



**AGH**

**AKADEMIA GÓRNICZO-HUTNICZA IM. STANISŁAWA STASZICA W KRAKOWIE**

**Wydział Elektrotechniki, Automatyki, Informatyki i Inżynierii Biomedycznej**

Projekt dyplomowy

*Metody selekcji cech w analizie dokumentów za pomocą  
przetwarzania języka naturalnego*

*Feature selection algorithms in natural language  
processing of documents*

Autor:

*Norbert Sak*

Kierunek studiów:

*Automatyka i Robotyka*

Opiekun pracy:

*dr hab. inż. Jerzy Baranowski*

Kraków, 2022

## Spis treści

<b>1. Wprowadzenie</b>	5
1.1. Cele pracy	5
1.2. Zawartość pracy	6
<b>2. Analiza sentymentów i uczenie maszynowe</b>	7
2.1. Przetwarzanie języka naturalnego	7
2.2. Analiza sentymentu	7
2.3. Uczenie maszynowe	8
2.4. Algorytmy klasyfikacyjne	9
<b>3. Metody selekcji cech</b>	12
3.1. Selekcja cech	12
3.2. Podział metod selekcji cech	12
3.3. $\chi^2$	14
3.4. ANOVA	14
3.5. RFE	15
3.6. Regularyzacja L1	15
<b>4. Użyte technologie</b>	16
<b>5. Implementacja</b>	17
5.1. Projekt uczenia maszynowego	17
5.1.1. Gromadzenie danych	17
5.1.2. Preprocessing danych	18
5.1.3. Ekstrakcja cech	18
5.1.4. Selekcja cech	18
5.1.5. Klasyfikacja	18
5.1.6. Ewaluacja	18
5.2. Aplikacja webowa	19
<b>6. Testy</b>	20

---

6.1. Testy.....	20
6.2. Wyniki.....	20
<b>7. Podsumowanie.....</b>	<b>21</b>
<b>Bibliografia .....</b>	<b>22</b>

# 1. Wprowadzenie

W rozdziale przedstawiono cel pracy oraz opis zawartości poszczególnych rozdziałów.

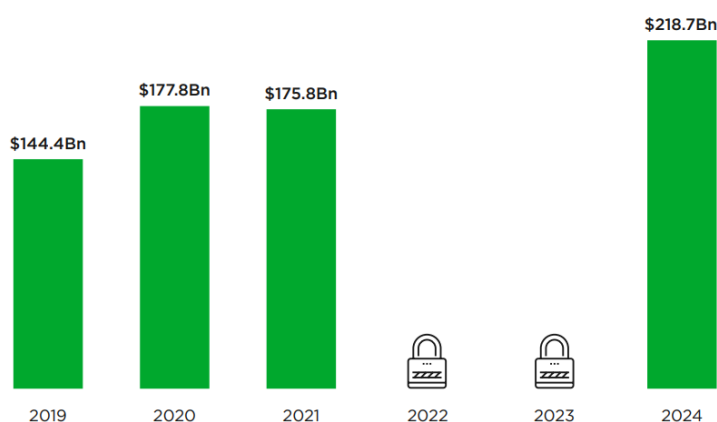
## 1.1. Cele pracy

Celem poniższej pracy jest porównanie różnych metod selekcji cech podczas procesu analizy sentymentu, czyli określenia nastawienia osoby piszącej przy pomocy przetwarzania języka naturalnego w uczeniu maszynowym.

W pracy jako zbiór danych zostały użyte recenzje gier w języku polskim napisane przez użytkowników jednej z największych platform dystrybucji cyfrowej gier na świecie, czyli Steam. Wybrany został taki typ danych, ponieważ rynek gier jest jednym z najbardziej rosnących na świecie. Wartość rynku gier wynosi obecnie około 175 miliardów dolarów i rośnie z roku na roku na rok. Przewiduje się, że w 2024 roku osiągnie wartość bliską 220 miliardów dolarów, co można zaobserwować na rysunku 1.1. Również liczba graczy na świecie rośnie w wysokim tempie. W Europie w 2021 roku wyniosła ona 408 milionów osób, co przekłada się na 4% wzrost w stosunku do roku poprzedniego co wynika z rysunku 1.2.

