POLITECHNIKA WROCŁAWSKA
WYDZIAŁ ELEKTRONIKI

KIERUNEK: INFORMATYKA
SPECJALNOŚĆ: Inżynieria systemów informatycznych

PRACA DYPLOMOWA MAGISTERSKA

Zastosowanie metod uczenia maszynowego w detekcji fałszywych informacji

Application of machine learning methods to fake news detection

AUTOR:
Inż. Dawid Mikowski

PROWADZĄCY PRACĘ:
Prof. Michał Woźniak

OCENA PRACY:

WROCŁAW, 2020

Spis treści

Spis rysunków

Wstęp

Rozdział 1

Informacje nieprawdziwe w dobie Internetu

Pojęcie “Fake news” odnosi się do informacji, które pomimo że nie posiadają pokrycia z rzeczywistością są przedstawiane jako prawdziwe w mediach takich jak np.: wiadomości, artykuły, portale społecznościowe itd.. Zwrot ten jest neologizmem i w języku angielskim oznacza dosłownie “Fałszywe wiadomości”. Tego rodzaju wiadomości często wykorzystywane są w tworzeniu treści humorystycznych np.: satyra, jednak ich główne przeznaczenie sprowadza się do spełniania jednego z dwóch zadań:

- oszukać odbiorcę i wpłynąć na jego poglądy w sposób żądany przez autora danej informacji (Propaganda)
- namówienie go na zakup czegoś czego w innym przypadku by on nie kupił (Reklama)

Niektóre fałszywe informacje łączą oba cele.

Fake news’y stały się bardzo popularnym zagadnieniem w ostatnich czasach ponieważ Internet a w szczególności media społecznościowe dają możliwość przekazywania informacji z niespotykaną wcześniej prędkością dzięki czemu rozprzestrzenianie dezinformacji stało się zadaniem stosunkowo prostym.

Zagadnienie to zyskało ogromny rozgłos podczas kampanii wyborczej oraz prezydentury Donalda Trumpa, który zasłynął z częstego wykorzystywania tego zwrotu podczas wywiadów debat oraz wypowiedzi na mediach społecznościowych tj. Twitter. Do roku 2020 pojęcie “Fake News” zostało umieszczone w słownikach języka angielskiego takich jak “Oxford English Dictionary”, “Macmillan Dictionary”.

Rys. 1.1: Post udostępniony przez Donalda Trumpa na portalu Twitter Źródło: https://twitter.com/

Według założonego przez dziewięć organizacji w skład których wchodzą Google, Facebook oraz Twitter projektu “First Draft News” możemy wyróżnić siedem typów Fake Newsów [1]:

1. Satyra bądź parodia - nie ma na celu wyrządzić krzywdy ale może oszukać,
2. Fałszywe połączenie - nagłówki oraz obrazy nie mają powiązania z zawartością,
3. Myląca zawartość - tekst napisany w mylący sposób,
4. Fałszywy kontekst - prawdziwa zawartość powiązana ze złym kontekstem,
5. Oszukana zawartość - źródła pochodzenia informacji są fałszywe,
6. Zmanipulowana zawartość - prawdziwa zawartość zmanipulowana w odpowiedni sposób by oszukać odbiorcę,
7. Sfabrykowana zawartość - całkowicie zmyślona zawartość mająca na celu wyrządzić krzywdę.

Jak podaje słownik "Merriam Webster" po raz pierwszy wykorzystano zwrot "Fake news" w roku 1890.

Jednym z najsłynniejszych typów fake newsów jest propaganda, czyli według definicji "technika sterowania poglądami i zachowaniami ludzi polegająca na celowym, natarczywym, połączonym z manipulacją oddziaływaniu na zbiorowość" [4] pomimo iż najczęściej propaganda ma charakter polityczny nie jest to jedyne jej zastosowanie. Najstarszym przykładem pisemnej propagandy są opisy podbojów Dariusza Wielkiego datowane na rok 515 p.n.e. Od tego czasu w historii ludzkości można znaleźć wiele przypadków wykorzystania tego typu dezinformacji w takich krajach jak Starożytny Rzym, Niemcy podczas drugiej wojny światowej a nawet w dzisiejszych czasach Korea Północna. Propagandę można podzielić na 3 różne typy:

1. Biała propaganda - źródło pochodzenia informacji jest prawdziwe i podane,
2. Szara propaganda - źródło pochodzenia informacji jest dla odbiorcy nieznane i może się on jedynie domyślać,
3. Czarna propaganda - źródło pochodzenia informacji jest umyślnie sfałszowane w celu wyrządzenia szkody.

