created by humans.

Basic literary terms were introduced in Chapter 3, explaining the concept of literary style and stylization. Subsequently, the profiles of the poets were presented along with a brief description of their works, which served as the research material for this thesis.

Dedicated to the quantitative analysis of the studied poems, for which vector representations were prepared using the herBERT language model, Chapters 4 and 5 delve into the topic. Chapter 4 presents cluster analysis employing linear and nonlinear dimensionality reduction algorithms. Meanwhile, Chapter 5 showcases the results of text classification using three types of classifiers: the nearest neighbor algorithm, the XGBoost machine learning method, and deep neural networks.

The conclusion encompasses a summary of the obtained results and analyses, along with outlining prospects for the application of neural natural language processing techniques in literary research.

Warszawa, 04.07.2023
miejscowość i data

Nikola Renata Janik
imię i nazwisko studenta

P-10535
numer albumu

Głębokie sieci neuronowe – zastosowania w mediach
kierunek studiów

OŚWIADCZENIE

Świadomy/-a odpowiedzialności karnej za składanie fałszywych zeznań oświadczam, że niniejsza praca dyplomowa została napisana przeze mnie samodzielnie, pod opieką kierującego pracą dyplomową.

Jednocześnie oświadczam, że:

- niniejsza praca dyplomowa nie narusza praw autorskich w rozumieniu ustawy z dnia 4 lutego 1994 roku o prawie autorskim i prawach pokrewnych (Dz.U. z 2006 r. Nr 90, poz. 631 z późn. zm.) oraz dóbr osobistych chronionych prawem cywilnym,
- niniejsza praca dyplomowa nie zawiera danych i informacji, które uzyskałem/-am w sposób niedozwolony,
- niniejsza praca dyplomowa nie była wcześniej podstawą żadnej innej urzędowej procedury związanej z nadawaniem dyplomów lub tytułów zawodowych,
- wszystkie informacje umieszczone w niniejszej pracy, uzyskane ze źródeł pisanych i elektronicznych, zostały udokumentowane w wykazie literatury odpowiednimi odnośnikami,
- znam regulacje prawne Politechniki Warszawskiej w sprawie zarządzania prawami autorskimi i prawami pokrewnymi, prawami własności przemysłowej oraz zasadami komercjalizacji.

Oświadczam, że treść pracy dyplomowej w wersji drukowanej, treść pracy dyplomowej zawartej na nośniku elektronicznym (płycie kompaktowej) oraz treść pracy dyplomowej w module APD systemu USOS są identyczne.

.....
czytelny podpis studenta

Podziękowania

Serdecznie dziękuję mojemu promotorowi prof. dr. hab. inż. Władysławowi Skarbkowi za cenne rady, wskazówki oraz inspirujące dyskusje. Dziękuję także moim bliskim, którzy wspierali mnie na każdym etapie studiowania i pracy nad projektem.

Spis treści

Streszczenie	2
Summary.....	3
Spis treści.....	6
1. Wprowadzenie	7
1.1. HISTORIA UCZENIA MASZYNOWEGO I GŁĘBOKICH SIECI NEURONOWYCH	7
1.2. HISTORIA PRZETWARZANIA JĘZYKA NATURALNEGO.....	8
1.3. ZASTOSOWANIE NLP W BADANIACH LITERATUROZNAWCZYCH	9
1.4. CEL PRACY.....	10
2. Wstęp do NLP	12
2.1. PRZETWARZANIE JĘZYKA NATURALNEGO – OD KOMUNIKACJI MIĘDZYLUDZKIEJ, DO KOMUNIKACJI CZŁOWIEK - MASZYNA	12
2.2. TOKENIZACJA.....	14
2.3. WEKTORY TF-IDF	15
2.4. WORD2VEC.....	16
2.5. REKURENCYJNE SIECI NEURONOWE W NLP	18
2.6. LONG SHORT-TERM MEMORY	18
2.7. ARCHITEKTURA TRANSFORMERS	20
3. Wstęp do analizy literaturoznawczej	23
3.1. CZYM JEST STYL W LITERATURZE	23
3.2. ANALIZA LITERATUROZNAWCZA WYBRANYCH POETÓW	24
3.2.1. Jan Kochanowski.....	24
3.2.2. Krzysztof Kamil Baczyński	25
3.2.3. Czesław Miłosz.....	26
3.2.4. Zbigniew Herbert	27
3.2.5. Wisława Szymborska	27
3.2.6. Halina Poświatowska.....	28
3.2.7. Maria Pawlikowska-Jasnorzewska.....	29
3.2.8. Ewa Lipska.....	30
4. Analiza zbioru danych. Metody nienadzorowane: redukcja wymiarowości i analiza skupień.....	31
4.1. ANALIZA ODLEGŁOŚCI REPREZENTACJI WEKTOROWYCH WERSZY	31
4.2. REDUKCJA WYMIAROWOŚCI I ANALIZA SKUPIEŃ WERSZY WSZYSTKICH KLAS	32
4.2.1. PCA.....	32
4.2.2. t-SNE	35
4.2.3. UMAP	36
4.3. REDUKCJA WYMIAROWOŚCI I ANALIZA SKUPIEŃ REPREZENTACJI WEKTOROWYCH WERSZY POETÓW	39
4.4. REDUKCJA WYMIAROWOŚCI I ANALIZA SKUPIEŃ REPREZENTACJI WEKTOROWYCH WERSZY POETEK.....	43
5. Uczenie nadzorowane i klasyfikacja wierszy	49
5.1. KLASYFIKACJA NA PODSTAWIE ODLEGŁOŚCI EUKLIDESOWYCH	49
5.2. KLASYFIKACJA PRZY POMOCY XGBOOST	52
5.3. KLASYFIKACJA PRZY POMOCY SIECI NEURONOWEJ.....	56
6. Podsumowanie	62
Spis ilustracji	64
Spis tabel.....	66
Referencje.....	67

1. Wprowadzenie

1.1. Historia uczenia maszynowego i głębokich sieci neuronowych

Uczenie maszynowe (ang. *Machine Learning, ML*) [1] jest dziedziną nauki na styku informatyki oraz matematyki, zajmującą się ilościową analizą danych w celu wyszukiwania zależności między nimi, które są trudne do dostrzeżenia dla człowieka. W szczególności ML nazywamy zbiór metod statystycznych oraz zbiór algorytmów numerycznych, które pozwalają na klasyfikację badanych danych. ML jest podzbiorem szerszej dziedziny nazywanej Sztuczną Inteligencją (ang. *Artificial Intelligence, AI*), której celem jest symulowanie ludzkiej inteligencji, naturalnych ludzkich zachowań, czy procesów uczenia się przez ludzi. Następnie, podzbiorem ML jest tzw. uczenie głębokie (ang. *Deep Learning, DL*) [2,3], w którym celem jest wyszukanie kluczowych relacji pomiędzy analizowanymi danymi przy pomocy wysokowymiarowych nieliniowych funkcji. Podstawowym narzędziem DL są tzw. sztuczne sieci neuronowe (ang. *Artificial Neural Networks, ANN*), które swoją konstrukcją początkowo miały przypominać połączenia pomiędzy neuronami w ludzkim mózgu. Główna idea ANN opiera się o tzw. Uniwersalne Twierdzenie Aproksymacyjne (ang. *Universal Approximation Theorem*) [4], które mówi, że sieć neuronowa o jednej warstwie ukrytej, wykorzystująca odpowiednią nieliniową funkcję aktywacji, potrafi przybliżyć dowolną funkcję ciągłą na określonym zbiorze danych z dowolną dokładnością, przy wystarczająco dużej liczbie neuronów. Tym samym, sieci neuronowe są w stanie „nauczyć się” skomplikowanych relacji między danymi pozwalając maszynie wyłuskać kluczowe cechy analizowanych danych. Proces trenowania ANN, czyli dobór odpowiednich parametrów w celu rozwiązania analizowanego problemu na podstawie dostarczonych danych jest, w dużym uproszczeniu, odpowiednikiem procesu uczenia się u ludzi. W celu wyszukiwania skomplikowanych relacji między danymi, wyłuskiwania kluczowych cech rozpoczęto badania nad skutecznością sieci neuronowych składających się z większej niż jednej warstwy ukrytej dając początek głębokim sieciom neuronowym (ang. *deep neural networks*) oraz uczeniu głębokiemu.

W ostatnich kilkunastu latach można obserwować ogromny skok w rozwoju DL oraz szerzej AI, głównie za sprawą ogromnej ilości danych dostępnych w sieci Internet (przewidywania sięgają 175 zetabajtów danych do 2025 roku¹), oraz możliwości wykorzystania ogromnych mocy obliczeniowych, co pozwala na trenowanie ANN składających się z miliardów parametrów (obecnie najpopularniejszy generatywny model językowy Chat-GPT3.5

¹ Źródło, dostęp zdalny (20.06.2023): <https://www.statista.com/statistics/871513/worldwide-data-created/>.

posiada około 175 miliardów parametrów). Jednak początki uczenia głębokiego sięgają już lat 40. XX wieku. Poniżej zwięzle przedstawiam historię rozwoju tej dziedziny w pierwszych dekadach jej istnienia:

- 1943 Walter Pitts i Warren McCulloch stworzyli model komputerowy, wykorzystujący wiedzę na temat działania neuronów w mózgu, nazywany logiką progową (ang. *threshold logic*)
- 1958 Frank Rosenblatt opublikował artykuł, w którym opisał działanie perceptronu (*Principles of Neurodynamics: Perceptrons and the Theory of BrainMechanisms*)
- 1962 David Hubel i Torsten Wiesel opublikowali pracę, gdzie przedstawiono właściwości pojedynczego neuronu biologicznego (*Receptive Fields, Binocular Interaction and Functional Architecture in the Cat's Visual Cortex*)
- 1969 Marvin Minsky i Seymour Papert opisali problemy w nowej odsłonie sztucznej inteligencji, które wynikały z ograniczeń obliczeniowych oraz wad pojedynczego sztucznego neuronu.
- 1979 Geoffrey Hinton i James Anderson zorganizowali konferencję poświęconą nowym podejściom do uczenia głębokiego (*Parallel Models of Associative Memory*, California)
- 1986 David E. Rumelhart opublikował artykuł, w którym opisał mechanizm propagacji wstecznej (*Backpropagation: The Basic Theory*)
- 1987 Pierwsza konferencja w pełni poświęcona głębokim sieciom neuronowym (*Neural Information Processing Systems (NIPS)*).

1.2. Historia przetwarzania języka naturalnego

Przetwarzanie języka naturalnego [5] (ang. *Natural Language Processing, NLP*) to dziedzina uczenia maszynowego i obecnie uczenia głębokiego, koncentrująca się na komputerowej analizie treści tekstowych, które z łatwością są rozumiane przez ludzi, ale stanowią wyzwanie dla komputerów. Rozwój NLP pozwala komputerom interpretować, generować teksty w języku naturalnym, czyli ludzkim oraz w ograniczonym stopniu „rozumieć” analizowane treści pozwalając komputerom np. na wyłuskiwanie faktów, lub ich ograniczonej analizy.

Pierwsze badania w kierunku przetwarzania języka naturalnego miały miejsce już w latach 50. XX wieku. Jednym z pierwszych projektów była próba przeprowadzenia tłumaczenia

przez maszynę (ang. *Machine Translation Project*), który był realizowany na Uniwersytecie Goergetown. Pomimo braku zadowalających wyników, przyczyniło się to do rozwoju badań nad NLP. Kolejne dekady przyniosły wiele prac dotyczących rozumienia tekstu przez maszyny. Wykorzystywano między innymi techniki analizy składniowej, analizę morfologiczną, składniową oraz semantyczną. W latach 80. i 90. w badaniach nad NLP wprowadzono statystyczne metody uczenia maszynowego, które umożliwiły analizę większych zbiorów tekstów. Duże znaczenie w tym czasie miało użycie w NLP algorytmów ukrytych modeli Markowa (ang. *Hidden Markov Models, HMM*). Dzięki wykorzystaniu metod statystycznych pojawił się nowy obszar badawczy – ekstrakcja informacji (ang. *information extraction*), którego celem było wydobywanie konkretnych informacji z tekstów. Przełomowe badania były prowadzone także nad techniką NER (ang. *Named Entity Recognition*), czyli przetwarzaniem języka w celu identyfikacji i klasyfikacji nazwanych jednostek w tekście np. osób, miejsc, organizacji, dat, liczb, lub walut.

W ubiegłych latach można było zaobserwować ogromny skok rozwojowy w dziedzinie przetwarzania języka naturalnego. Zaczęto wykorzystywać uczenie maszynowe oraz sieci neuronowe, które wraz z dostępnością ogromnych zbiorów danych spowodowały rozwój nowych modeli NLP, jak na przykład word2vec, LSTM czy najnowsze sieci o architekturze Transformers, które są głównie oparte na mechanizmie atencji (ang. *attention mechanism*).

1.3. Zastosowanie NLP w badaniach literaturoznawczych

Badania nad literaturą opierają się głównie na metodach jakościowych. Oznacza to, że badacze są w stanie jedynie w sposób opisowy scharakteryzować np. dany styl literacki, styl danego autora lub powiedzieć jak dwa teksty literackie są do siebie podobne lub jak się różnią. Brakuje w narzędziach badawczych literaturoznawców szerokiej gamy miar, którymi mogliby mierzyć obiekty swoich badań. W ostatnich latach rozwój technik NLP pozwala jednak na wykorzystanie modeli językowych do zadań z obszaru literaturoznawstwa nadając im charakter ilościowy, co otwiera bardzo wiele możliwości rozwoju w tej dziedzinie [6]. W przyszłości literaturoznawcy będą w stanie nie tylko powiedzieć, że np. dwa teksty literackie są do siebie podobne, ale dzięki modelom NLP będą dysponowali miarą numeryczną mierzącą podobieństwo analizowanych tekstów.

Najbardziej podstawowym zagadnieniem do jakiego można wykorzystać NLP w kontekście literatury, to analiza tematyczna. Metody NLP, takie jak modelowanie tematyczne, mogą pomóc w identyfikowaniu dominujących tematów w zbiorze tekstów literackich.

Badacze mogą stosować te metody do analizy zmian tematyki w czasie lub do porównywania różnych autorów i ich preferencji tematycznych.

Kolejnym zastosowaniem NLP w analizie literaturoznawczej jest badanie sentymantu tekstu. Rozwiązania proponowane przez NLP mogą być używane do klasyfikacji tekstów literackich pod kątem ich sentymantu. Badacze mogą oceniać emocjonalne nacechowanie tekstów, odkrywać tendencje emocjonalne w różnych okresach literackich, a nawet analizować reakcje czytelników na teksty.

Bardzo ciekawym wykorzystaniem przetwarzania języka naturalnego w literaturze jest analiza sieci społecznych [7]. Takie badania wymagają zaczerpnięcia wiedzy z więcej niż dwóch dziedzin naukowych, ponieważ w tym wypadku oprócz literaturoznawstwa i sztucznej inteligencji istotne jest także rozumienie podstaw socjologii. Analiza sieci społecznych przy wykorzystaniu metod NLP może pomóc w analizie relacji między różnymi postaciami literackimi, autorami czy czytelnikami. Może to obejmować badanie wpływu jednego autora na innych twórców, analizę interakcji między postaciami w fikcyjnym świecie literackim, czy też relację między odbiorcami tekstu.

Ostatnim obszarem, w którym można zastosować przetwarzanie języka naturalnego są badania nad stylem pisarza [8,9,10,11]. Najczęściej tego typu rozwiązania są używane w celu identyfikacji autora tekstu. Wykorzystanie NLP do określenia autorstwa pozwala np. klasyfikować dzieła z dawnych epok, które do tej pozostawały anonimowe. Innym interesującym zagadnieniem, w którym ta metoda mogłaby się sprawdzić jest odszyfrowywanie utworów tworzonych pod pseudonimami np. przez kobiety przed XX wiekiem, ponieważ był to okres, kiedy pisarzami byli głównie mężczyźni lub panie kryjące się pod męskim imieniem.

1.4. Cel pracy

Celem niniejszej pracy jest wykorzystanie metod uczenia maszynowego oraz nowoczesnych modeli językowych opartych o głębokie sieci neuronowe w celu klasyfikacji zbioru czterystu wierszy napisanych przez osmioro polskich poetów z różnych epok literackich oraz dokonanie analizy ilościowej podobieństw i różnic pomiędzy różnymi twórcami. Wybrane przeze mnie wiersze oraz ich autorzy charakteryzują się różnymi stylami literackimi, które są dobrze zrozumiane i wielokrotnie przeanalizowane przez literaturoznawców standardowymi metodami badawczymi. Moim celem jest rzucenie światła na badania literaturoznawcze przy pomocy metod NLP, próba uchwycenia ilościowych różnic, jak i podobieństw pomiędzy twórcami, próba scharakteryzowania analizowanych wierszy ze względu na płeć autorów, jak i czas ich twórczości.

W kolejnych rozdziałach omówię przełomowe rozwiązania wykorzystywane do przetwarzania języka naturalnego. Następnie wyjaśnię podstawowe pojęcie z dziedziny literaturoznawstwa, tj. styl literacki oraz przybliżę sylwetki i twórczość poetów wybranych do projektu.

2. Wstęp do NLP

W tym rozdziale zostaną przedstawione kolejne kamienie milowe w rozwoju metod przetwarzania języka naturalnego. W szerszy sposób zostaną omówione podstawowe operacje wykonywane na tekstuach oraz rozwiązań pozwalające uchwycić znaczenie słów: word2vec, LSTM oraz architektura Transformers.

2.1. Przetwarzanie języka naturalnego – od komunikacji międzyludzkiej, do komunikacji człowiek - maszyna

Języki naturalne są to języki jakimi ludzie się na co dzień porozumiewają, jednak nie są zrozumiałe dla maszyn liczących, czyli komputerów. Komunikacja lub bardziej precyzyjnie – wydawanie poleceń przez człowieka maszynie, oparte jest o algorytmy, do których polecenia są przekazywane poprzez pseudokod. Powstałe na bazie algorytmów języki programowania – czyli zbiór reguł i składni używanych do komunikacji i interakcji z komputerem – pozwalają na jeszcze sprawniejszą komunikację między ludźmi, a maszynami. Są to języki, które pozwalają programistom pisać kod komputerowy, który jest następnie komplikowany lub interpretowany przez komputer do kodu maszynowego a ostatecznie binarnego, aby wykonać określone zadania. Języki programowania mogą mieć różne poziomy abstrakcji i przeznaczenia. Mogą być niskopoziomowe, takie jak języki asemblerowe, które są bliskie kodowi maszynowemu i pozwalają na precyzyjne sterowanie sprzętem komputerowym, lub wysokopoziomowe, takie jak języki takie jak Python, lub Java które są potrafią operować na bardziej abstrakcyjnych strukturach danych ułatwiając pisanie kodu. Języki programowania są zaprojektowane tak, aby były precyzyjne i wykonują określone zadania, podczas gdy języki naturalne służą do przekazywania informacji, wyrażania uczuć i komunikacji społecznej.

Główne różnice między językami naturalnymi, a językami programowania w kontekście zagadnień przetwarzania języka naturalnego (NLP) są struktura i gramatyka: języki naturalne, którymi posługują się ludzie posiadają skomplikowaną strukturę gramatyczną i złożone reguły. Z drugiej strony, języki programowania mają ścisłe określona składnię i gramatykę, które są zwykle logiczne i precyzyjne. Następnie języki naturalne są często niejednoznaczne i mogą mieć wiele interpretacji. Zdania mogą być niejasne, zawierać przenośnie, idiomatyczne wyrażenia, homonimy, synonimy itp., co jest dokładnym przeciwieństwem języków programowania, które są zwykle zaprojektowane tak, aby unikać niejednoznaczności, a instrukcje są precyzyjne i zrozumiałe dla komputera. Oznacza to, że w językach naturalnych niezwykle istotne jest stosowanie semantyki, gdzie znaczenie wyrażeń

często zależy od kontekstu, intencji mówiącego i innych czynników. W językach programowania, znaczenie instrukcji jest zazwyczaj ściśle zdefiniowane i niezależne od kontekstu.

W kontekście przetwarzania języka naturalnego (NLP), głównym wyzwaniem jest rozumienie i przetwarzanie złożonych struktur językowych, niejednoznaczności, semantyki i metafor języków naturalnych. Widać zatem, że język służący do komunikacji między ludźmi, różni się fundamentalnie od tego do komunikacji pomiędzy człowiekiem a maszyną. Techniki NLP, takie jak analiza morfologiczna, analiza składniowa, ekstrakcja informacji, rozpoznawanie nazw własnych, tłumaczenie maszynowe i wiele innych, mają na celu rozwiązanie tych wyzwań w celu efektywnego przetwarzania i zrozumienia języka naturalnego przez komputery.

Przetwarzanie języka naturalnego (NLP) jako dział uczenia maszynowego, zajmujący się analizą języków naturalnych, wymaga przeprowadzenia odpowiednich procesów, które pozwolą na utworzenie numerycznych reprezentacji dla jednostek znaczących w językach naturalnych – słów. Za przetworzeniem języka naturalnego do postaci numerycznej stoją złożone algorytmy oparte na metodach statystycznych. Komputery dużo łatwiej radzą sobie z dużymi obliczeniami niż ludzie, więc takie podejście polegające na znalezieniu przez model językowy zależnościach statystycznych da zdecydowanie lepsze efekty niż tworzenie zbioru zasad gramatycznych, z których ludzie korzystają intuicyjnie.

Głównymi wyzwaniami w dziedzinie NLP jest reprezentacja numeryczna wyrazów, jak i całych dokumentów, tak aby nowa reprezentacja numeryczna potrafiła wiernie odzwierciedlić znaczenie, kontekst oraz relacje pomiędzy analizowanymi treściami. Najbardziej intuicyjną i naiwną metodą numerycznej reprezentacji słów jest tzw. kodowanie gorącej jedynki (*ang. one-hot-encoding, OHE*), która polega na przypisaniu każdemu wyrazowi ze słownika danego języka jednego wektora o długości odpowiadającej liczbie wyrazów w danym języku, lub rozważanym słowniku wyrazów. Metoda polega na przypisaniu liczby porządkowej, k, każdemu wyrazowi ze zbioru wszystkich słów zadanego języka (który ma K elementów) : k-temu słowu przypisany jest wektor K wymiarowy, który składa się z samych zer, oprócz k-tego elementu który ma wartość 1. Niestety ta prosta konstrukcja bardzo szybko napotyka na swoje ograniczenia, głównie sprzętowe: liczba wyrazów w języku polskim szacowana jest na 150 000 – 200 000, wraz z odmianami i nazwami własnymi, z tego względu reprezentacji OHE jest bardzo niepraktyczna.

Na przestrzeni ostatnich lat powstało kilka metod polegających na analizie tekstu. Pierwsza część z metod oparta jest o analizę statystyczną występowania słów (metoda TF-IDF),

a druga część metod polega na znalezieniu efektywnej reprezentacji numerycznej słów w przestrzeni skończonej wymiarowej – tzw. wektory zanurzeń (ang. *embeddings*) mają zazwyczaj kilkaset elementów. Wyzwaniami stojącymi przed NLP było znalezienie reprezentacji wektorowej (wektorów zanurzeń), które potrafiłyby uwzględniać kontekst oraz semantykę użycia wyrazów. W konsekwencji odpowiednie reprezentacje wektorowe słów pozwalają przygotowywać reprezentacje wektorowe całych treści. Przełomem w NLP było połączenie metod sieci neuronowych i uczenia głębokiego, które pozwoliły na znajdywanie odpowiednich reprezentacji numerycznych wyrazom używanym w języku codziennym.

W kolejnych podrozdziałach omówione zostaną po krótkie kolejne kroki pozwalające przeprowadzać analizę tekstu przy pomocy metod NLP.

2.2. Tokenizacja

Pierwszym procesem, jakiemu poddawane są słowa w przetwarzaniu języka naturalnego jest tokenizacja, czyli podział tekstu na mniejsze całości nazywane tokenami [5]. W NLP tokeny mogą reprezentować słowa, znaki interpunkcyjne lub fragmenty zdania. Pojęcie n-gram dotyczy większych tokenów, które składają się z więcej niż jednego słowa, gdzie n jest liczbą słów w danym tokenie. Celem tokenizacji jest utworzenie zbioru słów, znajdujących się w dokumencie oraz utworzenie dyskretnych elementów, zawierających informacje z nieustrukturyzowanych danych tekstowych.

Najprostszy sposób podziału tekstu to wykorzystanie białych znaków do wydzielenia poszczególnych słów. Natomiast najbardziej podstawową metodą wektoryzacji jest wspomniany wcześniej one-hot encoding. Uzyskane w ten sposób wektory wykorzystuje się w modelach NLP. Taka reprezentacja numeryczna zachowuje gramatykę i kolejność występowania słów w dokumencie. Na tokenach można wykonywać operacje, które pozwolą w późniejszych etapach usprawnić model. Jednym z takich zabiegów jest usunięcie z tekstu słów, które najczęściej się pojawiają, nazywanych *stop words*. Dzięki temu można pozbyć się informacji nieistotnych, które jedynie utrudniają pracę na kolejnych etapach. Zmniejszenie w ten sposób zbioru słów pozwala zmniejszyć prawdopodobieństwo przeuczenia się wybranego modelu NLP. Innymi sposobami na redukcję zbioru słów są normalizacje:

- *case folding* – ujednolicenie zbioru w taki sposób, aby wszystkie słowa zaczynały się małą literą, dzięki czemu unika się powtórzeń tych samych wyrazów, które różnią się jedynie wielkością liter. Problemem mogą okazać się nazwy własne zapisywane wielką literą, żeby nie stracić tych informacji warto poddawać temu zabiegowi jedynie

pierwsze słowo w zdaniu. Są także przypadki, gdy ten typ normalizacji spowoduje pogorszenie jakości modelu.

- *stemming* – polega na usuwaniu drobnych różnic pomiędzy słowami, które ulegają odmianie: na przykład rzeczowniki w liczbie pojedynczej i mnogiej. Ten sposób obniża precyzję modelu, ponieważ wywołuje dużo „fałszywych alarmów” (ang. *false positive*). Jednak pozwala usprawnić pracę wyszukiwarek, ponieważ poszerza zakres wyszukiwania wyników.
- *lemmatization* – jest to bardziej zaawansowana operacja, polegająca na połączeniu słów o tym samym temacie (każdy wyraz składa się ze stałego tematu, mówiącego o znaczeniu danego słowa oraz z końcówki, która zmienia się w zależności od odmiany). W tym procesie mogą zostać połączone nawet te słowa, których temat zmienia się w trakcie odmiany lub te, które odmieniają się w sposób nieregularny.

2.3. Wektory TF-IDF

Inną metodą tworzenia numerycznych reprezentacji tekstów są wektory TF-IDF (ang. *term frequency - inverse document frequency*), które składają się z częstości występowania słów (ang. *token frequency*, TF) oraz odwrotnej częstości dokumentu (ang. *inverse document frequency*, IDF) [5]. Pierwsza część wskazuje na liczbę razy, jakie dane słowo pojawia się w dokumencie. Natomiast IDF mierzy, jak rzadkie jest słowo w całym korpusie dokumentów. Słowa, które występują często we wszystkich dokumentach mają niższą wartość, a słowa występujące częściej wyższą. Wskaźnik IDF pokazuje jak istotne są dane słowa dla konkretnych dokumentów. Warto wynik zlogarytmizować, ponieważ różnice w ważności słów mogą się wydawać mało znaczące z powodu zbyt dużych lub zbyt małych liczb.

Wektory TF-IDF tworzy się poprzez przemnożenie częstości występowania słów przez odwrotną częstość dokumentu dla każdego słowa w dokumencie. W przypadku operacji *one-hot encoding* otrzymuje się liczby typu całkowitego, natomiast TF-IDF zwraca liczby zmiennoprzecinkowe co pozwala na bardziej zaawansowane operacje matematyczne.

Zastosowanie wektorów TF-IDF ma kilka korzyści. Pierwsza to zmniejszenie wpływu często występujących słów, które są mało znaczące dla analizy, jak na przykład spójniki czy przyimki. Kolejną jest fakt, że wektory TF-IDF są zdolne do wyodrębniania istotnych cech słów, które mogą pomóc w różnych zadaniach przetwarzania tekstu. Na przykład, w zadaniu klasyfikacji tekstu, wektory TF-IDF mogą pomóc w identyfikacji kluczowych słów, które wskazują na przynależność do danej klasy.

