Sprawozdanie

Metody i Narzędzia Big Data

Mikołaj Stępień

269274

INS, W4N

11.06.2023

cel ćwiczenia:

Badanie możliwości klasyfikacji gier w zależności od ich opisów.

Od wielu lat podstawowym narzędziem klienta przy wyborze produktu był jego opis. Pozwalał on na wyciągnięcie kluczowych informacji o potencjalnym nabytku. W badaniu tym planuje wykorzystując najbardziej podstawowe narzędzia NLP (z ang. Natural Language Processing – przetwarzanie języka naturalnego) udostępnione w 3 czołowych bibliotekach pythonowych do tego stworzonych sprawdzić czy za pomocą klasyfikatora jesteśmy w stanie określić cechy produktu takie jak cena, lub wiek , oraz do jakiego gatunku należy.

W badaniu postaramy się również określić jakie metoda:

- Klasyfikacji
- Lematyzacji
- Sprowadzenia do postaci macierzowej tekstu

sprawuje się najlepiej w zależności od zadania.

Założenia aprioryczne:

Najbardziej podstawowe i biblioteki do NLP powinny z łatwością wykrywać gatunek gry.

Cena gry również powinna być łatwa do sklasyfikowania, ze względu na tendencję w opisach do „chwalenia” się dużym studiem stojącym za nią, co z kolei przekłada się na cenę sklepową.

Ciężej natomiast będzie z wiekiem, oraz przyjęciem gry, tutaj opis nie powinien wskazać pożądanych wyników.

Z czym planuję się zapoznać:

Do przeprowadzenia takiego badania niezbędne będzie zapoznanie się z 3 wiodącymi bibliotekami pythonowymi do pracy na tekście:

- SpaCy
- Gensim
- NLTK

Oprócz tego, użyte zostaną 2 najbardziej podstawowe metody przekształcania tekstu w macierz słów:

- Bag Of Words
- TF – IDF

Oprócz tego używane będą metody klasyfikacji poznane na zajęciach:

- Naiwny klasyfikator Bayesa
- Drzewo decyzyjne
- SVM – maszynę wektorów nośnych
- KNN – algorytm k- najbliższych sąsiadów
- Regresja logistyczna
- Spacer losowy

oraz oceny jakości tychże:

- Dokładność
- Czułość
- Precyzja
- F1 - score

Opis przeprowadzonych badań:

Narzędzia używane przy wykonaniu badania:

- Kaggle.com – jako źródło moich danych
- Excel – Jako pomoc do wstępnej eksploracji, oraz wizualizacji wyników
- Python 3.10 wraz z bibliotekami:
- SciKit Learn – Do tworzenia, trenowania i oceniania modeli
- ScyPy – Jako jeden z badanych narzędzi lematyzcji
- Gensim - Jako jeden z badanych narzędzi lematyzcji
- NLTK - Jako jeden z badanych narzędzi lematyzcji
- Pandas – Do pracy nad badanym zbiorem danych
- JSON – Do zapisu i i odczytu wyników
- DATETIME – Do pracy nad formatami czasowymi w danych
- MatplotLib – Do wizualizacji wyników badań

Szczegółowy opis prac:

I. Zdobycie i eksploracja danych.

Dane pasujące do pytania badawczego pobrane zostały z platformy Kaggle za uprzejmością użytkownika Aleksandr Antonov, który tudostępnił „ Steam games complete dataset” w formacie CSV.

W ramach pierwotnej eksploracji próbowałem otworzyć dane w programie EXCEL, który to dobrze radzi sobie z obsługą plików CSV.

Tutaj napotkałem pierwsze problemy, mianowicie pliki CSV był błędnie skonstruowany przez co dane okazały się nieużywalne.

Po głębszym zapoznaniu się z plikiem, oraz konstrukcją pliku CSV, doszedłem do wniosku, że problemem były ukryte symbole CRLF (Cartridge Reset Line Feed) odpowiadające za przejście do nowej linijki, które to w odczycie pliku traktowane są na równi z separatorem, którego funkcję pełnił tutaj przecinek. Powodowało to „rozlanie" się danych ze stylizowanych opisów na wiele tabelek uniemożliwiając poprawną interpretację danych przez program.

W celu naprawy bazy danych napisałem szybki program który miał za zadanie „wyczyścić" bazę danych ze stylistycznych CRLF'ów. Wykorzystałem w niej fakt, że poprawne CRLF'y zawsze powinny następować przed przejściem do 1 komórki następnego rzędu, która będąc linkiem zawsze zaczynała się od „https":

Po przepuszczeniu pliku przez program „czyszczący" i zweryfikowaniu, dane gotowe były do pracy.

II. Przygotowanie danych do pracy.

Zestaw danych składał się z 42 576 gier zawierających 19 atrybutów:

Link – unikalny dla każdego adres prowadzący do witryny na platformie steam gdzie produkt można nabyć

typ – Czy produkt jest grą, narzędziem deweloperskim albo zestawem

nazwa – unikalna nazwa gry

opis – opis gry widniejący na stronie produktu

„snippet" – skrócona wersja opisu wyświetlana w katalogu

recenzje z ostatniego miesiące - % pozytywnych recenzji , oraz ich ilość w ostatnich 30 dniach.

wszystkie recenzje –% pozytywnych recenzji , oraz ich ilość od początku

data wydania – data wydania produktu w formacie DD nazwa miesiąca, YYYY.

Deweloper – nazwa dewelopera

Wydawca – nazwa wydawcy

popularne tagi – tagi przypisywane przez użytkowników

cechy gry – cechy takie jak „gra jednoosobowa", „gra wieloosobowa" itp.

Języki – lista języków w jakich gra jest dostępna

liczba osiągnięć do zdobycia – liczba „osiągnięć" do zdobycia w grze

gatunek – lista gatunków gry

Czy i czemu dla pełnoletnich – opis treści dla pełnoletnich

minimalne wymagania – lista sprzętu spełniającego minimalne wymagania

rekomendowane wymagania – lista sprzętu opisująca rekomendowane wymagania

cena premierowa – cena gry w dniu premiery

cena z przeceny – aktualna cena gry

Ze względu na obszar badania wyciągnąłem 8 kluczowych atrybutów:

Nazwa, typ, opis, snippet, wszystkie recenzje, data wydania, gatunek, cena premierowa.

Zacząłem od zawężenia interesujących nas produktów do samych gier. Użyłem do tego tabeli typ gdzie prostym filtrem usunąłem wszystkie wpisy nie będące w kategorii „app". Na tym skończyła się użyteczność tej kolumny, została więc usunięta.

Następnie połączyłem tabele „opis" i „snippet" w jedno używając konkatenacji stringów. Nowa tabela „opis" będzie podstawą naszej pracy, dlatego filtrowałem wszystkie pozycje gdzie pole to było puste.

Tabelę wszystkie recenzje sprowadziłem do 3 klas:

POSITIVE – Dla produktów które miały ponad 59% pozytywnych recenzji

MIXED – Dla produktów w przedziale od 41 do 59% pozytywnych recenzji.

NEGATIVE – Dla produktów poniżej 41% pozytywnych recenzji.

Produkcje z niewystarczającą próbką lub brakiem recenzji zostały wyfiltrowane.

Tabelę „data wydania" sprowadziłem do 5 klas:

RETRO – gry wydane do 2000 roku

OLD – gry wydane od 2001 do 2011 roku

YOUNG – gry wydane od 2012 do 2016 roku

NEW – gry wydane od 2016 do dziś

Produkcje bez daty wydania zostały wyfiltrowane.

Tabelę „cena premierowa" sprowadziłem do 4 klas:

F2P – gry darmowe

Small – gry do 20 $

AA - gry od 20 $ do 50 $

AAA - gry od 50 $ wzwyż

Tabelę „gatunek" rodzieliłem na 10 tabel boolowskich reprezentujących czy gra jest z danego gatunku:

'Action','Adventure', 'Massively Multiplayer', 'Strategy','RPG', 'Indie','Simulation', 'Racing', 'Casual'

i 'Sports'.

W ten sposób uzyskaliśmy zestaw danych będący podstawą do naszych badań.

III. Badanie danych

Podstawą przetwarzania języka naturalnego jest tak zwana „wektoryzacja” opisu, czyli przekształcenie tekstu w postać macierzową.

Ponieważ w tworzeniu takiej macierzy każde słowo będzie miało znaczenie dobrą praktykom jest pierwotne „oczyszczenie” tekstu przez:

a) Usunięcie słów z tzw. „stop listy”, czyli listy słów które nic nie wnoszą do wypowiedzi same w sobie. W języku angielskim, bowiem to na nim pracujemy przykłady takich słów to np.: „the”, „a” czy „this”.

b) Lematyzację, czyli sprowadzanie danego słowa do jego formy podstawowej (hasłowej), która reprezentuje dany wyraz, np. wiórkami → wiórek, jeżdżący → jeździć.