1.1. Sposoby rozprzestrzeniania fałszywych informacji

Wraz ze zmianami w sposobach rozprzestrzeniania informacji na świecie zmieniało się także podejście do tworzenia fake newsów w odpowiedni sposób oszukujących osoby, do których były one skierowane.

1.1.1. Gazety

Wykorzystanie fake newsów w gazetach miało głownie na celu przyciągnąć uwagę a co za tym idzie zwiększyć sprzedaż danej gazety. Stało się to na tyle popularne że spowodowało narodziny nowego pojęcia "żółtej prasy", czyli takiej, której działanie polega na zamieszczaniu w nagłówkach w pełni lub częściowo nieprawdziwych informacji aby przyciągnąć uwagę przechodnia za wszelką cenę, nawet jeśli wiąże się to z utratą wiarygodności.

Dziennikarz Frank Luther Mott wyróżnia 5 cech charakteryzujących żółtą prasę [7]:

- Napisane dużą czcionką straszące nagłówki na temat mniej ważnych wydarzeń,
- Nadmierna ilość zdjęć i rysunków,
- Zawarcie sfałszowanych wywiadów, mylących nagłowków, pseudonauki oraz nieprawdziwych informacji od ludzi podających się za ekspertów,
- Dodanie w pełni kolorowych dodatków do gazet w niedzielę,
- Stawianie siebie jako słabszego w walce przeciwko systemowi.

Rys. 1.2: Przykład żółtej prasy z roku 1993 Źródło: https://www.nytimes.com/

1.1.2. Telewizja

Wynalezienie telewizji na początku 20 wieku zmieniło całkowicie sposób w jaki ludzie pozyskiwali wiadomości ze świata. Aby pozyskać informacje na temat najnowszych wydarzeń nie było konieczne kupienie gazety a nawet wyjście z domu, w tym celu wystarczyło posiadać dostęp do telewizji i urządzenie do jej odbioru czyli telewizor. Połączenie zarówno obrazu jak i dzwięku zmusiło osoby chcące oszukać swoich odbiorów do stworzenia nowych technik pozwalających w wiarygodny sposób przedstawić kłamstwo.

Przykładem osoby, która w znakomity sposób wykorzystała siłę daną mu przez telewizję był Edward Bernays nazywany "Ojcem public relations". W roku 1929 został on zatrudniony aby wypromować papierosy firmy "Lucky Strike" stworzona przez niego reklama ukazywała kobiety kobiety palące papierosy podczas marszu. Ponieważ kobiety palące były uznawane w tamtych czasach tematem taboo autor reklamy nazwał ją w gazetach walką o prawa kobiet. Reklama ta spowodowała tak duże spopularyzowanie palenia papierosów, że własnie Edwardowi Bernays głownie przypisuje ich dużą sprzedaż przez kolejne lata. Był on także osobą dzięki której w dzisiejszych czasach diamenty są uznawane za symbol miłości po tym jak został zatrudniony do wypromowania diamentów firmy "De Beers" [6].

1.1.3. Internet

Pojawienie się Internetu wpłynęło na każdy element życia codziennego. Czynności takie jak komunikacja, rozrywka, a także rozprzestrzenianie informacji uległy zmianom tak wielkim, że ciężko wyobrazić sobie w jaki sposób działały one wcześniej.

Dzięki pojawieniu się takich portali społecznościowych jak Facebook, Twitter oraz Instagram każdy użytkownik Internetu może opowiedzieć o swoich przemyśleniach, wydarzeniach z życia każdej osobie zainteresowanej. Portale te pozwoliły nie tylko na przekazywanie informacji o sobie ale także opowiadanie o wydarzeniach ze świata przez wszystkie chętne osoby. Wraz z rozwojem Internetu stopniowo zwiększa się ilość osób czerpiących informacje na temat

wydarzeń właśnie z portali społecznościowych i do dnia dzisiejszego w Ameryce wynosi 68% dorosłych osób.

Udostępnienie każdej osobie możliwości wypowiedzenia się doprowadziło do sytuacji w której duża część informacji w internecie jest całkowicie fałszywa bądź w pewien sposób zmanipulowana poprzez osobę nieobiektywnie opisującą wydarzenia. Ogrom informacji można zauważyć na podstawie ilości potwierdzonych fake newsów związanych z wydarzeniami wokół wirusa COVID-19, których do 20-06-2020 jest aż 110 niektóre z nich to [2]:

- Szczepionka na koronawirusa jest ukrywana od marca,
- Aspiryna jest lekarstwem na COVID-19,
- Komary przenoszą koronawirusa,
- Kraje bez sieci 5G są wolne od koronawirusa ,
- Koronawirus nie zagraża uczestnikom zgromadzeń religijnych,
- Koronawirus to środek do zmniejszenia populacji Ziemi.