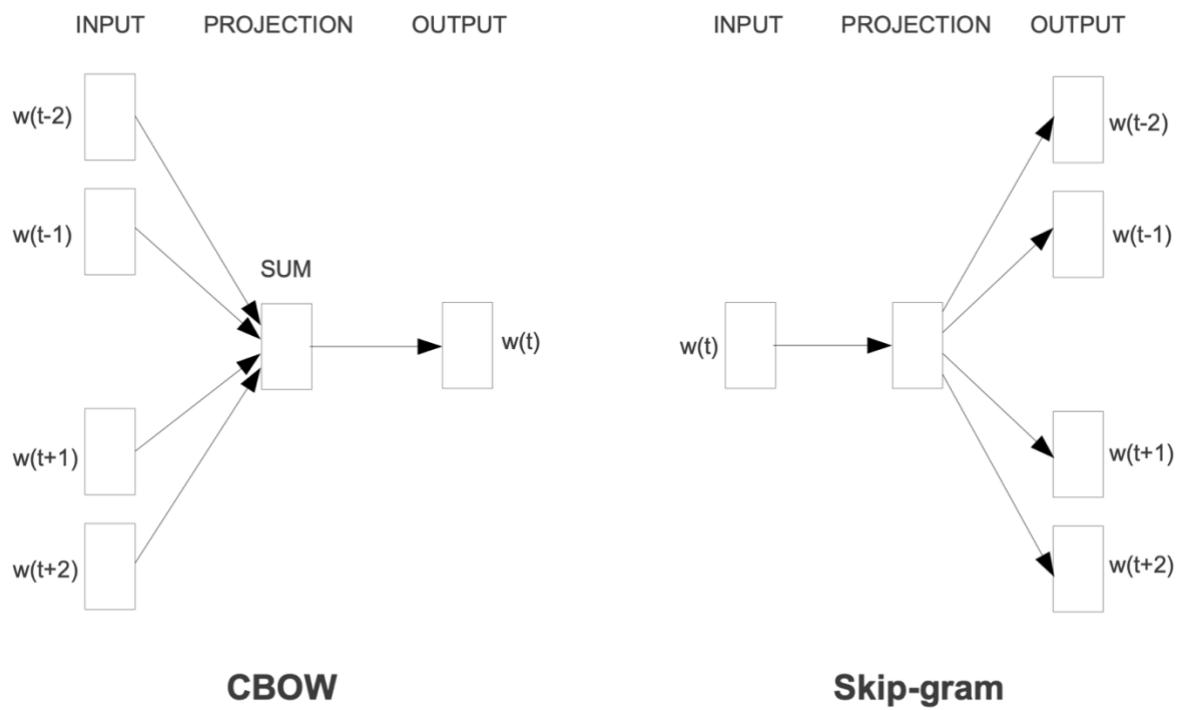
2.4. Word2vec

Przełomowa metoda word2vec została opracowana przez Tomasa Mikolova w 2012 roku [12] służy do reprezentacji słów jako wektorów o skończonej długości. Ta metoda umożliwia przekształcenie słów na punkty w przestrzeni o niskiej wymiarowości, które zawierają informacje o semantyce, oraz relacje między wyrazami. Word2vec jest nienadzorowanym algorytmem uczenia maszynowego i opiera się na hipotezie, według której słowa występujące w podobnych kontekstach mają podobne znaczenie. Oznacza to, że słowa, które są używane w podobny sposób, powinny być bliskoznaczne, a zatem ich reprezentacja wektorowa powinna być relatywnie podobne, wg zadanej miary, np. iloczynu skalarnego. Word2vec pozwala na wykrywanie semantycznych relacji między słowami, takich jak synonimy, antonimy czy kategorie semantyczne.

Forma reprezentacji numerycznej utworzona przez word2vec pozwala na przeprowadzanie działań dodawania i odejmowania na utworzonych wektorach, które dotyczą zależności semantycznych zwektoryzowanych tokenów. Jedną z ciekawych obserwacji jest podobna odległość pomiędzy rzeczownikami w liczbie pojedynczej i mnogiej w przypadku różnych takich zestawień. Dzięki temu rozwiązaniu model jest w stanie odpowiadać na pytania zadawane o analogie dotyczące przeróżnych zjawisk. Przykładem sukcesu word2vec jest wskazanie pewnej relacji pomiędzy wyrazami „mężczyzna”, „król”, „królowa” i „kobieta”. Mianowicie reprezentacja wektorowa word2vec pozwala zauważyc:

$\text{word2vec(„król”)} - \text{word2vec(„mężczyzna”)} + \text{word2vec(„kobieta”)} \sim \text{word2vec(„królowa”)}$,
gdzie word2vec(string) oznacza reprezentację numeryczną argumentu „string”.

Do wytrenowania modelu word2vec wykorzystuje się jeden z dwóch możliwych sposobów: *skip-gram* lub *Continuous Bag-of-Words* (CBOW). Pierwsze podejście polega na przewidywaniu kontekstu na podstawie danego słowa. Natomiast CBOW przewiduje wybrane słowo wykorzystując kontekst otaczających je wyrazów. Rys. 1 przedstawia porównawczy schemat działania metod *skip-gram* oraz CBOW.



Rys. 1. Prezentacja schematów działania metody *skip-gram* oraz CBOW. Źródło rysunku: [12]

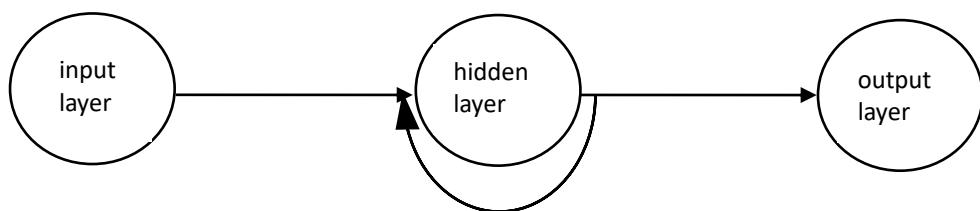
Obie metody są oparte na sieciach neuronowych, a wynikiem procesu uczenia jest zbiór wektorów reprezentujących słowa. W tabeli Tab. 1 został zaprezentowany przykład tworzenia par słów, które wykorzystywane w sposobie trenowania *skip-gram*. Pierwsze słowo z pary jest targetem, a następne kontekstem.

Liczba słów kontekstowych	Tekst	Skip-grams
1	dzisiaj jest piękna pogoda	(„dzisiaj”, „jest”)
	dzisiaj jest piękna pogoda	(„jest”, „dzisiaj”) („jest”, „piękna”)
	dzisiaj jest piękna pogoda	(„piękna”, „jest”) („piękna”, „pogoda”)
	dzisiaj jest piękna pogoda	(„pogoda”, „piękna”)
2	dzisiaj jest piękna pogoda	(„dzisiaj”, „jest”) („dzisiaj”, „piękna”)
	dzisiaj jest piękna pogoda	(„jest”, „dzisiaj”) („jest”, „piękna”) („jest”, „pogoda”)
	dzisiaj jest piękna pogoda	(„piękna”, „dzisiaj”) („piękna”, „jest”) („piękna”, „pogoda”)
	dzisiaj jest piękna pogoda	(„pogoda”, „jest”) („pogoda”, „piękna”)

Tab. 1 Przykładowe tworzenie par słów (target – kontekst) w algorytmie skip-gram wykorzystywanym do trenowanie modelu word2vec.

2.5. Rekurencyjne sieci neuronowe w NLP

Model word2vec opiera się na analizowaniu słów, jednak równie ważna jest praca z całym zdaniem, w sensie kontekstu i jego całościowego znaczenia. W tym celu zaczęto wykorzystywać rekurencyjne sieci neuronowe (ang. *Recurrent Neural Network*), które mają zdolność do analizowania sekwencji danych [5]. RNN zapamiętują i przechowują informację o poprzednich stanach i uwzględniają ją w analizie tekstu. Na Rys. 2 został przedstawiony schemat działania sieci rekurencyjnej.



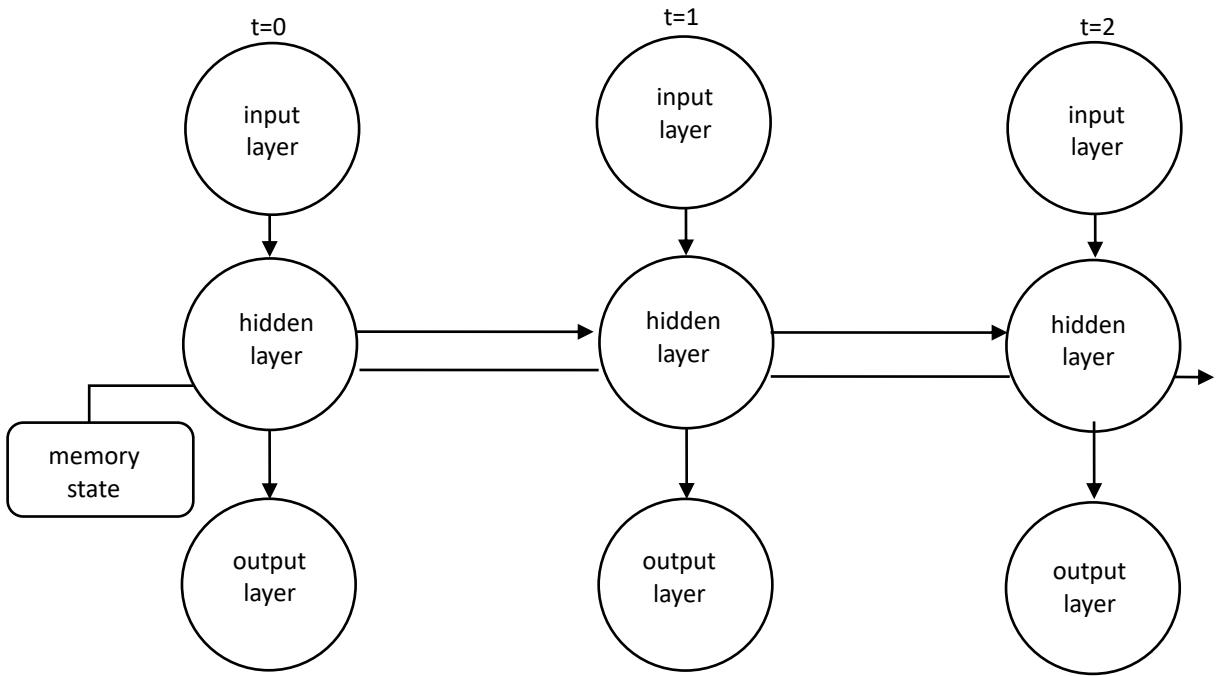
Rys. 2. Schemat neuronowej sieci rekurencyjnej. Dane zostają wprowadzone do sieci przez warstwę wejściową. Następnie są analizowane przez warstwy ukryte, a kierowane dalej informacje wyjściowe trafiają ponownie do opuszczanego komponentu, gdzie dołączają się do kolejnych danych dostarczanych przez warstwę wejściową.

Wykorzystanie rekurencyjnych sieci neuronowych w NLP polega na traktowaniu sekwencji słów jako wejścia, gdzie każde słowo jest podawane na wejście sieci, a stan wewnętrzny jest aktualizowany na podstawie informacji z poprzednich kroków czasowych. Dzięki temu RNN potrafi analizować tekst w kontekście, uwzględniając zależności między kolejnymi słowami. Sieć przy każdym kroku czasowym ignoruje dane wyjściowe, przekazując je jako element wejściowy w kolejnym kroku i koncentruje się jedynie na ostatnim wyjściu.

2.6. Long Short-Term Memory

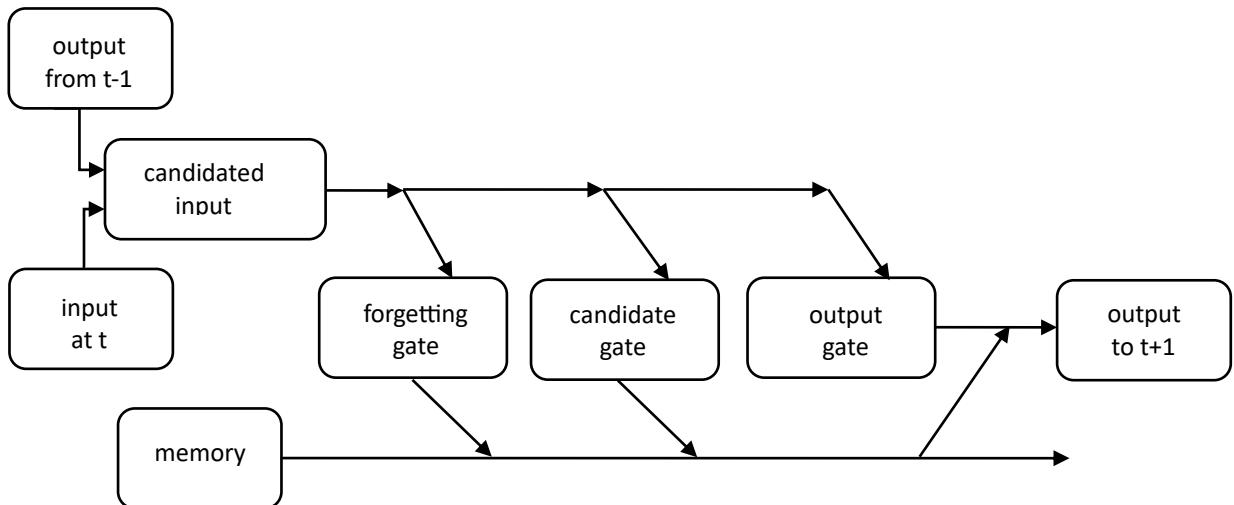
Szczególnym rodzajem rekurencyjnej sieci neuronowej wykorzystywanej do NLP jest Long Short-Term Memory (LSTM) [5]. LSTM został zaprojektowany w celu rozwiązania problemu zanikającego gradientu, który może wystąpić przy treningu głębokich sieci rekurencyjnych. Model ten był wykorzystywany do tłumaczenia języków naturalnych przy użyciu podejścia *seq2seq* (ang. *sequence to sequence*) [13].

LSTM wprowadza mechanizm bramkowy, który umożliwia kontrolowane przechowywanie, usuwanie i odczytywanie informacji z pamięci sieciowej. Składa się z komórek pamięci (ang. *memory cell*), bramek wejściowych (ang. *input gate*), bramek zapomnienia (ang. *forgetting gate*), bramek kandydata (ang. *candidate gate*) i bramek wyjściowych (ang. *output gate*). LSTM jest dodatkową warstwą w sieci rekurencyjnej, która działa równolegle do warstw ukrytych. Rys. 3 prezentuje schemat sieci LSTM.



Rys. 3. Schemat sieci neuronowej LSTM przedstawia działanie sieci dla trzech kroków czasowych. W każdym kroku dane są wprowadzane przez warstwę wejściową i kierowane do warstw ukrytych. Przez warstwy ukryte przepuszczane są informacje zdobyte przez model w poprzednich krokach czasowych, a proces zapamiętywania jest równoległy do uczenia się sieci.

Każda bramka LSTM jest warstwą typu *feedforward*, składającą się one z wag, które pozwolą usprawnić uczenie się sieci oraz z odpowiedniej funkcji aktywacji. W przypadku sekwencji wejściowej, każde słowo jest podawane na wejście LSTM w kolejnych krokach czasowych. Na początku komórka pamięci jest inicjalizowana wartościami początkowymi. Bramki zapomnienia określają, jakie informacje mają być usunięte z komórki pamięci, poprzez przemnożenie wektorów wag wejściowych i tych zapamiętanych z poprzedniego kroku. Następnie wektor trafia do bramki kandydata, gdzie odbywa się łączenie informacji pobranych na wejściu nowego kroku czasowego oraz tych z poprzedniego. W tej bramce mają miejsce dwie operacje. Do pierwsze jest używana funkcja aktywacji sigmoid i ma ona na celu wybranie elementów wektora wejściowego, które warto zapamiętać. Druga operacja wymaga funkcji aktywacji tanh i przekazuje ona zapamiętane elementy do odpowiedniego slotu. Na końcu dane trafiają do bramki wyjściowej, która kontroluje jakie informacje mają być odczytane z komórki pamięci i przekazane na wyjście LSTM. Rys. 4 przedstawia schemat bramek w sieci LSTM.



Rys. 4. Schemat bramek w sieci LSTM przedstawia proces wprowadzenia, selekcji i przekazania dalej informacji zapamiętanych w poprzednich krokach czasowych. Ten element sieci LSTM przyjmuje na wejściu dane z poprzedniego i obecnego kroku czasowego, które zostają przefiltrowane przez kolejne bramki: zapomnienia, kandydata oraz wyjściową.

Główną zaletą LSTM jest zdolność do przechowywania i analizowania długotrwałych zależności w sekwencjach danych. W tradycyjnych RNN problem zanikającego gradientu utrudnia nauczenie się takich zależności, ale LSTM potrafi utrzymywać informacje w pamięci przez dłuższą ilość sekwencji wejścia-wyjścia, co pozwala na analizę odległych kontekstów.

Sieć LSTM została wykorzystana jako podstawa modelu ELMO (ang. *Embeddings from Language Models*) opracowanego przez zespół naukowców z Uniwersytetu Stanforda. Innowacja tej technologii polega na wprowadzeniu do architektury LSTM dwukierunkowości. Dzięki takiemu zabiegowi model może analizować osadzenie kontekstualne słowa w przód oraz wstecz, co pozwala na zgromadzenie większej liczby informacji. ELMO znaczco poprawia wyniki zadań związanych z NLP, ponieważ jego główną zaletą jest wychwytywanie subtelnych znaczeń, które są uzależnione od kontekstu.

2.7. Architektura Transformers

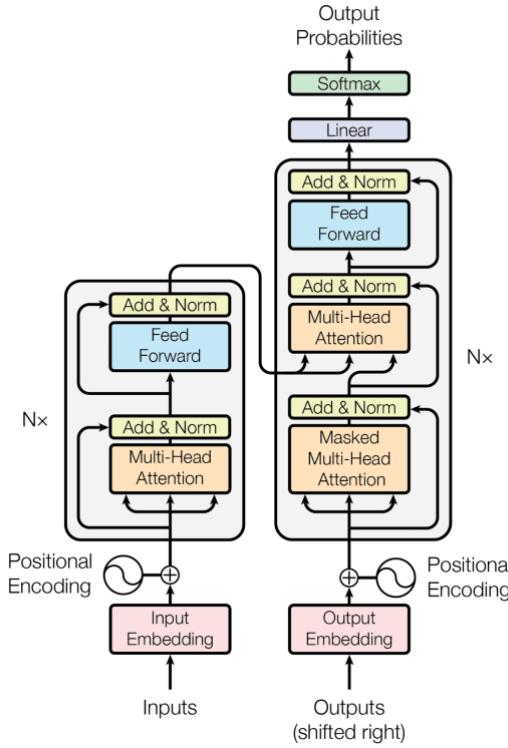
Kolejnym przełomowym krokiem w dziedzinie NLP był mechanizm atencji [14] oraz architektura Transformers. Mechanizm atencji w NLP pozwala wytrenować model tak aby skupiał swoją uwagę na odpowiednich częściach informacji podczas przetwarzania tekstu. Można go porównać do naszego umysłu, który skupia uwagę na pewnych elementach, a inne ignoruje. W architekturze Transformers, mechanizm atencji pozwala modelowi na skupienie się na różnych słowach w zdaniu i zrozumienie ich znaczenia w kontekście. Proces ten składa się

z trzech głównych kroków: *(i)* obliczenia ważności, *(ii)* obliczenia wyniku atencji oraz *(iii)* łączenia wyników [15,16].

Pierwszym krokiem jest obliczenie ważności, czyli określenie, które słowa w zdaniu są najważniejsze dla danego słowa. Jest to wykonywane poprzez obliczenie podobieństwa między słowami. Na przykład, jeśli model przetwarza zdanie "Pies goni kota", ważnością dla słowa "goni" mogą być słowa "pies" i "kot", ponieważ odnoszą się one bezpośrednio do akcji.

Kolejnym krokiem jest obliczenie wyniku atencji, który jest ważnym zsumowaniem wektorów reprezentujących słowa. Ważności określone w poprzednim kroku są używane jako wagi do przemnożenia odpowiednich wektorów. Wektor wynikowy reprezentuje koncentrację uwagi na poszczególnych słowach. Ostatnim krokiem jest łączenie wyników atencji z różnych warstw w modelu Transformers. Każda warstwa Transformers ma własny mechanizm atencji, a wyniki są łączone, aby zrozumieć kontekst całego zdania. Mechanizm atencji jest bardzo użyteczny w NLP, umożliwiając modelowi naukę relacji między słowami w zdaniu. Może to pomóc w tłumaczeniu, generowaniu podpisów obrazków, analizie sentymentu i wielu innych zastosowaniach, gdzie zrozumienie kontekstu jest kluczowe. Dzięki mechanizmowi atencji modele Transformers osiągają wysoką skuteczność w zadaniach językowych.

Transformery zostały wprowadzone w 2017 roku przez Ashish Vaswani, et al. [14] jako alternatywę dla rekurencyjnych sieci neuronowych (RNN) w celu radzenia sobie z problemem długotrwałych zależności w analizie sekwencji. W przeciwieństwie do RNN, które operują sekwencyjnie, transformery analizują całą sekwencję jednocześnie. Pozwala to na równoległą analizę, przez co model może być trenowany przy pomocy metod obliczeń równoległych, w przeciwieństwie do RNN, które z racji konstrukcji nie mogły być trenowane równolegle. Transformery składają się z dwóch głównych komponentów: enkodera i dekodera. Enkoder odpowiada za analizę wejściowej sekwencji słów i generuje reprezentacje kontekstowe dla każdego słowa. Dekoder z kolei używa tych reprezentacji do generowania wyjściowej sekwencji. Na Rys. 5 została przedstawiona w sposób graficzny architektura Transformers.



Rys. 5. Schemat architektury modelu Transformer. Źródło: [14]

Główną techniką w transformerach jest wspomniany wcześniej mechanizm atencji (ang. *attention*), który pozwala na przechowywanie i odwoływanie się do różnych części sekwencji podczas generowania reprezentacji. Umożliwia to bardziej globalne zrozumienie kontekstu i uwzględnianie długotrwałych zależności między słowami. Mechanizm uwagi składa się z kilku warstw. Na początku obliczana jest miara podobieństwa między każdym słowem a innymi słowami w sekwencji. Następnie obliczane są wagi, które określają, jak dużo uwagi należy przywiązać do różnych słów podczas generowania reprezentacji. Te wagi są stosowane do ważonego sumowania reprezentacji słów, tworząc ostateczną reprezentację kontekstową. Warstwy transformerów są stosowane zarówno w enkoderze, jak i dekoderze. Każda warstwa transformatora składa się z dwóch podwarstw: warstwy uwagi i warstwy sieciowej. Warstwa uwagi wykonuje mechanizm samoobserwacji, podczas gdy warstwa sieciowa wykonuje operacje liniowe i nieliniowe na wyjściu warstwy uwagi.

Przykładami znanych modeli opartych na transformatorach są np. BERT [17] (*Bidirectional Encoder Representations from Transformers*), czy GPT (*Generative Pre-trained Transformer*). Jest także kilka modeli wytrenowanych przy użyciu danych w języku polskim są to między innymi: polBERT oraz herBERT [18].

3. Wstęp do analizy literaturoznawczej

W rozdziale 3. zostaną wyjaśnione pojęcia z obszaru literaturoznawstwa: styl literacki oraz stylizacja. Przedstawione będą również sylwetki poetów, których wybrano do przeprowadzenia analizy w niniejszej pracy. Przytoczone zostaną fakty z biografii pisarzy, które miały znaczący wpływ na ich twórczość. W tym rozdziale zawarte są także informacje dotyczące charakterystyki stylów poszczególnych autorów.

3.1. Czym jest styl w literaturze

Styl tekstu literackiego jest zagadnieniem badanym zarówno przez literaturoznawców, jak i językoznawców. Aleksander Wilkoń definiuje styl jako element zawierający się w pojęciu języka, gdzie język jest pojęciem szerszym [19]. Można zatem stwierdzić, że styl to szczególny typ języka. Należy także wyjaśnić, czym jest tekst literacki. Podążając za analizą Wilkonia, który posługuje się teorią strukturalistyczną, dowiedzieć się można, że język utworu literackiego charakteryzuje się następującymi cechami:

- język tekstu jest spójną, zaplanowaną całością,
- wszystkie elementy utworu są nieprzypadkowe,
- język tekstu literackiego jest szczególnym przypadkiem języka funkcjonalnego,
- język ten posiada odróżniające go cechy, które mają podkreślać znaczenie tekstu.

Wilkoń wskazuje jednak zróżnicowane koncepcje językoznawców, którzy podejmują temat stylu. Cechy, które bezsprzecznie zostały uznane przez większość badaczy to:

- wyjątkowość płynąca za stylem, która ma świadczyć o niepospolitości tekstu,
- styl ma charakter funkcjonalny,
- styl tworzy się poprzez dobieranie środków językowych.

Językoznawcy spierają się między innymi, czy styl jest indywidualny, czy jest wspólny dla pewnej społeczności oraz czy można mówić o stylu w tekstach, które pełnią funkcję komunikatywną, a nie jedynie ekspresywną lub impresywną. W przypadku wybranych przeze mnie poetów, każdy z nich stosuje charakterystyczny dla siebie styl, który wynika z różnych aspektów, wśród których można wymienić epokę literacką, wydarzenia historyczne, płeć, wiek lub przeżycia osobiste.

Innym bardzo istotnym zjawiskiem literackim obok stylu jest stylizacja, czyli dobieranie słów, składni i środków stylistycznych w taki sposób, aby wypowiedź nabrala cech wybranego przez nadawcę stylu np. urzędowego lub młodzieżowego. Stylizacja może także dotyczyć

elementów literatury. Autorzy naśladują poetykę charakterystyczną dla danej epoki lub czerpią inspirację bezpośrednio z twórczości wybranego artysty.

Do projektu klasyfikacji wybrano czterech poetów oraz cztery poetki, wybór ten był kierowany odpowiednim zbalansowaniem danych. Ze względu na rozróżnienie rodzaju męskiego i żeńskiego w koniugacji w języku polskim, takie zestawienie mogło okazać się istotne w aspekcie klasyfikacji. Każdy z wybranych pisarzy wykreował charakterystyczny dla siebie styl. Wybrani poeci to: Jan Kochanowski, Krzysztof Kamil Baczyński, Czesław Miłosz, Zbigniew Herbert, Wisława Szymborska, Halina Poświatowska, Maria Pawlikowska-Jasnorzewska oraz Ewa Lipska. Poniżej przedstawiam po krótce charakterystykę tych postaci oraz ich twórczości.

3.2. Analiza literaturoznawcza wybranych poetów

W tym podrozdziale zostaną omówione sylwetki wybranych autorów. Przybliżę zwięźle w jakiej epoce tworzyli oraz jakie wydarzenia historyczne i osobiste są istotne w analizie ich dzieł. Krótko scharakteryzuję także styl każdego poety, zwracając uwagę na kluczowe elementy.