Global Game Revenue Forecast

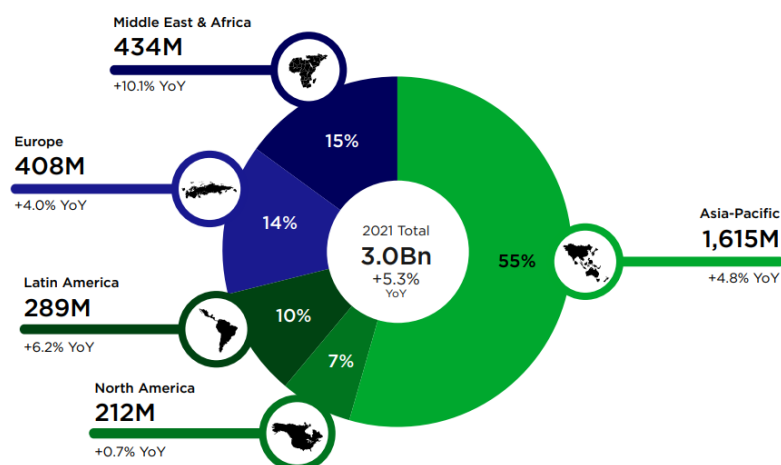
Toward 2024



**Rys. 1.1.** Wzrost wartości rynku gier na świecie w latach 2019-2024 [1].

## 2021 Global Players

Per Region



Rys. 1.2. Liczba graczy na świecie z podziałem na regiony w 2021 roku [1].

Proces analizy sentymentu jest bardzo istotny, szczególnie w dzisiejszych czasach, gdzie każda firma stara się jak najbardziej polepszać jakość swoich usług i produktów. Z tego właśnie powodu w poniższej pracy zostały przeanalizowane różne metody selekcji cech za pomocą przetwarzania języka naturalnego w celu usprawnienia całego procesu analizy sentymentu.

Finalnym efektem poniższej pracy jest porównanie wyników klasyfikacji sentymentu recenzji gier dla różnych metod selekcji cech i wskazanie metody optymalnej, czyli tej, która osiągnie najlepszy wynik dla zbioru testowego oraz stworzenie aplikacji webowej z zaimplementowanym najlepszym modelem.

## 1.2. Zawartość pracy

...[do uzupełnienia na końcu]

## **2. Analiza sentymentów i uczenie maszynowe**

W rozdziale tym zostanie omówiony proces przetwarzania języka naturalnego oraz analizy sentymentu. Zostanie także omówione uczenie maszynowe i klasyfikatory użyte w projekcie.

### **2.1. Przetwarzanie języka naturalnego**

Przetwarzanie języka naturalnego (ang. Natural Language Processing, NLP) jest to dziedzina sztucznej inteligencji, która daje maszynom możliwość wydobywania informacji z ludzkich języków. Każda wypowiedź ludzka zawiera bardzo dużą dawkę informacji, a dzięki uczeniu maszynowemu jesteśmy w stanie przekazać te informacje komputerom i wykorzystać je do analizy danych. Przetwarzanie języka naturalnego bardzo szybko się rozwija między innymi dzięki poprawie dostępu do danych oraz zwiększaniu się mocy obliczeniowej. NLP znajduje swoje zastosowanie między innymi w obszarach takich jak finanse, media czy opieka zdrowotna [2]. Przykładowe zastosowania przetwarzania języka naturalnego to:

- pomoc w wykrywaniu fałszywych wiadomości publikowanych w mediach,
- detekcja spamu oraz podejrzanych wiadomości mailowych,
- chat-boty pomagające klientom w sklepach internetowych,
- zbieranie informacji co klienci sądzą o produkcie, usłudze badając informacje ze źródeł takich jak media społecznościowe poprzez analizę sentymentu.

### **2.2. Analiza sentymentu**

Dzięki pojawieniu się Web 2.0 (serwisów internetowych, w których główną rolę odgrywa treść generowana przez użytkowników), użytkownicy mogą dzielić się swoimi opiniami i odczuciami na różne tematy w nowych formach, w których ludzie nie są tylko biernymi odbiorcami informacji. Termin analiza sentymenty (ang. sentiment analysis) jest używany w odniesieniu do automatycznego przetwarzania opinii i nastrojów. Analiza sentymentu jest bardzo istotną

dziedziną badań w przetwarzaniu języka naturalnego, ponieważ w połączeniu z uczeniem maszynowym pozwala nam na wydobycie opinii i nastroju z zestawu dokumentów [3]. Analiza sentymentu może być badana na różnych poziomach:

- analiza na poziomie całego dokumentu,
- analiza na poziomie jednego zdania,
- analiza na poziomie aspektu.

