

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

Praca dyplomowa

*Rozpoznawanie tzw. „fake newsów”*

*na wąskim obszarze tematycznym*

*Recognition of so called „fake news”*

*on specified topic scope*

Autor: *Konrad Daniel Golemo*

Kierunek studiów: *Automatyka i Robotyka*

Opiekun pracy: *dr hab. inż. Jerzy Baranowski*

Kraków, *2022*

**Spis treści**

[1. Wstęp 3](#_Toc92380753)

[1.1. Wprowadzenie 3](#_Toc92380754)

[1.2. Cel i zakres pracy 6](#_Toc92380755)

[1.3. Układ pracy 6](#_Toc92380756)

[2. Analiza aktualnego stanu wiedzy 7](#_Toc92380757)

[2.1. Uczenie maszynowe 7](#_Toc92380758)

[2.2. Przetwarzanie języka naturalnego 7](#_Toc92380759)

[2.3. Wyniki realizacji podobnych projektów 9](#_Toc92380760)

[2.3.1. Detekcja tzw. „fake newsów” w języku angielskim 9](#_Toc92380761)

[2.3.2. Detekcja tzw. „fake newsów” w języku niemieckim 11](#_Toc92380762)

[2.3.3. Detekcja tzw. „fake newsów” w języku hiszpańskim 12](#_Toc92380763)

[3. Realizacja projektu 13](#_Toc92380764)

[3.1. Wybór obszaru tematycznego i tworzenie zbiór danych 13](#_Toc92380765)

[3.2. Przetwarzanie zbioru danych 16](#_Toc92380766)

[3.2.1. Oczyszczanie danych 16](#_Toc92380767)

[3.2.2. Ekstrakcja i selekcja cech tekstu 16](#_Toc92380768)

[3.3. Wybrane algorytmy klasyfikacji 17](#_Toc92380769)

[3.3.1. Naiwny klasyfikator bayesowski 17](#_Toc92380770)

[3.3.2. Liniowa maszyna wektorów nośnych 18](#_Toc92380771)

[3.3.3. Las losowy 19](#_Toc92380772)

[3.4. Kompletny model klasyfikatora 20](#_Toc92380773)

[3.5. Aplikacja webowa 20](#_Toc92380774)

[4. Rezultaty 21](#_Toc92380775)

[4.1. Testy jednostkowe funkcji 21](#_Toc92380776)

[4.2. Ewaluacja algorytmów 22](#_Toc92380777)

[4.2.1. Naiwny klasyfikator bayesowski 22](#_Toc92380778)

[4.2.2. Liniowa maszyna wektorów nośnych 23](#_Toc92380779)

[4.2.3. Las losowy 24](#_Toc92380780)

[5. Podsumowanie 26](#_Toc92380781)

[5.1. Zrealizowane założenia projektowe 26](#_Toc92380782)

[5.2. Wnioski 26](#_Toc92380783)

[5.3. Kierunki dalszego rozwoju 26](#_Toc92380784)

[6. Bibliografia 27](#_Toc92380785)

# Wstęp

## Wprowadzenie

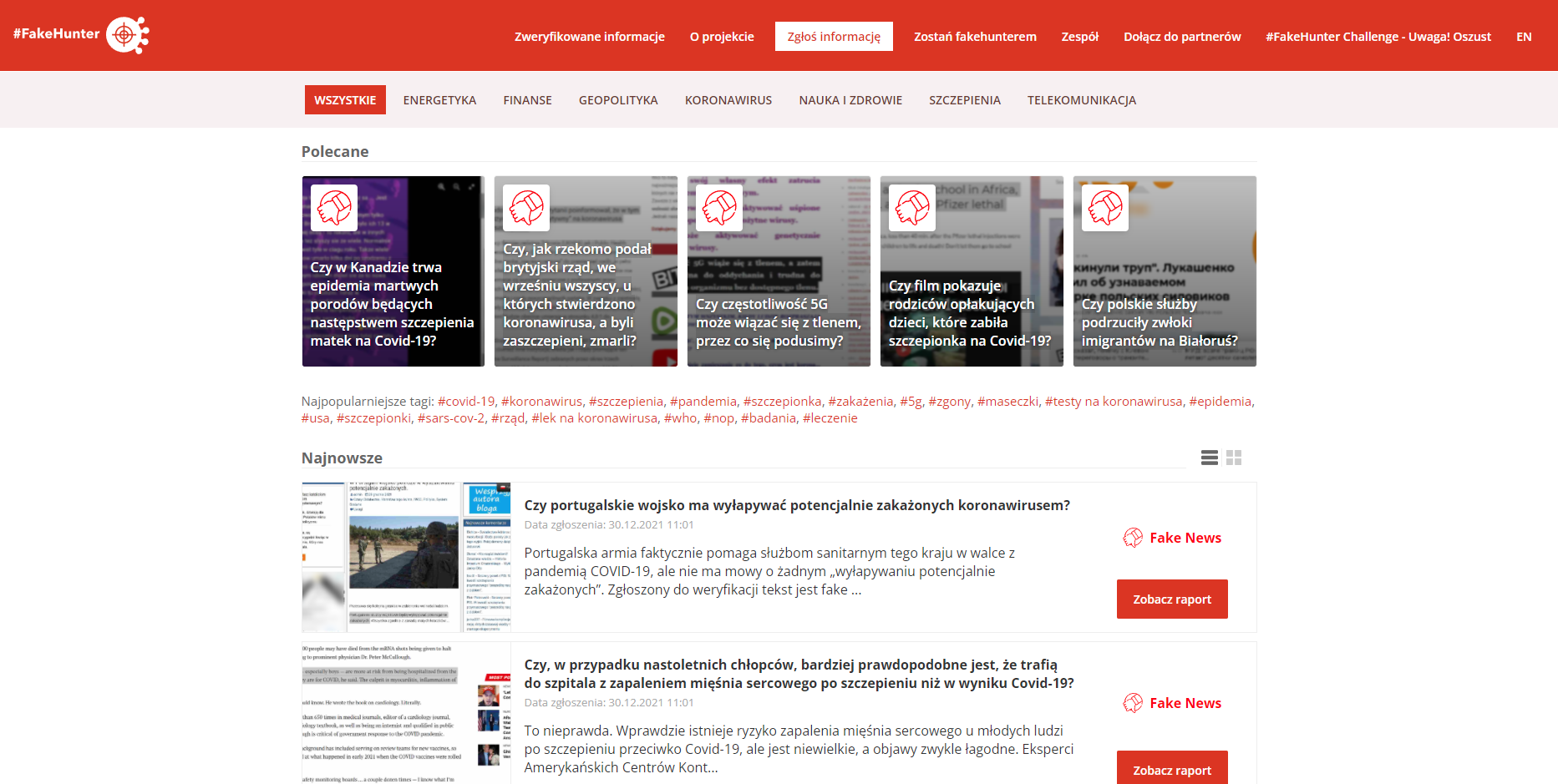
Pojęcie tzw. „fake newsów” odnosi się do fałszywych lub celowo mylących treści publikowanych jako prawdziwe wiadomości. Znaczny wzrost zainteresowania tym zagadnieniem nastąpił w roku 2016, po wyborach prezydenckich w Stanach Zjednoczonych i referendum w Wielkiej Brytanii w sprawie opuszczenia Unii Europejskiej. [1]

W celu odróżnienia informacji prawdziwych od fałszywych (tzw. „fake newsów”) przeprowadzany jest proces tzw. „fact-checkingu”, którego elementy to np. weryfikacja cytatów czy sprawdzanie wiarygodności źródeł. [2]

W Polsce funkcjonuje kilka organizacji weryfikujących informacje, m.in.:

* FakeHunter – serwis tworzony przez Polską Agencję Prasową i GovTech Polska [3],
* FakeNews.pl – portal fundacji „Przeciwdziałamy Dezinformacji” [4],
* AntyFAKE – projekt prowadzony przez Grupę Iberion [5],
* Stowarzyszenie Demagog – pierwsza w Polsce organizacja tego typu [6],
* Konkret24 – serwis należący do Grupy TVN [7].