3.2.1. Jan Kochanowski

Najstarszym twórcą spośród wybranych przeze mnie autorów jest Jan Kochanowski [20]. Poeta, dramaturg oraz tłumacz żyjący w Renesansie – wiek XVI. Pisarz kształcił się w Krakowie, Królewcu oraz w Padwie, dzięki czemu pomimo niezamożności jego rodziców, obracał się w środowisku najwybitniejszych humanistów. Karierę Kochanowskiego w Polsce wspierały zarówno zamożne rodziny arystokratyczne, ale także wysokie osobistości kościoła, co pozwoliło mu rozpocząć pracę na dworze królewskim. Jego twórczość rozwinęła się jeszcze bardziej, gdy zamieszkał w dziedziczonym po ojcu majątku w Czarnolesie, gdzie żył szczęśliwie ze swoją żoną i dziećmi. Jan Kochanowski był najwybitniejszym poetą polskiego Renesansu. Tworzył zgodnie z ówczesnym kanonem. Literatura renesansowa charakteryzowała się powrotem do antycznych wartości, które miały być odpowiedzią na okres średniowieczny. Główną zmianą jako zaszła w tej epoce było stawienie człowieka w centrum uwagi. Praktykowano hedonizm, afirmację życia i szczęście w trakcie życia ziemskiego. Bardzo ważnym aspektem twórczości Kochanowskiego jest także filozofia, która również swoje korzenie ma w Starożytności. Utwory poety można podzielić na te pisane po łacinie, w języku polskim lub tłumaczone na język polski. Do projektu wybrane zostały pieśni z *Książ pierwszych* i *Książ wtórych*, a także fraszki ze zbioru *Fraszki księgi pierwsze*, *Fraszki księgi drugie*, *Fraszki*

księgi trzecie oraz *Treny*. Utwory są zróżnicowane tematycznie. Wśród dzieł Kochanowskiego znajdują się takie, które opowiadają o miłości, zabawie, beztroskim życiu oraz te, w których podejmuje temat przemijalności, utraty wiary w Boga czy cierpienia po stracie dziecka. Pieśni podzielone na *Księgi pierwsze* i *Księgi wtóre* różnią się w aspekcie tematycznym. Drugi zbiór zawierał mniej utworów poświęconych miłości, natomiast pojawiło się w nim dużo więcej dzieł, które były inspirowane twórczością Horacego. Kompozycja pieśni jest spójna dla obydwu ksiąg. Kochanowski najczęściej pisał czterowersem rymowanym parzyście (aa bb). Fraszki są najbardziej charakterystycznym gatunkiem dla Kochanowskiego, ponieważ on jako pierwszy użył tego słowa określając utwór liryczny oraz wprowadził ten gatunek do literatury polskiej. Zazwyczaj są to wiesze krótkie i humorystyczne, ale mogą również mieć charakter refleksyjny. Często w tytule zawierają apostrofę, która wskazuje odbiorcę utworu. *Treny* są cyklem utworów poświęconych zmarłej córce autora, których celem było dopełnienie żałoby po stracie. Do czasów Renesansu tego typu utwory żałobne pisano jedynie dla ważnych osobistości lub bohaterów. Kochanowski był pierwszym polskim pisarzem, który uczynił dziecko bohaterem mowy pogrzebowej. Jan Kochanowski jest jednym poetą renesansowym wśród wybranych do projektu pisarzy, więc w tym zestawieniu również charakterystycznym dla niego będzie staropolski język.

3.2.2. Krzysztof Kamil Baczyński

Następnym autorem jest Krzysztof Kamil Baczyński [21], który żył i tworzył w pierwszej połowie XX wieku. Był jednym z pokolenia Kolumbów, czyli twórców, którzy urodzili się około 1920 roku, a czas wojny przypadał na ich okres wchodzenia w dorosłość - to właśnie II wojna światowa była dla nich wydarzeniem pokoleniowym, wielu z nich zmarło podczas działań wojennych. Baczyński zmarł wieku 23 lat w czasie Powstania Warszawskiego. Szanowane postaci literatury polskiej, jak Stanisław Pigoń czy Tadeusz Gajcy uważali Baczyńskiego za najwybitniejszego poetę okresu wojennego. Dla przedstawicieli pokolenia Kolumbów sztuka była możliwością na poradzenie sobie z trudną sytuacją, jaką była trwająca wojna. Baczyński również dużo miejsca w swojej poezji poświęcił wojnie. Charakterystycznym dla niego zabiegiem były elementy oniryczne oraz uniwersalny sposób opisywania rzeczywistości, co sprawiało, że jego dzieła nie odnosiły się jedynie do jemu współczesnych wydarzeń, ale w dużo bardziej ogólny sposób pisały o tragedii wywołanej wojną. Warto podkreślić, że okres twórczości Baczyńskiego był krótki, ponieważ wynosił około 5 lat. Pomimo ogromnego dorobku w postaci ponad 500 tekstów, jego twórczość nie miała możliwości ewoluować. Utwory pisane przez Baczyńskiego nie zmieniały się, ponieważ on nie miał możliwości doświadczyć życia w pełni, jako dorosły czy starzejący się człowiek.

Poeta często kreując pejzaże w swoich utworach osadzał je w krajobrazie zimowym, co współgrało z onirycznym charakterem wiersza. Młody autor pisał także o przeczuciu nadchodzącej śmierci, ale jego twórczość nie była jedynie katastroficzna, wiele razy wspominał o nadziei i odrodzeniu nowego pokolenia. Bardzo charakterystycznym elementem utworów Baczyńskiego jest podawana do końca data, kiedy wiersz został napisany. Do projektu wybrane zostały działa z tomów: *Zamkniętym echem*, *Dwie miłości*, *Modlitwa*, *Wiersze wybrane* oraz *Śpiew z pożogi*.

3.2.3. Czesław Miłosz

Trzecim poetą jest Czesław Miłosz [22], laureat Literackiej Nagrody Nobla, którą otrzymał w 1980 roku. Był poetą, eseistą, prozaikiem, tłumaczem oraz historykiem literatury. Urodził się na początku wieku XX, okres okupacji niemieckiej przeżył jako trzydziestokilkuletni mężczyzna, kilka lat po zakończeniu wojny wyjechał za granicę i na emigracji żył ponad 40 lat. W tym czasie tworzył i rozwijał karierę uniwersytecką we Francji, a później w Stanach Zjednoczonych. Twórczość Miłosza była cenzurowana w Polsce do lat 80. XX wieku. Na początku swojej drogi pisarskiej założył awangardową grupę literacką Żagary. Do wybuchu II wojny światowej jego twórczość miała charakter katastroficzny. Miłosz przewidywał upadek sztuki i ludzkości. Cechą, która wyróżniała pisarza od innych twórców zrzeszonych w Żagarach, było konsekwentne realizowanie przez Miłosza klasycystycznych założeń dotyczących budowy dzieła. Oznacza to, że wiersze są regularne, strofy zazwyczaj mają tą samą liczbę wersów, autor przykłada także uwagę do zachowania rytmu w rozłożeniu akcentów. Utwory z tego okresu charakteryzują się również bogatą metaforeką i poważnym tonem. Po okresie II wojny światowej Miłosz zmienił wydźwięk swojej twórczości, ponieważ nie była ona już tak pesymistyczna. Pisarz szukał sposobu na odrodzenie kultury i człowieczeństwa, które przez straszne wydarzenia wojenne przechodziły niewątpliwy kryzys. Według Noblisty odpowiedzi należało szukać właśnie w sztuce, a konkretnie w antyku i klasycznych zasadach tworzenia. Twórczość Miłosza jest zróżnicowana, ponieważ okres, w którym pisał to około 70 lat, więc w tym czasie wielokrotnie przechodził przez kolejne doświadczenia, które wpływały na ukształtowanie osobowości, a tym samym na jego sferę artystyczną. Wiersze wybrane zostały z różnych etapów życia Miłosza między innymi z tomów: *Ocalenie*, *Światło dzienne*, *Miasto bez imienia*, *Gdzie wschodzi słońce i kiedy zapada*, *Dalsze okolice*, *To*.

3.2.4. Zbigniew Herbert

Czwartym wybranym pisarzem jest Zbigniew Herbert [23]. Czas jego dorastania przypada po części na okres II wojny światowej oraz stalinizmu, który nastąpił zaraz po 1945. Herbert ukończył trzy kierunki studiów: ekonomię, prawo i filozofię. Jednak jego postawa moralna nie pozwalała mu na robienie kariery zawodowej, ponieważ musiałby opowiedzieć się po stronie komunizmu. Z powodu panującego w sztuce socrealizmu nie mógł także publikować swoich tekstów. Przez ten okres żył, więc bardzo skromnie. W roku 1956, gdy miała miejsce odwilż polityczna, wydał swój pierwszy tom *Struna światła*. Wtedy też warunki materialne pisarza uległy poprawie. Herbert za swoje pierwsze stypendium od Ministra Kultury i Sztuki wyjechał za granicę, żeby zobaczyć świat. Odkrył tym samym swoją nową pasję - podróżowanie, która napędzana była głównie niechęcią pisarza do rzeczywistości PRL. Najbardziej znanym cyklem poetyckim Herberta jest seria o Panu Cogito. Postać ta została wymyślona przez autora i stała się stałym bohaterem jego utworów. Za pośrednictwem Pana Cogito Herbert przekazywał swoją postawę moralną oraz emocjonalny stosunek wobec rzeczywistości, nierzadko jednak pisarz stosował zabieg ironii i wtedy słowa wypowiadane przez bohatera należało odczytać opacznie. Jeszcze innym sposobem na niebezpośrednie prezentowanie postawy pisarza była dyskusja pomiędzy Panem Cogito, a autorem tekstu. Przez wielu Herbert był uznawany za osobę, która odznaczała się przykładową moralnością. Głównym tematem poruszonym przez poetę w jego dziełach jest filozofia, rozważania nad moralnością oraz analiza rzeczywistości. Wielokrotnie w utworach Herberta pojawiają się postacie, miejsca lub określenia pochodzące z kultury Antyku, co również wskazuje na klasycyzujący charakter jego twórczości. Powrót do Starożytności widać jedynie w przytaczanych motywach mitologicznych, ale nie w budowie utworów, które zazwyczaj nie posiadają regularnej struktury, rymów, czy powtarzalnego rytmu akcentów. Herbert korzysta ze współczesnego podejścia do budowy wiersza. Strofy w jego utworach są, więc nieregularne, wersy są różnej długości i nie rymują się. Wybrane wiersze pochodzą z tomów: *Struna światła*, *Hermes*, *pies i gwiazda*, *Elegia na odejście*, *Raport z oblężonego Miasta*, *Pan Cogito*, *Napis*, *Studium przedmiotu*.

3.2.5. Wisława Szymborska

Pierwszą artystką jest Wisława Szymborska [24], laureatka Literackiej Nagrody Nobla z 1996 roku. Była wybitną poetką, eseistką, tłumaczką, felietonistką i krytyczką literacką. Jej debiut poetycki miał miejsce 1945 roku. Na początku swojej kariery tworzyła w stylu socrealizmu, za co w późniejszych latach bardzo ją krytykowano. Mimo tego jej pierwszy tom

poetycki został zablokowany przez cenzurę PRL-u. Pisarka była twórczynią kilku form lirycznych, jedną z nich jest np. Lepiej. Jest to jednozdaniowy, najczęściej dwuwersowy utwór, którego koncept opiera się na porównaniu dwóch nieprzyjemnych sytuacji, gdzie pierwszy wers zaczyna się od słowa „Lepiej” lub „lepszy”, „lepsza”, a drugi od wyrazu „niż”. Utwór miał charakter groteskowy. Szymborska w swoich utworach często korzystała z ironii oraz humoru. Teksty Noblistki nie zamykały się w konkretnych schematach, ale widoczne było w nich zainteresowanie rozwojem ludzkości oraz analityczne podejście do rzeczywistości. Szymborska była spokojną, stonowaną osobą i taka też była jej twórczość. Poetka nie stosowała patetycznych, górnolotnych stwierdzeń. Korzystała natomiast z bardziej subtelnych zabiegów stylistycznych jak kontrast, paradoks czy gry lingwistyczne. Mimo pozornej prostoty jej utwory są bardzo bogate w przemyślenia i krytykę świata. Szymborska w swoich tekstach zadaje często proste, naiwne pytania, ale robi to, żeby na nowo je zinterpretować i odkryć jakieś nowe znaczenie, które się za nimi kryje. Kontrast w twórczości Noblistki polega na zderzeniu ze sobą refleksji nad ogólnoludzkimi problemami, która wynika z analizy drobnego, niezauważalnego szczegółu. Bardzo ważnym elementem twórczości Szymborskiej są zwierzęta, którym poświęca dużo przestrzeni w swoich wierszach. Pisarka często stosuje lirykę maski, czyli prezentuje swoje przemyślenia przypisując je komuś lub czemuś innemu. W tym właśnie zabiegu często wykorzystuje zwierzęta lub nawet animizując przedmioty codziennego użytku. Utwory do pracy zostały wyselekcyjowane między innymi z tomów: *Dlatego żyjemy*, *Wszelki wypadek*, *Wielka liczba*, *Dwukropek*.

3.2.6. Halina Poświatowska

Kolejną poetką jest Halina Poświatowska [25]. Jest uznawana za przedstawicielkę pokolenia „Współczesności”, czyli pisarzy debiutujących po roku 1956. Na jej twórczość ogromnie wpłynęła choroba serca, z którą poetka zmagała się przez większość swojego życia i która również stała się przyczyną jej śmierci. W poezji Poświatowskiej łączą się ze sobą motywy śmierci i miłości, ponieważ artystka była świadoma niepewności swojego losu. Drugim bardzojącym elementem twórczości poetki jest śmiało wyrażana przez nią kobiecość. Poświatowska pisze o kobiecym ciele, uczuciach i emocjach, które są charakterystyczne dla twórczości kobiet. Virginia Woolf – pisarka i eseistka – wybitna działaczka na rzecz kobiet w XX wieku podjęła się opisania stylu kobiecego w literaturze. Teksty pisane przez kobiety są bardziej chaotyczne, mniej ustrukturyzowane niż te tworzone przez mężczyzn. Kobiety pozwalają sobie na swobodny przepływ myśli, które w takiej formie, w jakiej się pojawiają przelewają na papier wraz ze wszystkimi emocjami, jakie im przy tym towarzyszą. Somatyzm, czyli pisanie o ciele jest również znakiem rozpoznawczym dla

pisarstwa kobiecego. Wiele z tych cech charakteryzuje także twórczość Poświatowskiej. Pomimo bardzo trudnej sytuacji życiowej poetka nie popadała w pesymizm. Kruczość jej życia sprawiała, że jeszcze bardziej chciała tego życia doświadczać. Charakterystyczne w jej wierszach jest częste wyznawanie uczuć, gwałtowna emocjonalność i zmysłowość. Jednocześnie w teksthach Poświatowskiej nie brakuje filozoficznych refleksji na temat przemijania czy kondycji człowieka. Warto także podkreślić, że artystka zmarła bardzo młodo, bo w wieku 32 lat. Krótki okres twórczości sprawia, że pisarka nie miała czasu, aby rozwinąć swoją twórczość, przez co tematyka jej wierszy nie jest zbyt zróżnicowana, ale równocześnie bardzo dla niej samej charakterystyczna. Utwory wybrane zostały z tomów: *Hymn bałwochwalczy*, *Dzień dzisiejszy*, *Oda do rąk*.

3.2.7. Maria Pawlikowska-Jasnorzewska

Trzecią poetką jest Maria Pawlikowska-Jasnorzewska [26]. Tworzyła w okresie dwudziestolecia międzywojennego. Była w tym czasie związana z grupą literacką Skamander. Skamandryci inspirowali się przede wszystkim twórczością Leopolda Staffa. Panowała opinia, że są oni grupą sytuacyjną, czyli taką, która wiążą różne spotkania i aktywności okołoliterackie. Głównymi założeniami Skamandrytów była bezprogramowość, wolność literacka, język ma być raczej prosty, zawierający kolokwializmy lub wulgaryzmy, a poeta miał być częścią tłumu. Pawlikowska-Jasnorzewska jako artystka spotykająca się z członkami Skamandra przejęła od nich wybrane cechy poetyki. Jej twórczość cechuje witalizm oraz skupienie uwagi na codzienności. Pawlikowska-Jasnorzewska wypracowała charakterystyczną dla siebie formę miniatury poetyckiej. Są to krótkie, czterowersowe wiersze zakończone ironiczną puentą. Nie jest to nowy gatunek liryyczny, bo wcześniej często korzystał z niego Jan Kochanowski, ale poetka zaadaptowała go do współczesnych realiów. Warto również zwrócić szczególną uwagę na środki stylistyczne, często pojawiające się w lirykach artystki. Stosowanie oksymoronów, paradoksów i zainteresowanie brzydotą wskazuje na inspiracje sztuką barokową. Jednak zamiłowanie do tradycyjnej formy nie sprawia, że pisarka ogranicza się w sprawie doboru tematów. W kwestii tematyki Pawlikowska-Jasnorzewska jest bardzo nowoczesna. Otwarcie mówi o miłości fizycznej, aborcji czy relacjach pozamałżeńskich. Z tego powodu przez wielu krytyków zostaje okrzyknięta poetą skandalizującą. Również charakterystyczną dla artystki są krótkie formy poetyckie o tematyce miłosnej, czyli erotyki, w których uczucia przekazywała w sposób bezpośredni. Twórczość Pawlikowskiej-Jasnorzewskiej ewoluowała na przestrzeni czasu. Na początku jej wiersze były pełne humoru, ironii i beztroskiej radości życia. Natomiast w późniejszych latach, gdy autorka przekroczyła 35 rok życia, poezja staje się bardziej refleksyjna, tematyka kieruje się w stronę przemijania. W trakcie wojny twórczość pisarki staje

się jeszcze bardziej ponura, często omawianymi problemami jest istnienie i działanie zła oraz zagadnienia metafizyczne. Utwory wybrane do klasyfikacji pochodzą z tomów: *Pocałunki*, *dancing. karnet balowy*, *Profil białej damy*, *Szkicownik poetycki*.

3.2.8. Ewa Lipska

Ostatnią wybraną autorką jest Ewa Lipska [27]. Wśród wytypowanych pisarzy jest jedną obecnie żyjącą postacią. Na początki kariery była związana z pokoleniem Nowej Fali, czyli artystów, którzy na świat przyszli już po II wojnie światowej i debiutowali pod koniec lat 60. XX wieku, ale nie była członkinią ugrupowania. Nową Falę charakteryzuje refleksyjna, krytyczna postawa wobec rzeczywistości. Członkowie grupy jako młode osoby chcieli tworzyć realną opozycję wobec zastanego ustroju politycznego. Głównym narzędziem Nowej Fali był język zarówno prosty, a także pełny gier językowych i niejednoznaczności. Celem takich zabiegów jest obnażenie wszechobecnego zakłamania i przeinaczania prawdy. Ewa Lipska również reprezentuje idee wolnościowe. Poetka uważnie obserwuje rzeczywistość i dostrzega niepokojące zmiany, jakie zachodzą we świecie. Jednym z problemów, jakie pojawiają się w jej twórczości jest osamotnienie, które wiąże się z kondycją współczesnego człowieka. Autorka porusza także kwestie dotyczące niepewnej lub nieprawdziwej wolności. Lipska, jak większość twórców Nowej Fali stosuje w swoich wierszach różnorakie gry językowe. Poetka często korzysta także z metafor o charakterze ironicznym. Mimo tych zabiegów jej teksty są klarowne i dostępne w odbiorze. Ze względu na prosty język i osadzenie utworów w tematyce codzienności Lipska jest porównywana do Wisławy Szymborskiej, z którą łączyć może ją również wyciszający charakter wierszy. Poetka często sięga również do poetyki neoklasycznej. Z tej tradycji zainspirowała się formami gatunkowymi, a mianowicie aforyzmem i przypowieścią. Często wiersze Lipskiej mają charakter aforystyczny, gdy chce przekazać własne refleksje dotyczące świata. Z powodu wnikliwej analizy rzeczywistości poezję autorki cechuje nastrój niepokoju. Wynika on z tego, że pisarka dostrzega nieprawidłowości w świecie oraz fakt, że niemożliwym jest, aby je naprawić. Nie mówi o tym jednak w sposób patetyczny czy pesymistyczny. Poetka ma raczej spokojne podejście nawet do najtrudniejszych tematów, jak na przykład samobójstwo. Wiersze wybrane do pracy pochodzą między innymi z tomów: *Wiersze*, *Drugi zbiór wierszy*, *Trzeci zbiór wierszy*, *Żywa śmierć*, *Utwory wybrane*.

4. Analiza zbioru danych. Metody nienadzorowane: redukacja wymiarowości i analiza skupień

W tym rozdziale przedstawiona zostanie analiza zbioru danych. Do klasyfikacji została użyta autorska baza danych, składająca się z 400 wierszy autorów wymienionych w rozdziale 3., po 50 wierszy dla każdego z nich. Dobór został pokierowany różnorodnością w kwestii: czasu historycznego, płci, okresu twórczego (jak długi on był) oraz podejmowanej tematyki. Baza danych zawierała następujące informacje: tekst wiersza, klasę wiersza, czyli imię i nazwisko autora oraz tytuł utworu. Tekst dobrano z różnych etapów twórczości poszczególnych pisarzy w przypadkach, gdzie było to możliwe.

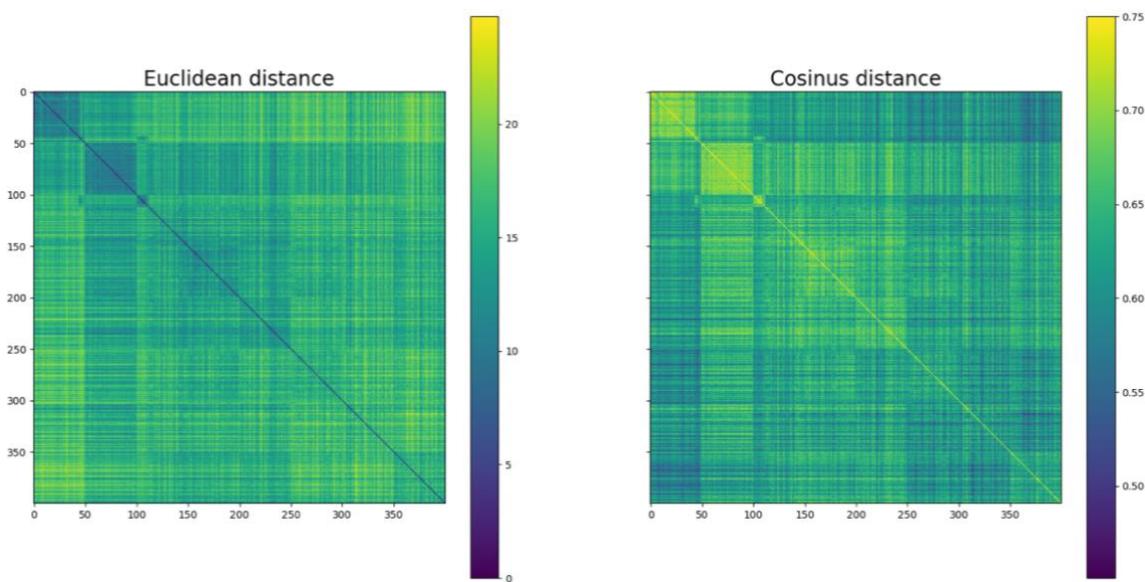
Analiza ilościowa rozważanego zbioru danych oparta jest na reprezentacji wektorowej każdego wiersza, która została uzyskana przy pomocy modelu herBERT w wersji KLEJ (Kompleksowa Lista Ewaluacji Językowych) [28]. Model herBERT działa na reprezentacji wektorowej 512 całostek (średnia liczba słów w wierszach wynosi 126). Model herBERT w wersji KLEJ pozwala na wygenerowanie reprezentacji wektorowej każdego wiersza, która składa się z 768 liczb rzeczywistych.

W kolejnych sekcjach skupię się na analizie odległości wektorów zanurzeń oraz analizie skupień po ówczesnej redukcji wymiarowości każdego z wierszy.

4.1. Analiza odległości reprezentacji wektorowych wierszy

W pierwszym kroku zostały obliczone odległości euklidesowe oraz kosinusowe pomiędzy każdą parą wierszy w zbiorze danych, Rys. 6. Klasy próbek dzielone są w zbiorach po 50 wierszy. Patrząc na osie są to kolejno: Jan Kochanowski (0-49), Krzysztof Kamil Baczyński (50-99), Czesław Miłosz (100-149), Zbigniew Herbert (150-199), Wisława Szymborska (200-249), Halina Poświatowska (250-299), Maria Jasnorzewska-Pawlikowska (300-349), Ewa Lipska (350-399).

W pierwszym kroku analizy zebranych wierszy widoczne są różnice pomiędzy twórcami płci męskiej i żeńskiej. Odległości euklidesowe pomiędzy wierszami pisany przez mężczyzn są mniejsze niż wierszy pisanych przez kobiety. Zatem poeci oraz Szymborska są pomiędzy klasami bardziej do siebie podobni, a poetki bardziej różnorodne. Dla kilku autorów są widoczne są wyraźne kwadraty w obrębie klasy o małej odległości euklidesowej, co wskazuje na podobieństwo wewnętrzklasowe. Duże podobieństwa (małe wartości odległości) są dobrze widoczne dla wierszy Jana Kochanowskiego, Krzysztofa Kamila Baczyńskiego, Zbigniewa Herberta, Wisławy Szymborskiej oraz Ewy Lipskiej.



Rys. 6. Graficzne przedstawienie odległości euklidesowych (lewy panel) oraz cosinusowych (prawy panel) dla wektorowej reprezentacji analizowanego zbioru danych składającego się z 400 wierszy. Pierwsze 200 wektorów to zbiór mężczyzn, a wektory od 201 do 400, to zbiór kobiet.

4.2. Redukcja wymiarowości i analiza skupień wierszy wszystkich klas

Analiza skupień polega na znalezieniu niskowymiarowej (dwu- lub trójwymiarowej) reprezentacji każdego wiersza i zaznaczeniu jego położenia w 2- lub 3- wymiarowej przestrzeni, a następnie próby wyodrębnienia skupisk punktów odpowiadających danej klasie. Redukcja wymiarowości dokonana jest przy użyciu trzech metod numerycznych – metody PCA (ang. *Principal Component Analysis*), oraz dwóch nieliniowych metod: t-SNE (ang. *t-distributed Stochastic Neighbor Embedding*), oraz UMAP (ang. *Uniform Manifold Approximation and Projection*), które są zaimplementowane w bibliotece *scikit-learn*.

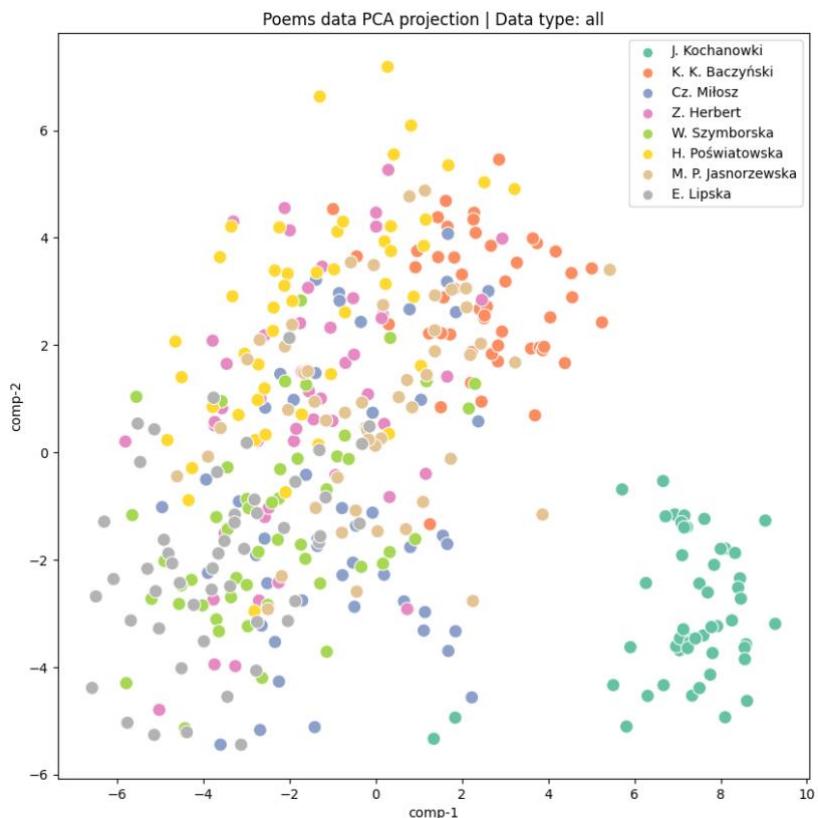
Ze względu na różnice w analizie odległości euklidesowych i kosinusowych dane zostały podzielone na twórczość mężczyzn i kobiet, żeby dokładniej zbadać, jak dane prezentują się w charakterystycznych dla siebie grupach. Przygotowano również analizę wspólną dla wszystkich autorów oraz z uwzględnieniem podziału na płeć.

4.2.1. PCA

Redukcja wymiarowości oparta na PCA polega na policzeniu iloczynu skalarnego analizowanego wektora z kilkoma wektorami własnymi macierzy korelacji danych, które mają największe wartości własne. W ten sposób, wybierając d wektorów własnych macierzy

korelacji, jesteśmy w stanie uzyskać reprezentację oryginalnego wektora w d wymiarach. W celu wizualizacji danych wykorzystuje się $d = 2$ oraz $d = 3$.