Osoby oszukane fake newsami grają istotną role w dalszym ich rozprzestrzenianiu poprzez udostępnianie informacji swoim znajomym lub rozmowy jest to cecha Internetu, która daje niespotykaną wcześniej efektywność oszukiwania dużej ilości ludzi w szybkim czasie.

Rys. 1.3: Przykład fałszywej informacji na portalu facebook Źródło: https://www.facebook.com/

Na powyższym obrazie ukazany jest post udostępniony na portalu Facebook z którego wynika, że jeżeli osoba jest w stanie wstrzymać oddech na 10 sekund to nie jest ona zarażona Sars-Cov2. Treść postu wskazywała że metoda miała być skuteczna według ekspertów z Japonii. Jak się później okazało informacja ta była całkowicie fałszywa nie powstrzymało to jednak ponad 2.4 tysiąca ludzi przed udostępnieniem jej swoim znajomym. [5]

1.2. Sposoby ochrony przed nieprawdziwymi informacjami

Rozwój tak potężnego narzędzia jak Internet spowodował że fałszywe informacje stały się poważnym zagrożeniem w dzisiejszym świecie. Aby zapobiec oszukaniu dużej ilości społeczeństwa znaleziono różne sposoby na ochronę przed nieprawdziwymi informacjami niektóre z nich to

1. IFLA "How to spot fake news"- jest to stworzona przez międzynarodową instytucję reprezentującą interesy bibliotekarzy i pracowników informacji o nazwie IFLA infografika przedstawiające listę rzeczy, które należy zrobić podejrzewając że jesteśmy oszukiwani.

Rys. 1.4: Infografika stworzona przez IFLA Źródło: https://www.ifla.org/

Czynności które są na niej zawarte to

- Sprawdzenie źródła informacji,
- Dokładne przeczytanie treści,
- Sprawdzenie autora,
- Analiza odnośników ,
- Sprawdzenie dat związanych ,
- Upewnienie się że informacja nie jest formą żartu,
- Obiektywna ocena informacji,
- Zapytanie ekspertów.

2. Strony Internetowe stworzone w celu walki z dezinformacją - istnieje wiele witryn gdzie eksperci sprawdzają nadesłane przez użytkowników informacje pod względem ich zgodności z prawdą. Przykładami takich portali są: *fakenews.pl, snopes.com, FactCheck.org, factchecker.in*
3. Grupy osób powołanych przez rząd do walki z fałszywymi informacjami - najlepszym przykładem takiej grupy są tzw. *Litewskie Elfy* jest to grupa ochotników, którzy w wolnym czasie przeglądają fora Internetowe oraz media społecznościowe sprawdzając znajdujące

się na nich informacje. W przypadku znalezienia nieprawdziwej informacji osoby te informują administratora strony o problemie co, najczęściej prowadzi do jego rozwiązania. Nazwa grupy wzięła się z faktu że walczą oni z grupami powszechnie zwanymi *trollami*. Grupa ta zyskała na popularności w takim stopniu że rozpoczęto organizację kolejnych oddziałow w innych krajach między innymi na Łotwie. [3]

Pomimo istnienia wielu sposobów ochrony przed fałszywymi informacjami ich ilość powoduje, że bardzo łatwo zostać oszukanym. Z tego powodu bardzo pomocne byłoby stworzenie oprogramowania pozwalającego na automatyczne rozpoznanie czy informacja jest prawdziwa. Implementacja takiego systemu w mediach społecznościowych jak *facebook, twitter* pozwoliłoby na usunięcie kłamstw zanim trafiłyby one przed oczy użytkowników. Popularne w ostatnich czasach algorytmy uczenia maszynowego *ML* oraz analizy języka naturalnego *NLP* osiągają zaskakująco dobrą efektywność w klasyfikacji róznego rodzaju tekstów więc ich wykorzystanie w rozwiązaniu takiego problemu mogłoby pozwolić na rozwiązanie go albo zmniejszenie.