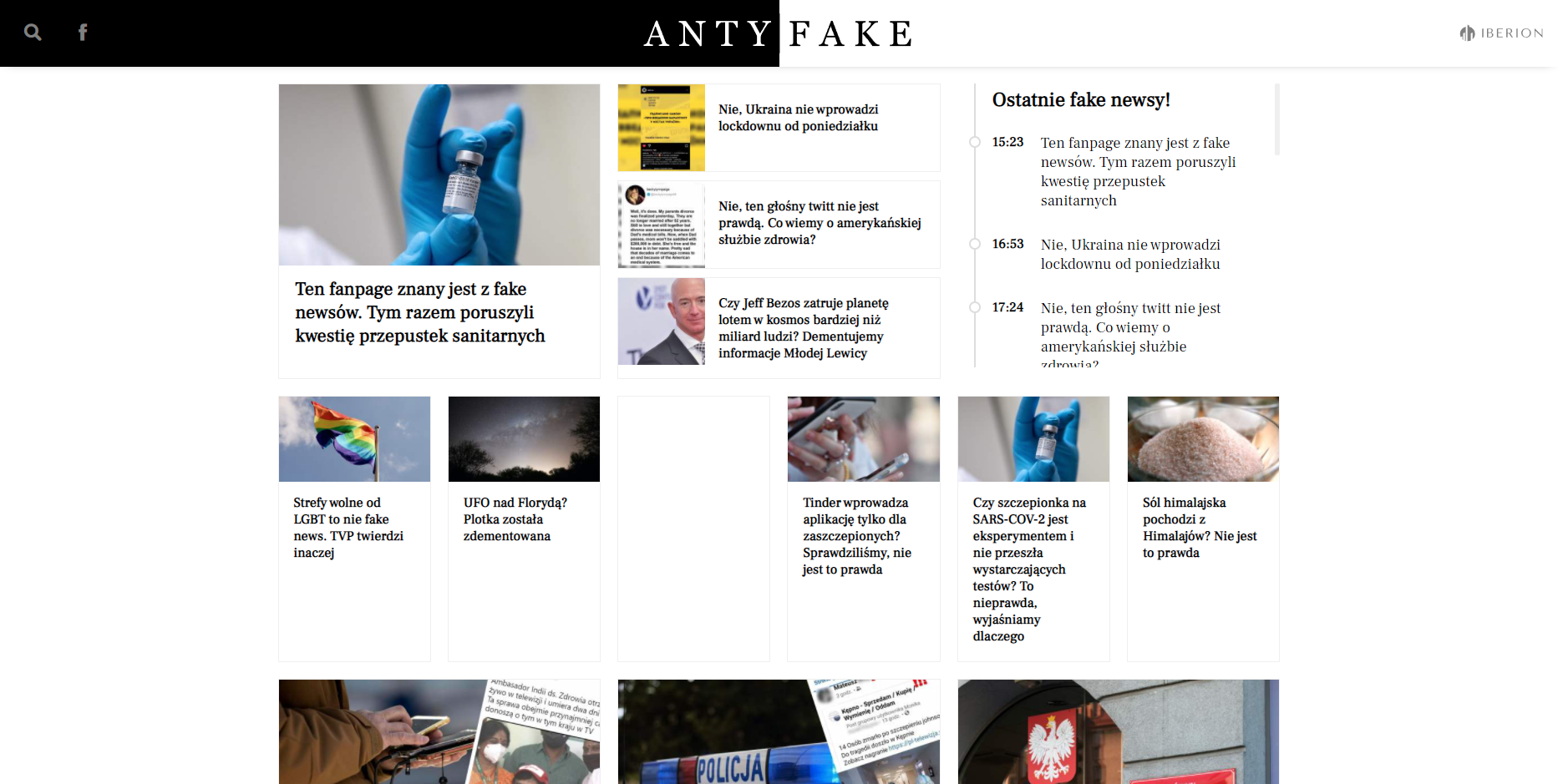
Na Rysunkach 1.1–1.5 przedstawiono strony główne wymienionych portali.



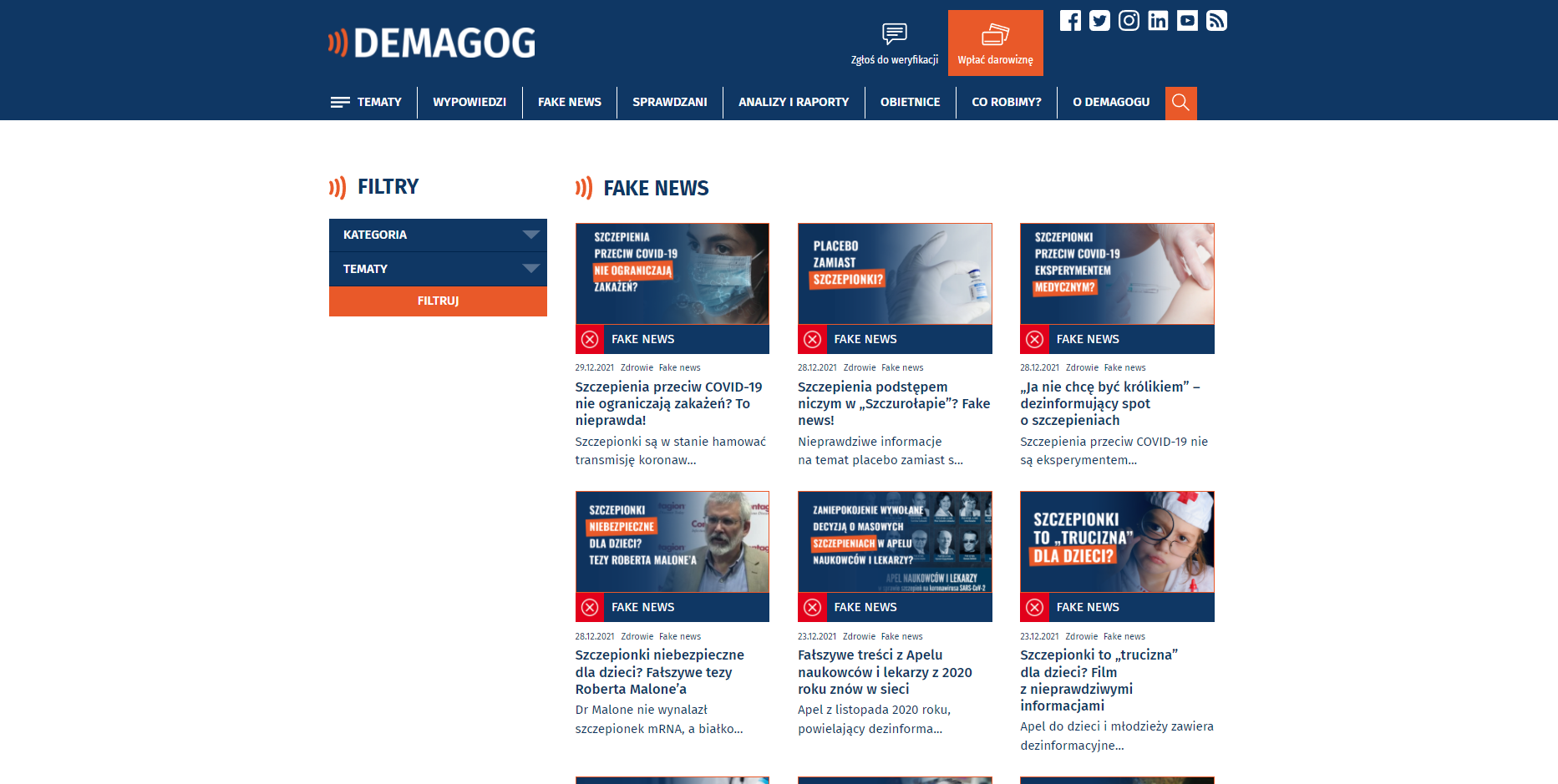
**Rysunek 1.1.** Strona główna portalu FakeHunter [3]



**Rysunek 1.2.** Strona główna portalu FakeNews.pl [4]



**Rysunek 1.3.** Strona główna portalu AntyFAKE [5]



**Rysunek 1.4.** Strona główna portalu Demagog [6]



**Rysunek 1.5.** Strona główna portalu Konkret24 [7]

## Cel i zakres pracy

# TODO

Celem niniejszej pracy jest zastosowanie podstawowych metod klasyfikacji tekstu do detekcji tzw. „fake newsów” dotyczących pewnego wybranego obszaru tematycznego. W procesie tworzenia oprogramowania

Rozbudowanie tego co poniżej…

Celem pracy jest przetestowanie podstawowych metod klasyfikacji tekstu do wykrywania tzw. „fake newsów” pochodzących z pewnego wybranego wąskiego obszaru.

Od opracowanego oprogramowania oczekuje się:

- wykorzystania systemu kontroli wersji (git, najchętniej Github)

- pokrycia projektu testami jednostkowymi i wykorzystania mechanizmu Continuous Integration (np. Travis CI)

- zapewnienia dokumentacji (dowolne narzędzie)

- minimalnego front endu, np. w postaci aplikacji webowej w Dashu opublikowanej na Heroku.