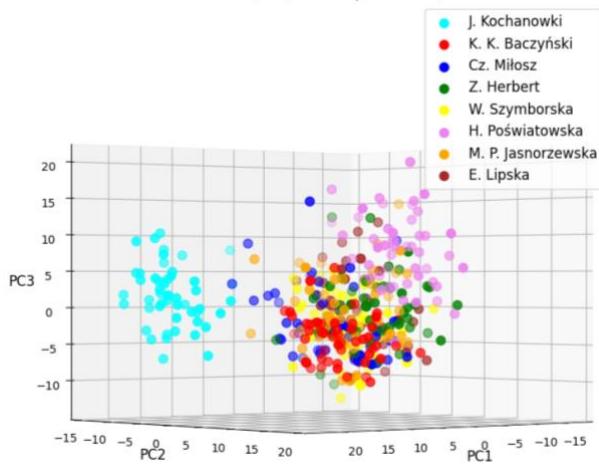
Rys.7 prezentuje dane dla wszystkich autorów przy użyciu PCA dla $d = 2$. Można zaobserwować tylko jeden wyraźny klaster, którym są utwory Jana Kochanowskiego. Najpewniej odróżnia się on od reszty pisarzy przez używanie staropolskiego języka.



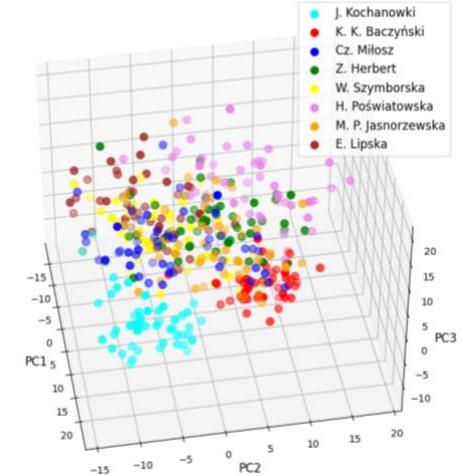
Rys. 7. Dwuwymiarowa analiza skupień przy użyciu PCA dla wszystkich klas. Jedynie wiersze J. Kochanowskiego stanowią dobrze określony klaster, natomiast pozostały twórcy nie tworzą wyraźnych klastrów.

Redukcja wymiarowości do 3-wymiarów, pozwala na bardziej wyraźną analizę skupień przy pomocy PCA. Analiza PCA dla wszystkich klas w trzech wymiarach przestrzeni pozwala na obserwację większej ilości wyraźnych klastrów np. dla Haliny Poświatowskiej czy Krzysztofa Kamila Baczyńskiego. Wciąż jednak większość punktów jest skumulowanych w jednym obszarze wykresu. Oznacza to, że redukcja wymiarowości wektorów zanurzeń analizowanych wierszy przy pomocy metody liniowej PCA nie potrafi poprawnie wyszczególnić wyraźnych klastrów autorów, a zatem relacje między wierszami są z natury nieliniowe i należy sięgnąć po bardziej zaawansowane metody redukcji wymiarowości. Graficzna prezentacja trójwymiarowej analizy skupień jest przedstawiona na Rys. 8.

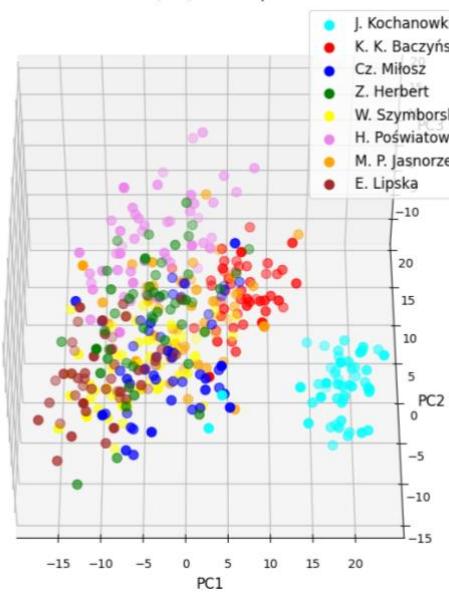
Poems data 3D PCA projection | Author: J. Kochanowski



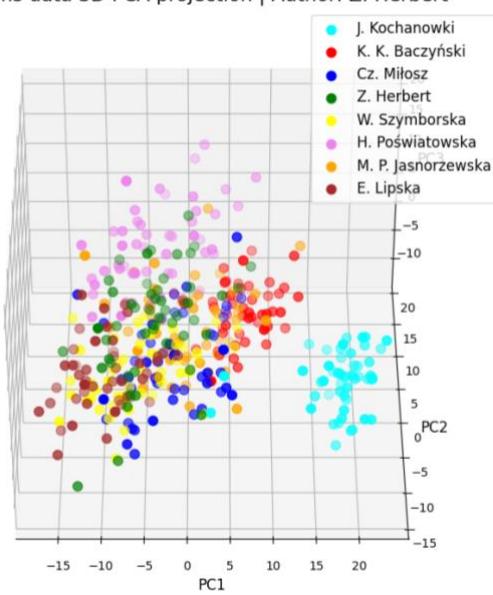
Poems data 3D PCA projection | Author: K. K. Baczyński



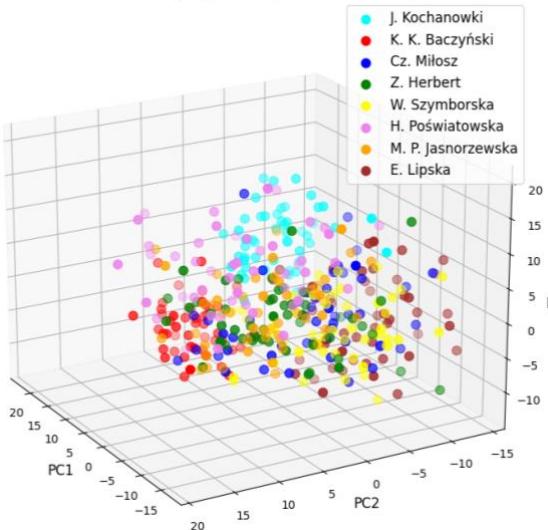
Poems data 3D PCA projection | Author: Cz. Miłosz



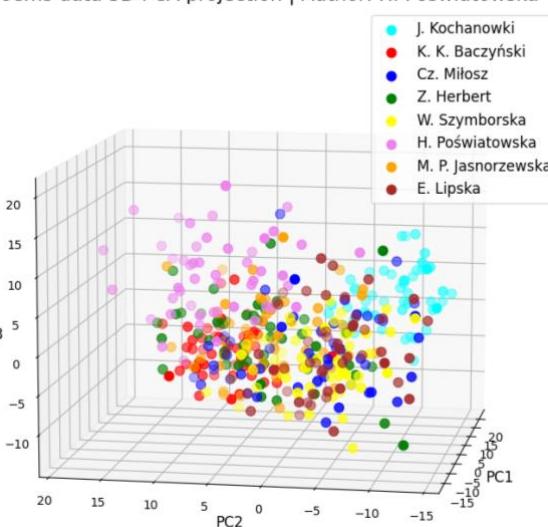
Poems data 3D PCA projection | Author: Z. Herbert



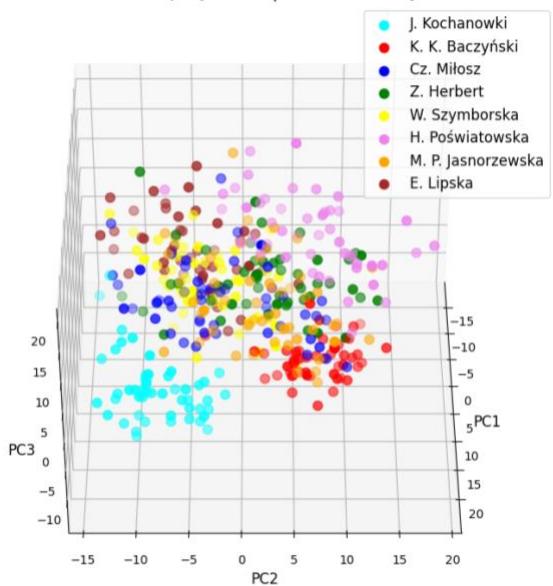
Poems data 3D PCA projection | Author: W. Szymborska



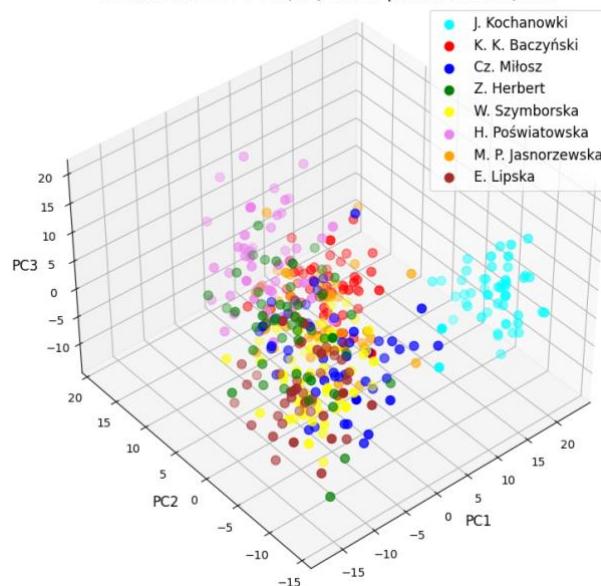
Poems data 3D PCA projection | Author: H. Poświatowska



Poems data 3D PCA projection | Author: M. P. Jasnorzewska



Poems data 3D PCA projection | Author: E. Lipska



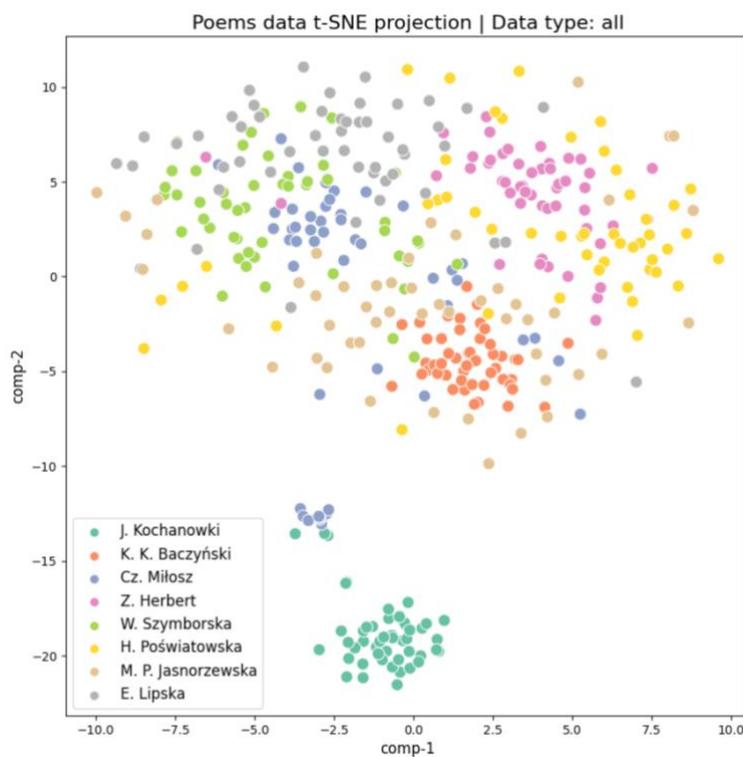
Rys. 8. Trójwymiarowa analiza skupień przy użyciu PCA dla wszystkich klas, prezentowana z różnych kątów w celu głębszej analizy wyników. Każdy widok jest dedykowany przedstawieniu danego autora. Wiersze J. Kochanowskiego stanowią dobrze określony klaster. Widoczne są także nieco słabiej zdefiniowane klastry utworów K. K. Baczyńskiego oraz H. Poświatowskiej.

4.2.2. t-SNE

Redukcja wymiarowości wektorów zanurzeń wierszy przy pomocy t-SNE pozwala na wyodrębnienie klastrów danych, które nie są liniowo separowalne. Algorytm t-SNE działa na zasadzie przekształcania odległości między punktami w przestrzeni oryginalnej na odległości w przestrzeni docelowej, która jest przestrzenią o niższej wymiarowości od oryginalnej. Metoda ta opiera się na podobieństwie lokalnym, co oznacza, że punkty, które są blisko siebie w oryginalnym zbiorze danych, będą znajdować się blisko siebie po redukcji wymiarowości. Ważnym elementem t-SNE jest zachowanie w sąsiedztwie rozkładu prawdopodobieństwa.

Jednocześnie, odległości między wyszczególnionymi klastrami nie odzwierciedlają w żadnym wypadku odległości pomiędzy grupami danych w oryginalnej przestrzeni. Algorytm t-SNE jest algorytmem nieliniowym pozwalającym uchwycić bardziej złożone relacje między danymi, niedostępnyimi dla metody PCA.

Rys. 9 przedstawia dwuwymiarową reprezentację danych uzyskaną przy pomocy t-SNE (perplexity=35). Widać wyraźne klastry danych dla większości klas. Utwory Czesława Miłosza tworzą kilka skupisk punktów. Uwagę zwraca kilka punktów usytuowanych blisko klastra Jana Kochanowskiego, które pojawiają się również w kolejnych analizach. Najbardziej rozrzucone punkty dotyczą twórczości Haliny Poświatowskiej oraz Marii Powlikowskiej-Jansnorzewskiej.



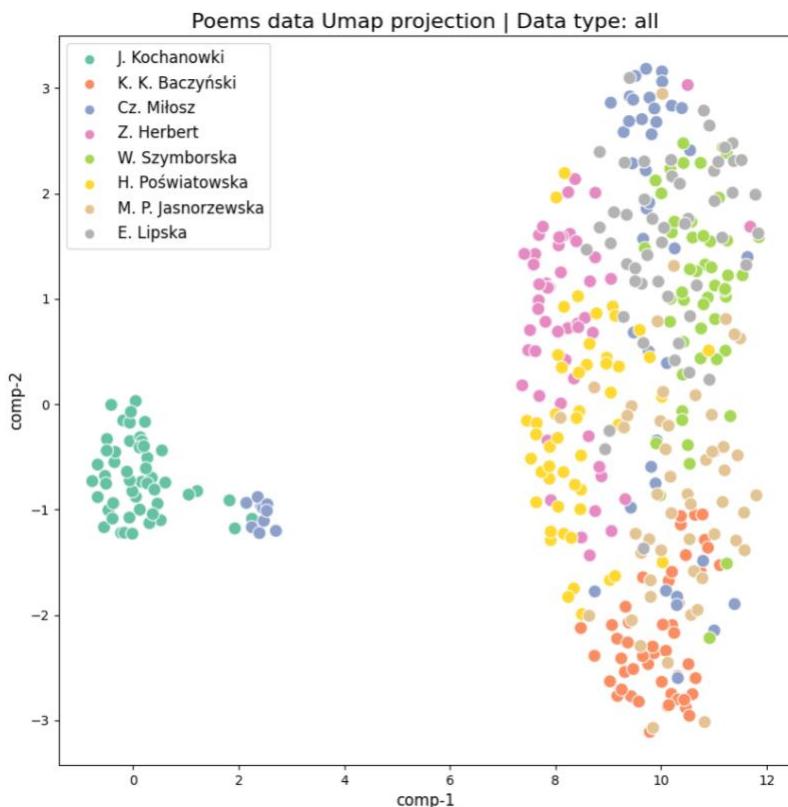
Rys. 9. Analiza skupień przy pomocy redukcji wymiarowości t-SNE dla wszystkich klas. Większość klas tworzy zwarte klastry.

4.2.3. UMAP

UMAP to inna popularna metoda redukcji wymiarowości, podobna do t-SNE, która jest wykorzystywana do wizualizacji i analizy wysokowymiarowych danych. UMAP próbuje zachować podobieństwo odległości między punktami w wysokowymiarowej przestrzeni, przekształcając je na odległość w przestrzeni docelowej o niższym wymiarze. UMAP opiera się na dwóch kluczowych krokach: konstrukcji grafu sąsiedztwa i optymalizacji mapowania. W pierwszym kroku UMAP oblicza lokalne sąsiedztwo dla każdego punktu danych, biorąc pod uwagę ich odległości w oryginalnej przestrzeni. Następnie, na podstawie tych sąsiedztw,

budowany jest graf, w którym punkty są połączone krawędziami o wagach, które odzwierciedlają podobieństwo między nimi. W drugim kroku UMAP optymalizuje mapowanie przestrzeni wysokowymiarowej na przestrzeń docelową o niższym wymiarze. Optymalizacja ta polega na minimalizowaniu różnicy między odległościami na grafie w oryginalnej przestrzeni a grafie w przestrzeni docelowej.

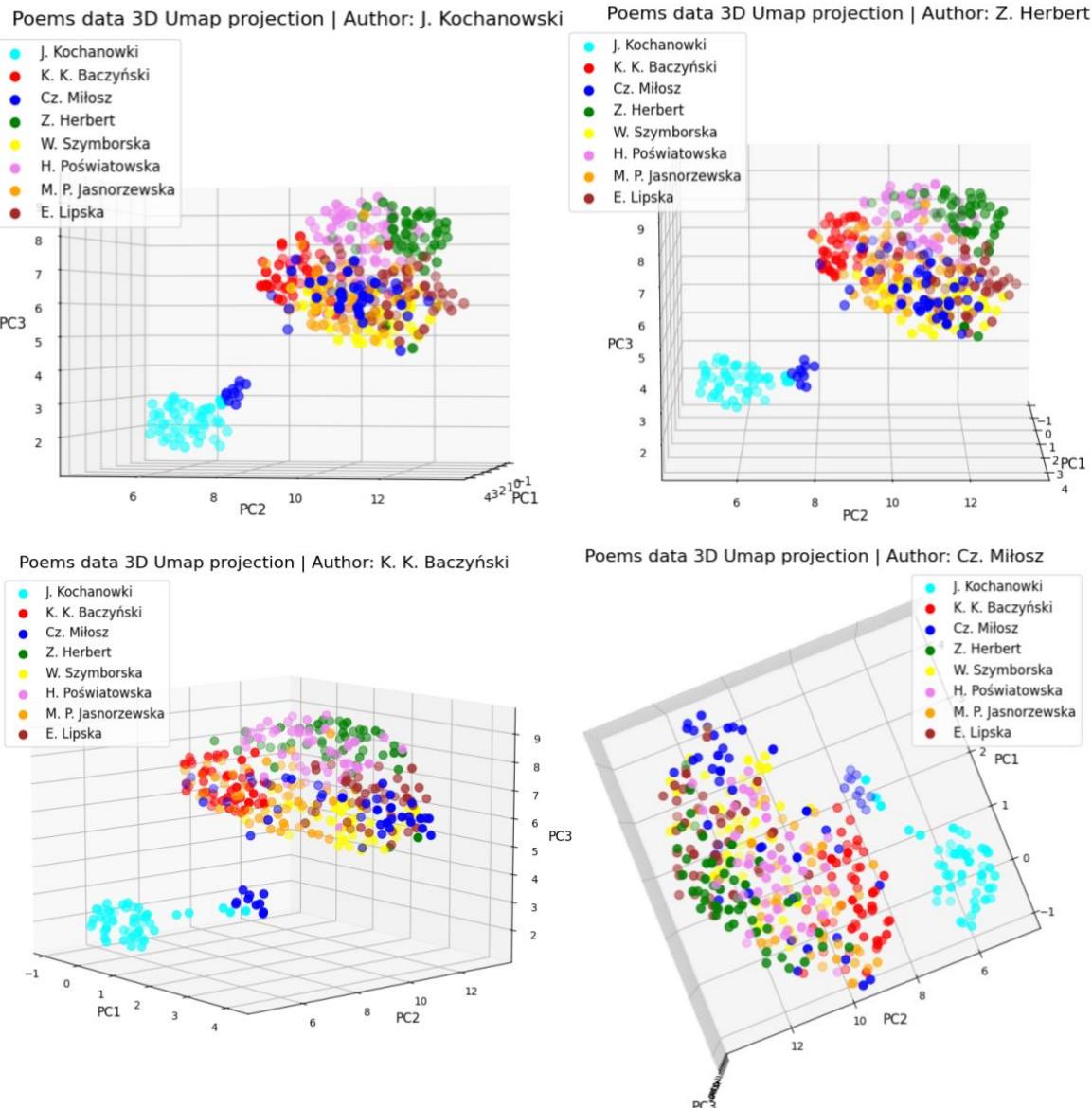
Rys. 10 przedstawia reprezentację dwuwymiarową przy użyciu algorytmu UMAP (`n_neighbors = [30]`, `min_distnecs= [0.3]`, `n_components=2`, `metric='euclidean'`). UMAP umożliwia obserwację dobrze zdefiniowanych klastrów dla większości klas.



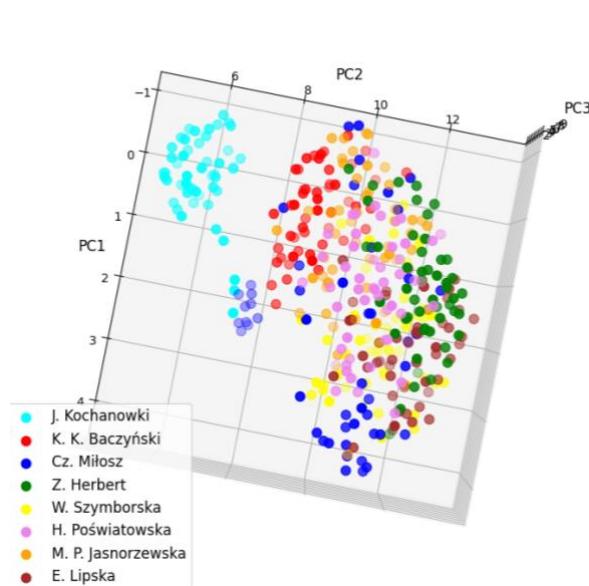
Rys. 10. Dwuwymiarowa analiza skupień przy pomocy redukcji wymiarowości UMAP dla wszystkich klas. Większość klas tworzy zwarte klastry. Wyjątek stanowią wiersze Cz. Miłosza, M. Pawlikowskiej-Jasnorzewska oraz Ewy Lipskiej.

Na Rys. 10 widocznie oddziela się klasa Jana Kochanowskiego oraz kilka próbek z klasy Czesława Miłosza. W poprzednich akapitach zwróciłem uwagę na część twórczość Miłosza, która na wykresach znajduje się blisko klastra Kochanowskiego. Miłosz kierował się na zasadami klasycznej poetyki, z tego powodu właśnie kilka jego utworów mogło znaleźć się blisko Kochanowskiego – poety epoki klasyczmu. Pomiędzy skupiskami punktów dobrze zdefiniowanych klas znalazły się dwie, które są rozłożone w chaotyczny sposób. Pisarze mniej wyraźni to Czesław Miłosz i Maria Pawlikowska-Jasnorzewska. Rys. 11 przedstawia

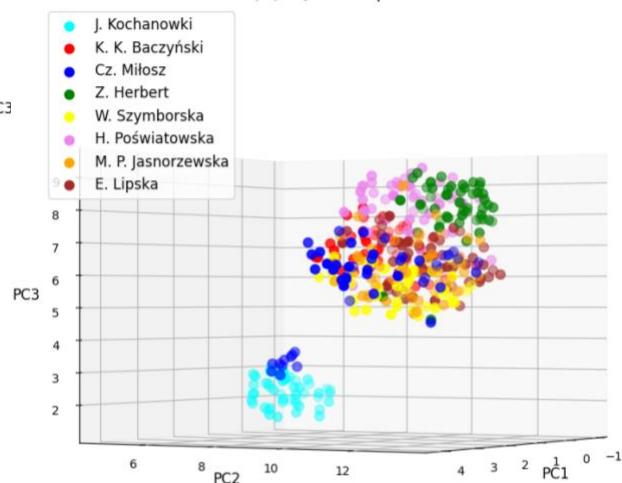
reprezentację w przestrzeni trójwymiarowej pozwalającą na lepszą obserwację poszczególnych klas.



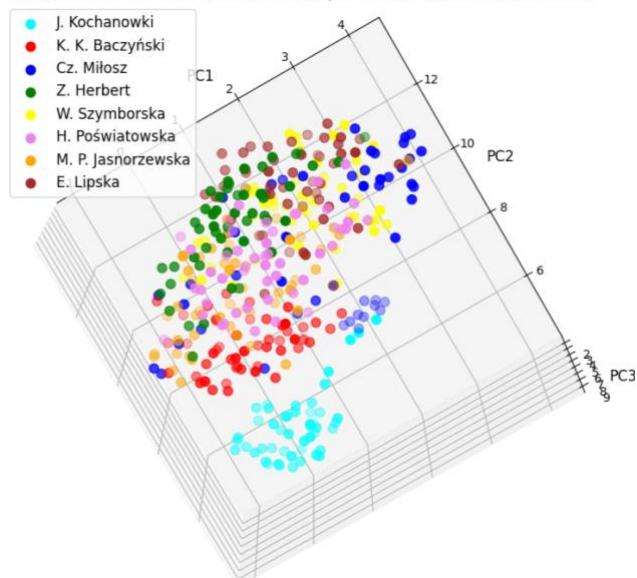
Poems data 3D Umap projection | Author: W. Szymborska



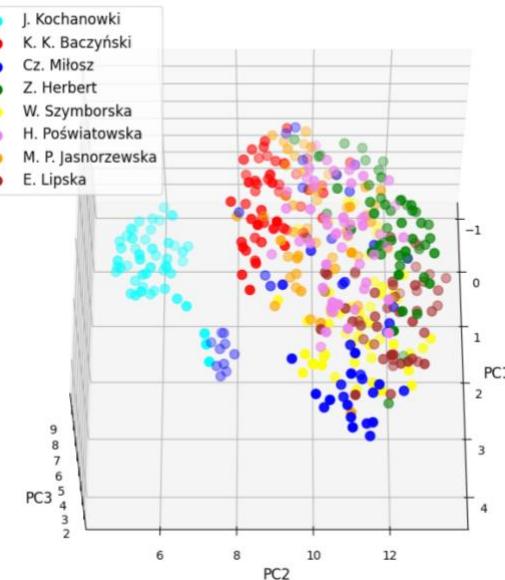
Poems data 3D Umap projection | Author: H. Poświatowska



Poems data 3D Umap projection | Author: M. P. Jasnorzewska



Poems data 3D Umap projection | Author: E. Lipska

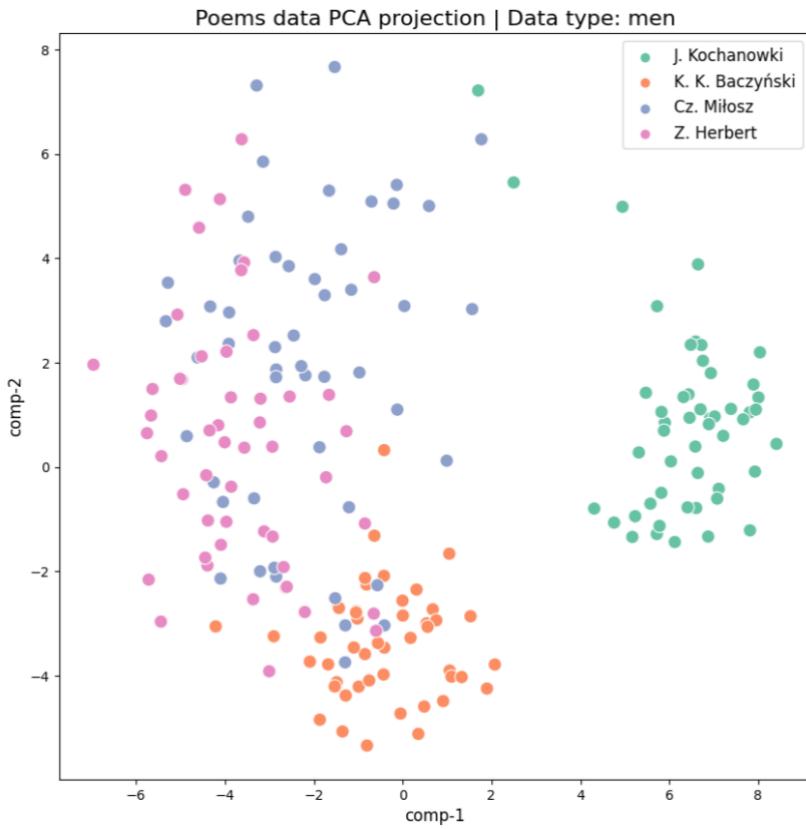


Rys. 11. Trójwymiarowa analiza skupień przy użyciu UMAP dla wszystkich klas, prezentowana z różnych kątów w celu głębszej analizy wyników. Każdy widok jest dedykowany przedstawieniu danego autora. Widoczna jest oddzielona klasa J. Kochanowskiego. Większość klas tworzy dobrze zdefiniowane klastry. Pisarze mniej wyraźni to Czesław Miłosz i Maria Pawlikowska-Jasnorzewska.