W tej pracy poruszony został problem analizy sentymentu na poziomie całego dokumentu, czyli ustalenie całościowej opinii, nastroju w dokumencie. Proces ten niesie ze sobą jednak wiele wyzwań i problemów, a niektóre z nich to:

- kontekstowa niejednoznaczność polaryzacji (opinia zawarta w zdaniu często zależy od kontekstu),
- wykrywanie sarkazmu,
- obsługa negacji.

## **2.3. Uczenie maszynowe**

Uczenie maszynowe to proces, w którym program komputerowy lub system jest w stanie uczyć się i stawać coraz mądrzejszy. Na bardzo podstawowym poziomie uczenie maszynowe wykorzystuje algorytmy do znajdowania wzorców. Innymi słowami jest to proces komputerowego modelowania ludzkiej inteligencji, który z czasem ulega automatycznej poprawie. Maszyny są w stanie przewidywać na podstawie tego, co zaobserwowały i nauczyły się [4]. Uczenie maszynowe możemy podzielić na trzy typy:

- uczenie nadzorowane - zbiór danych dostarczonych maszynie do uczenia zawiera również wyjście (etykiety),
- uczenie nienadzorowane - dostarczone dane posiadają jedynie dane wejściowe (nie zawierają etykiety),
- uczenie przez wzmacnianie - system działa w środowisku zupełnie mu nie znanym, brak zarówno danych wejściowych jak i wyjściowych.

Natomiast algorytmy uczenia maszynowego możemy podzielić na trzy główne rodzaje. Algorytmy klasyfikacyjne, czyli takie, które pozwalają przypisać dane do odpowiednich kategorii (na przykład określenie rasy psa na podstawie różnych jego cech). Kolejnym typem są algorytmy regresyjne, w których zmienna wyjściowa przyjmuje wartości ciągłe (na przykład

przewidywanie ceny domu). Istnieją także algorytmy grupujące, czyli takie, których zadaniem jest podzielenie danych na grupy na podstawie ich podobieństw (na przykład podział klientów sklepu na grupy).

W pracy tej użyte zostało uczenie nadzorowane wraz z algorytmami klasyfikacyjnymi. Projekt uczenia maszynowego zawierający przetwarzanie języka naturalnego zwykle składa się z następujących kroków.

1. Zebranie danych.
2. Preprocessing danych.
3. Ekstrakcja cech.
4. Selekcja cech.
5. Uczenie modelu.
6. Klasyfikacja
7. Ewaluacja wyników.

Wymienione powyżej etapy zostały dokładnie omówione w rozdziale 5.

## 2.4. Algorytmy klasyfikacyjne

W poniższej pracy użyte zostały następujące algorytmy klasyfikacyjne:

- naiwny klasyfikator Bayesa,
- maszyna wektorów nośnych,
- regresja logistyczna,
- las losowy.

Naiwny klasyfikator Bayesa (ang. Naive Bayes Classifier) to probabilistyczny model uczenia maszynowego używany do problemu klasyfikacji. Opiera się on na twierdzeniu Bayesa:

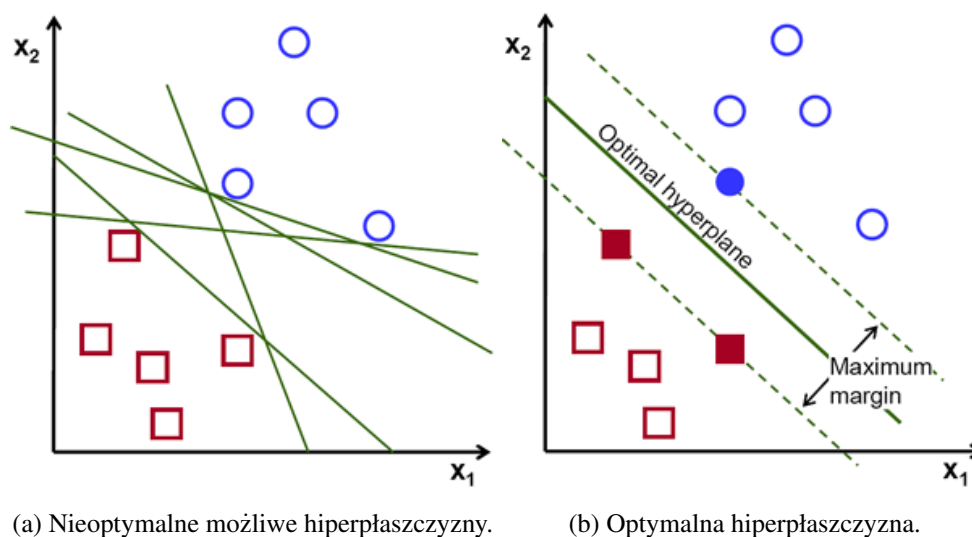
$$P(y|X) = \frac{P(X|y)P(y)}{P(X)}, \quad (2.1)$$