- kod po zakończeniu pracy ma być publiczny, licencja dowolna asdas

## Układ pracy

Niniejsza praca zorganizowana jest w następujący sposób: Rozdział 1 zawiera wprowadzenie w tematykę, cel i zakres pracy wraz z jej układem. W Rozdziale 2 przeprowadzono analizę aktualnego stanu wiedzy – opisano rozważane zagadnienia i dokonano przeglądu realizacji podobnych projektów. Rozdział 3 dotyczy wykonania projektu – uzasadniono wybór obszaru tematycznego, omówiono proces gromadzenia i przygotowania danych, przedstawiono wybrane algorytmy klasyfikacji a także model kompletnego klasyfikatora. W Rozdziale 4 zaprezentowano rezultaty testów jednostkowych stworzonego oprogramowania oraz wyniki ewaluacji zaprojektowanego klasyfikatora. Rozdział 5 zawiera opis zrealizowanych celów i założeń pracy, wnioski oraz propozycję kierunków możliwego rozwoju.

# Analiza aktualnego stanu wiedzy

## Uczenie maszynowe

Zgodnie z [8] uczenie maszynowe (ang. *Machine Learning, ML*) to ogół metod, za pomocą których na podstawie dostępnych danych dokonywana jest pewna predykcja. Wyróżnia się trzy główne grupy uczenia maszynowego:

* uczenie nadzorowane (ang. s*upervised learning*) – w przypadku zbioru danych wejściowych ze znanymi oczekiwanymi wyjściami (etykietami w klasyfikacji, wartościami liczbowymi w regresji),
* uczenie nienadzorowane (ang. *unsupervised learning*) – w przypadku zbioru danych wejściowych bez określonych oczekiwanych wyjść,
* uczenie przez wzmacnianie (ang. *reinforcement learning*) – w przypadku, gdy dane wejściowe pobierane są ze środowiska, w którym funkcjonuje model, w celu maksymalizacji pewnej nagrody.

Schemat typowego postępowania w realizacji nadzorowanego uczenia maszynowego, składa się z następujących kroków:

1. Gromadzenie danych.
2. Wykorzystanie danych w algorytmie uczenia maszynowego.
3. Zastosowanie stworzonego modelu do wyznaczenia predykcji w oparciu o nowe dane.

Procedura ta, w kontekście niniejszej pracy, została szerzej opisana w Rozdziale 3.

## Przetwarzanie języka naturalnego

Przetwarzanie języka naturalnego (ang. *Natural Language Processing, NLP*) to dziedzina badań zajmująca się wykorzystaniem komputerów w analizie tekstu i mowy ludzkiej, łącząca elementy m.in. informatyki, lingwistyki i sztucznej inteligencji. [9]

Do zastosowań przetwarzania języka naturalnego można zaliczyć [10] [11]:

* klasyfikację tekstu (np. detekcja tzw. „fake newsów”, filtracja spamu),
* tłumaczenie automatyczne,
* pozyskiwanie informacji z tekstu (np. rozpoznawanie nazw),
* podsumowania tekstów, systemy odpowiadania na pytania,
* systemy dialogowe (tzw. „czatboty”).

Zgodnie z [12] w przypadku klasyfikacji tekstu, a więc zadania nadzorowanego uczenia maszynowego, schemat typowego postępowania składa się z następujących kroków:

1. Wstępne przetwarzanie danych.
2. Ekstrakcja cech tekstu.
3. Selekcja cech tekstu.
4. Wykorzystanie danych w algorytmie uczenia maszynowego.

Na etapie wstępnego przetwarzania danych do najczęściej stosowanych metod należą:

* tokenizacja – podział sekwencji znaków na tokeny (poszczególne wyrazy lub zdania),
* filtracja – usunięcie wyrazów należących do listy tzw. „stop-words”, czyli często pojawiających się w tekście, lecz nie przekazujących informacji,
* lematyzacja – sprowadzenie wyrazów do ich podstawowych form (czasowników do formy bezokolicznika, rzeczowników do mianowników liczby pojedynczej),
* tzw. „stemming” – usunięcie formantów fleksyjnych wyrazów.

Ekstrakcja cech tekstu zazwyczaj przeprowadzana jest za pomocą konwersji dokumentów do wektorów w przestrzeni cech, w których z każdym wyrazem (tokenem) związana jest pewna wartość liczbowa (waga tego wyrazu). Wyróżnia się dwa modele doboru wag:

* model boolowski – do każdego wyrazu należącego do dokumentu przypisana jest pewna dodatnia waga, a do każdego nienależącego – wartość 0,
* metoda TF-IDF (ang. *term frequency-inverse document frequency*) – wagi wyznaczane są na podstawie liczby wystąpień wyrazów.

Do stosowanych algorytmów klasyfikacji zalicza się:

* naiwny klasyfikator bayesowski (ang. *Naive Bayes Classifier, NB*) – klasyfikator probabilistyczny, zakładający niezależność występowania wyrazów w tekście, szerzej opisany w Rozdziale 3.3.1,
* liniową maszynę wektorów nośnych (ang. *Linear* *Support Vector Machine, LSVM*) –algorytm separujący dostępne klasy za pomocą pewnej hiperpłaszczyzny, szerzej opisany w Rozdziale 3.3.2,
* drzewo decyzyjne (ang. *Decision Tree, DT*) – bazujące na rekurencyjnym podziale danych, szerzej opisane w Rozdziale 3.3.3,
* algorytm -najbliższych sąsiadów (ang. *K-Nearest Neighbour Classfier, KNN*) – klasa wyznaczana jest w oparciu o podobieństwo dokumentu do jego sąsiedztwa.

Całość wspomnianej procedury, w kontekście niniejszej pracy, została szerzej opisana w Rozdziale 3.

## Wyniki realizacji podobnych projektów

### Detekcja tzw. „fake newsów” w języku angielskim

W publikacji [13] wykorzystano zbiór danych zgromadzony przez serwis BuzzFeed News, zawierający 2282 artykułów o tematyce politycznej przypisanych ręcznie do jednej z czterech kategorii: „głównie prawda”, „głównie fałsz”, „mieszanina prawdy i fałszu” oraz „brak treści merytorycznej”. Podczas implementacji klasyfikatora zredukowano liczebność zestawu poprzez odrzucenie artykułów nienależących do grupy „głównie prawda” lub „głownie fałsz” oraz rekordów niezawierających żadnego tekstu. Finalnie otrzymano zbiór zawierający łącznie 1771 artykułów, spośród których ok. 5% stanowiły tzw. „fake newsy”.

Do wykrywania tekstów zawierających nieprawdziwe informacje wykorzystano naiwny klasyfikator bayesowski (Rozdział 2.2 i 3.3.1). Dane podzielono na trzy zestawy: zbiór uczący, walidacyjny i testowy. W zaproponowanej procedurze klasyfikacji iterowano po każdym wyrazie w analizowanym artykule, a o przydzieleniu do jednej z dwóch grup (prawdziwy albo fałszywy) decydowało prawdopodobieństwo warunkowe wyznaczone na podstawie zawartych w tekście pewnych konkretnych wyrazów. Po wytrenowaniu klasyfikatora i dostrojeniu jego parametrów przystąpiono do testów i ewaluacji działania algorytmów za pomocą dokładności (Rozdział 4.2.1), w wyniku której ogółem uzyskano wynik ok. 75% dokładności klasyfikacji (Tabela 2.1).

W celu poprawy działania klasyfikatora zaproponowano m.in. usunięcie z artykułów wyrazów należących do listy tzw. „stop-words” (Rozdział 2.2), wykorzystanie tzw. „stemmingu” (Rozdział 2.2), czy zmianę sposobu wyznaczania prawdopodobieństwa.

**Tabela 2.1.** Rezultaty ewaluacji klasyfikacji przeprowadzonej na zbiorze testowym [13]

|  |  |
| --- | --- |
| **Rodzaj artykułu** | **Dokładność []** |
| Prawdziwy |  |
| Fałszywy |  |
| Ogółem |  |

W pracy [14] zgromadzono 25 200 artykułów o tematyce politycznej, spośród których połowa stanowią wiadomości prawdziwe, a długość każdego z tekstów jest nie mniejsza niż 200 znaków. Dane podzielono na zbiór treningowy i testowy. Do określenia przestrzeni cech użyto modelu -gramowego, który na podstawie wyrazów przewiduje kolejny [11], a następnie zredukowano liczbę tych cech przy pomocy metody TF (Rozdział 3.2.2) oraz TF-IDF (Rozdział 2.2 i 3.2.2).