Do czyszczenia tekstu posłużą nam 3 biblioteki:

SpaCy, Gensim, oraz NLTK.

Stworzymy 3 osobne atrybuty zawierające kolumnę „opis” przepuszczoną przez algorytmy czyszczące wbudowane w każdą z tych bibliotek i porównamy, która sprawiła się najlepiej.

Przykład działania lematyzacji i usunięcia stop listy z tekstu:

Oryginalny opis:

Portal™ is a new single player game from Valve. Set in the mysterious Aperture Science Laboratories,

Portal has been called one of the most innovative new games on the horizon and will offer gamers hours of unique gameplay.

About This Game Portal™ is a new single player game from Valve. Set in the mysterious Aperture Science Laboratories,

Portal has been called one of the most innovative new games on the horizon and will offer gamers hours of unique gameplay.

The game is designed to change the way players approach, manipulate, and surmise the possibilities in a given environment;

similar to how Half-Life® 2\'s Gravity Gun innovated new ways to leverage an object in any given situation.

Players must solve physical puzzles and challenges by opening portals to maneuvering objects, and themselves, through space.

Opis obrobiony z SpaCy:

portal new single player game valve set mysterious aperture science laboratories portal call innovative new game horizon offer gamer hour unique gameplay game portal new single player game valve set mysterious aperture science laboratories portal call innovative new game horizon offer gamer hour unique gameplay game design change way player approach manipulate surmise possibility give environment similar half life gravity gun innovate new way leverage object give situation player solve physical puzzle challenge open portal maneuver object space

Opis obrobiony z Gensim:

portal new single player game valve set mysterious aperture science laboratories portal called innovative new games horizon offer gamers hours unique gameplay about this game portal new single player game valve set mysterious aperture science laboratories portal called innovative new games horizon offer gamers hours unique gameplay the game designed change way players approach manipulate surmise possibilities given environment similar half life gravity gun innovated new ways leverage object given situation players solve physical puzzles challenges opening portals maneuvering objects themselves space

Opis obrobiony z NLTK:

Portal™ new single player game Valve . Set mysterious Aperture Science Laboratories , Portal called one innovative new game horizon offer gamers hour unique gameplay . Game Portal™ new single player game Valve . Set mysterious Aperture Science Laboratories , Portal called one innovative new game horizon offer gamers hour unique gameplay . game designed change way player approach , manipulate , surmise possibility given environment ; similar Half-Life® 2\ 's Gravity Gun innovated new way leverage object given situation . Players must solve physical puzzle challenge opening portal maneuvering object , , space .

Wersja tekstu:	Średnia liczba słów w opisie:
SpaCy	141.75
Gensim	154.16
NLTK	194.39
Oryginał	261.14

Teraz mając już dostęp do 3 różnych sposobów lematyzacji, skupłem się na porównywaniu wyników dla 2 sposobów zamieniania tekstu w macierze:

1. BOW (Bag of Words) - BOW to jedna z podstawowych technik przetwarzania języka naturalnego, polegająca na reprezentacji tekstu jako zbioru wszystkich słów w dokumencie, bez uwzględniania kolejności i gramatyki. Metoda ta tworzy słownik słów z korpusu tekstowego i tworzy wektory liczące częstość wystąpień poszczególnych słów w danym dokumencie. Słowa są traktowane jako "worek" (stąd nazwa) bez uwzględniania ich znaczenia.

2. TF-IDF (Term Frequency-Inverse Document Frequency) - TF-IDF to metoda statystyczna wykorzystywana do oceny ważności słów w dokumencie w kontekście całego korpusu. Składa się z dwóch czynników: częstości występowania słowa w dokumencie (TF) i odwrotności częstości występowania słowa we wszystkich dokumentach (IDF). Wysoka wartość TF-IDF dla danego słowa oznacza, że jest ono istotne dla danego dokumentu, ale jednocześnie występuje rzadko we wszystkich dokumentach korpusu.

Na każdej powstałej macierzy testowałem 5 klasyfikatorów dla domyślnej konfiguracji w bibliotece scikit-learn:

1. Naiwny klasyfikator Bayesa - Algorytm zakłada niezależność między cechami, a w kontekście NLP, te cechy to zazwyczaj słowa lub n-gramy. Na podstawie prawdopodobieństw warunkowych wyliczanych na podstawie wystąpień słów w korpusie treningowym, algorytm przypisuje dokument do odpowiedniej klasy.

2. Drzewo decyzyjne w NLP -. Drzewa decyzyjne w kontekście NLP opierają się na cechach tekstowych, takich jak słowa, n-gramy, cechy morfologiczne itp., które są używane jako testy decyzyjne w węzłach drzewa. Algorytm podejmuje decyzje na podstawie sekwencji testów, aby przyporządkować dokument do odpowiedniej klasy lub kategorii.

3. SVM (maszyna wektorów nośnych) - W kontekście NLP, algorytm SVM stara się znaleźć optymalne płaszczyzny, które rozdzielają dokumenty tekstowe należące do różnych klas. Wektory cech reprezentujące dokumenty tekstowe są mapowane do przestrzeni o wysokim wymiarze, a SVM znajduje granicę decyzyjną, która maksymalizuje margines oddzielający różne klasy.

4. KNN (algorytm k-najbliższych sąsiadów) - Algorytm KNN opiera się na podobieństwie między dokumentami tekstowymi. Przy pomocy funkcji podobieństwa, takiej jak odległość kosinusowa, KNN znajduje k najbliższych sąsiadów dla danego dokumentu tekstowego w przestrzeni cech. Na podstawie większości etykiet k najbliższych sąsiadów algorytm przyporządkowuje dokument do odpowiedniej klasy. W używanym przeze mnie klasyfikatorze wartość k ustawiona została na 5.

5. Regresja logistyczna - Algorytm estymuje parametry funkcji logistycznej, która przekształca kombinację cech tekstowych na prawdopodobieństwa przynależności do danej klasy.

6. Spacer losowy - Algorytm polega na losowych ruchach po przestrzeni tekstowej na podstawie prawdopodobieństw przejścia z jednego słowa na drugie. Może być używany do generowania nowych sekwencji tekstowych, takich jak generowanie tekstu, tworzenie nowych zdani itp.

Wygenerowane w ten sposób modele były ocenianie na podstawie:

1. Dokładności – odsetek poprawnie spredykowanych do wszystkich.

2. Precyzji – to stosunek poprawnie sklasyfikowanych elementów z B (TN) do wszystkich, które nasz klasyfikator oznaczył

3. Czułości – stosunek poprawnie rozpoznanych elementów z B (TN) do wszystkich, które powinien rozpoznać

4. F1- Score - jest miarą harmoniczną precyzji i czułości. Łączy te dwie miary w jedną liczbę, która uwzględnia zarówno precyzję, jak i czułość.

Po wykonaniu wszystkich modeli całość zapisałem w pliku JSON na podstawie którego wyciągałem wnioski i wyniki.

Wyniki i analiza

Fragment danych po wytrenowaniu i wyćwiczeniu ponad 600 modeli zamieściłem pod spodem w Tabeli 1:

Tabela 1

Atrybut	Tekst	Metoda	Kryterium Jakości	Klasyfikator	Wynik
age	SpaCy	BOW	Dokladnosc	Naive Bayes	0,585479
age	SpaCy	BOW	Dokladnosc	SVM	0,602546
age	SpaCy	BOW	Dokladnosc	Logistic Regression	0,573619
age	SpaCy	BOW	Dokladnosc	KNN	0,508823
age	SpaCy	BOW	Dokladnosc	Decision Tree	0,536592
age	SpaCy	BOW	Dokladnosc	Random Forest	0,597917

age	SpaCy	BOW	Precyzja	Naive Bayes	0,574965
age	SpaCy	BOW	Precyzja	SVM	0,62885
age	SpaCy	BOW	Precyzja	Logistic Regression	0,571947
age	SpaCy	BOW	Precyzja	KNN	0,493368
age	SpaCy	BOW	Precyzja	Decision Tree	0,530434
age	SpaCy	BOW	Precyzja	Random Forest	0,618395
age	SpaCy	BOW	Czulosc	Naive Bayes	0,585479
age	SpaCy	BOW	Czulosc	SVM	0,602546
age	SpaCy	BOW	Czulosc	Logistic Regression	0,573619
age	SpaCy	BOW	Czulosc	KNN	0,508823
age	SpaCy	BOW	Czulosc	Decision Tree	0,536592
age	SpaCy	BOW	Czulosc	Random Forest	0,597917
age	SpaCy	BOW	F1-Score	Naive Bayes	0,570661
age	SpaCy	BOW	F1-Score	SVM	0,582686
age	SpaCy	BOW	F1-Score	Logistic Regression	0,570726
age	SpaCy	BOW	F1-Score	KNN	0,47154
age	SpaCy	BOW	F1-Score	Decision Tree	0,533122
age	SpaCy	BOW	F1-Score	Random Forest	0,577231
age	SpaCy	TF-IDF	Dokladnosc	Naive Bayes	0,578826
age	SpaCy	TF-IDF	Dokladnosc	SVM	0,605438
age	SpaCy	TF-IDF	Dokladnosc	Logistic Regression	0,598207