gdzie  $y$  to klasa którą przewidujemy, a  $X$  reprezentuje poszczególne cechy. Korzystając z powyższego twierdzenia, możemy znaleźć prawdopodobieństwo wystąpienia klasy  $y$ , zakładając, że wystąpiła cecha  $X$ . Przyjmuje się tutaj założenie, że cechy są niezależne (obecność



jednej nie wpływa na obecność drugiej). Dlatego właśnie klasyfikator ten nazywany jest naiwnym [5].

Maszyna wektorów nośnych (ang. Support Vector Machine, SVM) jest to algorytm uczenia maszynowego, którego celem jest znalezienie hiperpłaszczyzny w przestrzeni  $n$ -wymiarowej, gdzie  $n$  to liczba cech, która wyraźnie klasyfikuje punkty danych [6].



**Rys. 2.1.** Możliwe hiperpłaszczyzny [6].

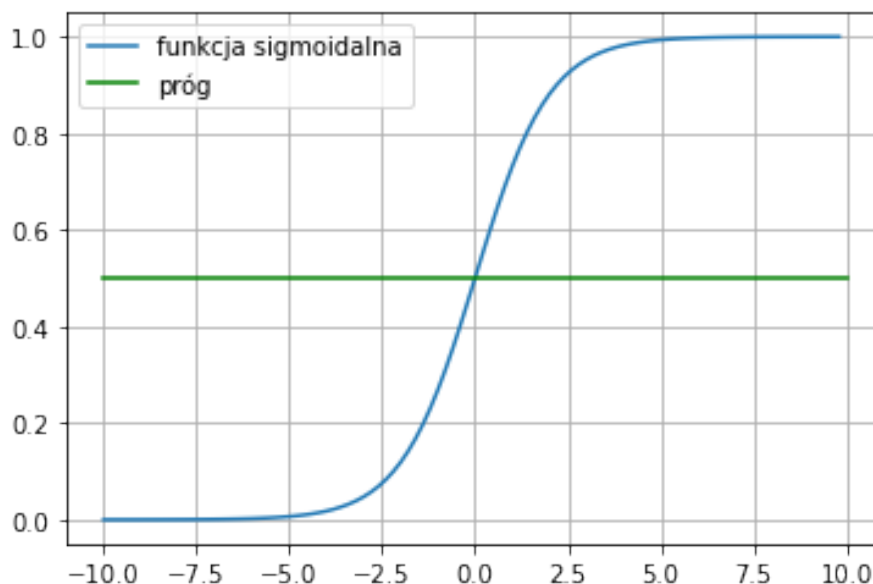
Na rysunku 2.1 można zauważyć przykład wyznaczania hiperpłaszczyzny dla dwóch cech  $x_1$  i  $x_2$ . Celem jest znalezienie takiej płaszczyzny, która ma maksymalny margines (ang. maximum margin), czyli maksymalną odległość pomiędzy punktami z obu klas. Dzięki tej maksymalizacji przyszłe punkty będą klasyfikowane z większą pewnością.

Regresja logistyczna (ang. logistic regression) jest algorytmem klasyfikacji używanym do przypisania obserwacji do dyskretnego zestawu klas. W przeciwieństwie do regresji liniowej, która generuje ciągłe wartości, regresja logistyczna przekształca swoje dane wyjściowe za pomocą logistycznej funkcji sigmoidalnej, w celu otrzymania prawdopodobieństwa, z którego następnie możemy określić do której z klas przynależy dana obserwacja [7]. Funkcja sigmoidalna określona jest następującym wzorem:

$$f(z) = \frac{1}{1 + e^{-z}}, \quad (2.2)$$

gdzie  $f(z)$  to rezultat z zakresu od 0 do 1 (oszacowane prawdopodobieństwo),  $z$  to dane wejściowe do funkcji (na przykład funkcja liniowa), a  $e$  to podstawa logarytmu naturalnego. Dla problemu binarnego ustalany próg (ang. threshold) z zakresu od 0 do 1. Jeśli wartość

funkcji sigmoidalnej jest większa od progu to przewidywana klasą jest 1, a jeśli jest mniejsza to 0. Funkcja  $f(z)$  została przedstawiona na rysunku 2.2 kolorem niebieskim, natomiast kolorem zielonym zaznaczony został próg.