Klasyfikację przeprowadzono m.in. za pomocą liniowej maszyny wektorów nośnych (Rozdział 2.2 i 3.3.2), algorytmu -najbliższych sąsiadów (Rozdział 2.2) oraz drzewa decyzyjnego (Rozdział 2.2 i 3.3.3). W trakcie trenowania algorytmów wykorzystano 5-krotny sprawdzian krzyżowy (ang. *k-fold cross-validation*, metoda polegająca na -krotnym losowym podziale oryginalnego zbioru treningowego na mniejszy podzbiór treningowy i testowy, na których uczony jest algorytm uczenia maszynowego [11]).

W eksperymencie porównano dokładność klasyfikacji (Rozdział 4.2.1) dla obu metod ekstrakcji cech przy zmiennym rozmiarze -gramu dla każdego z zastosowanych algorytmów. Największą dokładność na poziomie 92% osiągnięto dla algorytmu LSVM (Tabela 2.2).

**Tabela 2.2.** Rezultaty ewaluacji klasyfikacji przeprowadzonej na zbiorze testowym   
z wykorzystaniem algorytmu LSVM [14]

|  |  |  |  |
| --- | --- | --- | --- |
| **Rozmiar**  **-gramu** | **Liczba**  **cech** | **Dokładność []** | |
| **TF-IDF** | **TF** |
| Unigram |  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
| Bigram |  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

W artykule [15] wykorzystano ten sam zbiór danych, co w pozycji [13] (2282 tekstów o tematyce politycznej), lecz zaproponowano inną metodę podziału na treści wiarygodne i nieprawdziwe. Odrzucono wiadomości oznaczone etykietą „brak treści merytorycznej”, a te należące do grup „głównie fałsz” i „mieszanina prawdy i fałszu” połączono i potraktowano jako teksty fałszywe (342 pozycje). Pozostałe, należące do klasy „głównie prawda”, przydzielono do kategorii artykułów prawdziwych (1666 pozycje).

Klasyfikacji dokonano poprzez zastosowanie m.in: naiwnego klasyfikatora bayesowskiego (Rozdział 2.2 i 3.3.1), lasu losowego (ang. *Random Forest, RF*) (Rozdział 3.3.3), oraz XGBoost (metoda wzmacniająca gradient, bazująca na drzewach [16]). Cechy artykułów wyznaczono ręcznie, wyróżniając trzy kategorie: cechy tekstu (m.in. właściwości językowe), cechy źródła danych (m.in. uprzedzenia polityczne czy wiarygodność) i cechy otoczenia (liczba interakcji użytkowników z postem).

Ewaluację klasyfikacji przeprowadzono za pomocą współczynnika -score (Rozdział 4.2.1) oraz obliczonego pola pod krzywą ROC (ang. *Receiver Operating Characteristic*) (Rozdział 4.2.1). Najlepsze rezultaty uzyskano dla algorytmu RF i XGBoost (Tabela 2.3).

**Tabela 2.3.** Rezultaty ewaluacji klasyfikacji przeprowadzonej na zbiorze testowym   
przy pomocy pola pod krzywą ROC oraz współczynnika -score [15]

|  |  |  |
| --- | --- | --- |
| **Algorytm  klasyfikacji** | **Pole pod  krzywą ROC []** | **Współczynnik  -score []** |
| NB |  |  |
| RF |  |  |
| XGB |  |  |

### Detekcja tzw. „fake newsów” w języku niemieckim

W artykule [17] zgromadzono zbiór danych zawierający 490 artykułów przekazujących nieprawdziwe informacje oraz 4500 wiadomości opublikowanych przez wiarygodnych wydawców głównego nurtu. Zastosowano dwie metody klasyfikacji: jedną opierającą się na maszynie wektorów nośnych (Rozdział 2.2 i 3.3.2), drugą na konwolucyjnej sieci neuronowej (ang. *Convolutional Neural Network, CNN*) (wielowarstwowa sieć neuronowa wykorzystująca operacje konwolucji, często stosowana w przypadku dużej ilości danych [18]). Do wytrenowania klasyfikatorów wykorzystano teksty zawierające co najmniej 200 słów, co zredukowało liczbę tzw. „fake newsów” do 300.

W przypadku klasyfikacji z wykorzystaniem maszyny wektorów nośnych połowę zbioru artykułów wykorzystano do treningu, drugą połowę przeznaczając na testy. Na podstawie zestawu danych, za pomocą metody TF-IDF (Rozdział 2.2 i 3.2.2) określono wagę słów w tekście i wyznaczono odpowiadające im wektory cech. W ramach klasyfikatora opierającego się na CNN dane podzielono na trzy zestawy – treningowy (50% zbioru), walidacyjny (25%) i testowy (25%), a w celu wytrenowania architektury sieci neuronowej wykorzystano wcześniej przygotowane tzw. „word embeddings” – pewną wektorową reprezentację słów.

W wyniku ewaluacji za pomocą współczynnika -score (Rozdział 4.2.1), dla klasyfikatora z algorytmem SVM otrzymano , a dla metody wykorzystującej CNN (Tabela 2.4).

**Tabela 2.4.** Rezultaty ewaluacji klasyfikacji przeprowadzonej na zbiorze testowym   
przy pomocy dokładności oraz współczynnika -score [17]

|  |  |  |
| --- | --- | --- |
| **Algorytm  klasyfikacji** | **Pole pod  krzywą ROC []** | **Współczynnik  -score []** |
| SVM |  |  |
| CNN |  |  |

### Detekcja tzw. „fake newsów” w języku hiszpańskim

W pracy [19] zebrano łącznie 971 artykułów z 9 kategorii tematycznych, w tym 491 oznaczonych jako prawdziwe i 480 jako fałszywe, które następnie podzielono na zestawy treningowe (70% zbioru) i testowe (30%). Jako reprezentację cech wykorzystano m.in.:

* tzw. „bag-of-words” (BOW) – metoda reprezentowania dokumentów za pomocą wektora, w którym ciąg 0 i 1 określa występowanie wyrazów z pewnego słownika [11],
* tzw. „part-of-speech tagging” (POS) – metoda przypisywania do każdego wyrazu odpowiadającej mu części mowy [11],
* znakowe -gramy – metoda podziału każdego wyrazu na -znakowe części [20],

a do klasyfikacji wykorzystano algorytmy: maszynę wektorów nośnych (Rozdział 2.2 i 3.3.2), regresję logistyczną (ang. *Logistic Regression, LR*) (metoda umożliwiająca wyznaczenie prawdopodobieństwa należenia do klasy [11]), las losowy (Rozdział 3.3.3) oraz tzw. „boosting” (metoda polegająca na stworzeniu dobrego algorytmu na podstawie słabych [21]).

Korzystając z dokładności (Rozdział 4.2.1) porównano działanie klasyfikatorów trenowanych na różnych zbiorach cech tekstu – BOW, POS oraz połącznie obu (Tabela 2.5).

**Tabela 2.5.** Rezultaty ewaluacji klasyfikacji przeprowadzonej na zbiorze testowym [19]

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Reprezentacja cech** | **Dokładność klasyfikacji []** | | | |
| **SVM** | **LR** | **RF** | **BO** |
| BOW |  |  |  |  |
| POS |  |  |  |  |
| BOW+POS |  |  |  |  |

Zestawiono także dokładność klasyfikacji modeli bazujących na znakowych -gramach z usuwaniem wyrazów należących do listy tzw. „stop-words” i nie (Tabela 2.6).

**Tabela 2.6.** Rezultaty ewaluacji klasyfikacji przeprowadzonej na zbiorze testowym   
z wykorzystaniem modelów -gramowych [19]

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Rozmiar**  **znakowego**  **-gramu** | **Dokładność klasyfikacji []** | | | | | | | |
| **usuwanie wyrazów** | | | | **nieusuwanie wyrazów** | | | |
| **SVM** | **LR** | **RF** | **BO** | **SVM** | **LR** | **RF** | **BO** |
| -gram |  |  |  |  |  |  |  |  |
| -gram |  |  |  |  |  |  |  |  |
| -gram |  |  |  |  |  |  |  |  |

W publikacji lepsze rezultaty uzyskano w przypadku korzystania z reprezentacji -gramowej – 77% dokładności dla znakowego -gramu o rozmiarze 4 i metody „boosting” i 77% dokładności dla kombinacji BOW z POS i algorytmu RF).