itd……

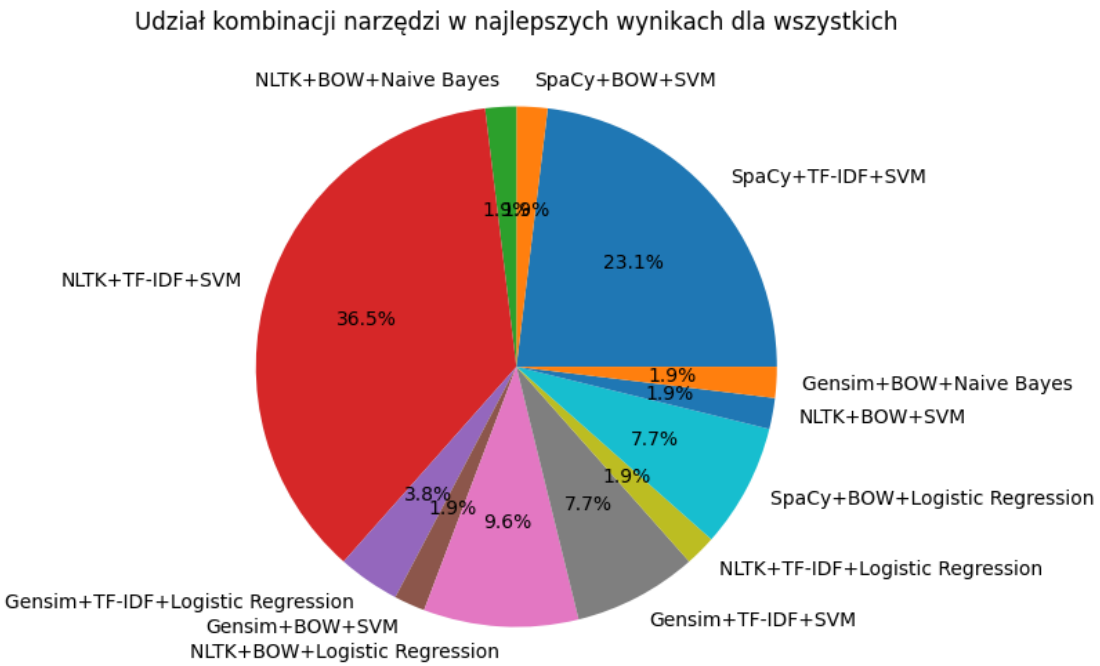
Żeby ustalić która metoda klasyfikacji jest najlepsza wyciągnąłem z tabeli najlepsze wyniki dla poszczególnych atrybutów umieszczone w tabeli 2:

Tabela 2

atrybut	kryterium	tekst	metoda	klasyfikator	wynik
reception	Dokladnosc	SpaCy	TF-IDF	SVM	0,68
	Precyzja	SpaCy	BOW	SVM	0,64
	Czulosc	SpaCy	TF-IDF	SVM	0,68
	F1-Score	NLTK	BOW	Naive Bayes	0,62
age	Dokladnosc	NLTK	TF-IDF	SVM	0,62
	Precyzja	NLTK	TF-IDF	SVM	0,63
	Czulosc	NLTK	TF-IDF	SVM	0,62
	F1-Score	NLTK	TF-IDF	SVM	0,60
price	Dokladnosc	Gensim	TF-IDF	Logistic Regression	0,76
	Precyzja	Gensim	BOW	SVM	0,76
	Czulosc	Gensim	TF-IDF	Logistic Regression	0,76
	F1-Score	NLTK	BOW	Logistic Regression	0,70
action	Dokladnosc	Gensim	TF-IDF	SVM	0,84
	Precyzja	Gensim	TF-IDF	SVM	0,84
	Czulosc	Gensim	TF-IDF	SVM	0,84
	F1-Score	Gensim	TF-IDF	SVM	0,84
adventure	Dokladnosc	SpaCy	TF-IDF	SVM	0,79

	Precyzja	SpaCy	TF-IDF	SVM	0,79
	Czulosc	SpaCy	TF-IDF	SVM	0,79
	F1-Score	SpaCy	TF-IDF	SVM	0,79
casual	Dokladnosc	SpaCy	TF-IDF	SVM	0,81
	Precyzja	SpaCy	TF-IDF	SVM	0,80
	Czulosc	SpaCy	TF-IDF	SVM	0,81
	F1-Score	NLTK	TF-IDF	Logistic Regression	0,79
indie	Dokladnosc	NLTK	TF-IDF	SVM	0,81
	Precyzja	NLTK	TF-IDF	SVM	0,81
	Czulosc	NLTK	TF-IDF	SVM	0,81
	F1-Score	NLTK	TF-IDF	SVM	0,80
massively multiplayer	Dokladnosc	SpaCy	BOW	Logistic Regression	0,98
	Precyzja	SpaCy	BOW	Logistic Regression	0,97
	Czulosc	SpaCy	BOW	Logistic Regression	0,98
	F1-Score	SpaCy	BOW	Logistic Regression	0,97
rpg	Dokladnosc	NLTK	TF-IDF	SVM	0,89
	Precyzja	NLTK	TF-IDF	SVM	0,89
	Czulosc	NLTK	TF-IDF	SVM	0,89
	F1-Score	NLTK	TF-IDF	SVM	0,88
racing	Dokladnosc	SpaCy	TF-IDF	SVM	0,98
	Precyzja	NLTK	BOW	SVM	0,98
	Czulosc	SpaCy	TF-IDF	SVM	0,98
	F1-Score	SpaCy	TF-IDF	SVM	0,98
simulation	Dokladnosc	NLTK	TF-IDF	SVM	0,86
	Precyzja	NLTK	TF-IDF	SVM	0,86
	Czulosc	NLTK	TF-IDF	SVM	0,86
	F1-Score	Gensim	BOW	Naive Bayes	0,85
sports	Dokladnosc	NLTK	BOW	Logistic Regression	0,97
	Precyzja	NLTK	BOW	Logistic Regression	0,97
	Czulosc	NLTK	BOW	Logistic Regression	0,97
	F1-Score	NLTK	BOW	Logistic Regression	0,97
strategy	Dokladnosc	NLTK	TF-IDF	SVM	0,89
	Precyzja	NLTK	TF-IDF	SVM	0,90
	Czulosc	NLTK	TF-IDF	SVM	0,89
	F1-Score	NLTK	TF-IDF	SVM	0,89

Umieściłem wszystkie występujące kombinację narzędzi wraz z ich procentem udziału na rysunku 1:



Rysunek 1 – diagram kołowy udziału kombinacji narzędzi w kategorii „najlepszych” klasyfikatorów.

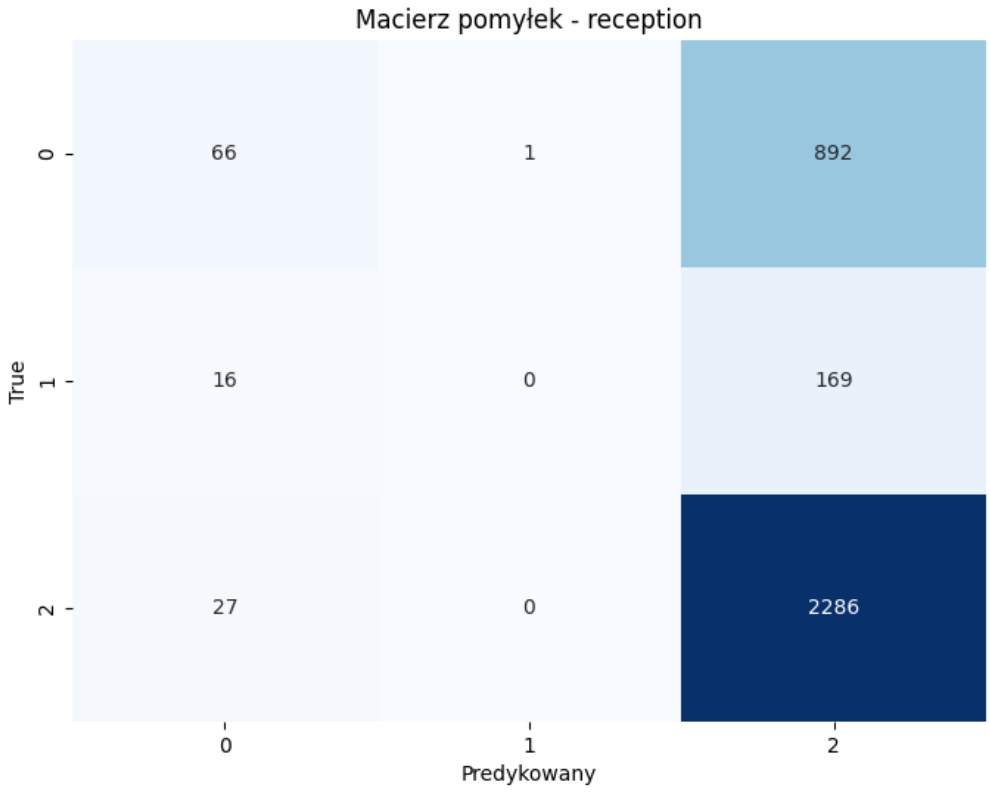
Tutaj podjąłem decyzję o tym, że metodą którą zastosuję do projektu będzie:

Lematyzator NLTK, przekształcony w macierz TF-IDF sklasyfikowany maszyną wektorów nośnych.