**Rys. 2.2.** Funkcja sigmoidalna.

Las losowy (ang. random forest) składa się z dużej liczby pojedynczych drzew decyzyjnych, które działają jako zespół. Każde pojedyncze drzewo dokonuje predykcji, a klasa z największą liczbą głosów jest prognozą modelu. Drzewa chronią się nawzajem przed indywidualnymi błędami, ponieważ, gdy kilka drzew się pomyli wiele innych może mieć racje. Kluczem jest niska korelacja pomiędzy drzewami [8].

## 3. Metody selekcji cech

W poniższym rozdziale został przedstawiony problem selekcji cech, podział oraz omówione poszczególne metody selekcji cech wykorzystane w projekcie.

### 3.1. Selekcja cech

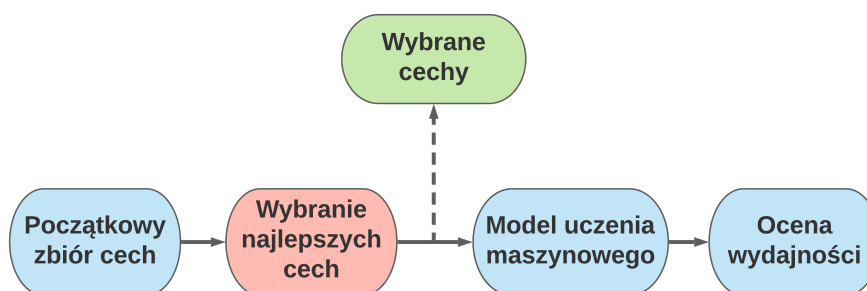
Cecha jest to indywidualna, mierzalna właściwość obserwowanego procesu. Algorytmy uczenia maszynowego wykorzystują zbiór cech, aby dokonać klasyfikacji. Selekcja cech pomaga w:

- zredukowaniu złożoności modelu, przez co jest on łatwiejszy w interpretacji,
- zmniejszeniu czasu uczenia modelu,
- uniknięcie zjawiska nadmiernego dopasowania (ang. overfitting),
- poprawy wydajności klasyfikatora, gdy zostały wybrane właściwe cechy.

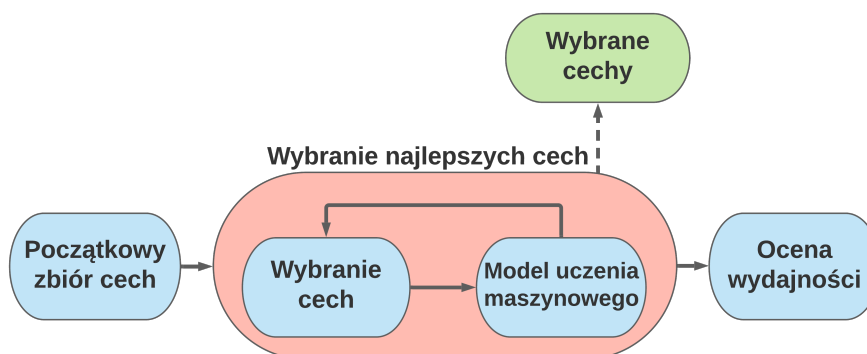
Celem selekcji cech jest wybranie, z danych wejściowych, podzbioru cech, które skutecznie opisują te dane i jednocześnie redukują nieistotne cechy zapewniając porównywalnie dobre wyniki predykcji. Aby usunąć nieistotne cechy, potrzebne jest kryterium wyboru, które jest w stanie zmierzyć powiązanie każdej cechy z klasą wyjściową. Z perspektywy uczenia maszynowego jeśli model używa mało znaczących, to wykorzysta je również dla nowych danych, co prowadzi do zbytniego uogólnienia modelu [9].