# Realizacja projektu

## Wybór obszaru tematycznego i tworzenie zbiór danych

Na początku roku 2020 w związku z pandemią COVID-19 zaczęto obserwować znaczną liczbę tzw. „fake newsów” podważających istnienie wirusa czy jego śmiertelność [22]. Z uwagi na mnogość wiadomości dotyczących koronawirusa, zarówno prawdziwych, jak i fałszywych, w niniejszej pracy zdecydowano się skupić właśnie na tym obszarze tematycznym.

Do stworzenia korpusu mogącego posłużyć jako dane w procesie klasyfikacji, konieczne było zgromadzenie odpowiedniej liczby tekstów zawierających prawdziwe albo fałszywe informacje. W tym celu wykorzystano tzw. „web scraping”, czyli technikę pozyskiwania informacji z witryn internetowych [23].

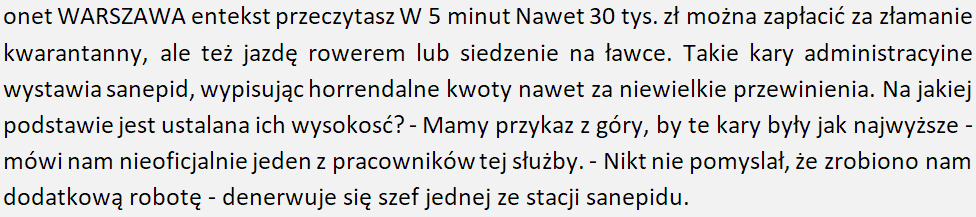
Nieprawdziwe wiadomości pobrano z portalu FakeHunter [3], na którego łamach, na podstawie wiarygodnych źródeł, weryfikowana jest poprawność informacji przesłanych przez użytkowników. Przeanalizowane doniesienia są następnie publikowane w formie artykułów zawierających werdykt („Fake News” albo „Prawda”) oraz raport eksperta wraz ze zdjęciem całości zgłoszonego tekstu lub jego fragmentu, a także link do jego źródła. Przykładowa analiza opublikowana w serwisie została przestawiona na Rysunku 3.1.



**Rysunek 3.1.** Raport z weryfikacji opublikowany na łamach portalu FakeHunter [24]

W celu pobrania danych wykorzystano tzw. „API” (ang. *application programming interface*) portalu, umożliwiające pobranie zawartości strony w postaci słownika składającego się z par klucz-wartość (np. do klucza „werdykt” przypisana jest wartość „fałszywy”). Treść weryfikowanych artykułów została wyodrębniona z zamieszczonych w analizie zdjęć fragmentów tekstów za pomocą oprogramowania rozpoznającego znaki, tzw. „OCR” (ang. o*ptical character recognition*), którego przykład działania widoczny jest na Rysunku 3.2.





**Rysunek 3.2.** Rezultat ekstrakcji tekstu za pomocą oprogramowania OCR z artykułu [24]

Pseudokod procedury pobierania danych z portalu FakeHunter przedstawiono poniżej:

**for** strona **in** zakres\_stron:

weryfikacje = pobierz\_weryfikacje(strona)

**for** weryfikacja **in** weryfikacje:

tekst = pobierz\_tekst\_z\_obrazu(weryfikacja[adres\_obrazu])

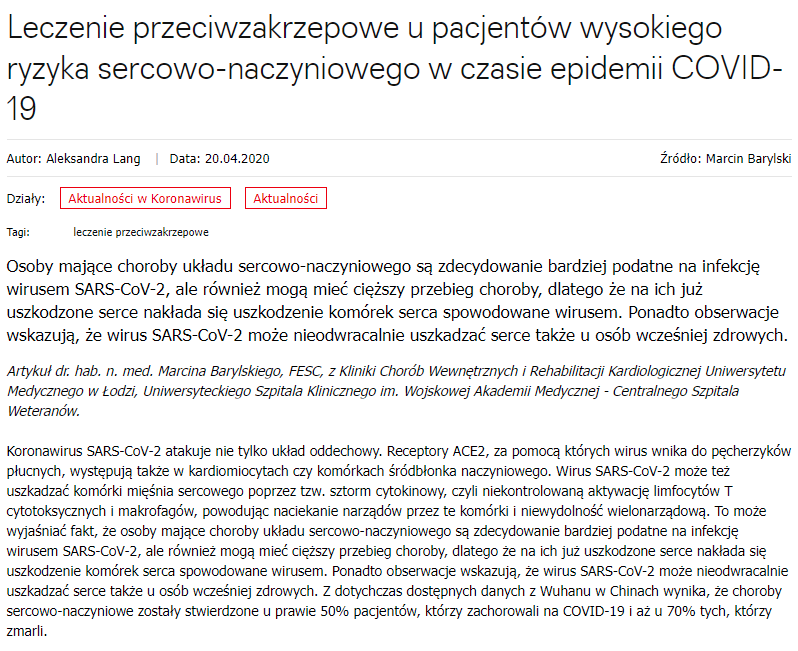
zapisz\_do\_tabeli(weryfikacja[werdykt], weryfikacja[tytuł],

tekst, weryfikacja[adres\_artykułu])

Z portalu wyodrębniono 939 tekstów określonych jako tzw. „fake news”, 151 prawdziwych i 74 bez werdyktu. Artykuły te umieszczono w pliku *fakehunter\_dataset.xlsx*, stanowiącym załącznik do pracy*.* Na Wykres X przedstawiono rozkład długości tekstów.



Jako bazę wiarygodnych wiadomości przyjęto serwis informacyjny wydawnictwa Termedia [25], publikujący artykuły o tematyce medycznej, w tym dotyczące wirusa SARS-CoV-2. Przykładowy tekst dostępny na stronie został przedstawiony na Rysunku 3.3.



**Rysunek 3.3.** Fragment artykułu zamieszczonego na portal Termedia [26]

Z uwagi na brak możliwości skorzystania z tzw. „API” serwisu, zawartość została wyodrębniona ze strony z wykorzystaniem parsowania kodu HTML witryny internetowej.

Pseudokod procedury pobierania danych z portalu Termedia przedstawiono poniżej:

**for** strona **in** zakres\_stron:

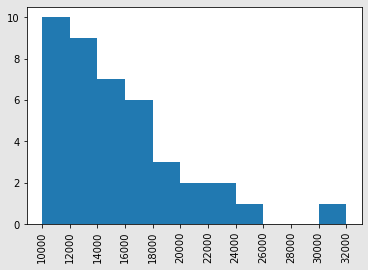
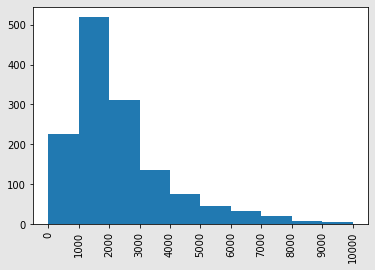
adresy\_artykulów = pobierz\_adresy(strona)

**for** adres **in** adresy\_artykułów:

tytuł, tekst, adres\_artykułu = pobierz\_tekst\_z\_artykulu(adres)

zapisz\_do\_tabeli(‘prawda’, tytuł, tekst, adres\_artykułu)

Z portalu wyodrębniono 1429 artykułów, które umieszczono w pliku *termedia\_dataset.xlsx*, stanowiącym załącznik do pracy*.* Na Wykres X przedstawiono rozkład ich długości.