Wyniki pokazane będą więc z perspektywy tego modelu.

Dla tego podejścia modelowi udało się pozyskać podane wyniki:

Klasyfikacja odbioru gry przez użytkowników:



Rysunek 2 - macierz pomyłek odbioru gry, 0 – oznacza grę z recenzjami negatywnymi, 1 – grę z recenzjami mieszanymi, 2 natomiast z pozytywnymi.

Klasyfikator odbioru gier przez użytkowników:

Dokładność: 0.68

Precyzja: 0.62

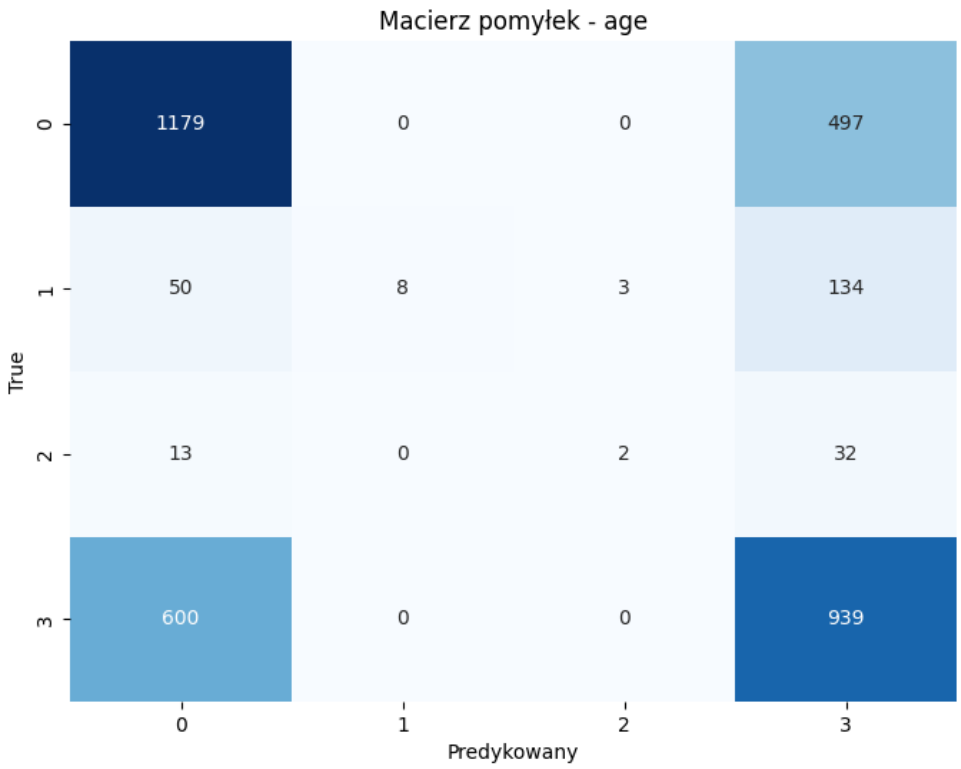
Czułość: 0.68

F1-Score: 0.57

Obserwacja:

Już na pierwszy rzut oka widać, że model ma tendencję do klasyfikowania zdecydowanie za dużej ilości gier jako „oceniona pozytywnie”. Nie potrafi on wychwycić praktycznie żadnej gry o mieszanych recenzjach, oraz rzadko kiedy uznaje grę za „ocenianą negatywnie”. Dokładność 0.68 pozostawia wiele do życzenia.

Klasyfikacja wieku gry:



Rysunek 3 – macierz pomyłek wieku gry, 0 – oznacza grę RETRO, 1 – oznacza grę starą, 2 – „młodą”, 3 – nowo wydaną grę

Klasyfikator wieku gier:

Dokładność: 0.62

Precyzja: 0.63

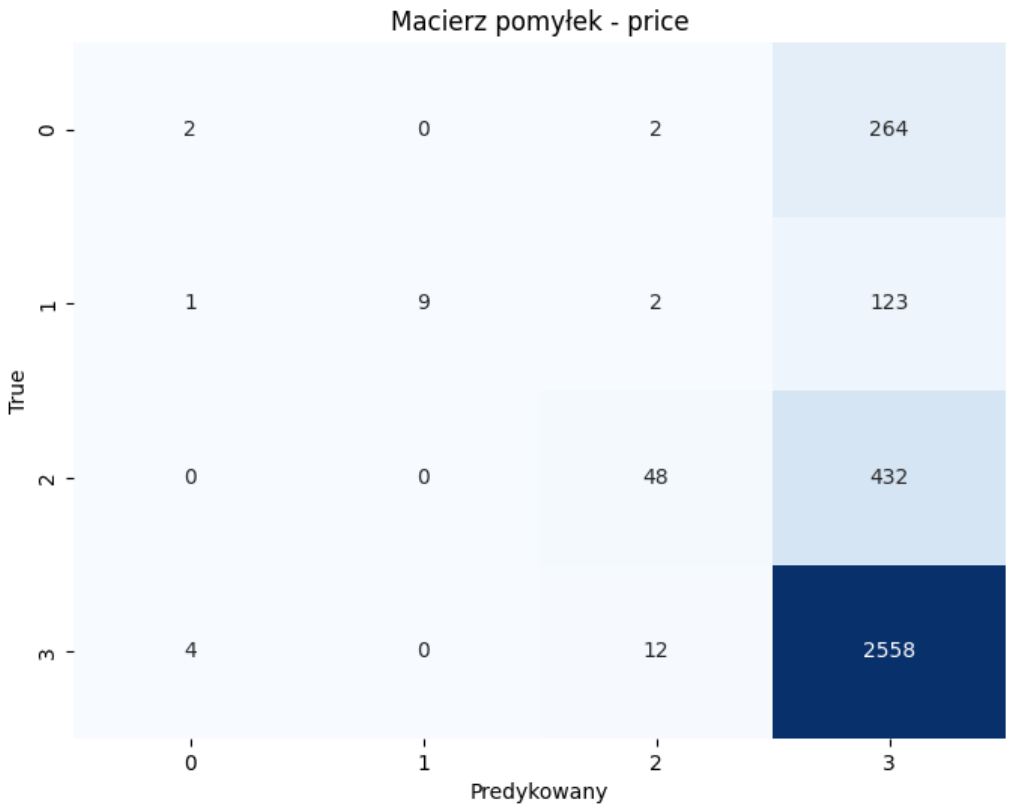
Czułość: 0.62

F1-Score: 0.60

Obserwacja:

W przypadku klasyfikatora wieku gry model bardzo rzadko uznaje kategorie „przechodnie” sprowadzając wszystko do klas „skrajnych”. Dokładność 0.62 w przypadku wysyłania zdecydowanej większości gier do tylko 2 kategorii jest bardzo niskim wynikiem.