4.3. Redukcja wymiarowości i analiza skupień reprezentacji wektorowych wierszy poetów

Podział danych umożliwia bardziej dokładną analizę poszczególnych klas. Dane dla tekstów pisanych przez mężczyzn prezentowane przy pomocy PCA w przestrzeni dwuwymiarowej dzielą się na dwa wyraźne skupiska punktów należących do Jana Kochanowskiego oraz Krzysztofa Kamila Baczyńskiego. Natomiast dwie pozostałe klasy są

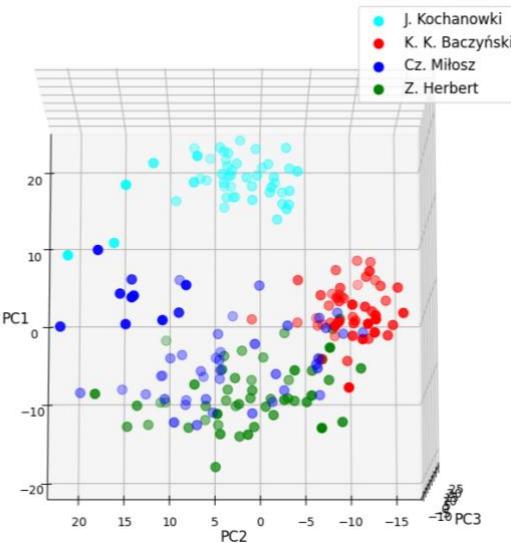
mniej zgrupowane. Rys. 12 prezentuje rozłożenie wektorów poezji mężczyzn w przestrzeni dwuwymiarowej.



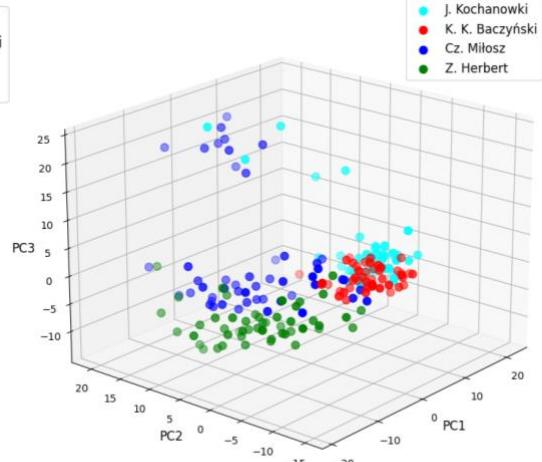
Rys. 12. Dwuwymiarowa analiza skupień przy pomocy redukcji wymiarowości PCA dla poezji mężczyzn. Można zaobserwować dwa wyraźne klastry utworzone przez wiersze J. Kochanowskiego oraz K. K. Baczyńskiego.

Na Rys. 13 prezentuję twórczość mężczyzn w przestrzeni trójwymiarowej przy użyciu PCA. Z różnej perspektywy sąauważalne trzy dobrze oddzielone klasy: Jan Kochanowski, Krzysztof Kamil Baczyński oraz Zbigniew Herbert. Dane dotyczące utworów Czesława Miłosza rozkładają się w dużej odległości od siebie. Jak w poprzednich przypadkach grupa wierszy Miłosza znajduje się w pobliżu Kochanowskiego.

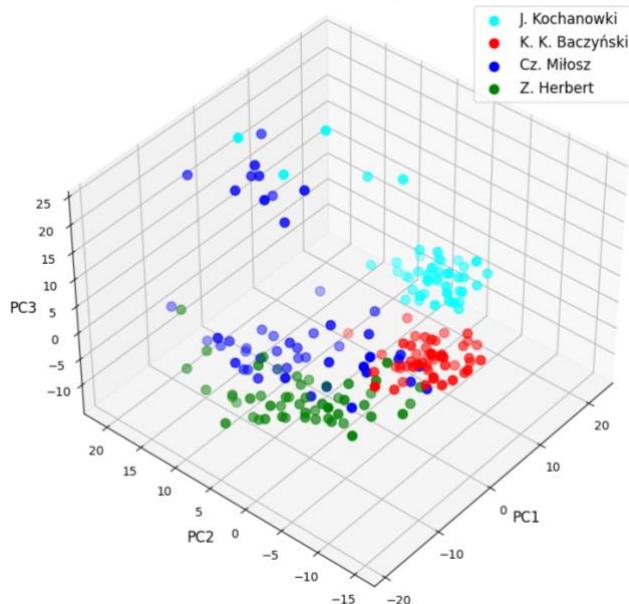
Poems data 3D PCA projection | Author: J. Kochanowski



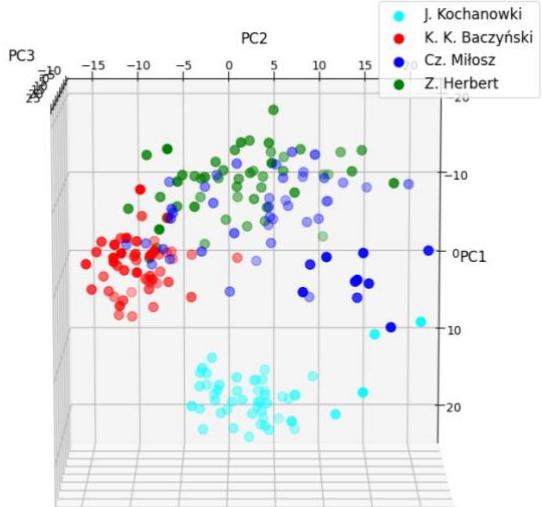
Poems data 3D PCA projection | Author: Cz. Miłosz



Poems data 3D PCA projection | Author: Z. Herbert

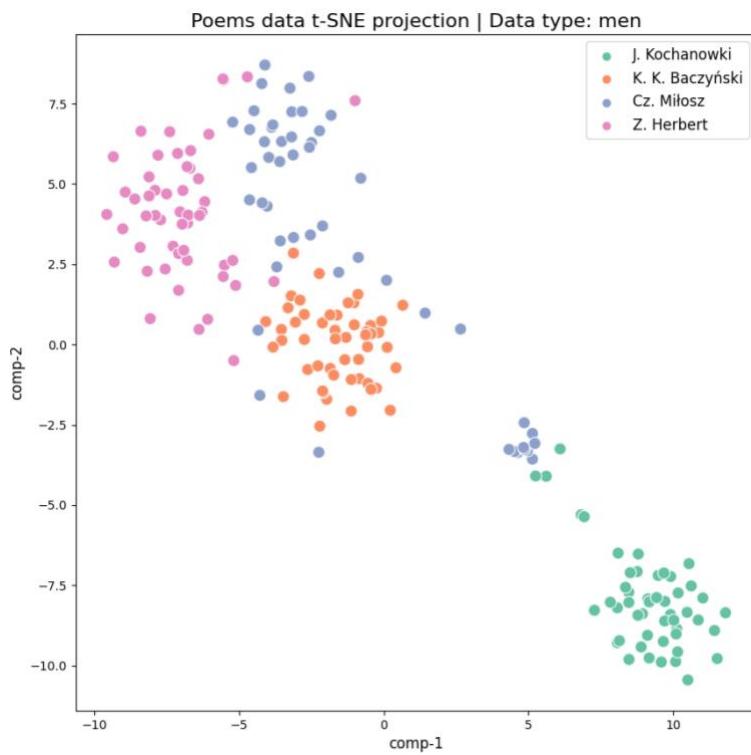


Poems data 3D PCA projection | Author: K. K. Baczyński



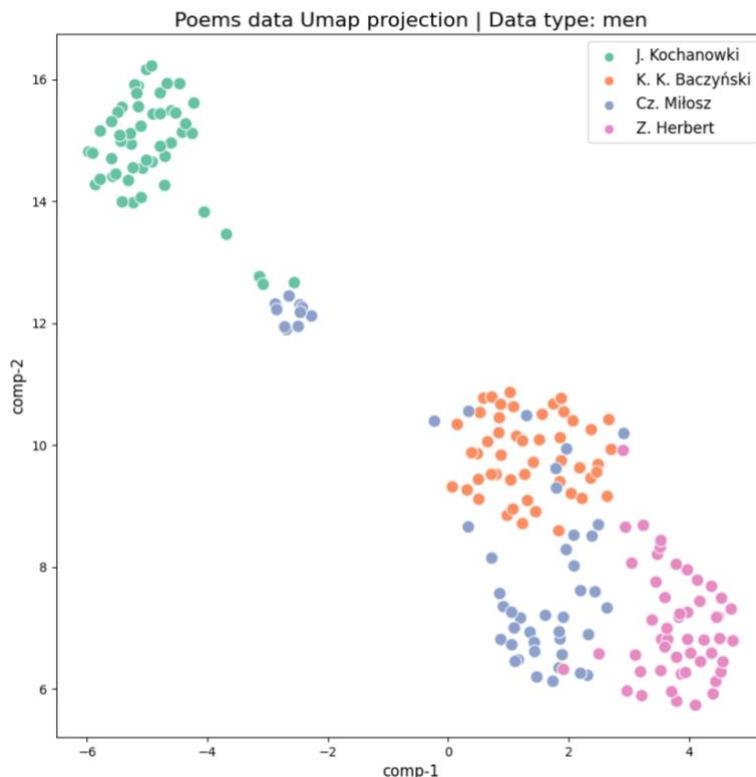
Rys. 13. Trójwymiarowa analiza skupień przy użyciu PCA dla poezji mężczyzn, prezentowana z różnych kątów w celu głębszej analizy wyników. Każdy widok jest dedykowany przedstawieniu danego autora. Zauważalne są trzy dobrze oddzielone klastry: J. Kochanowski, K. K. Baczyński oraz Z. Herbert.

Dane zaprezentowane na Rys. 14 są wynikiem analizy za pomocą techniki t-SNE. W bardzo dobrym stopniu pokazują się klastry dla wszystkich klas. Teksty Czesława Miłosza tym razem bardziej skupiły się w jednej grupie oprócz kilku rozproszonych punktów oraz grupy w pobliżu Jana Kochanowskiego.



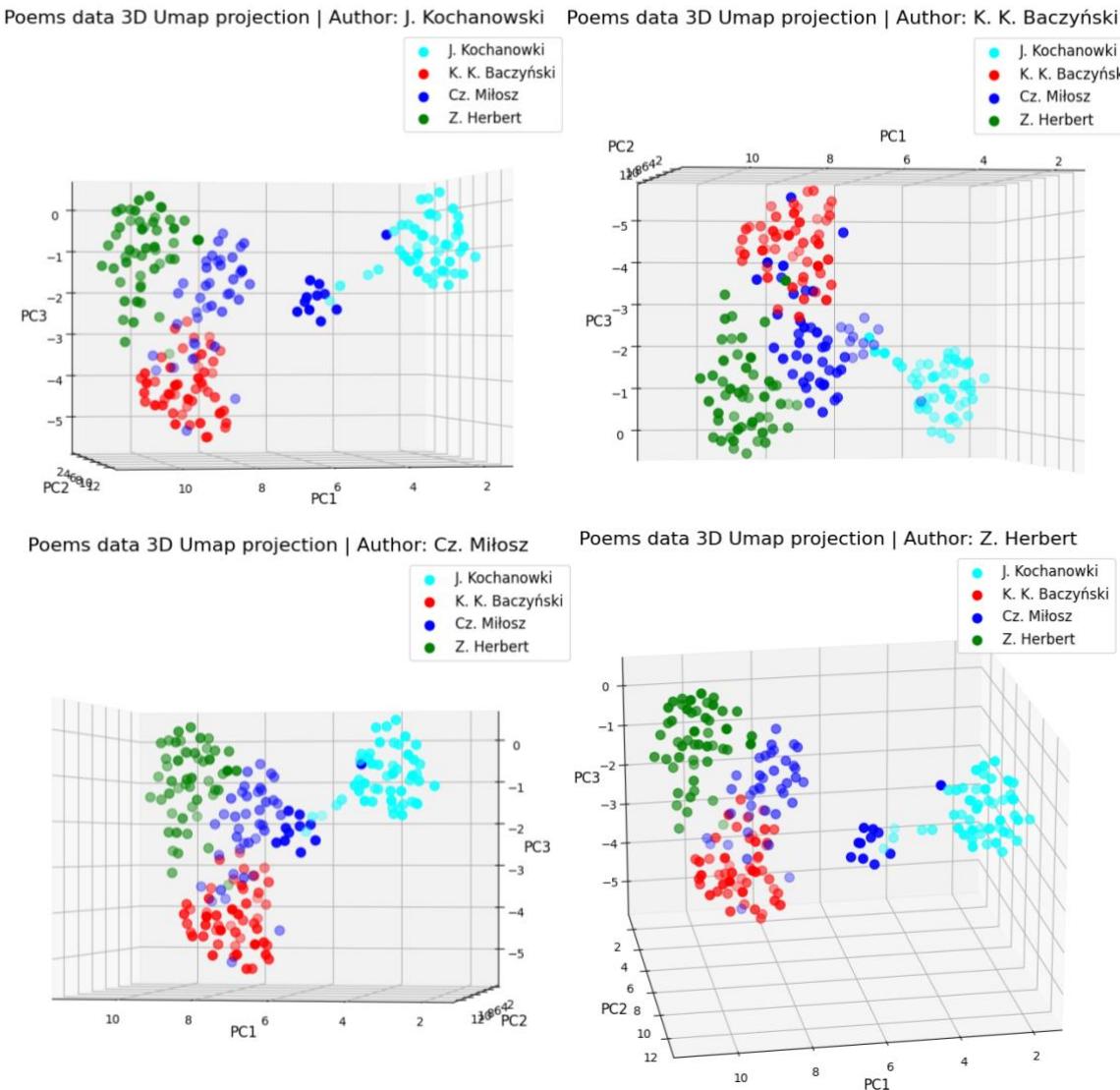
Rys. 14. Dwuwymiarowa analiza skupień przy pomocy redukcji wymiarowości t-SNE dla poezji mężczyzn. Można zaobserwować wyraźne klastry dla wszystkich klas.

Rys. 15. prezentuje teksty przy pomocy UMAP. Dane układają się podobnie do t-SNE.



Rys. 15. Dwuwymiarowa analiza skupień przy pomocy redukcji wymiarowości UMAP dla poezji mężczyzn. Widoczne są wyraźne klastry dla wszystkich klas.

Wykorzystanie UMAP w przestrzeni 3D pozwala jeszcze lepiej zaobserwować ułożenie danych. Na Rys. 16 w różnych rzutach można zobaczyć dobrze odseparowaną klasę Czesława Miłosza z wyłączeniem kilku punktów oddalonych od centrum klastra i małą grupą zbliżoną do przestrzeni, gdzie sytuują się wiersze Jana Kochanowskiego.



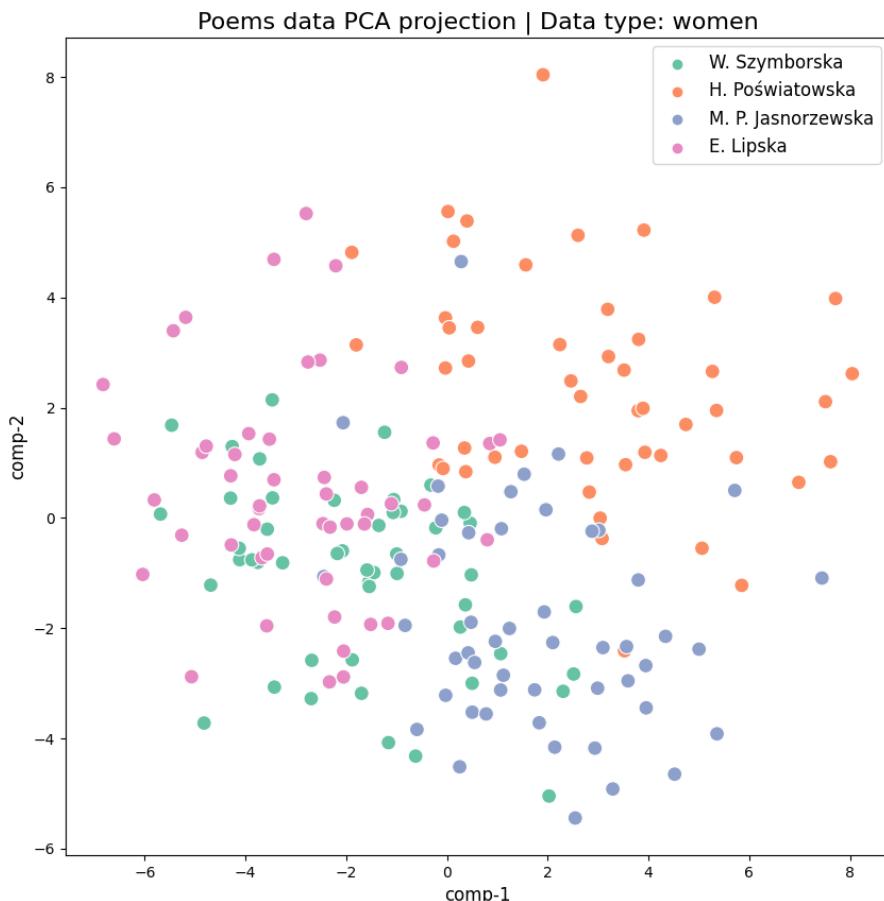
Rys. 16. Trójwymiarowa analiza skupień przy użyciu UMAP dla poezji mężczyzn, prezentowana z różnych kątów w celu głębszej analizy wyników. Każdy widok jest dedykowany przedstawieniu danego autora. Zauważalne są trzy dobrze oddzielone klastry: J. Kochanowski, K. K. Baczyński, Z. Herbert oraz większość wierszy Cz. Miłosza.

4.4. Redukcja wymiarowości i analiza skupień reprezentacji wektorowych wierszy poetek

Poezja kobiet jest trudniejsza do zaprezentowania na wykresach dwuwymiarowych.

Rys. 17 przedstawia dane dla poezji kobiet przy pomocy PCA. Wszystkie klasy skupiają się

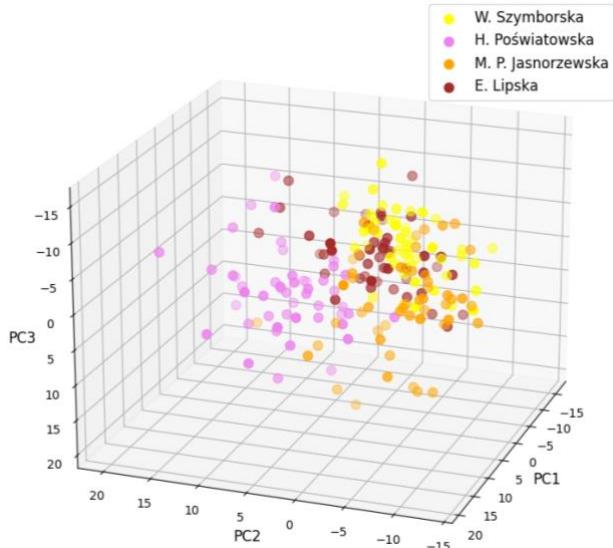
blisko siebie, przy czym punkty należące do poszczególnych klas są rozproszone w dużych odległościach. Można zauważyć, że zajęty obszar dzieli się na trzy części zajmowane kolejno przez Halinę Poświatowską, Marię Pawlikowską-Jasnorzewską oraz Ewę Lipską. Pomiędzy danymi Lipskiej i Pawlikowskiej-Jasnorzewskiej rozłożone są punkty należące do Wisławy Szymborskiej.



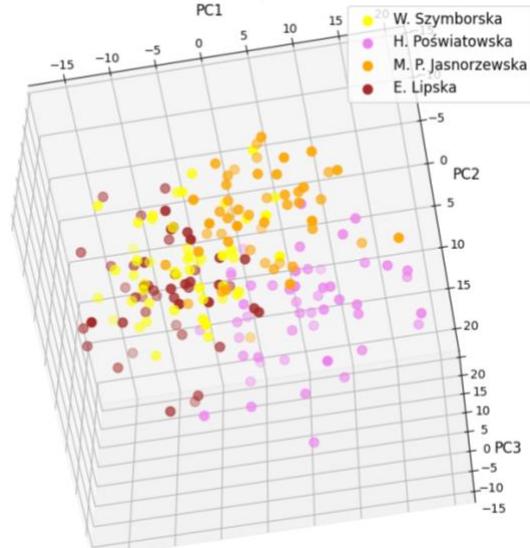
Rys. 17. Dwuwymiarowa analiza skupień przy pomocy redukcji wymiarowości PCA dla poezji kobiet. Dane są rozmieszczone w sposób niezgrupowany. Można zaobserwować podział przestrzeni na trzy części zajmowane przez H. Poświatowską, M. Pawlikowską-Jasnorzewską oraz E. Lipską.

Prezentacja danych przy pomocy PCA w przestrzeni trójwymiarowej na Rys. 18 pozwala nieznaczenie poprawić możliwości obserwacji danych. Można zobaczyć dobrze wyznaczone klastry dla Haliny Poświatowskiej oraz Marii Pawlikowskiej-Jasnorzewskiej. Widać jednak nakładające się na siebie punkty należące do Ewy Lipskiej i Wisławy Szymborskiej. Zbieżność tych dwóch autorek można opierać na podobnym, prostym języku oraz refleksjach na temat codzienności, o czym wspominałam w rozdziale 3.9.

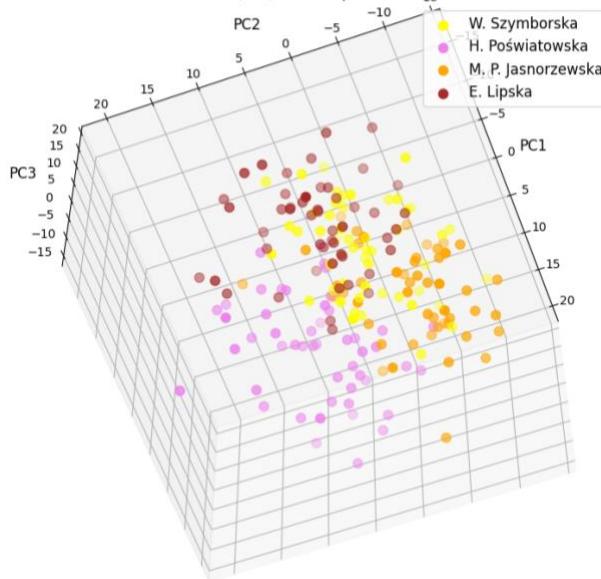
Poems data 3D PCA projection | Author: W. Szymborska



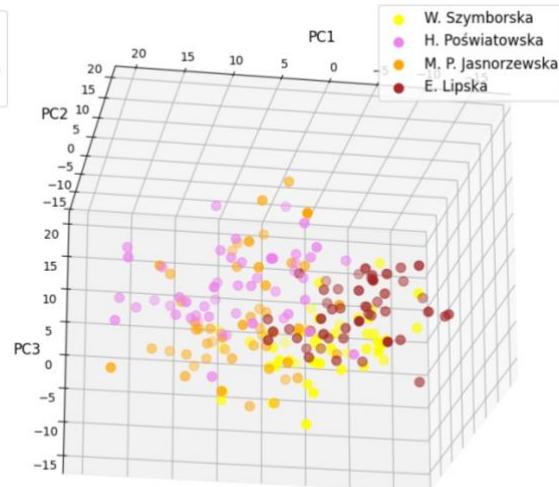
Poems data 3D PCA projection | Author: M. P. Jasnorzewska



Poems data 3D PCA projection | Author: H. Poświatowska

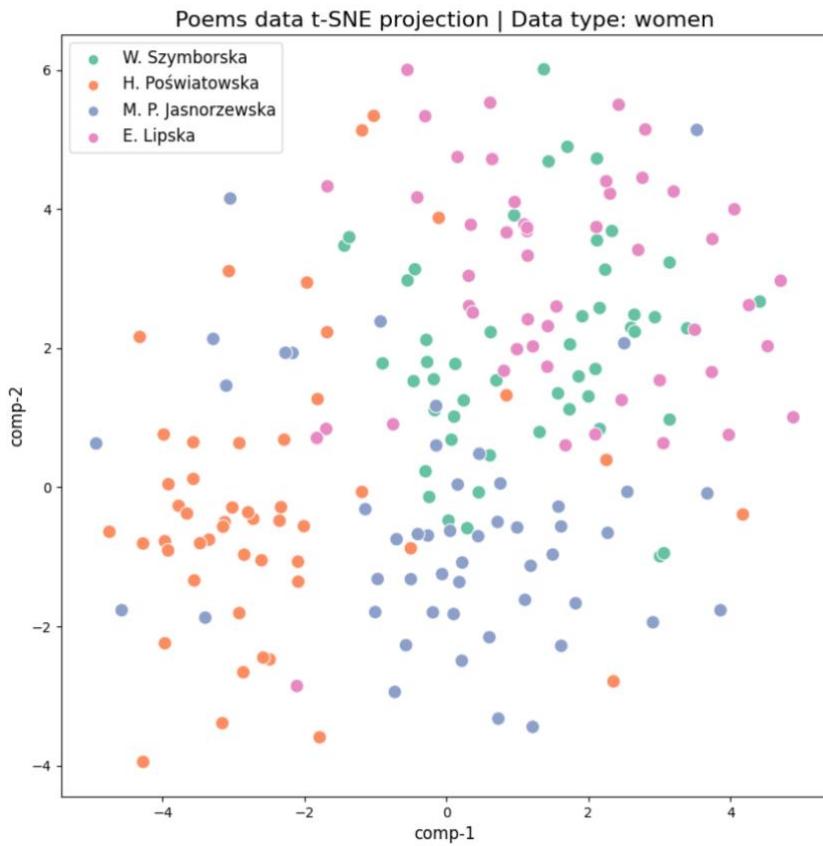


Poems data 3D PCA projection | Author: E. Lipska



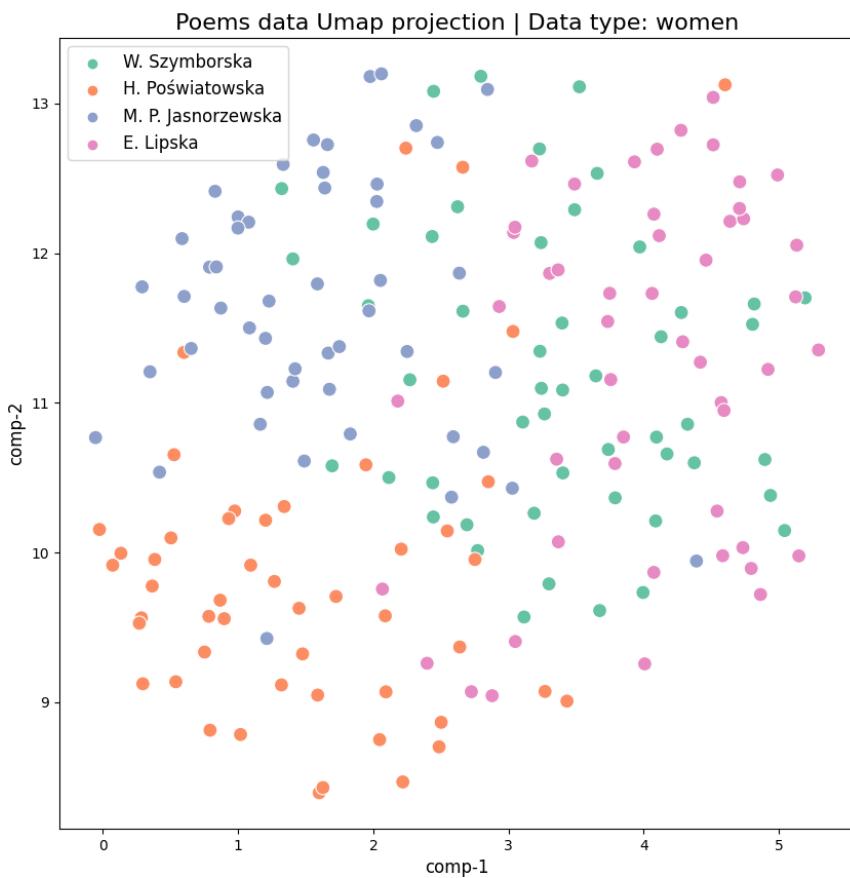
Rys. 18. Trójwymiarowa analiza skupień przy użyciu PCA dla poezji kobiet, prezentowana z różnych kątów w celu głębszej analizy wyników. Każdy widok jest dedykowany przedstawieniu danej autorki. Zauważalne są klastry H. Poświatowskiej oraz M. Pawlikowskiej-Jasnorzewskiej. Jednak punkty należące do E. Lipskiej i W. Szymborskiej nakładają się na siebie

Wykres t-SNE dla poezji kobiet przedstawiony na Rys. 19 pokazuje bardziej niż w przypadku PCA rozrzucone punkty. Można zaobserwować dwa większe skupiska danych dla Haliny Poświatowskiej oraz dla Marii Pawlikowskiej-Jasnorzewskiej. Jednak większa część punktów wewnętrz klas jest rozmieszczona w dużych odległościach od siebie i miesza się pomiędzy klasami na obszarze całego wykresu.