## Przetwarzanie zbioru danych

### Oczyszczanie danych

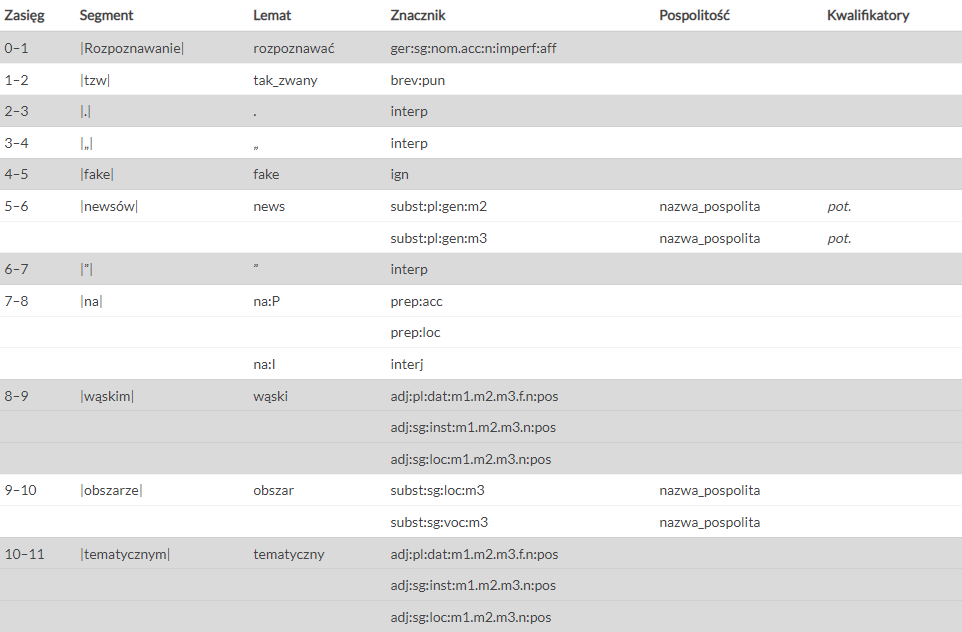
W ramach przygotowania zbioru danych z portalu FakeHunter do konwersji do przestrzeni cech tekstu, konieczne było wykonanie poniższych operacji:

1. Usunięcie rekordów niezawierających treści artykułu – przypadki, w których oprogramowanie wykrywające znaki nie odnalazło tekstu na zdjęciu.
2. Usunięcie tekstów niepolskojęzycznych – przypadki, w których artykuł pochodził z zagranicznego, najczęściej anglojęzycznego źródła.
3. Usunięcie tekstów bez werdyktu – przypadki, w których zgłoszona wiadomość przekazywała informacje nieweryfikowalne.
4. Usunięcie tekstów błędnie wyodrębnionych z serwisu Twitter – przypadki, w których oprogramowanie wykrywającego znaki niepoprawnie wyodrębniło niektóre teksty pochodzące z serwisu Twitter.

W przypadku artykułów z portalu Termedia, należało jedynie usunąć puste rekordy (przypadki, w których treść artykułu była niedostępna).

W kolejnym kroku na obu wstępnie oczyszczonych zestawach przeprowadzono następujące operacje:

1. Usunięcie znaków ucieczki – znaki tabulacji, nowej linii itp. stanowią jedynie utrudnienie analizy tekstu.
2. Usunięcie znaków interpunkcyjnych i nienależących do polskiego alfabetu – w przetwarzaniu języka naturalnego stanowią one szum, nieprzekazujący żadnej informacji.
3. Zastąpienie wielokrotnych spacji pojedynczą – tokenizacja przeprowadzana jest na podstawie pojedynczych spacji rozdzielających wyrazy, wielokrotne spacje stanowiłby wiec utrudnienie.
4. Zamiana wielkich liter w wyrazach na małe – wielkość liter nie ma znaczenia w przeprowadzanej analizie tekstów.
5. Podział tekstów na tokeny – każdy tekst rozdzielany jest na pojedyncze wyrazy.
6. Usunięcie wyrazów należących do listy tzw. „stop-words” – wyrazy nieprzekazujące informacji często występujące w języku polskim lub będące spójnikami.
7. Przeprowadzenie lematyzacji – w tym celu wykorzystano analizator i generator fleksyjny dla języka polskiego Morfeusz2 [27], którego przykład działania (z pominięciem ww. kroków przygotowawczych) przedstawiono na Rysunku 3.4.



**Rysunek 3.4.** Demonstracja analizy morfologicznej programu Morfeusz2 przeprowadzonej   
dla zdania „Rozpoznawanie tzw. ≫fake newsów≪ na wąskim obszarze tematycznym” [28]

1. Usunięcie rekordów niezawierających treści artykułu – w rezultacie ww. kroków pojawiły się do tej pory niewykryte rekordy bez użytecznego tekstu.
2. Usunięcie tekstów zbyt krótkich – w celu zmniejszenia szumu, zdecydowano się na analizowanie jedynie tekstów zawierających 30 lub więcej znaków.

Finalnie uzyskano zbiór danych składający się z 1497 artykułów przekazujących prawdziwe informacje i 753 fałszywych wiadomości – łącznie 2240 pozycji.

### Ekstrakcja i selekcja cech tekstu

W celu konwersji zgromadzonych dokumentów do wektorów cech wykorzystano metodę TF-IDF (ang. *term frequency-inverse document frequency*), która zgodnie z [11] przypisuje tokenom wagi według następującej zależności:

|  |  |  |
| --- | --- | --- |
|  |  | (2.1.1) |

gdzie to częstość występowania wyrazu w dokumencie określona jako stosunek liczby wystąpień wyrazu w dokumencie do liczby wszystkich wyrazów w dokumencie :

|  |  |  |
| --- | --- | --- |
|  |  | (2.1.2) |

a to logarytm dziesiętny ze stosunku liczby wszystkich dokumentów do liczby dokumentów, w których występuje wyraz :

|  |  |  |
| --- | --- | --- |
|  |  | (2.1.3) |

W realizacji projektu wykorzystano metodę TF-IDF z biblioteki scikit-learn, w której składnik IDF wyznaczany jest w następujący sposób [29]:

|  |  |  |
| --- | --- | --- |
|  |  | (2.1.4) |

Wyznaczone wektory TF-IDF są następnie normalizowane z wykorzystaniem normy euklidesowej [29]. Funkcja tworząca wektory cech umożliwia także wyznaczanie ich w oparciu o wyrazowe -gramy, w analizie wykorzystano więc uni- i bigramy.

Zestaw danych został podzielony na zbiór treningowy (70% danych, 1568 tekstów) i testowy (30% danych, 672 teksty), a następnie dokonano wektoryzacji, w wyniku której uzyskano 27416 cech w przypadku unigramów i 193552 dla reprezentacji bigramowej. W kolejnym kroku zastosowano wymienione w Rozdziale 3.3 algorytmy klasyfikacji.

## Wybrane algorytmy klasyfikacji

### Naiwny klasyfikator bayesowski

Naiwny klasyfikator bayesowski (ang. *Naive Bayes Classifier*) to probabilistyczna metoda klasyfikacji, której zadaniem jest wyznacznie zgodnie z twierdzeniem Bayesa najbardziej prawdopodobnej klasy dla danego wektora zmiennych. Istotnym uproszczeniem znacznie usprawniającym proces uczenia (zwłaszcza przy dużej liczbie zmiennych) jest założenie o wzajemnej niezależność predyktorów, umożliwiające wyznaczanie parametrów dla każdej ze zmiennych oddzielnie. [30]

Zgodnie z [31] prawdopodobieństwo przynależności wektora cech tekstu do klasy jest równe:

|  |  |  |
| --- | --- | --- |
|  |  | (3.1.1) |

Uwzględniając założenie o niezależności predyktorów:

|  |  |  |
| --- | --- | --- |
|  |  | (3.1.2) |

Zgodnie z [32], z uwagi na fakt, że mianownik jest stały, reguła klasyfikacji określona jest zależnością:

|  |  |  |
| --- | --- | --- |
|  |  | (3.1.3) |

a więc najbardziej odpowiednią klasę dla danego wektora cech tekstu stanowi:

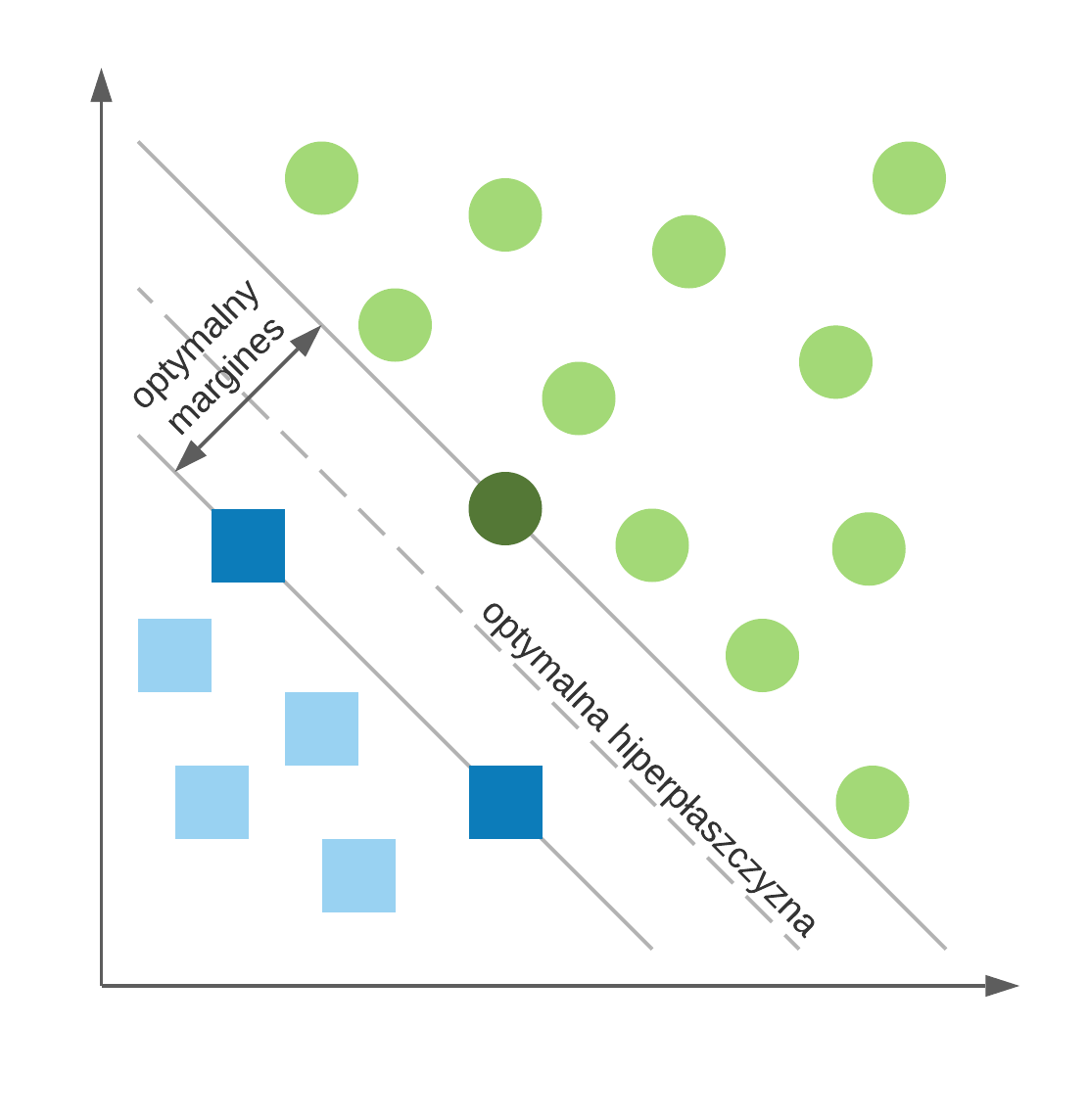
|  |  |  |
| --- | --- | --- |
|  |  | (3.1.4) |

gdzie wartości częstość występowania klasy w zbiorze danych () oraz częstość występowania cechy w tekstach należących do klasy () są szacowane na podstawie zestawu danych uczących.

W naiwnym klasyfikatorze bayesowskim z biblioteki scikit-learn, wykorzystanym w realizacji projektu, wartość jest estymowana z uwzględnieniem wygładzania addytywnego, co ma na celu wyeliminowanie zerowania się tego składnika w przypadkach, gdy dana para cecha-klasa nie wystąpiła w zbiorze treningowym. [33]

### Liniowa maszyna wektorów nośnych

Liniowa maszyna wektorów nośnych (ang. *Linear* *Support Vector Machine, LSVM*) to algorytm stosowany w problemach klasyfikacji binarnej, którego celem jest określenie hiperpłaszczyzny rozdzielającej zestaw danych pomiędzy dwie klasy. Wymiar tej hiperpłaszczyzny zależy od liczebności zbioru cech, na podstawie którego będzie ona wyznaczana. Margines wokół granicy (funkcji) decyzyjnej maksymalizowany jest na podstawie par wektorów cech i odpowiadających im etykiet . [34].



**Rysunek 3.5.** Przykład funkcji decyzyjnej dla danych liniowo separowalnych,   
ciemniejszym odcieniem oznaczono wektory nośne;

opracowanie własne na podstawie [35]

Zgodnie z [35] zestaw danych uczących , jest liniowo separowalny, jeśli nierówność:

|  |  |  |
| --- | --- | --- |
|  | , | (3.2.1) |

jest spełniona dla wszystkich jego elementów.

Wektory, dla których to tzw. wektory nośne, na podstawie których w przestrzeni cech konstruowana jest optymalna hiperpłaszczyzna określona zależnością:

|  |  |  |
| --- | --- | --- |
|  | , | (3.2.2) |

rozdzielająca dane uczące z największym możliwym marginesem, opisanym wzorem:

|  |  |  |
| --- | --- | --- |
|  | . | (3.2.3) |

Wyznaczanie hiperpłaszczyzny (2.2) sprowadza się do rozwiązania problemu programowania kwadratowego – minimalizacji iloczynu przy ograniczeniach (2.1).

Maszyna wektorów nośnych z biblioteki scikit-learn, wykorzystana w realizacji projektu, przy obliczaniu współczynników hiperpłaszczyzny uwzględnia dodatkowo funkcję straty typu hinge loss, kontrolującą wpływ predykcji ze zbyt małym marginesem lub błędnych. [36]

### Las losowy

Las losowy (ang. *Random Forest, RF*) to metoda uczenia maszynowego działająca w oparciu o grupę drzew decyzyjnych (ang. *Decision Tree*).

Zgodnie z [37], algorytm drzewa decyzyjnego dokonuje rekurencyjnego, binarnego podziału przestrzeni cech według poszczególnych zmiennych objaśniających (predyktorów), aż do momentu spełnienia kryterium zatrzymania. W rezultacie otrzymywane są niepodzielone węzły, zwane końcowymi, na podstawie których określana jest predykcja algorytmu. Do zalet drzew decyzyjnych należą m.in.:

* możliwość stosowania w problemach klasyfikacji i regresji,
* możliwość opisania zależności między zmiennymi objaśniającymi,
* odporność na braki danych, odstające wartości i predyktory bez znaczenia.

Istotną wadę drzew decyzyjnych stanowi m.in. niestabilność (nieznaczna zmiana w przestrzeni zmiennych skutkuje innym drzewem), a także mniejsza dokładność względem nowszych metod (np. sieci neuronowych).

W celu poprawienia rezultatów, drzewa, z których składa się las losowy, tworzone są na podstawie wylosowanego ze zwracaniem podzbioru dostępnych danych, a podział w węzłach tych drzew odbywa się z wykorzystaniem pewnej wylosowanej liczby predyktorów – poprzez wybór najlepszego podziału dla tego podzbioru zmiennych objaśniających. W przypadku klasyfikacji prognozę algorytmu stanowi wynik głosowania wśród drzew (klasa wybrana przez większość), a przy wyznaczaniu wartości liczbowej zmiennej objaśnianej w regresji, średnia wyników uzyskanych przez poszczególne drzewa. [38]

W modelu lasu losowego z biblioteki scikit-learn, wykorzystanym w realizacji projektu, prognozę stanowi średnia predykcji wszystkich drzew, na podstawie której określana jest przynależność do klasy, a nie klasa wybrana przez większość. [39]

## Kompletny model klasyfikatora

# TODO

## Aplikacja webowa

# TODO



# Rezultaty

## Testy jednostkowe funkcji

# TODO

## Ewaluacja algorytmów klasyfikacji

### Zastosowane metryki

Precyzja

Czułość

F1 score

Dokładność

Krzywa ROC

Macierz pomyłek

### Naiwny klasyfikator bayesowski

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Rozmiar**  **-gramu** | **Liczba**  **cech** | **Precyzja []** | **Czułość []** | **Współczynnik -score []** | **Dokładność []** |
| Unigram |  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
| Bigram |  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |

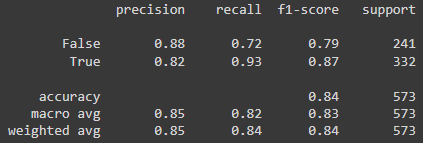
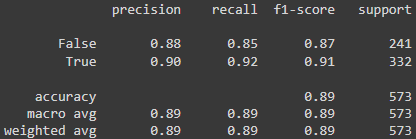
### Liniowa maszyna wektorów nośnych

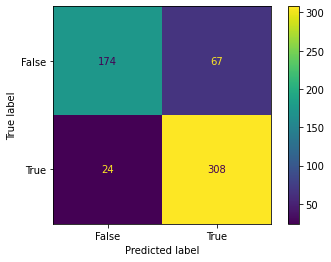
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Rozmiar**  **-gramu** | **Liczba**  **cech** | **Precyzja []** | **Czułość []** | **Współczynnik -score []** | **Dokładność []** |
| Unigram |  | 92 | 94 | 93 | 91 |
|  | 94 | 94 | 94 | 92 |
|  | 96 | 95 | 95 | 94 |
|  | 96 | 94 | 95 | 94 |
| Bigram |  | 90 | 91 | 90 | 87 |
|  | 91 | 93 | 92 | 89 |
|  | 91 | 94 | 94 | 90 |
|  | 90 | 94 | 92 | 89 |
|  | 87 | 95 | 91 | 88 |
|  | 84 | 97 | 90 | 86 |