Klasyfikacja ceny gry:



Rysunek 4 - macierz pomyłek ceny gry, 0- oznacza grę darmową, 1- oznacza grę do 20$, 2- oznacza grę od 20 do 50$, 3- oznacza grę 50$+

Klasyfikator ceny gier:

Dokładność: 0.76

Precyzja: 0.73

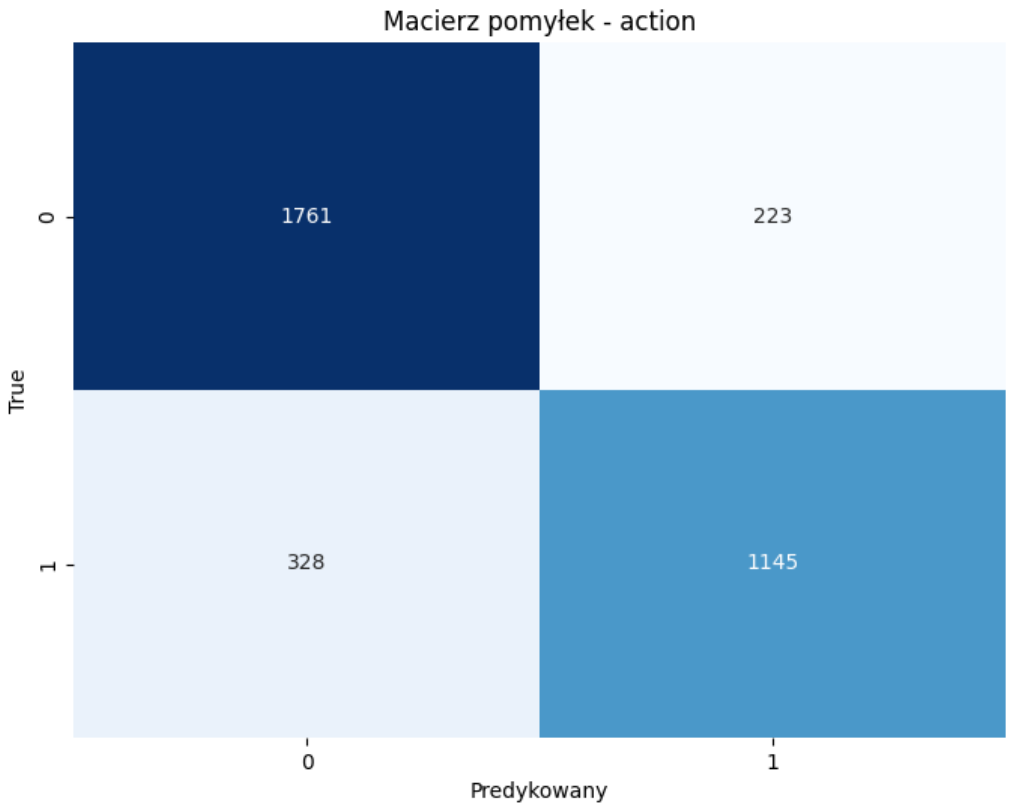
Czułość: 0.76

F1-Score: 0.67

Obserwacja:

Po macierzy pomyłek klasyfikatora ceny gry bardzo łatwo zauważyć można tendencję do sprowadzania każdej gry do jednej kategorii, wszystko klasyfikowane jest jako gry kosztująca powyżej 50$ dolarów. Pozostałe kategorie są niedostatecznie reprezentowane. Dokładność 0.76 jest tu wynikiem tylko przewagi tej kategorii nad resztą.

Klasyfikacja „Czy gra należy do gatunku akcji”:



Rysunek 5 - macierz pomyłek gatunku akcji, 0 - gra nie należy do gatunku, 1 - gra należy do gatunku

Klasyfikator gatunkowy gier akcji:

Dokładność: 0.84

Precyzja: 0.84

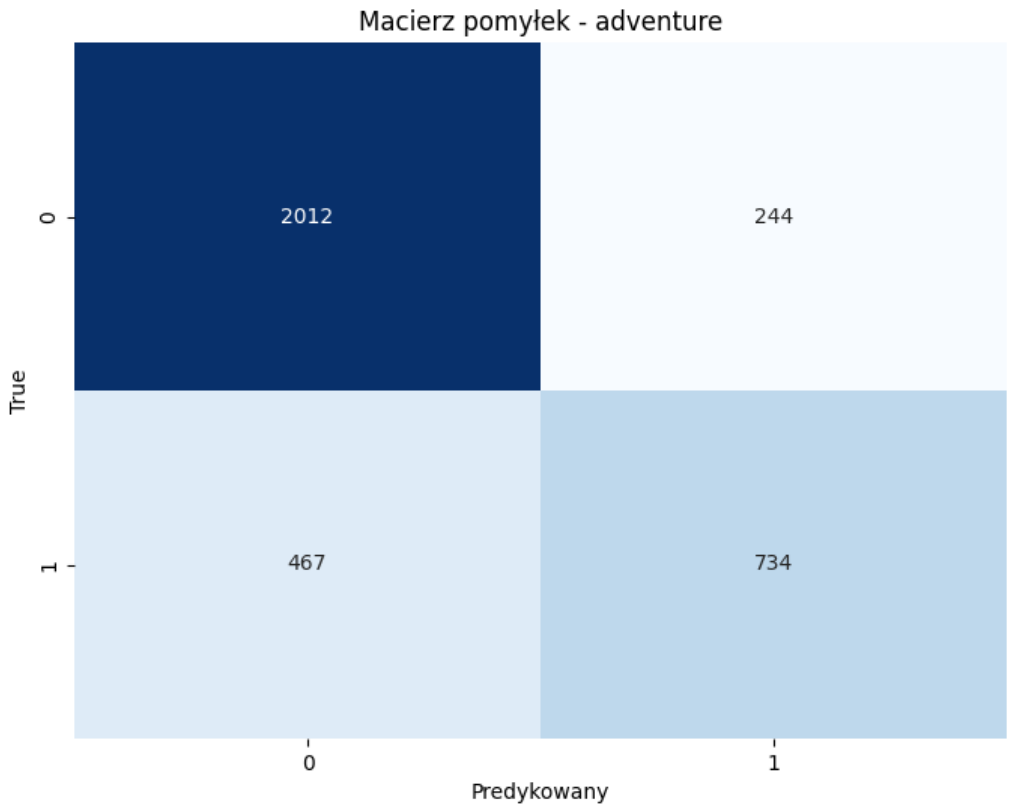
Czułość: 0.84

F1-Score: 0.84

Obserwacja:

Dla tej macierzy pomyłek klasyfikator radzi sobie już o wiele lepiej. Klasyfikator jest bardziej skłonny założyć, że coś nie jest grą akcji, niż na odwrót. Dokładność 0.84 daje nadzieje na w miarę dobry model przy większym nakładzie pracy.

Klasyfikacja „Czy gra należy do gatunku przygodowych”:



Rysunek 6 - macierz pomyłek gatunku gier przygodowych, 0 - gra nie należy do gatunku, 1 - gra należy do gatunku

Klasyfikator gatunkowy gier przygodowych:

Dokładność: 0.79

Precyzja: 0.79

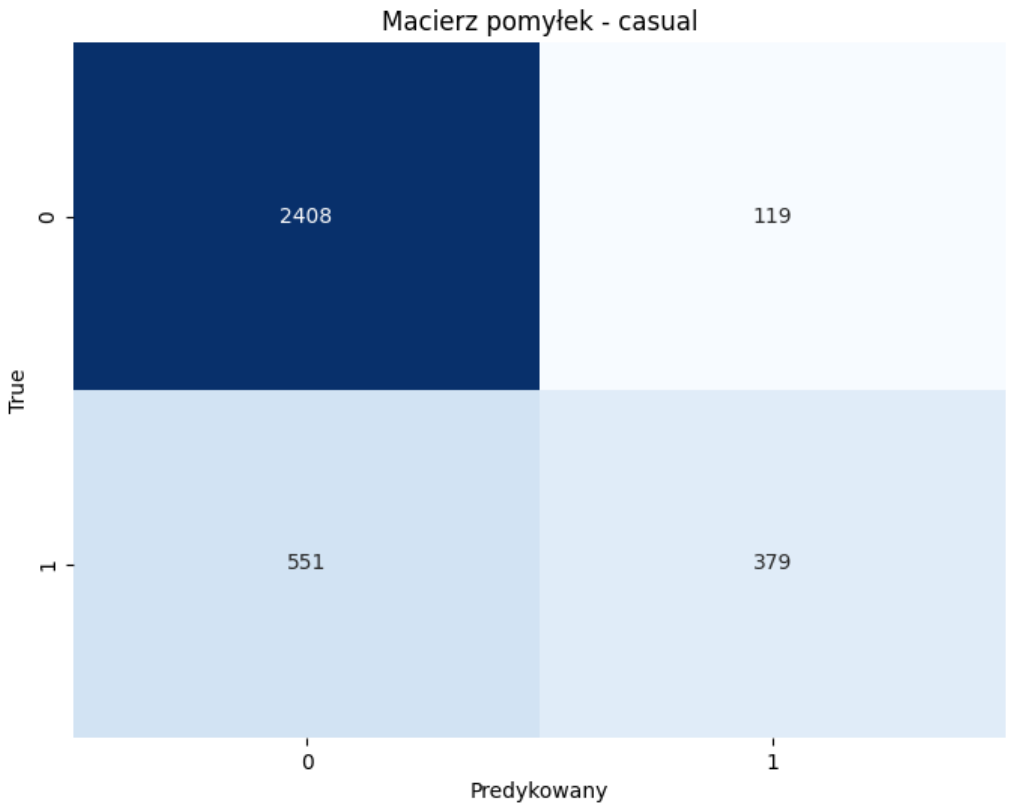
Czułość: 0.79

F1-Score: 0.79

Obserwacja:

W przypadku klasyfikatora gier przygodowych model ma w zwyczaju częściej zakładać o tym, że badany opis nie jest grą przygodową. Dokładność 0.79 również daje nadzieję na poprawę przy dalszych iteracjach.