### 3.2. Podział metod selekcji cech

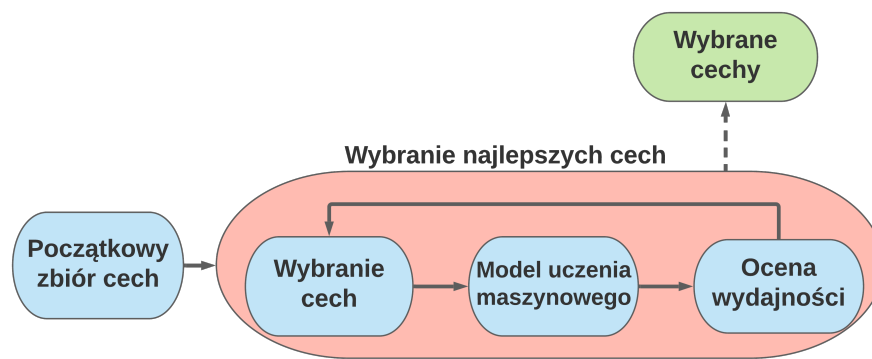
Metody selekcji cech możemy podzielić na trzy główne grupy. Pierwszą z nich są filtry, które autonomicznie podejmują decyzje na temat wybranych cech. Często korzystają z różnych metod statystycznych. Dla każdej z cech obliczany jest pewien współczynnik na podstawie danego kryterium i dla ustalonego progu wybierana jest odpowiednia liczba cech. Metody te są niezależne od klasyfikatora, szybkie i względnie proste [10, 11]. Graficzna wizualizacja metod filtrowych została przedstawiona na Rys. 3.1.

**Rys. 3.1.** Selekcja cech z użyciem filtrów.

Kolejnym typem metod selekcji cech są wrappery, czyli metody opakowane. Wykorzystują one sprzężenie zwrotne pomiędzy modelem uczenia maszynowego, a elementem decyzyjnym. Wybieramy podzbiór cech i uczymy model z ich wykorzystaniem, a następnie korzystając z wyników dodajemy lub odejmujemy cechy z danego podzbioru. Metody te są zazwyczaj bardzo kosztowne obliczeniowo [10, 11]. Graficzna wizualizacja wrapperów została przedstawiona na Rys. 3.2.

**Rys. 3.2.** Selekcja cech z użyciem wrapperów.

Do metod selekcji cech należą również metody wbudowane. Są one zaimplementowane za pomocą algorytmów, które posiadają wbudowane metody selekcji cech. Selekcja dokonywana jest na etapie uczenia modelu. Jedną z metod wbudowanych jest regularyzacja wykorzystująca funkcje celu, które minimalizują błędy dopasowania a także obniżają wagę nieistotnych cech. Elementy z wagą bliską zeru są eliminowane [10, 11]. Graficzna wizualizacja metod wbudowanych została przedstawiona na Rys. 3.3.



**Rys. 3.3.** Selekcja cech z użyciem metod wbudowanych.

### 3.3. $\chi^2$

Metoda  $\chi^2$  jest statystyczną metodą filtrową, która mierzy powiązanie pomiędzy cechą a jej klasą. Określa jak bardzo zależne są od siebie dana cecha i klasa. Im bardziej zależna jest ich relacja, tym bardziej dana cecha jest użyteczna w predykcji danej klasy. Na poniższym wzorze A to liczba wystąpień f (ang. feature - cecha) i c (ang. class - klasa) razem, B to liczba wystąpień f bez c, C to liczba wystąpień c bez f, D to liczba przypadków, gdy f i c nie występują a N to całkowita liczba analizowanych dokumentów [12].

$$X^2(f, c) = \frac{N \times (AD - CB)^2}{(A + C) \times (B + D) \times (A + B) \times (C + D)} \quad (3.1)$$

$X^2$  jest wartością znormalizowaną, której wartość można porównywać z z cechami z tej samej kategorii. Jednak, gdy cecha występuje z niską częstotliwością w danym dokumencie to wspomniana normalizacja przestaje występować. Stąd możemy wywnioskować, iż ta metoda nie jest odpowiednia dla cech o niskiej częstotliwości. Jednak pomimo tego metoda  $\chi^2$  jest uważana za jedną z dwóch najlepszych filtrowych metod selekcji cech wraz z metodą Information Gain [12].

### 3.4. ANOVA

Analiza wariancji (ang. analysis of variance, ANOVA) jest kolejną z metod filtrowych. Polega na sprawdzeniu istotności różnic pomiędzy dwiema lub większej liczbie próbek. Metodę tę można podsumować jako obliczenie wielu średnich oraz wariancji i finalne określenie istotności ich stosunku. F-test służy do porównywania czynników sumy odchylenia, a formuła wygląda następująco:

$$F_{\text{values}} = \frac{MSG}{MSE} = \frac{SSG/df_g}{SSE/df_e}, \quad (3.2)$$

gdzie MSG to różnica pomiędzy zabiegami, MSE to wariancja w obrębie zabiegów, SSG oznacza grupę sumy kwadratów, SSE sumę błędu kwadratowego, a  $df_g$  i  $df_e$  są ich stopniami swobody [13].