Rys. 19. Dwuwymiarowa analiza skupień przy pomocy redukcji wymiarowości t-SNE dla poezji kobiet. Dane w większości są rozmieszczone w sposób niezgrupowany.

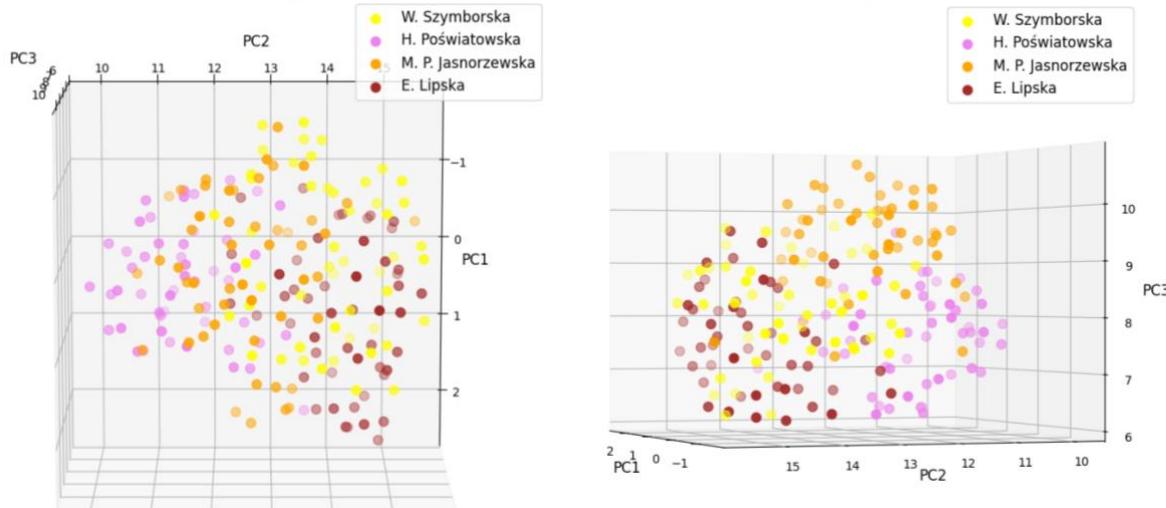
Wykres UMAP dla poezji kobiet, Rys. 20, pozwala dostrzec w odrobinę lepszym stopniu wyznaczone grupy danych dla Haliny Poświatowskiej oraz Marii Pawlikowskiej-Jansorzewskiej. Jednak wciąż zauważalne są pojedyncze punkty oddalone od centralnego obszaru poszczególnych klas. Dane dla Wisławy Szymborskiej i Ewy Lipskiej mieszają się ze sobą, podobnie jak miało to miejsce w analizie przeprowadzonej przy pomocy PCA, nie tworząc większych skupisk charakterystycznych dla pojedynczych klas. Bliskość wektorów punktów Szymborskiej i Lipskiej pojawia się na większości przedstawianych wykresów. Pomimo, że techniki t-SNE i UMAP zasadniczo nie wskazują zależności pomiędzy klastrami bazując na odległościach między nimi, ta powtarzająca się zależność jest warta podkreślenia.



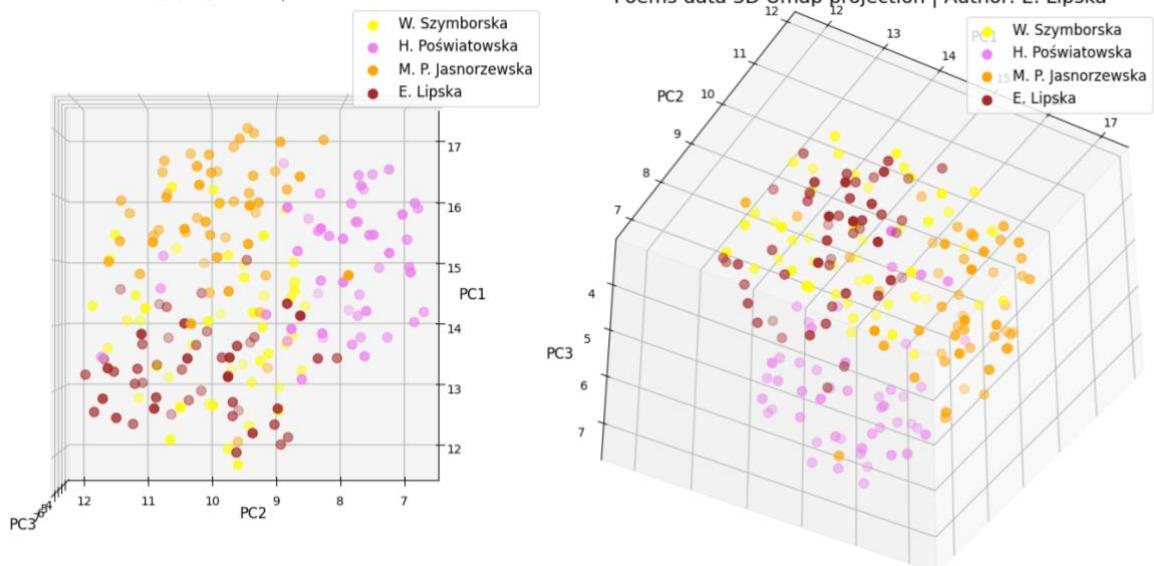
Rys. 20. Dwuwymiarowa analiza skupień przy pomocy redukcji wymiarowości UMAP dla poezji kobiet. Widoczne są dwa klastry utworzone przez wiersze M. Pawlikowskiej-Jasnorzewskiej oraz H. Poświatowskiej.

Prezentacja danych w przestrzeni trójwymiarowej przy pomocy UMAP, przedstawiona na Rys. 21 pokazuje w bardzo dobrym stopniu odseparowane dane dla Haliny Poświatowskiej oraz Marii Pawlikowskiej-Jasnorzewskiej. Punkty należące do tych klas w wyraźny sposób oddzielają się od pozostałych wektorów. W przypadku Wisławy Szymborskiej i Ewy Lipskiej prezentacja poprawiła się w nieznacznym stopniu. Jednak wciąż dane się na siebie nakładają i mieszają ze sobą.

Poems data 3D Umap projection | Author: W. Szymborska Poems data 3D Umap projection | Author: M. P. Jasnorzewska



Poems data 3D Umap projection | Author: H. Poświatowska



Rys. 21. Trójwymiarowa analiza skupień przy użyciu UMAP dla poezji kobiet, prezentowana z różnych kątów w celu głębszej analizy wyników. Każdy widok jest dedykowany przedstawieniu danej autorki. Widoczne są oddzielone klastry H. Poświatowskiej oraz M. Pawlikowskiej-Jasnorzewskiej. Utwory E. Lipskiej i W. Szymborskiej grupują się w nieznacznym stopniu.

5. Uczenie nadzorowane i klasyfikacja wierszy

Nadzorowana klasyfikacja danych, czyli przypisanie klasy (nazwiska autora) do danego wiersza została przeprowadzona przy pomocy trzech różnych technik. Pierwsza metoda opiera się na porównywaniu odległości euklidesowych dla każdej próbki. Druga to algorytm klasycznego uczenia maszynowego opartego o modele drzewiaste i uczenie zbiorowe (ang. *ensemble learning*) – XGBoost, a trzecia jest siecią neuronową typu *feedforward* o dwóch warstwach ukrytych.

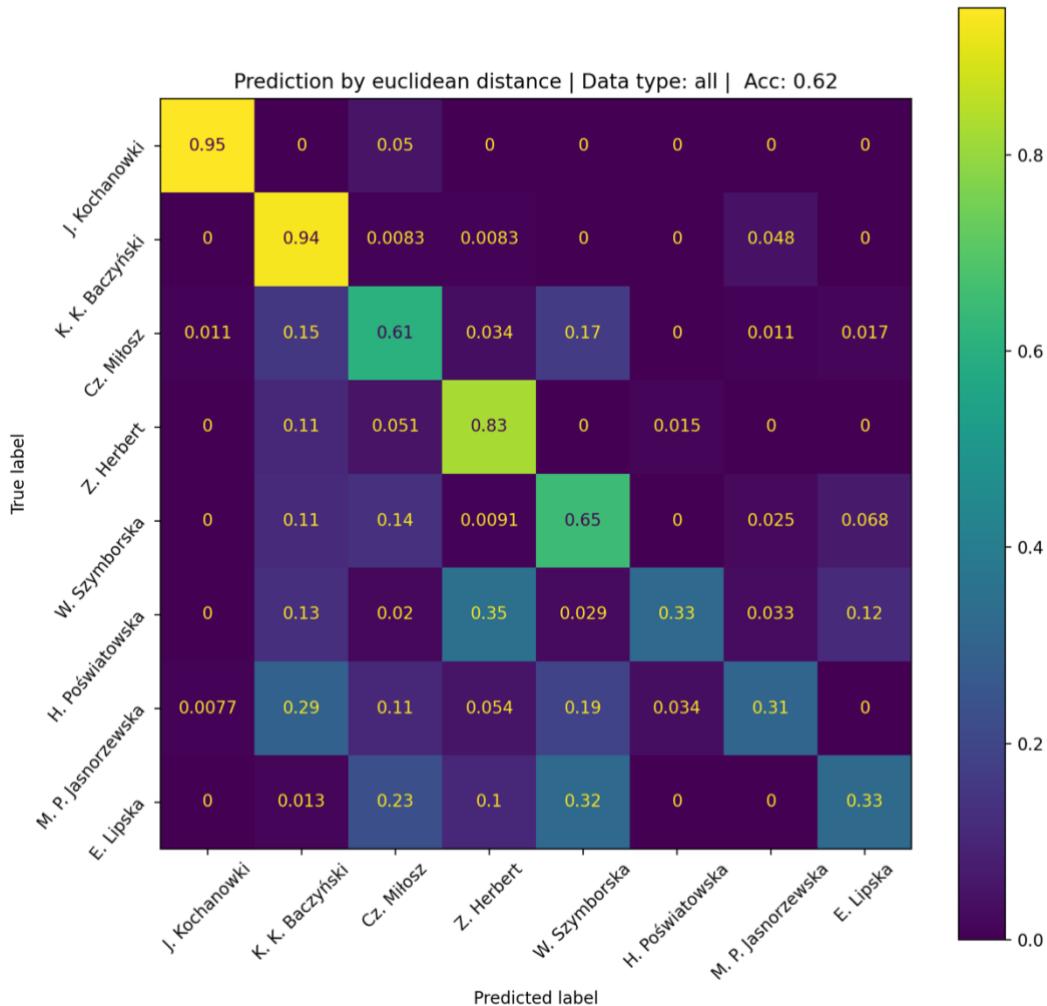
W celu uzyskania jak najbardziej rzetelnych wyników dla każdego eksperymentu z wykorzystaniem wyżej wymienionych metod zostało przeprowadzonych niezależnych 10 procesów trenowania i predykcji, w trakcie których na nowo wyznaczano zbiór treningowy, testowy i walidacyjny. Ponieważ zbiór danych jest zbalansowany, tj. każda klasa ma tyle samo danych treningowych, jako miarę jakości przewidywań używam dokładności (ang. *accuracy*), czyli stosunek dobrze sklasyfikowanych wiersz, do wszystkich wierszy w zbiorze testowym. Przedstawione wyniki w kolejnych podrozdziałach pokazują uśrednioną dokładność predykcji dla 10 realizacji wraz z uśrednioną macierzą konfuzji.

5.1. Klasyfikacja na podstawie odległości euklidesowych

Przeprowadzona analiza skupień w poprzednim rozdziale wskazuje, że reprezentacja wektorowa wierszy tworzy klastry grupowane wg etykiety wiersza, czyli nazwisk autorów. Zatem pierwszy rozważany klasyfikator wierszy, oparty o poczynione obserwacje klasteryzacji danych, opiera się na porównywaniu odległości euklidesowych reprezentacji wektorowych. Dla zadanego wektora zbioru testowego \mathbf{x}_{test} przypisuję klasę \mathbf{y}_{test} otrzymaną w następujący sposób: dla wektora \mathbf{x}_{test} liczona jest odległość euklidesowa pomiędzy każdym wektorem ze zbioru treningowego $\mathbf{x}_{\text{train}}$ z przypisaną klasą $\mathbf{y}_{\text{train}}$. Przypisana klasa \mathbf{y}_{test} jest klasą wektora odpowiadającą wektorowi $\mathbf{x}_{\text{train}}$, dla którego odległość euklidesowa $\|\mathbf{x}_{\text{train}} - \mathbf{x}_{\text{test}}\|$ jest najmniejsza.

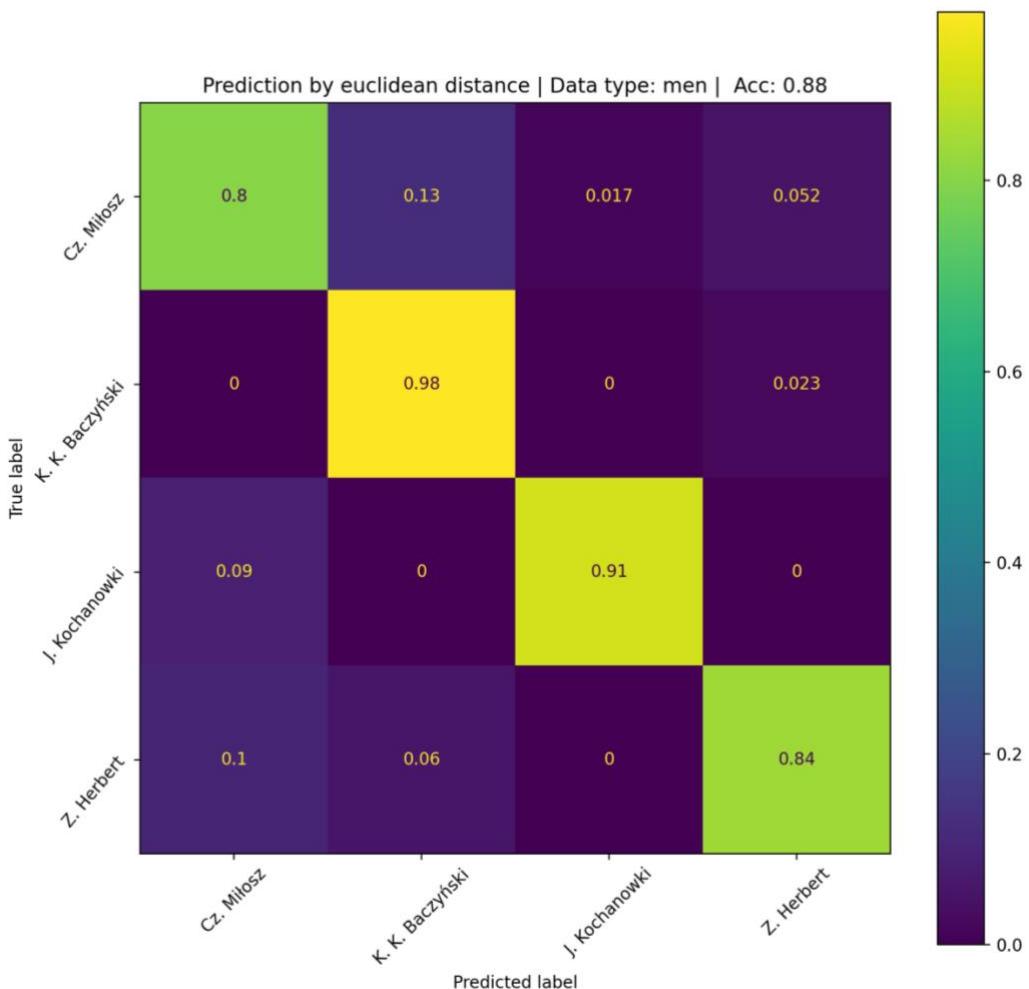
Ten niezwykle prosty algorytm klasyfikacji pozwolił osiągnąć wynik dokładności na poziomie 62%, Rys. 22. Najlepiej rozpoznawalnymi klasami były Jan Kochanowski oraz Krzysztof Kamil Baczyński, dokładność ich predykcji wynosi około 95%. Z nieco niższym wynikiem 83% został rozpoznany Zbigniew Herbert. Wśród poetek najlepszy wynik osiągnęły utwory Wisławy Szymborskiej. Widać jednak, że poezja kobiet jest przewidywana na dużo niższym poziomie przy wykorzystaniu tej metody. Najczęściej pomyłki modelu to predykcja: dla utworów Haliny Poświatowskiej klasy Herberta, dla wierszy Marii Pawlikowskiej-

Jasnorzewskiej klasy Baczyńskiego i Szymborskiej, a dla tekstów Ewy Lipskiej klasy Miłosza oraz Szymborskiej.



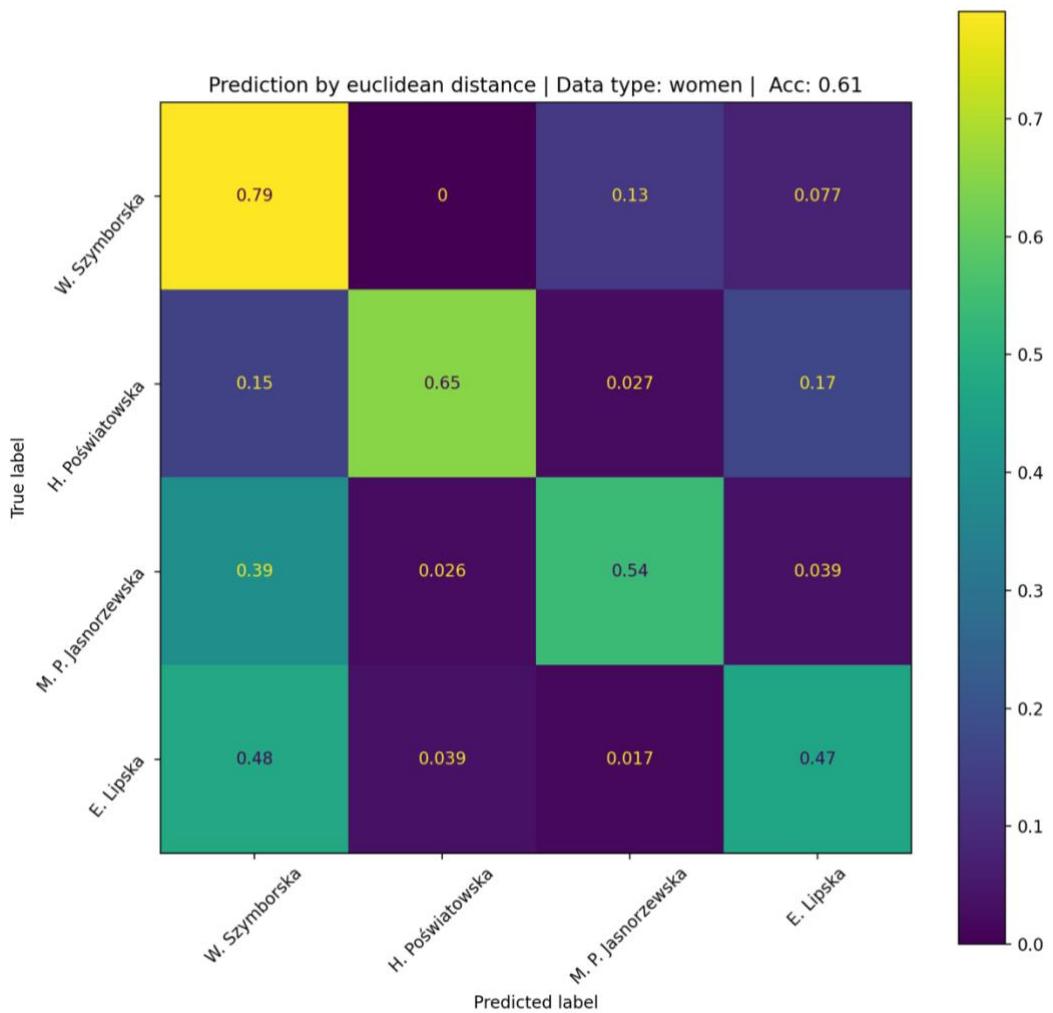
Rys. 22. Macierz konfuzji dla predykcji na podstawie odległości euklidesowych dla wszystkich klas. Dokładność predykcji klasy z wykorzystaniem tej metody osiąga 62%. Twórczość mężczyzn uzyskuje dużo wyższe wyniki niż w przypadku kobiet.

Predykcja na podstawie odległości euklidesowych dla samych poetów osiąga bardzo zadowalający rezultat 88% dokładności, Rys. 23. Najwyższy wynik należy do Krzysztofa Kamila Baczyńskiego, a najniższy do Czesława Miłosza. Na macierzy konfuzji można zaobserwować kilka pomyłek w predykcji. Widoczne są pomyłki pomiędzy Miłoszem z trzema pozostałymi poetami. Dla tekstów Jana Kochanowskiego oraz Baczyńskiego model przewiduje klasę Miłosza.



Rys. 23. Macierz konfuzji dla predykcji na podstawie odległości euklidesowych dla poezji mężczyzn. Dokładność predykcji klasy z wykorzystaniem tej metody osiąga 88%. Teksty K. K. Baczyńskiego i J. Kochanowskiego uzyskały wynik powyżej 90%. Model najsłabiej rozpoznaje utwory Cz. Miłosza.

W przypadku poezji jedynie kobiet model osiągną wynik 61% dokładności, Rys. 24. Najlepiej rozpoznaną twórczynią jest Wisława Szymborska, a najsłabiej Ewa Lipska. Na macierzy konfuzji zauważalne są wielokrotne pomyłki w przewidywaniu klasy Szymborskiej dla tekstów pozostałych poetek. Niemal połowa utworów Lipskiej została sklasyfikowana jako dzieła Noblistki. W rozdziale 3.9, gdzie omawiam charakterystykę twórczości Lipskiej można znaleźć informację, że styl poetki jest bliski temu, jaki wypracowała Szymborska, co potwierdza obserwacja macierzy konfuzji. Podobna sytuacja ma miejsce w przypadku tekstów Marii Pawlikowskiej-Jasnorzewskiej.



Rys. 24. Macierz konfuzji dla predykcji na podstawie odległości euklidesowych dla poezji kobiet. Dokładność predykcji klasy z wykorzystaniem tej metody osiąga 61%. Najlepiej rozpoznawaną przez model poetką jest W. Szymborska. Ciekawym zjawiskiem jest także duża liczba predykcji klasy W. Szymborskiej dla utworów E. Lipskiej.

5.2. Klasyfikacja przy pomocy XGBoost

Kolejnym użytym algorytmem do klasyfikacji jest XGBoost - algorytm uczenia maszynowego, oparty o drzewa decyzyjne i lasy losowe. Kolejne paragrafy po krótkie opisują zastosowaną metodę XGBoost oraz uzyskane wyniki klasyfikacji.

Drzewo decyzyjne (ang. *decision tree*) to struktura hierarchiczna, która reprezentuje sekwencję testów na cechach danych w celu podjęcia decyzji, składająca się z węzłów i krawędzi. Węzły reprezentują testy na cechach danych, a krawędzie reprezentują wynik testu. Drzewo rozpoczyna się od korzenia, który jest najwyższym węzłem, a następnie rozgałęzia się w kolejne węzły na podstawie wyników testów, aż osiągnie się liście, które reprezentują wyniki klasyfikacji lub predykcji. Proces konstruowania drzewa decyzyjnego polega na podziale

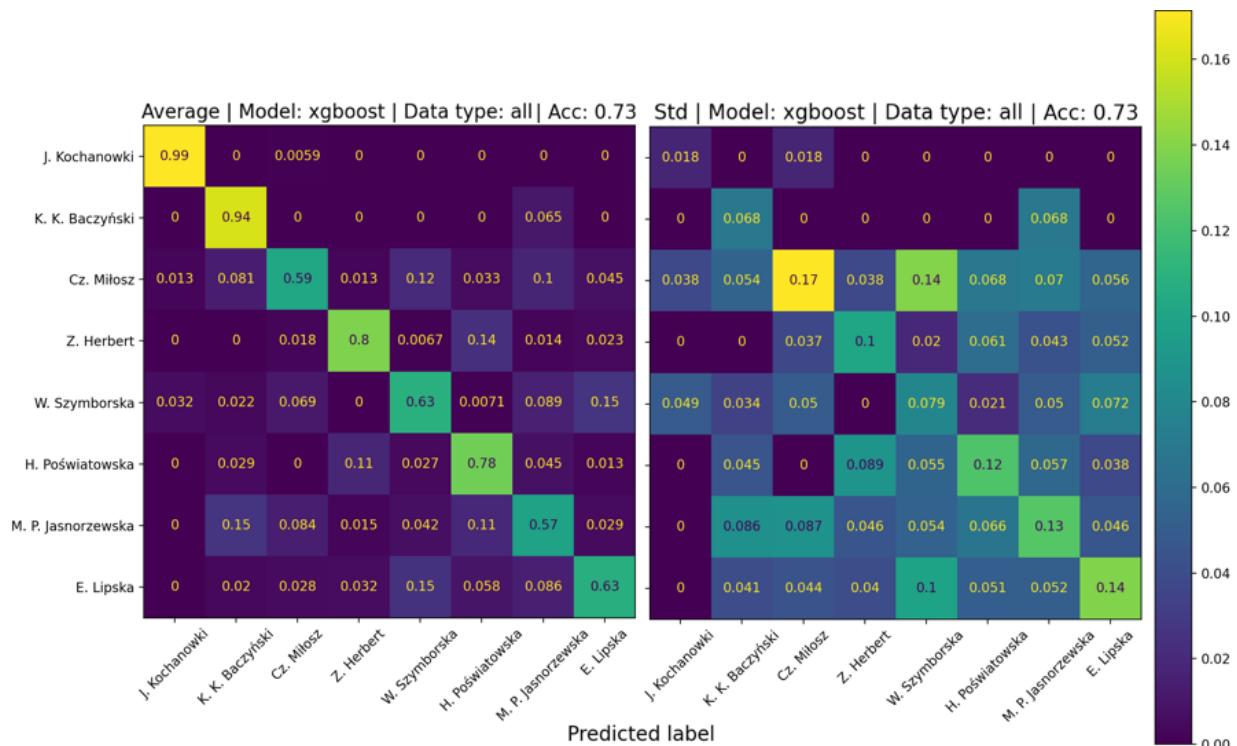
danych na podstawie różnych cech, w taki sposób, aby maksymalizować jednorodność klas lub minimalizować błąd predykcji. Algorytm wybiera najlepszą cechę i próg podziału, który dzieli dane na podzbiory o jak największej jednorodności (w przypadku klasyfikacji) lub minimalnym błędzie (w przypadku regresji). Proces ten jest powtarzany rekurencyjnie dla każdego podzbioru, aż osiągnie się warunek stopu, na przykład osiągnięcie maksymalnej głębokości drzewa lub osiągnięcie minimalnej liczby próbek w liściach.

Modelem opartym o drzewa decyzyjne są tzw. lasy losowe (ang. *random forest*) wykorzystujące wiele drzew decyzyjnych w celu uzyskania bardziej stabilnego i dokładnego modelu. Każde drzewo jest trenowane na losowym podzbiorze danych treningowych oraz na losowym podzbiorze cech. Podczas trenowania drzewa, losuje się podzbiór cech, które mają być uwzględnione w każdym węźle. Ta technika nazywana jest baggingiem (ang. *bootstrap aggregating*). Chcąc dokonać predykcji na nowych danych, każde drzewo w lesie generuje własną predykcję – w przypadku klasyfikacji, wynik końcowy jest determinowany przez głosowanie większościowe drzew, podczas gdy w przypadku regresji, wynik końcowy jest uśrednianiem predykcji drzew. Główną zaletą lasów losowych jest eliminacja problemu przetrenowania, ponieważ drzewa są trenowane na różnych podzbiorach danych i cech, oraz odporność na szum i nieregularności w danych, ponieważ wynik jest uzyskiwany na podstawie konsensusu wielu drzew. W celu polepszenia predykcji opartych o lasy losowe stosuje się tzw. *Gradient Boosting* – metodę polegającą na sekwencyjnym dodawaniu prostych modeli do istniejącego modelu w celu minimalizacji funkcji kosztu. *Gradient Boosting* rozpoczyna się od inicjalizacji modelu bazowego, który może być prostym modelem, na przykład pojedynczym drzewem decyzyjnym. Następnie, iteracyjnie dodaje się kolejne modele do modelu bazowego w celu poprawienia wyników predykcji. Każdy nowy model jest trenowany w oparciu o przewidywanie poprzednich modeli (różnice między rzeczywistymi etykietami a predykcjami modelu). Podczas trenowania nowego modelu, *Gradient Boosting* skupia się na znalezieniu optymalnych wartości wag dla cech, które minimalizują zadaną funkcję kosztu. Wykorzystuje w tym celu gradient funkcji kosztu, który wskazuje kierunek najszybszego spadku wartości funkcji kosztu. Nowy model jest trenowany w taki sposób, aby redukować reszty poprzednich modeli, podążając w kierunku przeciwnym do gradientu funkcji kosztu. Proces dodawania kolejnych modeli jest kontynuowany tak długo, aż zostaną spełnione pewne warunki zatrzymania, takie jak osiągnięcie maksymalnej liczby iteracji, osiągnięcie minimalnej wartości reszt lub brak dalszego poprawiania wyników.