Klasyfikacja „Czy gra należy do gatunku ”casualowych””:



Rysunek 7 - macierz pomyłek gatunku gier ”casualowych”, 0 - gra nie należy do gatunku, 1 - gra należy do gatunku

Klasyfikator gatunkowy gier ”casualowych”:

Dokładność: 0.81

Precyzja: 0.80

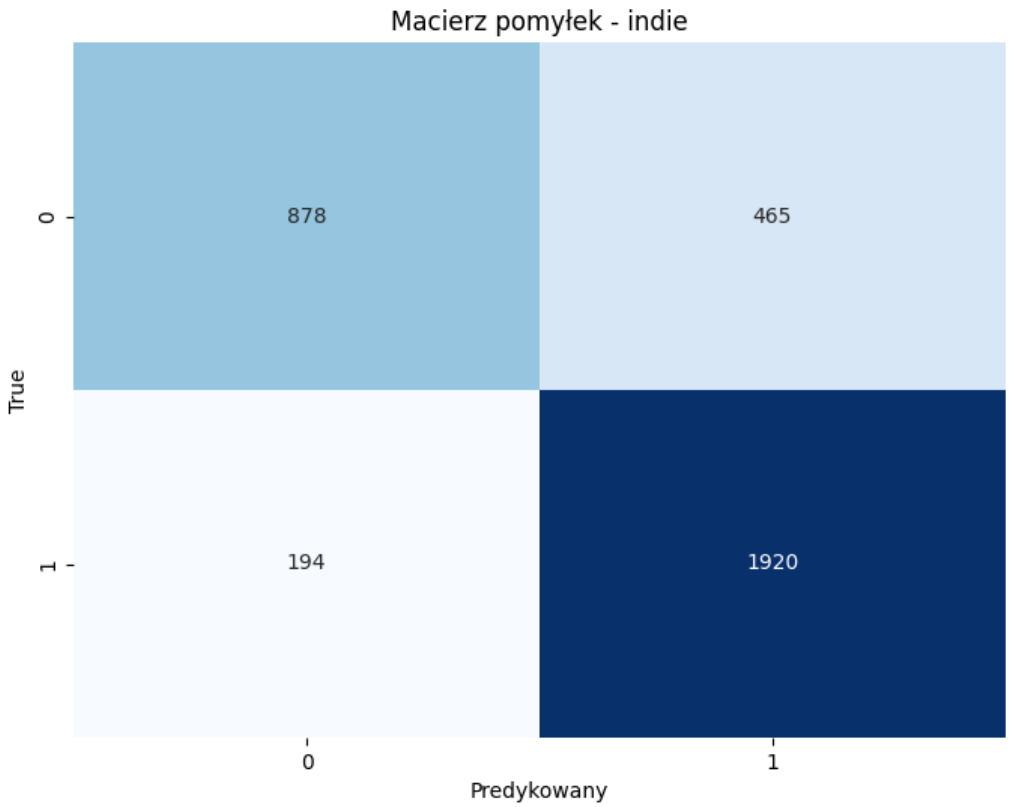
Czułość: 0.81

F1-Score: 0.78

Obserwacja:

Dla tego modelu również widać tendencję do zakładania, że gra nie jest „casualowa”, dzięki czemu mamy mało przypadków False Positive. Dokładność 0.81 w połączeni z niskim stosunkiem faktycznie rozpoznanych gier nie daje dużych nadziei na dalszą pracę.

Klasyfikacja „Czy gra należy do gatunku „indie””:



Rysunek 8 - macierz pomyłek gatunku gier „indie”, 0 - gra nie należy do gatunku, 1 - gra należy do gatunku

Klasyfikator gatunkowy gier ”indie”:

Dokładność: 0.81

Precyzja: 0.81

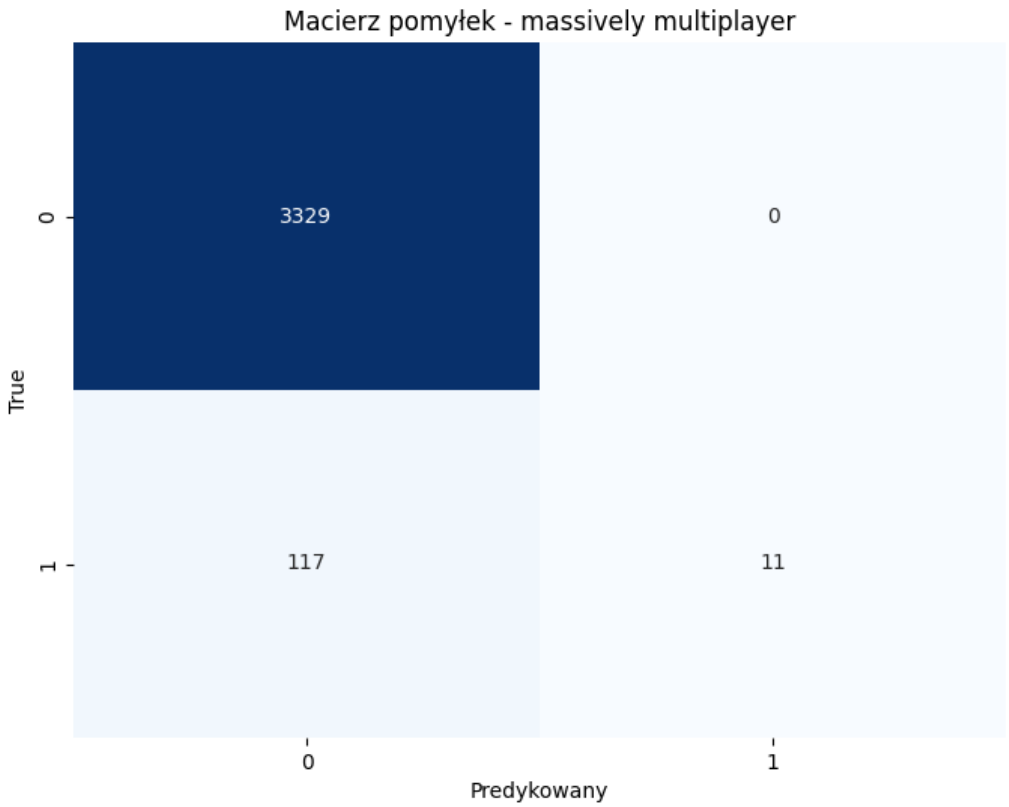
Czułość: 0.81

F1-Score: 0.80

Obserwacja:

Dla tego klasyfikatora macierz pomyłek „rozkłada” się w miarę równo, co może oznaczać, że model faktycznie zaczął „wyłapywać” jakąś zasadę, oraz ją aplikować. Razem z dokładnością na poziomie 0.81 tworzy to dobrze rokujący model, gotowy na poprawy w dalszych iteracjach.

Klasyfikacja „Czy gra należy do gatunku „Massively Multiplayer”:



Rysunek 9 - macierz pomyłek gatunku gier ”Massively Multiplayer”, 0 - gra nie należy do gatunku, 1 - gra należy do gatunku

Klasyfikator gatunkowy gier ”Massively Multiplayer”:

Dokładność: 0.97

Precyzja: 0.97

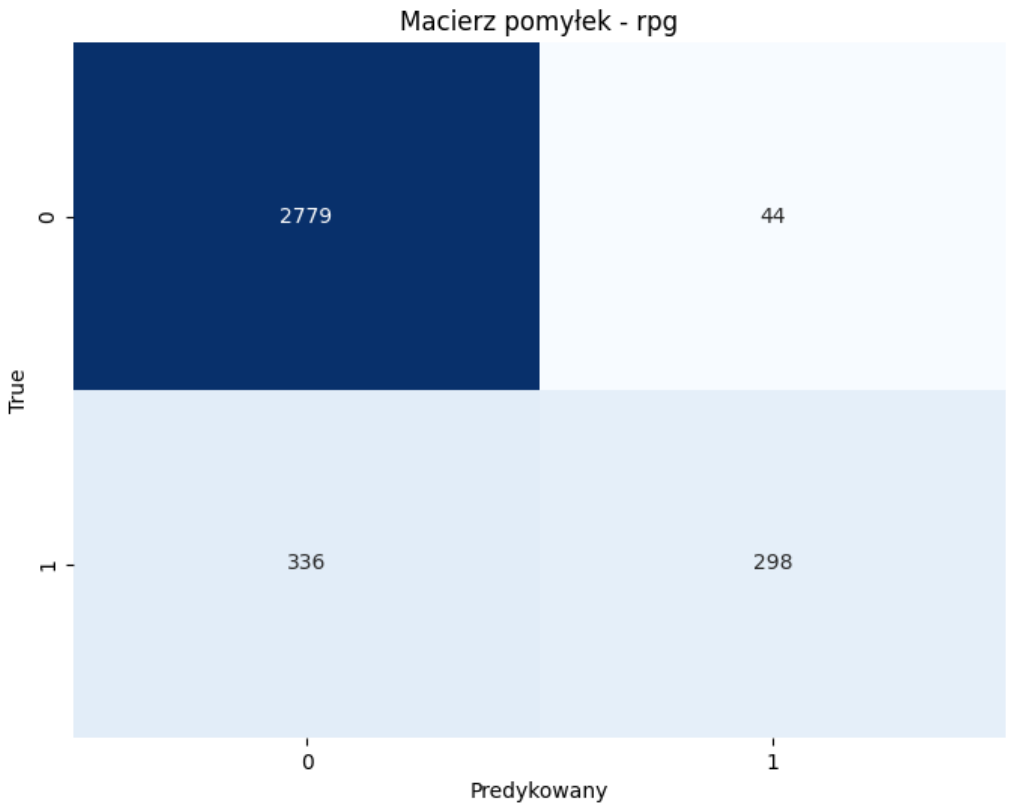
Czułość: 0.97

F1-Score: 0.95

Oberwacja:

Kolejny przypadek niezbalansowanej macierzy wskazującej na dużą tendencję do zakładania, że coś nie jest grą MMO. Sprawia to, że pomimo ogromnej dokładności model ten nie jest wiarygodny ze względu na ogromną liczbę False Negative.

Klasyfikacja „Czy gra należy do gatunku RPG”:



Rysunek 10 - macierz pomyłek gatunku gier RPG, 0 - gra nie należy do gatunku, 1 - gra należy do gatunku

Klasyfikator gatunkowy gier RPG:

Dokładność: 0.89

Precyzja: 0.89

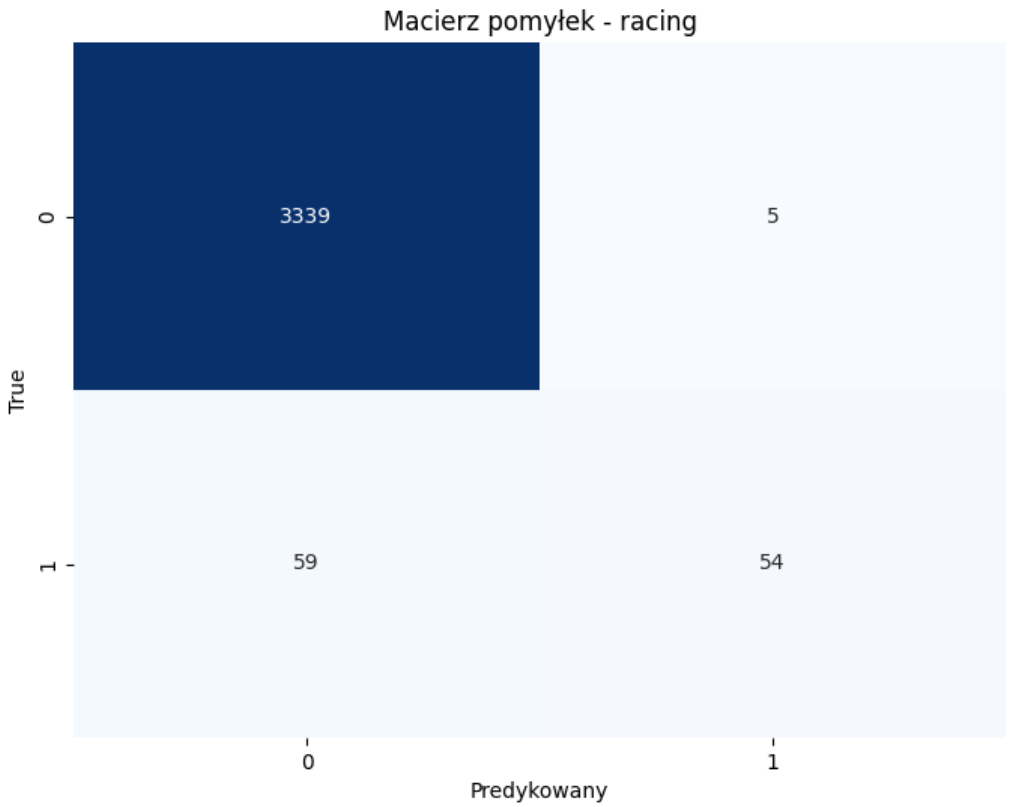
Czułość: 0.89

F1-Score: 0.88

Obserwacja:

Macierz pomyłek dla modelu rozpoznawania gatunku RPG również wskazuje na to, że model ma problemy ze wskazaniem co faktycznie jest grą RPG. Dużo lepiej radzi sobie ze wskazywaniem co nią nie jest. Pomimo dużej dokładności nie jest to wiarygodny model.

Klasyfikacja „Czy gra należy do gatunku gier wyścigowych”:



Rysunek 11 - macierz pomyłek gatunku gier wyścigowych, 0 - gra nie należy do gatunku, 1 - gra należy do gatunku

Klasyfikator gatunkowy gier wyścigowych:

Dokładność: 0.98

Precyzja: 0.98

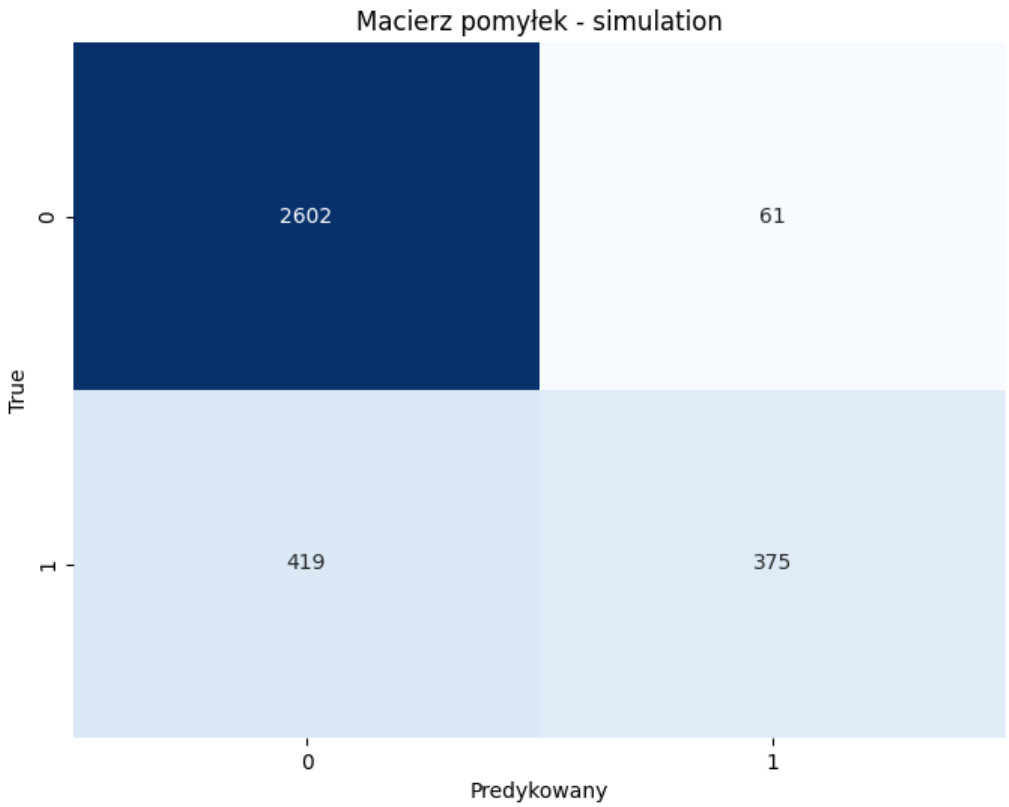
Czułość: 0.98

F1-Score: 0.98

Obserwacja:

Tutaj również model jest wyraźnie niezbalansowany. Ma on trudności ze wskazaniem co faktycznie jest grą wyścigową, co sprawia, że mimo bardzo wysokiej dokładności jest on mało użyteczny.