### 3.5. RFE

Rekurencyjna eliminacja cech (ang. recursive feature elimination, RFE) jest wrapperem i polega na użyciu zewnętrznego estymatora, który każdej z cech przypisuje wagę (na przykład współczynniki modelu liniowego). Celem rekurencyjnej jest wybranie cech poprzez rekurencyjne uwzględnienie coraz mniejszych podzbiorów cech. Estymator szkolony jest na początkowym zestawie cech, a waga każdej z cech uzyskiwana jest poprzez dowolny określony atrybut. Następnie najmniej istotne cechy są usuwane, aż osiągniemy pożądaną liczbę cech [14].

### 3.6. Regularyzacja L1

Regularyzacja L1 jest wbudowaną metodą selekcji cech, która korzysta z regresji Lasso, czyli modelu liniowego wykorzystującego następującą funkcję kosztu:

$$\frac{1}{2N} \sum_{i=1}^N (y_{\text{real}}^{(i)} - y_{\text{pred}}^{(i)})^2 + \alpha \sum_{j=1}^N |a_j|, \quad (3.3)$$

gdzie  $N$  jest liczebnością zbioru treningowego,  $y_{\text{real}}^{(i)}$  i  $y_{\text{pred}}^{(i)}$  to odpowiednio klasa prawdziwa i klasa przewidziana przez model. Ostatni składnik powyższej funkcji nazywany jest karą L1,  $\alpha$  jest hiperparametrem, który pozwala dostosować intensywność kary, a  $a_j$  jest współczynnikiem  $j$ -tej cechy. Im wyższy współczynnik cechy, tym wyższa wartość funkcji kosztu [15].

Podczas selekcji cech regresja Lasso, w celu zminimalizowania funkcji kosztu, automatycznie wybierze przydatne cechy, a zbędne wyeliminuje. Odrzucenie obiektu powoduje, że jego współczynnik będzie równy zero. Regresja Lasso jest dopasowywana do zbioru danych i brane są pod uwagę tylko cechy, które mają współczynnik różny od zera. Oczywiście do poprawnego działania selekcji wymagane jest odpowiednie dostrojenie parametru  $\alpha$ .

## **4. Użyte technologie**

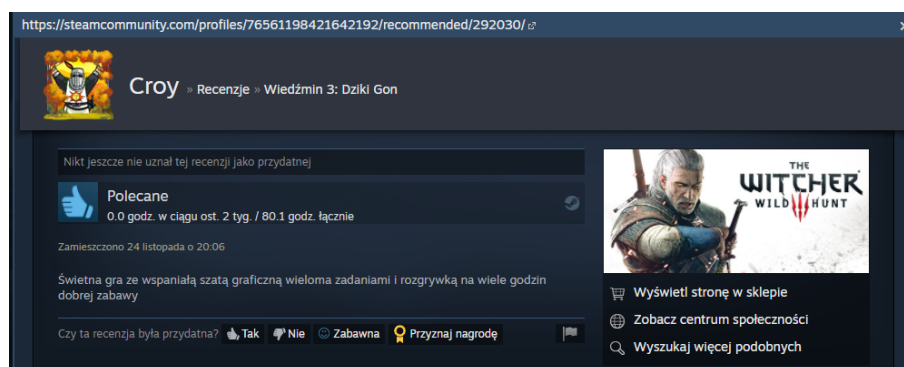
## 5. Implementacja

W poniższym rozdziale zostały przedstawione kolejne kroki implementacji realizowanego projektu.

### 5.1. Projekt uczenia maszynowego

#### 5.1.1. Gromadzenie danych

Dane, czyli recenzje gier użytkowników platformy Steam zostały zebrane przy pomocy API powyższej platformy. Wybrane zostały recenzje 25 aktualnie najbardziej popularnych gier w języku polskim z kategorii RPG (ang. role-playing game), czyli gier fabularnych. Recenzje zostały zawężone do jednej kategorii aby użyte w recenzjach słownictwo nie różniło od siebie w znaczący sposób. Na rysunku 5.1 można zaobserwować przykładowy wygląd recenzji z platformy Steam. Pod uwagę bierzemy jedynie treść recenzji oraz etykietę, czyli informację czy gra jest polecana przez użytkownika.