W końcu XGBoost (ang. *Extreme Gradient Boosting*) to metoda uczenia maszynowego, która stanowi rozwinięcie techniki *Gradient Boosting*. Podobnie jak lasy losowe, XGBoost

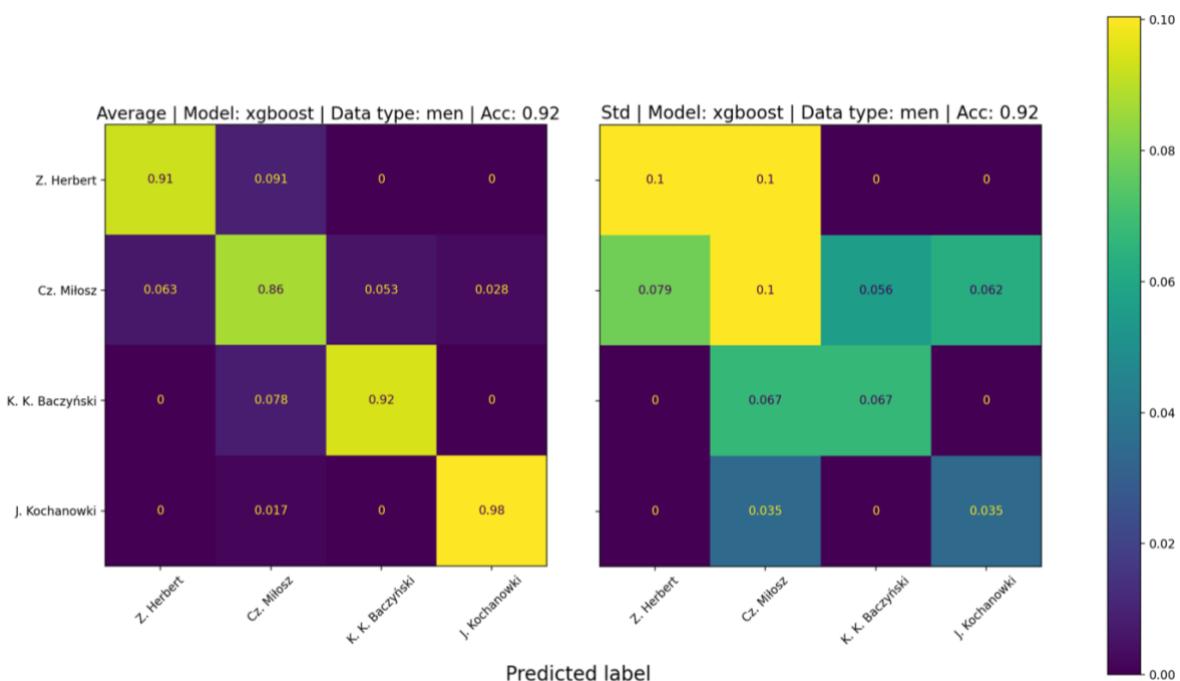
zamiast tworzyć niezależne drzewa decyzyjne, działa sekwencyjnie, dodając kolejne drzewa do modelu w celu minimalizacji funkcji kosztu poprzez gradientowe zstępowanie. Jest to proces iteracyjny, który dopasowuje drzewa do różnic między prawdziwymi etykietami a predykcjami modelu i łączy je z wcześniejszymi drzewami w taki sposób, aby zminimalizować błąd predykcji. Metoda XGBoost została wybrana jako *state of the art* w aspekcie wydajności algorytmów uczenia maszynowego.

Rys. 25 przedstawia macierz konfuzji predykcji przy pomocy XGBoost. Metoda ta pozwoliła osiągnąć dokładność przewidywań na poziomie 73%. Poniższe macierze konfuzji przedstawiają uśrednioną dokładność modelu oraz odchylenie standardowe dla 10 realizacji. XGBoost w bardzo dobrym stopniu radzi sobie z rozpoznawaniem tekstów Jana Kochanowskiego (99%), Krzysztofa Kamila Baczyńskiego (94%), Zbigniewa Herberta (80%) oraz Haliny Poświatowskiej (78%). Na części macierzy, gdzie są prezentowane predykcje klas kobiet można zaobserwować więcej pomyłek modelu niż dla klas mężczyzn. Analiza odchylenia standardowego pozwala zwrócić uwagę na błędą predykcję modelu klasy Wiślawy Szymborskiej dla tekstów Czesława Miłosza. Zauważalne też jest bardzo wysokie odchylenie dla Poświatowskiej, Marii Pawlikowskiej-Jasnorzewskiej oraz Ewy Lipskiej.



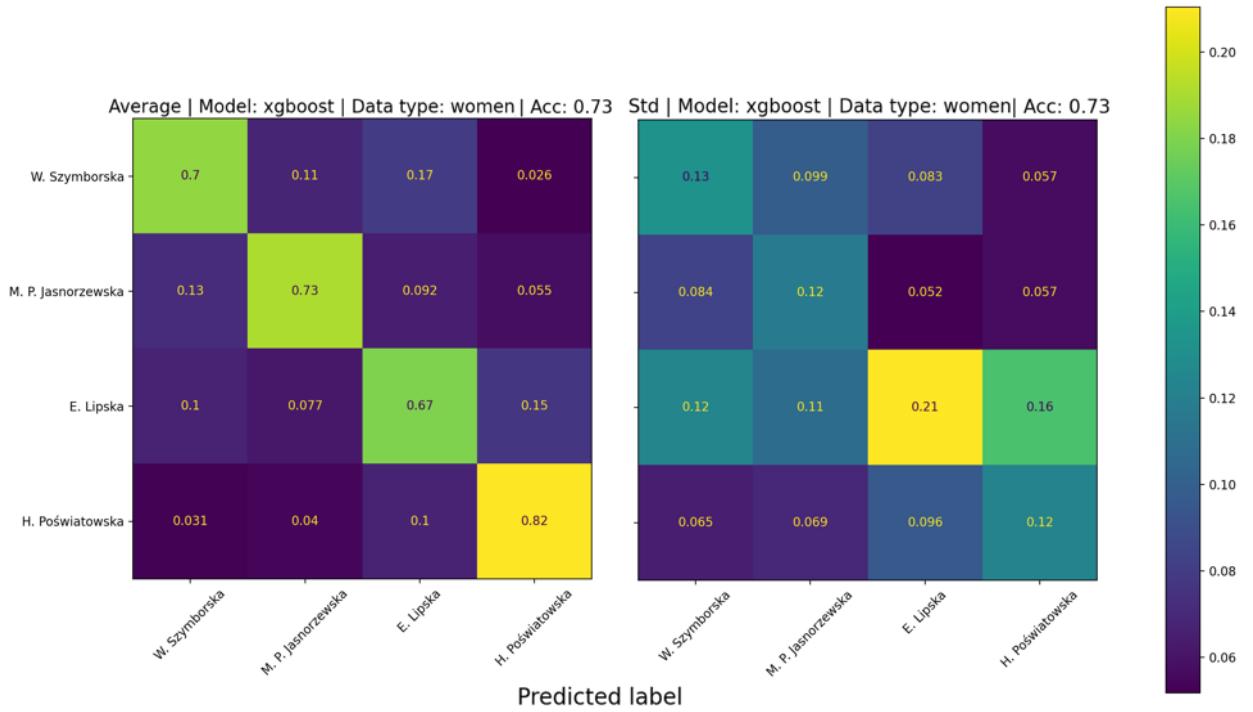
Rys. 25. Macierz konfuzji klasyfikacji opartej o XGBoost dla wszystkich klas. Model bardzo dobrze rozpoznaje teksty J. Kochanowskiego (99%), K. K. Baczyńskiego (94%), Z. Herberta (80%) oraz H. Poświatowskiej (78%). Na części macierzy, gdzie są prezentowane predykcje klas kobiet można zaobserwować więcej pomyłek modelu niż dla klas mężczyzn.

Predykcja modelu XGBoost tylko dla poezji mężczyzn osiąga dokładność równą 92%, Rys. 26. Większość poetów uzyskało wynik powyżej 90%, jedynie teksty Czesława Miłosza zostają rozpoznawane na poziomie dokładności 86%. Na macierzach konfuzji można zaobserwować drobne pomyłki modelu, które są związane z najsłabiej przewidywaną klasą. Miłosz jest błędnie przewidywany dla innych klas oraz sam jest też niepoprawnie identyfikowany. Pomyłki tekstów Miłosza i Zbigniewa Herberta, na które wskazuje wysokie odchylenie standardowe, mogą wynikać ze wspólnego czasu historycznego, w jakim tworzyli. Drugim argumentem, który może wyjaśnić taki wynik jest podobieństwo filozofii i sposobu prowadzenia refleksji Herberta i Miłosza w jego późniejszej fazie twórczości.



Rys. 26. Macierz konfuzji klasyfikacji opartej o XGBoost dla poezji mężczyzn. Predykcja modelu osiąga dokładność równą 92%. Większość poetów uzyskało wynik powyżej 90%. Teksty Cz. Miłosza zostają rozpoznawane na poziomie dokładności 86% i najczęściej są dla nich przewidywane klasy Z. Herberta oraz K. K. Baczyńskiego.

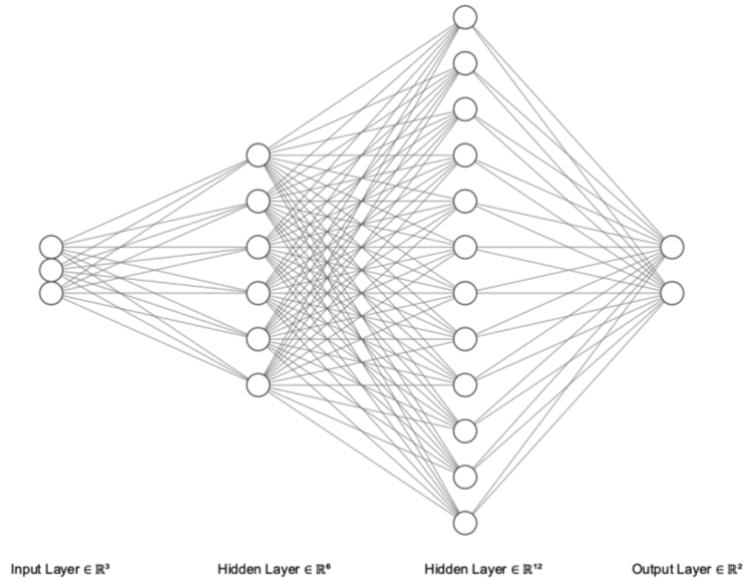
XGBoost uzyskuje dokładność równą 73% dla wierszy pisanych przez kobiety, Rys. 27. Najwyższy wynik osiąga twórczość Haliny Poświatowskiej (82%), natomiast pozostałe poetyki są rozpoznawane na poziomie dokładności około 70%. Model dla utworów Ewy Lipskiej często przewidywał klasę Wisławy Szymborskiej oraz Haliny Poświatowskiej. Natomiast teksty Szymborskiej były etykietowane jako twórczość Lipskiej lub Marii Pawlikowskiej-Jasnorzewskiej. Analiza wyników odchylenia standardowego dla tekstów Lipskiej pozwala zauważać dużą zbieżność pomiędzy jej twórczością, a utworami Poświatowskiej. Obie poetki często nadawały swoim utworom pesymistyczny charakter i wybierały tematy związane ze śmiercią, co mogło wpływać na tak liczne pomyłki.



Rys. 27. Macierz konfuzji klasyfikacji opartej o XGBoost dla poezji kobiet. Model uzyskuje dokładność równą 73%. Najwyższy wynik klasa H. Poświatowskiej (82%), natomiast pozostałe poetki są rozpoznawane na poziomie dokładności około 70%.

5.3. Klasyfikacja przy pomocy sieci neuronowej

Ostatnim klasyfikatorem jest kilkuwarstwowa sieć neuronowa o architekturze zaprezentowanej na Rys. 28 i Rys. 29.



Rys. 28. Grafika przedstawia uproszczony schemat budowy głębokiej sieci neuronowej wykorzystanej do klasyfikacji wierszy. Rozmiar warstwy wejściowej to 768, a następujące warstwy zwiększały liczbę neuronów dwukrotnie na każdej kolejnej warstwie. Na wyjściu sieci jest przewidywana jedna z ośmiu możliwych klas.

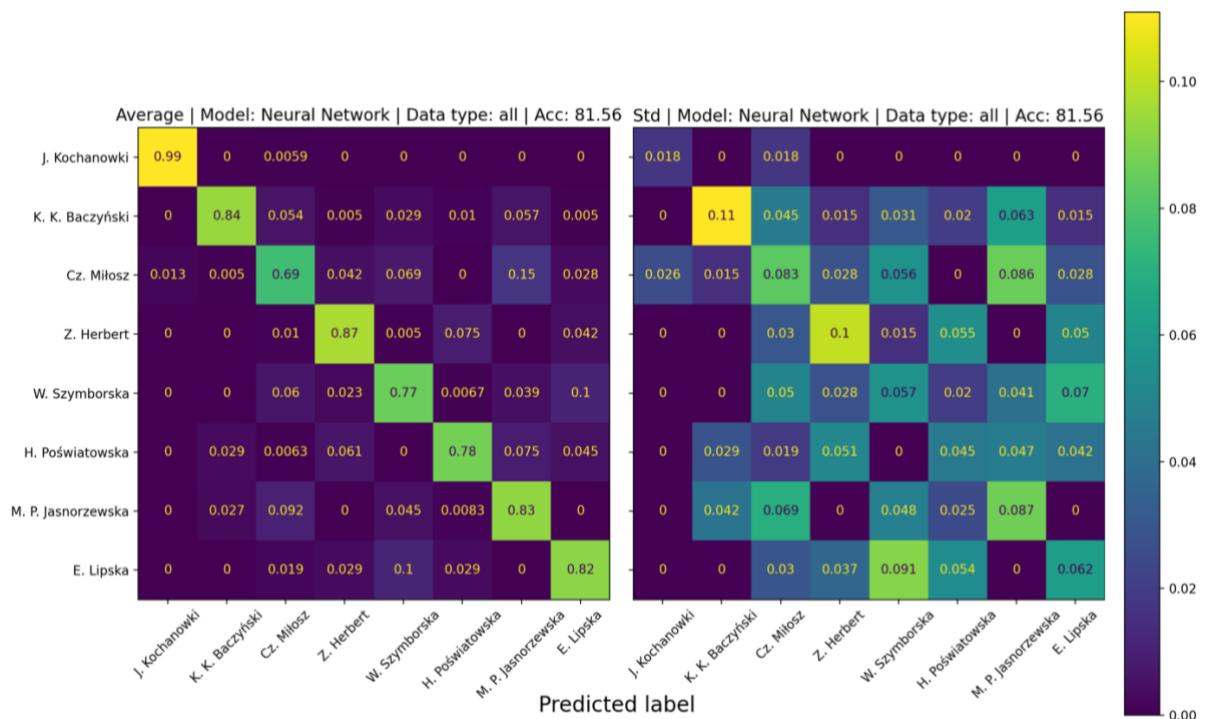
Layer (type)	Output Shape	Param #
Linear-1	[-1, 1536]	1,181,184
Linear-2	[-1, 3072]	4,721,664
Dropout-3	[-1, 3072]	0
Linear-4	[-1, 8]	24,584
<hr/>		
Total params:	5,927,432	
Trainable params:	5,927,432	
Non-trainable params:	0	
<hr/>		
Input size (MB):	0.00	
Forward/backward pass size (MB):	0.06	
Params size (MB):	22.61	
Estimated Total Size (MB):	22.67	

Rys. 29. Zapis modelu wykorzystanego do trenowania i predykcji przedstawia ilość oraz typ warstw zastosowanych w sieci. Liczba parametrów użytych do trenowania jest bliska 6000000.

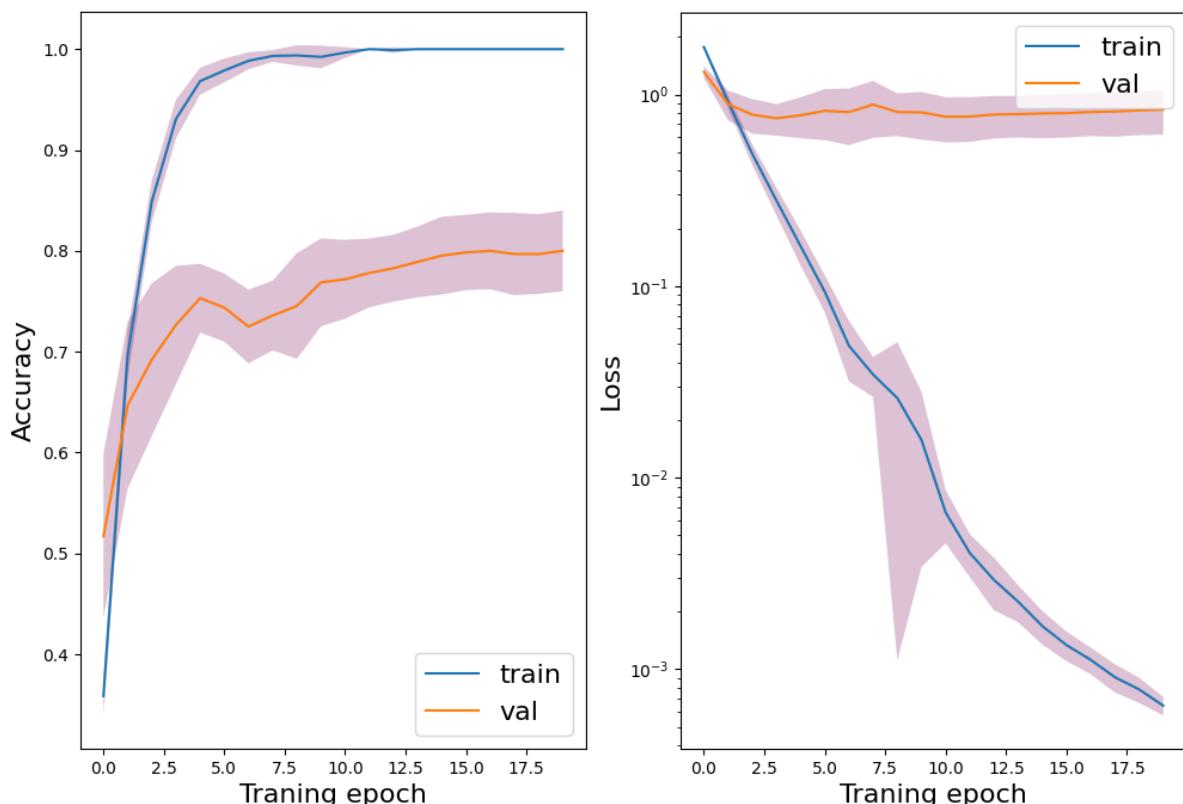
Do analizy używany jest model złożony z dwóch warstw ukrytych i jednej warstwy regularyzującej *dropout*, która odrzuca 0.2 neuronów przed warstwą wyjściową. Optymalizator użyty do trenowania sieci to Adam, którego parametr szybkości uczenia (ang. *learning rate*) wynosi 0.0001. Funkcja aktywacji zastosowana w warstwach ukrytych to ReLU (ang. *rectified linear unit*). Sieć jest trenowana przez 20 epok.

Rys. 20 przedstawia macierz konfuzji uśrednioną po 10 realizacjach (lewy panel) oraz odchylenie standardowe (prawy panel). Dokładność predykcji modelu osiąga wynik 82%. Najlepiej rozpoznawane klasy to: Jan Kochanowski (99%) i Zbigniew Herbert (87%). Reszta wyników oscyluje w okolicy 80%. Jedynym wyjątkiem są teksty Czesława Miłosza, które są poprawnie rozpoznawane jedynie w 69% przypadków. Model kilkakrotnie wskazał klasę Marii Pawlikowskiej-Jasnorzewskiej dla tekstów Miłosza.

Krzywa uczenia się sieci neuronowej dla wszystkich klas zaprezentowana na Rys. 31 przedstawia uśredniony proces trenowania i walidacji z oznaczeniem odchylenia standardowego. Z wykresu można wyczytać, że model się przeucza (ang. *overfitting*) już po drugiej epoce trenowania. Jest to niewątpliwie rezultatem małej ilości danych. Jednak operacje mające na celu augmentację danych nie poprawiały jakości modelu. Obecny w sieci *dropout* pozwala zredukować przeuczenie do prezentowanego poziomu.

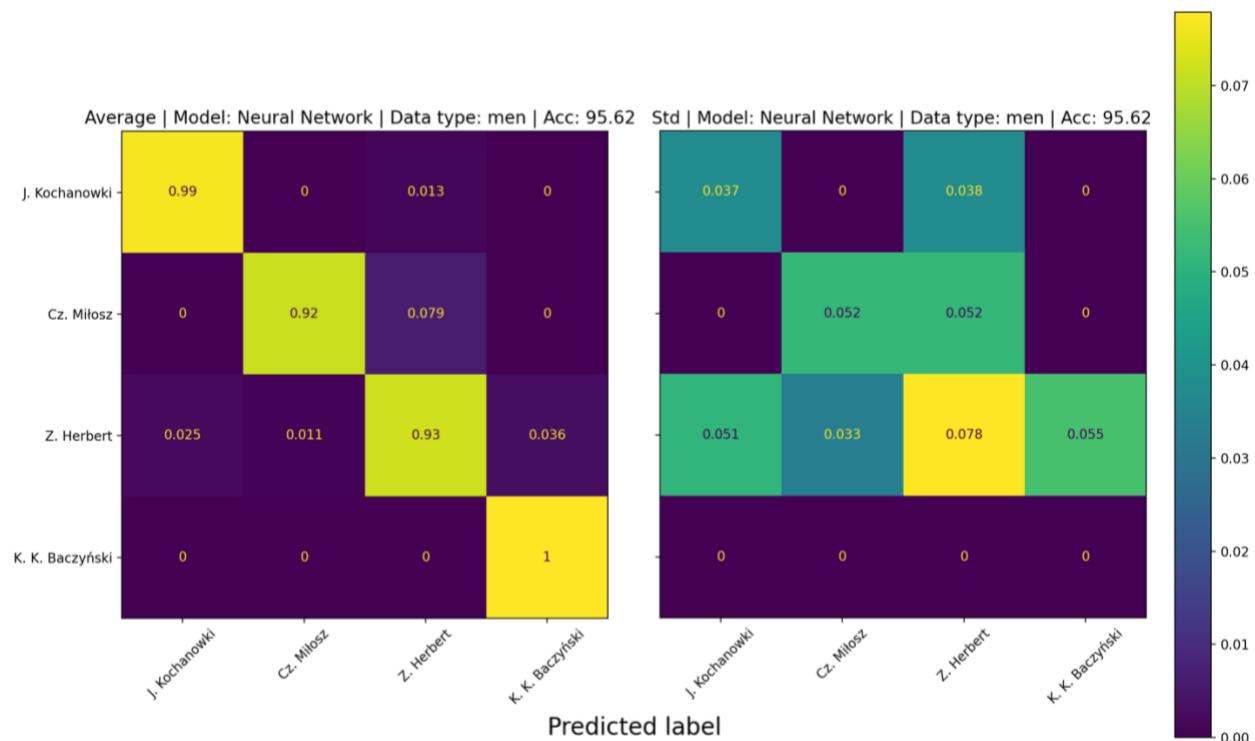


Rys. 30. Macierz konfuzji klasyfikacji przy użyciu modelu sieci neuronowej dla wszystkich klas. Dokładność predykcji modelu osiąga wynik 82%. Najlepiej rozpoznawane klasy to: J. Kochanowski i Z. Herbert. Pozostałe klasy osiągają wyniki w okolicy 80%.



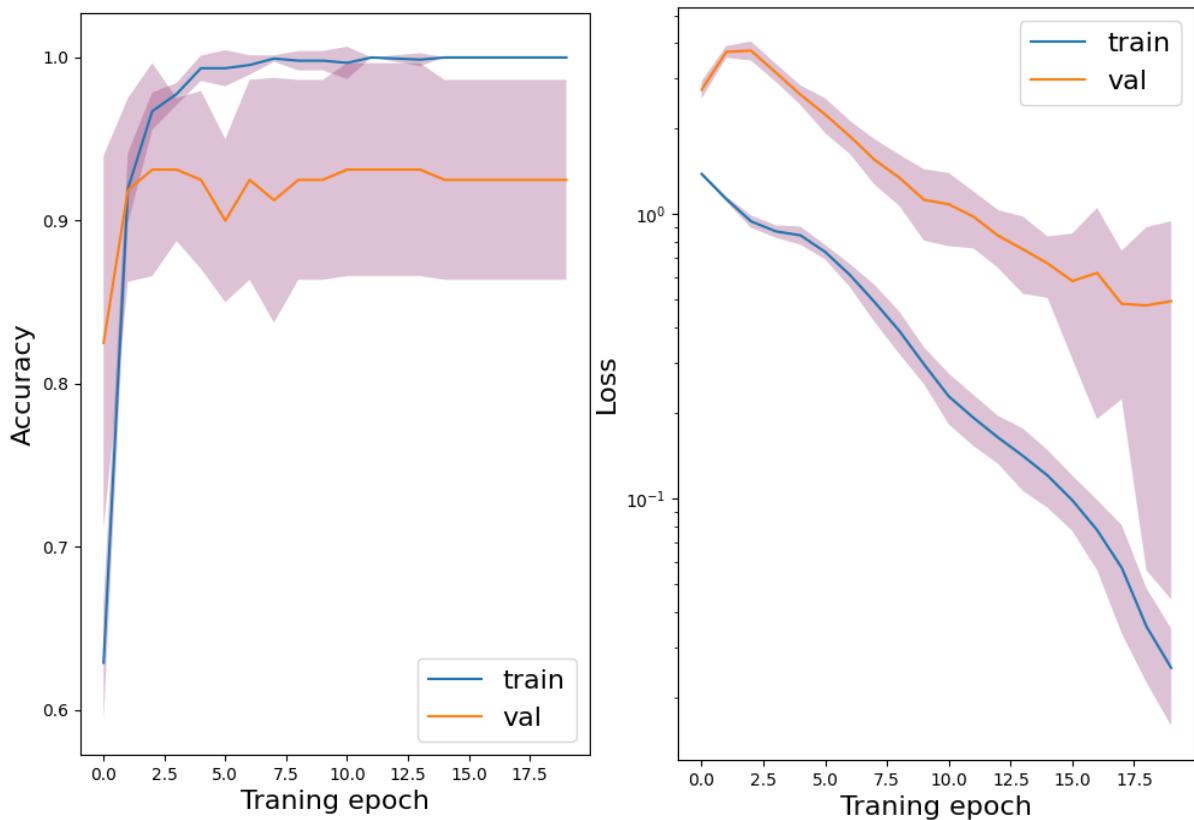
Rys. 31. Krzywa uczenia się modelu dla wszystkich klas. Sieć ulega przeuczeniu już po 2 epokach. Wzrost uczenia się modelu zatrzymuje się po 10 epoce i pozostaje na tym poziomie.

Wynik dokładności predykcji dla poezji mężczyzn wynosi 96%, Rys. 32. Najlepiej identyfikowanym twórcą jest Krzysztof Kamil Baczyński (100%). Drugi najwyższy wynik to 99% dla klasy Jana Kochanowskiego. Pozostali dwaj poeci są rozpoznawani z dokładnością powyżej 90%. Odchylenie standardowe utrzymuje się na bardzo niskim poziomie, co wskazuje na stabilność wyników.



Rys. 32. Macierz konfuzji dla modelu sieci neuronowej dla poezji mężczyzn. Dokładność predykcji wynosi 96%. Teksty K. K. Baczyńskiego zostały poprawnie rozpoznane w 100% przypadków, a J. Kochanowskiego w 99%.