Klasyfikacja „Czy gra należy do gatunku gier symulacyjnych”:



Rysunek 12 - macierz pomyłek gatunku gier symulacyjnych, 0 - gra nie należy do gatunku, 1 - gra należy do gatunku

Klasyfikator gatunkowy gier symulacyjnych:

Dokładność: 0.86

Precyzja: 0.86

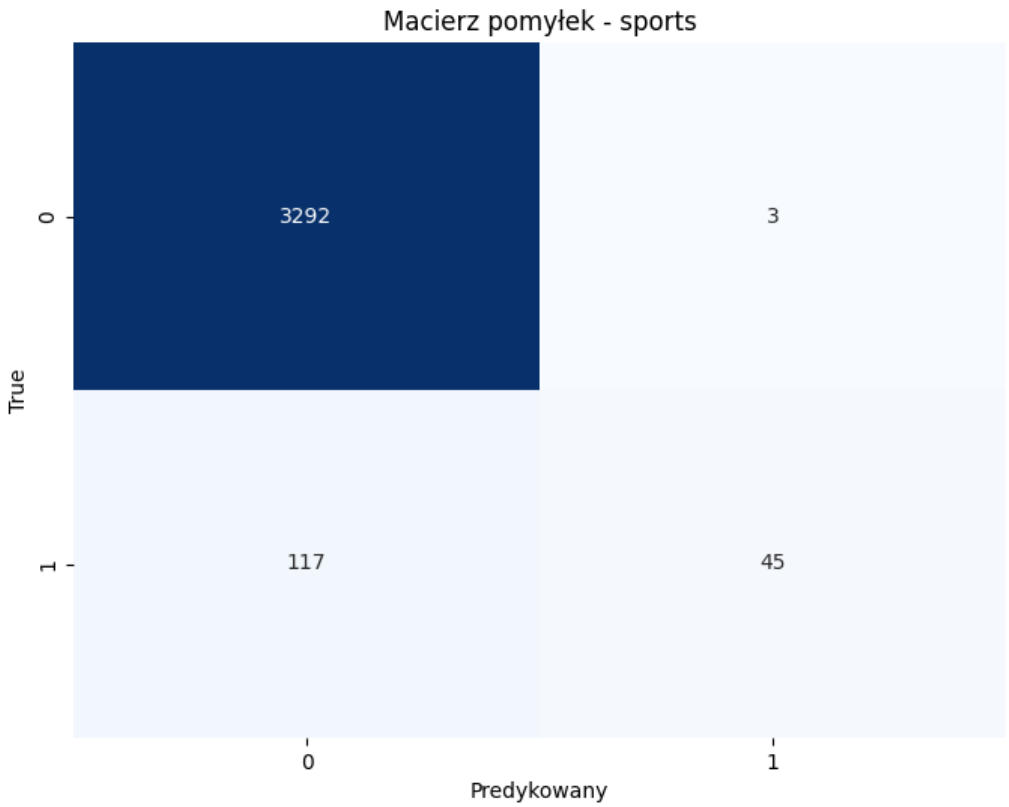
Czułość: 0.86

F1-Score: 0.85

Obserwacja:

Tutaj model również wydaje się być niezbalansowany, jednak widać tendencję do większej ilości False Negative od bratnich „niezbalansowanych” modeli. Po raz kolejny wysoka dokładność traci na wartości za okazaniem macierzy pomyłek.

Klasyfikacja „Czy gra należy do gatunku gier sportowych”:



Rysunek 13 – macierz pomyłek gatunku gier sportowych, 0 – gra nie należy do gatunku, 1 – gra należy do gatunku

Klasyfikator gatunkowy gier sportowych:

Dokładność: 0.97

Precyzja: 0.96

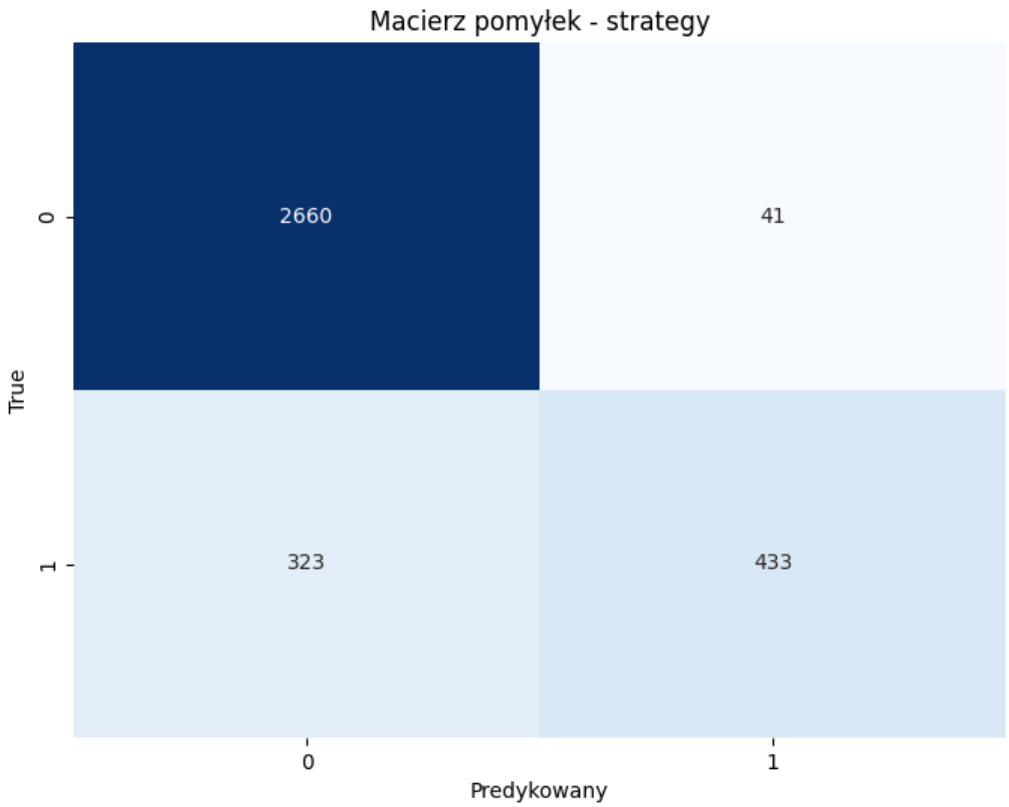
Czułość: 0.97

F1-Score: 0.96

Obserwacja:

Widać tutaj bardzo dużą tendencję do zakładania, że coś nie jest grą sportową, słabo radzi sobie z rozpoznawaniem faktycznych gier sportowych. Sprawia to, że mimo zadowalającej dokładności 0.97, model ten nie ma zbyt dużej wartości.

Klasyfikacja „Czy gra należy do gatunku gier strategicznych”:



Rysunek 14 – macierz pomyłek gatunku gier strategicznych, 0 – gra nie należy do gatunku, 1 – gra należy do gatunku

Klasyfikator gatunkowy gier strategicznych:

Dokładność: 0.90

Precyzja: 0.90

Czułość: 0.89

F1-Score: 0.89

Obserwacja:

Macierz pomyłek wskazuje na tendencję modelu do zakładania, że coś nie jest grą strategiczną, jednak na tle poprzednich modeli, widać tutaj potencjał ze względu na w miarę dobry stosunek skutecznie rozpoznanych gier strategicznych. Daje to nadzieje na rozwój tego aspektu modelu.

Wnioski:

Odpowiadając na postawione na początku pytanie:

„Jakie cechy gry można sklasyfikować na podstawie opisu?”

Zgodnie z oczekiwaniami, nie udało się stworzyć wiarygodnego modelu oceniającego jak „dobra” będzie gra na podstawie opisu.

Zawiodłem się jednak na wynikach innych modeli:

Ze względu na dużą „filtrację” danych na etapie czyszczenia i wprowadzania klas zostałem z stosunkowo małą próbką opisów do trenowania danych, przez co algorytmy NLP nie mogły „dojrzeć” i dostrzegać mało reprezentowane gatunki takie jak „sport”, czy gry „casualowe”.

Dzięki dużej reprezentacji, modele takie rozpoznawanie, czy gra należy do gatunku akcji lub Indie zdążyły się „rozwinąć” tworząc nieco bardziej zbalansowane macierze pomyłek.

Wybierając podejście do generowania modeli dałem zwieść się pułapce „dokładności” i dopiero macierz pomyłek pozwoliła mi zobaczyć, że modele z największą dokładności często są tymi najmniej „wyrafinowanymi”, szufladkującymi wszystko do jednej „najbardziej reprezentowanej” kategorii.

W sytuacjach nie binarnych mój wybrany model TF-IDF + SVM, miał tendencję do sprowadzania wszystkiego do 2 skrajności, pozostawiając kategorie pośrednie bez reprezentacji.

Pomimo tych pomyłek, uważam, że projekt ten był sukcesem pod względem edukacyjnym.

Zapoznałem się z podstawowym schematem działań przy pracy z modelami NLP. Dowiedziałem się o zastosowaniach tej dziedziny, podstawowych sposobach reprezentacji tekstu za postacią macierzy takich jak TF-ITF, czy BOW, oraz jak zaimplementować nowo poznaną wiedzę.