Rys. 5.1. Przykładowa recenzja z platformy Steam. [16]

Do analizy zostały wyselekcjonowane recenzje o długości minimum 100 znaków, aby wyeliminować opinie krótkie, których treść często jest niezwiązana z wystawioną oceną i posiadają wyłącznie charakter żartobliwy co mogłoby negatywnie wpłynąć na jakość modelu uczenia maszynowego.



### **5.1.2. Preprocessing danych**

### **5.1.3. Ekstrakcja cech**

### **5.1.4. Selekcja cech**

### **5.1.5. Klasyfikacja**

### **5.1.6. Ewaluacja**

## 5.2. Aplikacja webowa

## **6. Testy**

### **6.1. Testy**

### **6.2. Wyniki**

## **7. Podsumowanie**

## Bibliografia

- [1] Newzoo. „*Newzoo Global Games Market Report 2021 | Free Version*”. <https://newzoo.com/insights/trend-reports/newzoo-global-games-market-report-2021-free-version>. 2021.
- [2] Diego Lopez Ys. „*Your Guide to Natural Language Processing (NLP)*”. 2019. URL: <https://towardsdatascience.com/your-guide-to-natural-language-processing-nlp-48ea2511f6e1> (term. wiz. 2021).
- [3] Salima Behdenna, Fatiha Barigou i Ghalem Belalem. „*Document Level Sentiment Analysis: A survey*”. W: *EAI Endorsed Transactions on Context-aware Systems and Applications* (2018).
- [4] WGU. „*Machine learning: definition, explanation, and examples*.” URL: <https://www.wgu.edu/blog/machine-learning-definition-explanation-examples2007> (term. wiz. 2021).
- [5] Rohith Gandhi. „*Naive Bayes Classifier*”. 2018. URL: <https://towardsdatascience.com/naive-bayes-classifier-81d512f50a7c> (term. wiz. 2021).
- [6] Rohith Gandhi. „*Support Vector Machine — Introduction to Machine Learning Algorithms*”. 2018. URL: <https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47> (term. wiz. 2021).
- [7] ML Glossary documentation. „*Logistic Regression*”. URL: [https://ml-cheatsheet.readthedocs.io/en/latest/logistic\\_regression.html](https://ml-cheatsheet.readthedocs.io/en/latest/logistic_regression.html) (term. wiz. 2021).
- [8] Tony Yiu. „*Understanding Random Forest*”. 2019. URL: <https://towardsdatascience.com/understanding-random-forest-58381e0602d2> (term. wiz. 2021).
- [9] Ferat Sahin. „*A survey on feature selection methods*”. W: *Computers and Electrical Engineering* (2014), s. 16–28.
- [10] Lai Hung, Rayner Alfred i Mohd Hijazi. „*A Review on Feature Selection Methods for Sentiment Analysis*”. W: *Advanced Science Letters* (2015), s. 2952–2956.

- [11] Sauravkaushik8 Kaushik. „*Introduction to Feature Selection methods with an example (or how to select the right variables?)*” URL: <https://www.analyticsvidhya.com/blog/2016/12/introduction-to-feature-selection-methods-with-an-example-or-how-to-select-the-right-variables> (term. wiz. 2021).
- [12] Lai Hung, Rayner Alfred i Mohd Hijazi. „*A Review on Feature Selection Methods for Sentiment Analysis*”. W: *Advanced Science Letters* (2015), s. 2952–2956.
- [13] Shida He i in. „*MRMD2.0: A Python tool for machine learning features ranking and reduction*”. W: *Current Bioinformatics* (2020).
- [14] scikit-learn. „*sklearn.feature\_selection.RFE*”. URL: [https://scikit-learn.org/stable/modules/generated/sklearn.feature\\_selection.RFE](https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.RFE) (term. wiz. 2021).
- [15] Gianluca Malato. „*Feature selection in machine learning using Lasso regression*”. URL: <https://towardsdatascience.com/feature-selection-in-machine-learning-using-lasso-regression-7809c7c2771a> (term. wiz. 2021).
- [16] Steam. „*Wiedźmin 3: Dziki Gon - recenzje*”. URL: [https://store.steampowered.com/app/292030/Wiedmin\\_3\\_Dziki\\_Gon/?l=polish](https://store.steampowered.com/app/292030/Wiedmin_3_Dziki_Gon/?l=polish) (term. wiz. 2021).