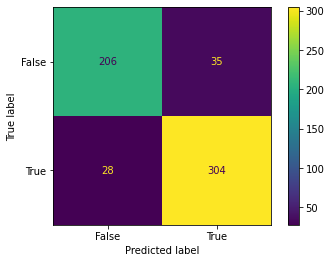
### Las losowy

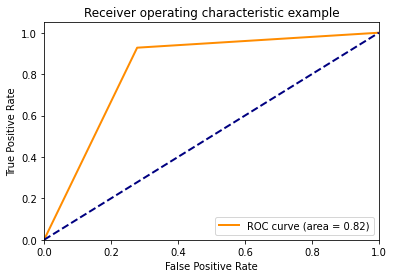
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Rozmiar**  **-gramu** | **Liczba**  **cech** | **Precyzja []** | **Czułość []** | **Współczynnik -score []** | **Dokładność []** |
| Unigram |  | 90 | 93 | 91 | 89 |
|  | 91 | 93 | 92 | 89 |
|  | 91 | 93 | 92 | 89 |
|  | 90 | 93 | 92 | 89 |
| Bigram |  | 92 | 89 | 90 | 87 |
|  | 93 | 88 | 90 | 88 |
|  | 94 | 88 | 91 | 88 |
|  | 96 | 86 | 91 | 89 |
|  | 97 | 86 | 91 | 89 |
|  | 98 | 84 | 91 | 89 |

# TODO Przykładowe wyniki









# Podsumowanie

## Zrealizowane założenia projektowe

# TODO

## Wnioski

# TODO

## Kierunki dalszego rozwoju

# TODO

# Bibliografia

x

|  |  |
| --- | --- |
| 1. | Gelfert A. Fake News: A Definition. Informal Logic. 2018 marzec: p. 84-117. |
| 2. | Molina MD, Sundar SS, Quang Le T, Lee D. “Fake News” Is Not Simply False Information: A Concept Explication and Taxonomy of Online Content. American Behavioral Scientist. 2019 październik. |
| 3. | FakeHunter. [Online]. [dostęp: 1 styczeń 2022]. <https://fakehunter.pap.pl>. |
| 4. | FakeNews.pl. [Online]. [dostęp: 1 styczeń 2022]. <https://fakenews.pl/>. |
| 5. | AntyFAKE. [Online]. [dostęp: 1 styczeń 2022]. <https://www.antyfake.pl/>. |
| 6. | Demagog. [Online]. [dostęp: 1 styczeń 2022]. <https://demagog.org.pl/>. |
| 7. | Konkret24. [Online]. [dostęp: 1 styczeń 2022]. <https://konkret24.tvn24.pl/>. |
| 8. | Molnar C. Interpretable machine learning: lulu.com; 2020. |
| 9. | Chowdhury G. Natural language processing. Annual Review of Information Science and Technology. 2005 styczeń: p. 51-89. |
| 10. | Khurana D, Koli A, Khatter K, Singh S. Natural Language Processing: State of Art, Current Trends and Challenges. 2017. |
| 11. | Jurafsky D, Martin J. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. 2nd ed.: Prentice Hall; 2008. |
| 12. | Allahyari M, Pouriyeh S, Assefi M, Safaei S. A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques. 2017 lipiec. |
| 13. | Granik M, Mesyura V. Fake news detection using naive Bayes classifier. In IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON); 2017; Kijów. p. 900-903. |
| 14. | Ahmed H, Traore I, Saad S. Detection of Online Fake News Using N-Gram Analysis and Machine Learning Techniques. In International Conference on Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments; 2017; Vancouver. p. 127-138. |
| 15. | Reis JCS, Correia A, Murai F, Veloso A, Benevenuto F. Supervised Learning for Fake News Detection. IEEE Intelligent Systems. 2019 marzec-kwiecień: p. 76-81. |
| 16. | Dokumentacja algorytmu XGBoost. [Online]. [dostęp: 5 styczeń 2022]. <https://xgboost.readthedocs.io/en/latest/index.html>. |
| 17. | Vogel I, Jiang P. Fake News Detection with the New German Dataset “GermanFakeNC”. In Digital Libraries for Open Knowledge; 2019; Oslo. p. 288-295. |
| 18. | Albawi S, Mohammed TA, Al-Zawi S. Understanding of a convolutional neural network. In International Conference on Engineering and Technology (ICET); 2017. p. 1-6. |
| 19. | Sidorov G, Moreno J, Adorno HG, Posadas-Durán JP. Detection of fake news in a new corpus for the Spanish language. Journal of Intelligent & Fuzzy Systems. 2019 maj; 36(5): p. 4869-4876. |
| 20. | Cavnar WB, Trenkle JM. N-gram-based text categorization. In Proceedings of SDAIR-94, 3rd annual symposium on document analysis and information retrieval; 1994. |
| 21. | Schapire RE. The Boosting Approach to Machine Learning: An Overview. In Denison DD, Holmes CC, Hansen MH, Mallick B, Yu B. Nonlinear Estimation and Classification. Lecture Notes in Statistics. Nowy Jork: Springer; 2003. p. 149-171. |
| 22. | van der Linden S, Roozenbeek J, Compton J. Inoculating Against Fake News About COVID-19. Frontiers in Psychology. 2020 październik. |
| 23. | Diouf R, Sall O, Birregah B, Bousso M, Sarr EN, Bousso M, et al. Web Scraping: State-of-the-Art and Areas of Application. In IEEE International Conference on Big Data; 2019; Los Angeles. p. 6040-6042. |
| 24. | Raport z portalu FakeHunter. [Online]. [dostęp: 3 styczeń 2022]. <https://fakehunter.pap.pl/raport/2143bb38-fa97-4a76-a9f1-40a322d5e29b>. |
| 25. | Termedia. [Online]. [dostęp: 12 grudzień 2021]. <https://www.termedia.pl>. |
| 26. | Artykuł z portalu Termedia. [Online]. [dostęp: 3 styczeń 2022]. <https://www.termedia.pl/koronawirus/Leczenie-przeciwzakrzepowe-u-pacjentow-wysokiego-ryzyka-sercowo-naczyniowego-w-czasie-epidemii-COVID-19,37561.html>. |
| 27. | Kieraś W, Woliński M. Morfeusz 2 – analizator i generator fleksyjny dla języka polskiego. Język Polski. 2017: p. 75-83. |
| 28. | Wersja demonstracyjna programu Morfeusz2. [Online]. [dostęp: 5 styczeń 2022]. <http://morfeusz.sgjp.pl/demo/>. |
| 29. | Dokumentacja metody TF-IDF (scikit-learn). [Online]. [dostęp: 5 styczeń 2022]. <https://scikit-learn.org/stable/modules/feature_extraction.html#text-feature-extraction>. |
| 30. | McCallum A, Nigam K. A Comparison of Event Models for Naive Bayes Text Classification. 2001 maj. |
| 31. | Zhang H. The Optimality of Naive Bayes. In Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference; 2004; Floryda. |
| 32. | Raghavan P, Schütze H, Manning C. Introduction to Information Retrieval Cambridge: Cambridge University Press; 2009. |
| 33. | Dokumentacja naiwnego klasyfikatora bayesowskiego (scikit-learn). [Online]. [dostęp: 18 grudzień 2021]. <https://scikit-learn.org/stable/modules/naive_bayes.html#multinomial-naive-bayes>. |
| 34. | Boser B, Guyon I, Vapnik V. A Training Algorithm for Optimal Margin Classifier. Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory. 1992: p. 144-152. |
| 35. | Cortes C, Vapnik V. Support-vector networks. Machine Learning. 1995: p. 273-297. |
| 36. | Dokumentacja liniowej maszyny wektorów nośnych (scikit-learn). [Online]. [dostęp: 20 grudzień 2021]. <https://scikit-learn.org/stable/modules/svm.html#id15>. |
| 37. | Cutler A, Cutler DR, Stevens JR. Random Forests. Machine Learning. 2011 styczeń: p. 157-176. |
| 38. | Breiman L. Random Forests. Machine Learning. 2001 październik: p. 5-32. |
| 39. | Dokumentacja algorytmu lasu losowego (scikit-learn). [Online]. [dostęp: 24 grudzień 2021]. <https://scikit-learn.org/stable/modules/ensemble.html#forest>. |

x