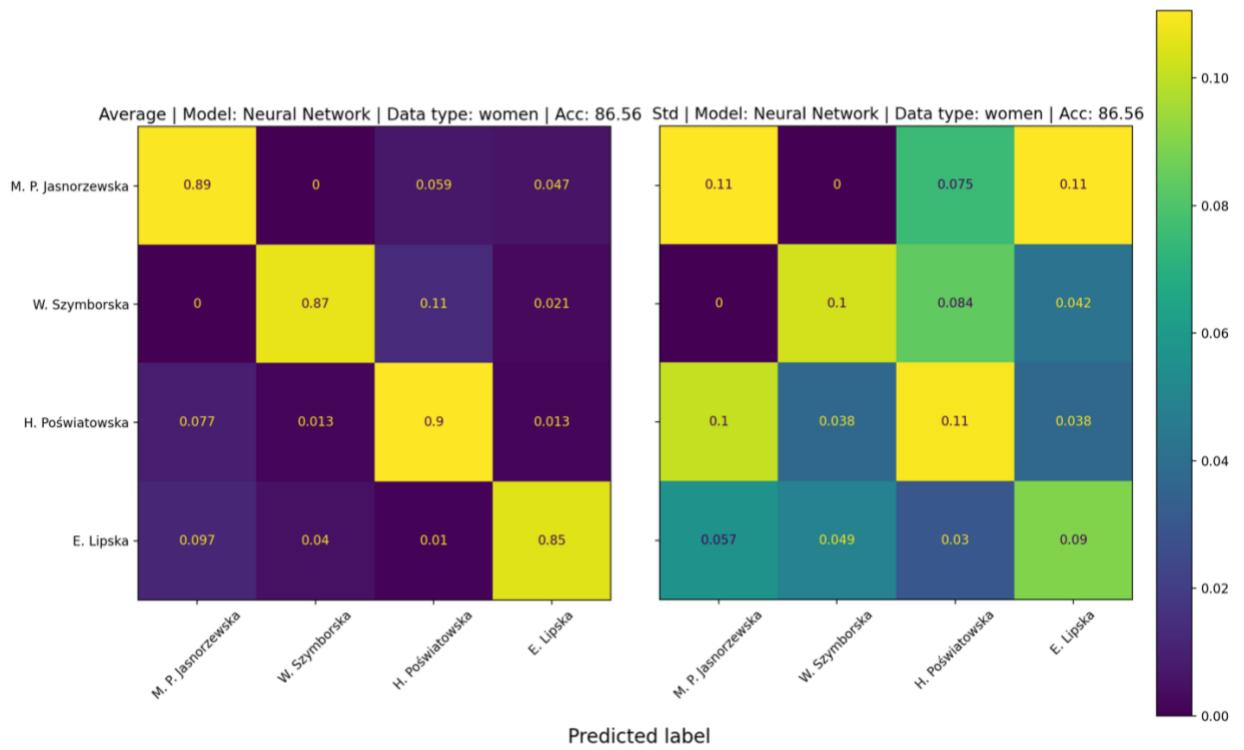
Krzywa uczenia dla tekstów pisanych przez mężczyzn prezentowana na Rys. 33 wciąż wskazuje na przeuczenie już po drugiej epoce trenowania, jednak jest ono mniejsze niż w przypadku klasyfikacji wszystkich klas. Odchylenie standardowe pokazuje, że kilkukrotnie dane walidacyjne były klasyfikowane z podobną dokładnością, co próbki treningowe. Krzywa stabilizuje się po około siedmiu epokach trenowania.



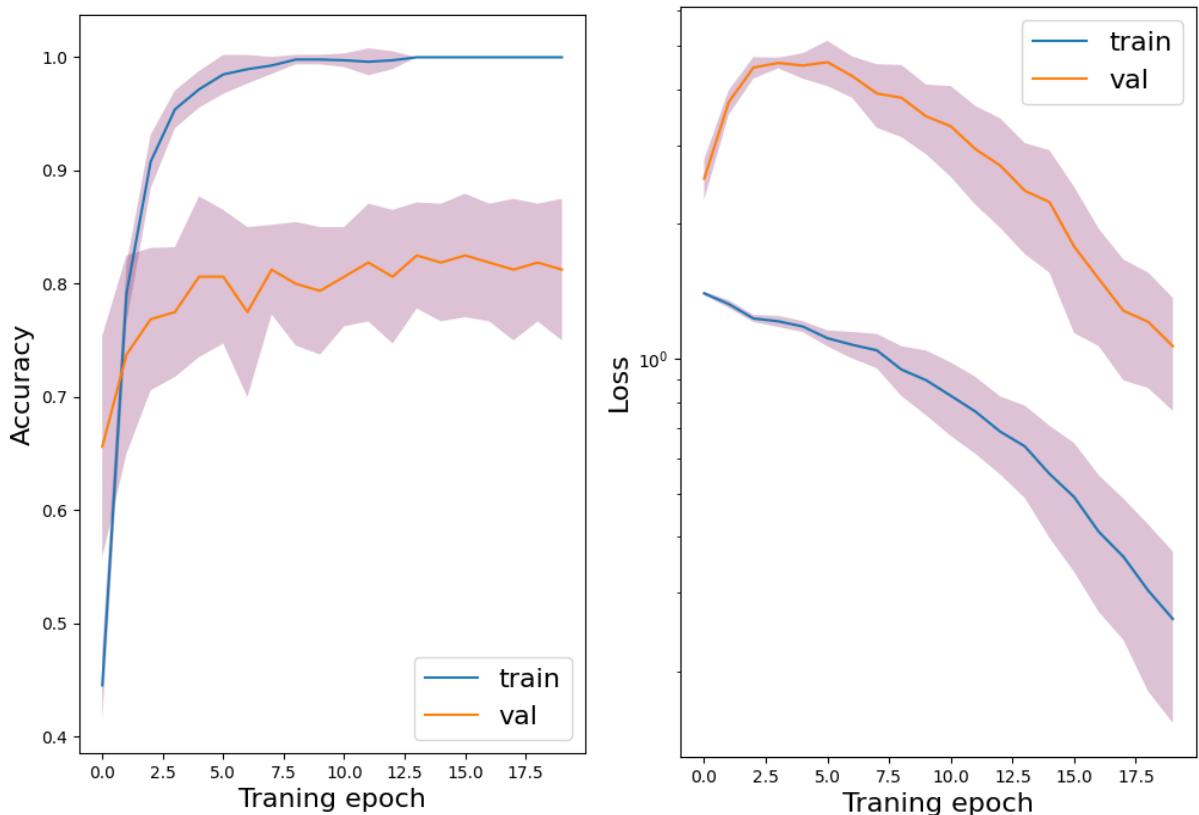
Rys. 33. Krzywa uczenia się modelu dla poezji mężczyzn. Sieć ulega przeuczeniu już po 2 epokach. Wzrost uczenia się modelu zatrzymuje się w okolicy siódmej epoki i pozostaje na tym poziomie.

Predykcja poezji kobiet osiąga dokładność równą 87%. Najlepiej jest rozpoznawana twórczość Haliny Poświatowskiej (90%) oraz Marii Pawlikowskiej-Jasnorzewskiej (89%). Pozostałe dwie klasy poetek są przewidywane z dokładnością powyżej 80%. Analizując macierz konfuzji przedstawianą na Rys. 34 można dostrzec, że Pawlikowska-Jasnorzewska i Wisława Szymborska jako jedyne nie są ze sobą mylone. Natomiast kilkakrotnie model przewidywał dla tekstów Szymborskiej klasę Haliny Poświatowskiej. Wysokie odchylenie standardowe w przypadku predykcji klasy Ewy Lipskiej dla tekstów Pawlikowskiej-Jasnorzewskiej jest niespotykanym wcześniej zjawiskiem. Jest to otwarcie nowych możliwości analizy jakościowej tych dwóch poetek w ramach badań komparatystycznych nad literaturą.

Krzywa uczenia dla tekstów pisanych przez kobiety prezentowana na Rys. 35 ponownie wskazuje na przeuczenie się modelu już po drugiej epoce uczenia. Krzywa osiąga maksymalny wynik w dwunastej epoce, jednak nie stabilizuje się, ponieważ wciąż są widoczne momenty pogorszenia lub poprawy wyników.



Rys. 34. Macierz konfuzji dla modelu sieci neuronowej dla poezji kobiet. Dokładność predykcji wynosi 87%. Twórczość H. Poświatowskiej jest rozpoznawana poprawnie w 90%, a M. Pawlikowskiej-Jasnorzowskiej 89% przypadków.



Rys. 35. Krzywa uczenia się modelu dla poezji kobiet. Wykres wskazuje na przeuczenie się modelu już po drugiej epoce uczenia. Krzywa osiąga maksymalny wynik w dwunastej epoce trenowania.

6. Podsumowanie

Niniejsza praca jest poświęcona problemowi klasyfikacji tekstów poetyckich polskich autorów. We wstępnie zostało omówiona pokrótko historia rozwoju głębokich sieci neuronowych oraz przetwarzanie języka naturalnego jako jeden z obszarów badawczych uczenia głębokiego. W kolejnym rozdziale wyjaśniono następujące po sobie innowacje w dziedzinie NLP oraz zwróciło uwagę na możliwe zastosowania przetwarzania języka naturalnego do badań literaturoznawczych. W rozdziale 3. Przybliżono, czym jest styl w literaturze oraz krótko omówiono charakterystykę twórczości poetów i poetek, których wybrano do zadania klasyfikacji. Następnie przy użyciu algorytmów pozwalających zredukować wymiarowość wektorowych reprezentacji wierszy (PCA, t-SNE oraz UMAP) przeprowadzono dogłębną analizę danych opartą o analizę odległości między wierszami i analizę skupień. W ostatnim rozdziale pracy zamieszczono wyniki dotyczące przeprowadzonej klasyfikacji tekstów poetyckich. Modele jakie zostały użyte do predykcji to: algorytm porównywania odległości euklidesowych pomiędzy wektorowymi reprezentacjami wierszy, model uczenia maszynowego XGBoost oraz sieć neuronowa. Klasyfikatorem o najwyższej dokładności okazała się sieć neuronowa, która dla wszystkich klas uzyskała wynik 82%. Natomiast grupą klas najłatwiejszą do klasyfikacji została poezja pisana wyłącznie przez mężczyzn, która osiągnęła wynik dokładności predykcji na poziomie 96%.

Model	Wszystkie klasy	Mężczyźni	Kobiety
Odległości Euklidesowe	62%	88%	61%
XGBoost	73%	92%	73%
Sieć Neuronowa	82%	96%	87%

Tab. 2. Porównanie wyników predykcji dla poszczególnych modeli. Najlepiej klasyfikującym modelem jest sieć neuronowa, która osiągnęła wynik dokładności predykcji 82%.

Modele wykorzystane do klasyfikacji były w stanie wychwycić zależności i podobieństwa opisywane także w badaniach literaturoznawczych. Ewolucja twórczości Czesława Miłosza powodowała liczne pomyłki w predykcji. Podobnych charakter wierszy i prowadzenie refleksji na temat codzienności podejmowane przez Wisławę Szymborską i Ewę Lipską, które również były ze sobą mylone. Ciekawym przypadkiem było bardzo wysokie odchylenie standardowe dla Ewy Lipskiej i Marii Pawlikowskiej Jasnorzewskiej, co mogłoby zachęcić literaturoznawców do głębszej analizy porównawczej tych dwóch poetek.

Przetwarzanie języka naturalnego ma przed sobą wiele wyzwań związanych z niejednoznacznością języka ludzkiego. Rozwój modeli językowych w taki sposób, aby były w stanie rozumieć i interpretować bardziej złożone środki językowe używane w literaturze pięknej jak: metafora, ironia czy oksymoron, jest bardzo ważne. Badania w tym celu oprócz znaczenia czysto literaturoznawczego mogą mieć znaczny wpływ na życie codzienne. Takie rozwiązania mogą pomóc np. w wyszukiwaniu mowy nienawiści (ang. *hate speech*) w serwisach Internetowych, czy zapobieganiu przestępstwom.

Spis ilustracji

Rys. 1. Prezentacja schematów działania metody <i>skip-gram</i> oraz CBOW. Źródło rysunku: [12]	17
Rys. 2. Schemat neuronowej sieci rekurencyjnej. Dane zostają wprowadzone do sieci przez warstwę wejściową. Następnie są analizowane przez warstwy ukryte, a kierowane dalej informacje wyjściowe trafiają ponownie do opuszczanego komponentu, gdzie dołączają się do kolejnych danych dostarczanych przez warstwę wejściową.	18
Rys. 3. Schemat sieci neuronowej LSTM przedstawia działanie sieci dla trzech kroków czasowych. W każdym kroku dane są wprowadzane przez warstwę wejściową i kierowane do warstw ukrytych. Przez warstwy ukryte przepuszczane są informacje zdobyte przez model w poprzednich krokach czasowych, a proces zapamiętywania jest równoległy do uczenia się sieci.....	19
Rys. 4. Schemat bramek w sieci LSTM przedstawia proces wprowadzenia, selekcji i przekazania dalej informacji zapamiętanych w poprzednich krokach czasowych. Ten element sieci LSTM przyjmuje na wejściu dane z poprzedniego i obecnego kroku czasowego, które zostają przefiltrowane przez kolejne bramki: zapomnienia, kandydata oraz wyjściową.	20
Rys. 5. Schemat architektury modelu Transformer. Źródło: [14].....	22
Rys. 6. Graficzne przedstawienie odległości euklidesowych (lewy panel) oraz cosinusowych (prawy panel) dla wektorowej reprezentacji analizowanego zbioru danych składającego się z 400 wierszy. Pierwsze 200 wektorów to zbiór mężczyzn, a wektory od 201 do 400, to zbiór kobiet.	32
Rys. 7. Dwuwymiarowa analiza skupień przy użyciu PCA dla wszystkich klas. Jedynie wiersze J. Kochanowskiego stanowią dobrze określony klaster, natomiast pozostały twórcy nie tworzą wyraźnych klastrów.	33
Rys. 8. Trójwymiarowa analiza skupień przy użyciu PCA dla wszystkich klas, prezentowana z różnych kątów w celu głębszej analizy wyników. Każdy widok jest dedykowany przedstawieniu danego autora. Wiersze J. Kochanowskiego stanowią dobrze określony klaster. Widoczne są także nieco słabiej zdefiniowane klastry utworów K. K. Baczyńskiego oraz H. Poświatowskiej.	35
Rys. 9. Analiza skupień przy pomocy redukcji wymiarowości t-SNE dla wszystkich klas. Większość klas tworzy zwarte klastry.	36
Rys. 10. Dwuwymiarowa analiza skupień przy pomocy redukcji wymiarowości UMAP dla wszystkich klas. Większość klas tworzy zwarte klastry. Wyjątek stanowią wiersze Cz. Miłosza, M. Pawlikowskiej-Jasnorzewskiej oraz Ewy Lipskiej.	37
Rys. 11. Trójwymiarowa analiza skupień przy użyciu UMAP dla wszystkich klas, prezentowana z różnych kątów w celu głębszej analizy wyników. Każdy widok jest dedykowany przedstawieniu danego autora. Widoczna jest oddzielona klasa J. Kochanowskiego. Większość klas tworzy dobrze zdefiniowane klastry. Pisarze mniej wyraźni to Czesław Miłosz i Maria Pawlikowska-Jasnorzewska.	39
Rys. 12. Dwuwymiarowa analiza skupień przy pomocy redukcji wymiarowości PCA dla poezji mężczyzn. Można zaobserwować dwa wyraźne klastry utworzone przez wiersze J. Kochanowskiego oraz K. K. Baczyńskiego.	40
Rys. 13. Trójwymiarowa analiza skupień przy użyciu PCA dla poezji mężczyzn, prezentowana z różnych kątów w celu głębszej analizy wyników. Każdy widok jest	

dedykowany przedstawieniu danego autora. Zauważalne są trzy dobrze oddzielone klastry: J. Kochanowski, K. K. Baczyński oraz Z. Herbert.	41
Rys. 14. Dwuwymiarowa analiza skupień przy pomocy redukcji wymiarowości t-SNE dla poezji mężczyzn. Można zaobserwować wyraźne klastry dla wszystkich klas.	42
Rys. 15. Dwuwymiarowa analiza skupień przy pomocy redukcji wymiarowości UMAP dla poezji mężczyzn. Widoczne są wyraźne klastry dla wszystkich klas.	42
Rys. 16. Trójwymiarowa analiza skupień przy użyciu UMAP dla poezji mężczyzn, prezentowana z różnych kątów w celu głębszej analizy wyników. Każdy widok jest dedykowany przedstawieniu danego autora. Zauważalne są trzy dobrze oddzielone klastry: J. Kochanowski, K. K. Baczyński, Z. Herbert oraz większość wierszy Cz. Miłosza.	43
Rys. 17. Dwuwymiarowa analiza skupień przy pomocy redukcji wymiarowości PCA dla poezji kobiet. Dane są rozmieszczone w sposób niezgrupowany. Można zaobserwować podział przestrzeni na trzy części zajmowane przez H. Poświatowską, M. Pawlikowską-Jasnorzewską oraz E. Lipską.	44
Rys. 18. Trójwymiarowa analiza skupień przy użyciu PCA dla poezji kobiet, prezentowana z różnych kątów w celu głębszej analizy wyników. Każdy widok jest dedykowany przedstawieniu danej autorki. Zauważalne są klastry H. Poświatowskiej oraz M. Pawlikowskiej-Jasnorzewskiej. Jednak punkty należące do E. Lipskiej i W. Szymborskiej nakładają się na siebie	45
Rys. 19. Dwuwymiarowa analiza skupień przy pomocy redukcji wymiarowości t-SNE dla poezji kobiet. Dane w większości są rozmieszczone w sposób niezgrupowany.	46
Rys. 20. Dwuwymiarowa analiza skupień przy pomocy redukcji wymiarowości UMAP dla poezji kobiet. Widoczne są dwa klastry utworzone przez wiersze M. Pawlikowskiej-Jasnorzewskiej oraz H. Poświatowskiej.	47
Rys. 21. Trójwymiarowa analiza skupień przy użyciu UMAP dla poezji kobiet, prezentowana z różnych kątów w celu głębszej analizy wyników. Każdy widok jest dedykowany przedstawieniu danej autorki. Widoczne są oddzielone klastry H. Poświatowskiej oraz M. Pawlikowskiej-Jasnorzewskiej. Utwory E. Lipskiej i W. Szymborskiej grupują się w nieznacznym stopniu.	48
Rys. 22. Macierz konfuzji dla predykcji na podstawie odległości euklidesowych dla wszystkich klas. Dokładność predykcji klasy z wykorzystaniem tej metody osiąga 62%. Twórczość mężczyzn uzyskuje dużo wyższe wyniki niż w przypadku kobiet.	50
Rys. 23. Macierz konfuzji dla predykcji na podstawie odległości euklidesowych dla poezji mężczyzn. Dokładność predykcji klasy z wykorzystaniem tej metody osiąga 88%. Teksty K. K. Baczyńskiego i J. Kochanowskiego uzyskały wynik powyżej 90%. Model najsłabiej rozpoznaje utwory Cz. Miłosza.	51
Rys. 24. Macierz konfuzji dla predykcji na podstawie odległości euklidesowych dla poezji kobiet. Dokładność predykcji klasy z wykorzystaniem tej metody osiąga 61%. Najlepiej rozpoznawaną przez model poetką jest W. Szymborska. Ciekawym zjawiskiem jest także duża liczba predykcji klasy W. Szymborskiej dla utworów E. Lipskiej.	52
Rys. 25. Macierz konfuzji klasyfikacji opartej o XGBoost dla wszystkich klas. Model bardzo dobrze rozpoznaje teksty J. Kochanowskiego (99%), K. K. Baczyńskiego (94%), Z. Herberta (80%) oraz H. Poświatowskiej (78%). Na części macierzy, gdzie są prezentowane predykcje klas kobiet można zaobserwować więcej pomyłek modelu niż dla klas mężczyzn.	54

Rys. 26. Macierz konfuzji klasyfikacji opartej o XGBoost dla poezji mężczyzn. Predykcja modelu osiąga dokładność równą 92%. Większość poetów uzyskało wynik powyżej 90%. Teksty Cz. Miłosza zostają rozpoznawane na poziomie dokładności 86% i najczęściej są dla nich przewidywane klasy Z. Herberta oraz K. K. Baczyńskiego.....	55
Rys. 27. Macierz konfuzji klasyfikacji opartej o XGBoost dla poezji kobiet. Model uzyskuje dokładność równą 73%. Najwyższy wynik klasa H. Poświatowskiej (82%), natomiast pozostałe poetyki są rozpoznawane na poziomie dokładności około 70%.....	56
Rys. 28. Grafika przedstawia uproszczony schemat budowy głębokiej sieci neuronowej wykorzystanej do klasyfikacji wierszy. Rozmiar warstwy wejściowej to 768, a następujące warstwy zwiększą liczbę neuronów dwukrotnie na każdej kolejnej warstwie. Na wyjściu sieci jest przewidywana jedna z ośmiu możliwych klas.....	56
Rys. 29. Zapis modelu wykorzystanego do trenowania i predykcji przedstawia ilość oraz typ warstw zastosowanych w sieci. Liczba parametrów użytych do trenowania jest bliska 6000000.....	57
Rys. 30. Macierz konfuzji klasyfikacji przy użyciu modelu sieci neuronowej dla wszystkich klas. Dokładność predykcji modelu osiąga wynik 82%. Najlepiej rozpoznawane klasy to: J. Kochanowski i Z. Herbert. Pozostałe klasy osiągają wyniki w okolicy 80%.....	58
Rys. 31. Krzywa uczenia się modelu dla wszystkich klas. Sieć ulega przeuczeniu już po 2 epokach. Wzrost uczenia się modelu zatrzymuje się po 10 epoce i pozostaje na tym poziomie.....	58
Rys. 32. Macierz konfuzji dla modelu sieci neuronowej dla poezji mężczyzn. Dokładność predykcji wynosi 96%. Teksty K. K. Baczyńskiego zostały poprawnie rozpoznane w 100% przypadków, a J. Kochanowskiego w 99%.....	59
Rys. 33. Krzywa uczenia się modelu dla poezji mężczyzn. Sieć ulega przeuczeniu już po 2 epokach. Wzrost uczenia się modelu zatrzymuje się w okolicy siódmej epoki i pozostaje na tym poziomie.....	60
Rys. 34. Macierz konfuzji dla modelu sieci neuronowej dla poezji kobiet. Dokładność predykcji wynosi 87%. Twórczość H. Poświatowskiej jest rozpoznawana poprawnie w 90%, a M. Pawlikowskiej-Jasnorzewskiej 89% przypadków.....	61
Rys. 35. Krzywa uczenia się modelu dla poezji kobiet. Wykres wskazuje na przeuczenie się modelu już po drugiej epoce uczenia. Krzywa osiąga maksymalny wynik w dwunastej epoce trenowania.....	61

Spis tabel

Tab. 1 Przykładowe tworzenie par słów (target – kontekst) w algorytmie skip-gram wykorzystywanym do trenowanie modelu word2vec.....	17
Tab. 2. Porównanie wyników predykcji dla poszczególnych modeli Najlepiej klasyfikującym modelem jest sieć neuronowa, która osiągnęła wynik dokładności predykcji 82%.....	62

Referencje

- [1] A. Geron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow. Concepts, Tools, and Techniques to Build Intelligent Systems*, O'Reilly, 2019.
- [2] T. J. Sejnowski, *The Deep Learning Revolution: Machine Intelligence Meets Human Intelligence*, The MIT Press, 2018.
- [3] J. D. Kelleher, *Deep Learning*, The MIT Press, 2019.
- [4] K. Hornik, M. Stinchcombe, H. White, *Multilayer Feedforward Networks are Universal Approximators*, Pergamon Press 1989, t. 2, s. 359-366.
- [5] H. Lane, C. Howard, H. M. Hapke, *Natural Language Processing in Action. Understanding, analyzing, and generating text with Python*, Manning Shelter Island, 2019.
- [6] A. Trzoss, *Komputerowo wspomagane metody badania tekstów w polskiej perspektywie*, 2022. Dostęp zdalny (20.06.2023):
https://www.researchgate.net/publication/359602168_Komputerowo_wspomagane_metody_badania_tekstow_w_polskiej_perspektywie
- [7] A. Jarynowski, S. Boland, *Rola analizy sieci społecznych w odkrywaniu narracyjnej struktury fikcji literackiej*, „Biuletyn Instytutu Systemów Informatycznych” 2013, 12, s. 32-42. Dostęp zdalny (20.06.2023):
https://isi.wat.edu.pl/sites/default/files/Biuletyn/Nr_12_2013/35_42_rolaanalizy_ajarynowski_sboland_12_2013.pdf
- [8] M. Baj, T. Walkowiak, *Computer Based Stylometric Analysis of Texts in Polish Language*, 2017. Dostęp zdalny (20.06.2023):
https://www.researchgate.net/publication/318017532_Computer_Based_Stylometric_Analysis_of_Texts_in_Polish_Language
- [9] Eder M., *In Search of the Author of Chronica Polonorum Ascribed to Gallus Anonymus: A Stylometric Reconnaissance*, „Acta Poloniae Historica” 2015, 112, s. 5–23.
- [10] Eder M., *Metody ścisłe w literaturoznawstwie i pułapki pozornego obiektywizmu – przykład stylometrii*, „Teksty Drugie” 2014, 2, s. 90–105.
- [11] Eder M., *Mind your corpus: systematic errors in authorship attribution*, „Literary and Linguistic Computing” 2013, 28(4), s. 603–614.
- [12] T. Mikolov, K. Chen, G. Corrado, J. Dean, *Efficient Estimation of Word Representations in Vector Space*, Dostęp zdalny (20.06.2023): <https://arxiv.org/abs/1301.3781>

- [13] I. Sutskever, O. Vinyals, Q. V. Le, *Sequence to Sequence Learning with Neural Networks*, 2014. Dostęp zdalny (20.06.2023): <https://arxiv.org/abs/1409.3215>
- [14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, *Attention Is All You Need*, 2017. Dostęp zdalny (20.06.2023): <https://arxiv.org/abs/1706.03762>.
- [15] L. Tunstall, L. von Werra, T. Wolf, *Natural Language Processing with Transformers. Building Language Applications with Hugging Face*, O'Reilly, 2022.
- [16] D. Rothman, *Transformers for Natural Language Processing. Build, train, and fine-tune deep neural network architectures for NLP with Python, PyTorch, TensorFlow, BERT, and GPT-3*, Packt, 2022.
- [17] J. Devlin, M.W. Chang, K. Lee, K. Toutanova, *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*, 2019. Dostęp zdalny (20.06.2023): <https://arxiv.org/abs/1810.04805>
- [18] R. Mroczkowski, P. Rybak, A. Wróblewska, I. Gawlik, *HerBERT: Efficiently Pretrained Transformer-based Language Model for Polish*, 2021. Dostęp zdalny (20.06.2023): <https://aclanthology.org/2021.bsnlp-1.1.pdf>
- [19] A. Wilkoń, *Język a styl tekstu literackiego*, „Język Artystyczny” 1978, 1, s. 11-21. Dostęp zdalny (20.06.2023): https://bazhum.muzhp.pl/media/files/Jazyk_Artystyczny/Jazyk_Artystyczny-r1978-t1/Jazyk_Artystyczny-r1978-t1-s11-21/Jazyk_Artystyczny-r1978-t1-s11-21.pdf
- [20] J. Jęśko, *Jan Kochanowski*, Biblioteka Narodowa, 1985.
- [21] J. Święch, *Wybór poezji / Krzysztof Kamil Baczyński*, Zakład Narodowy im. Ossolińskich, 1998.
- [22] Z. Łapiński, *Poezje wybrane / Czesław Miłosz*, Zakład Narodowy im. Ossolińskich, 2013.
- [23] M. Mikołajczak, *Wybór poezji / Zbigniew Herbert*, Zakład Narodowy im. Ossolińskich, 2018.
- [24] W. Ligęza, *Wybór poezji / Wisława Szymborska*, Zakład Narodowy im. Ossolińskich, 2016.
- [25] K. Karaskiewicz, *Halina Poświatowska w zwierciadle swej kobiecości*, Ryt, 2008.
- [26] E. Hurnikowa, *Maria Pawlikowska-Jasnorzewska: (zarys monograficzny)*, Śląsk, 1999.
- [27] K. Skibski, *Antropologia wierszem: język poetycki Ewy Lipskiej*, Wydawnictwo Poznańskie, 2008.
- [28] Dostęp zdalny (20.06.2023): <https://github.com/allegro/HerBERT>