1.3. Deep fake

Data pojawienia się technologii deep fake nie jest dokładnie znana jednak zaczęła zyskiwać na popularności w grudniu 2017 roku, kiedy użytkownik o pseudonimie “deepfakes” umieścił na portalu *Reddit* film pornograficzny w którym główną rolę grała aktorka Gal Gadot znana z filmu “Wonder Woman”. Aktorka jednak nigdy nie brała udziału w tego typu filmach a całe nagranie zostało stworzone wykorzystując algorytmy sztucznej inteligencji, która pozwoliła na zamianę twarzy dowolnego aktora z nagrania twarzą Gal Gadot w bardzo realistyczny sposób.

Sytuacja ta przyciągnęła zainteresowanie wielu użytkowników i już w styczniu 2018 roku pojawiła się aplikacja o nazwie “FakeApp” dzięki której każdy może zamienić twarze znajdujące się na nagraniu z kim tylko zechce. Powszechny dostęp oraz prostota w użytkowaniu powodują, że filmy “deep fake” stanowią niespotykane wcześniej zagrożenie mogąc oskarżyć dowolną osobę o wykonanie czynności, z którymi nie miała ona żadnego związku.

Aby zwrócić uwagę na niebezpieczeństwo firma buzzfeed w kwietniu 2018 roku umieściła na swoim kanale *Youtube* film, w którym Barack Obama opowiada o zagrożeniu płynącym z dezinformacji jednak słowa które wypowiada nie są naprawdę mowione przez niego, pochodzą one z nagrania aktora Jordana Peele, które następnie zostało podrobione techniką “deep fake”.

Rys. 1.5: Klatka z filmu firmy Buzzfeed Źródło: https://www.youtube.com/

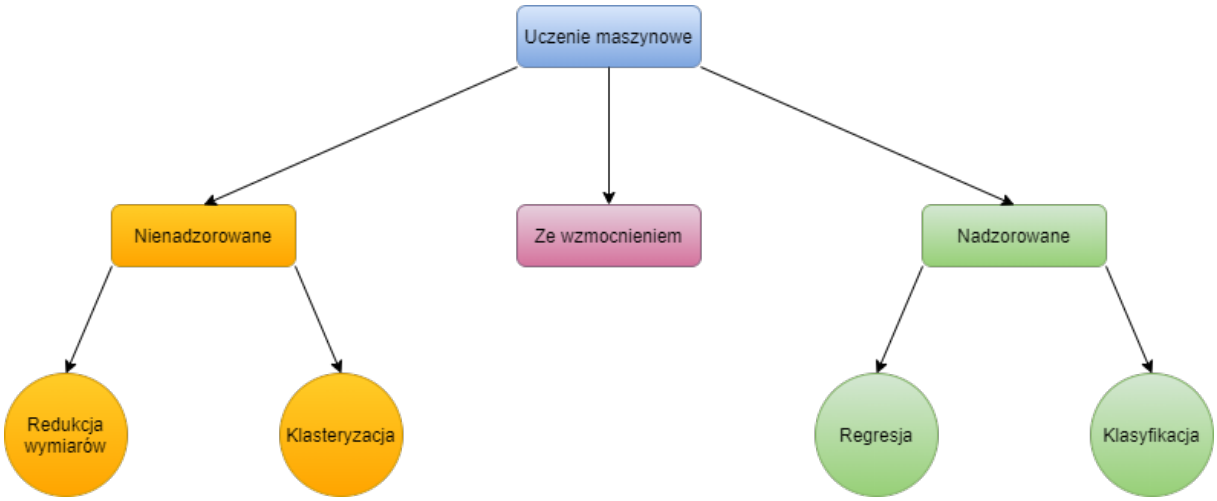
Filmy “deep fake” doprowadziły do sytuacji gdzie nie istnieje sposób przekazu, który daje stu procentową wiarygodność co w jeszcze większym stopniu zwiększa znaczenie odnalezienia uniwersalnej i efektywnej metody walki z dezinformacją.

Rozdział 2

Uczenie maszyn

Uczenie maszynowe jest dziedzina algorytmów komputerowych, które automatycznie poprawiają swoją efektywność poprzez doświadczenie. Określa się je jako poddziedzinę sztucznej inteligencji. Algorytmy te pozwalają na zbudowanie matematycznego modelu na podstawie przykładowych danych nazywanych danymi treningowymi co pozwala im na wykonywanie predykcji lub decyzji bez potrzeby ich dokładnego zaimplementowania przez programistę. Uczenie maszynowe stosuje się do wielu zadań jak filtrowanie poczty mailowej ze spamu, reklamy Internetowe, wykrywanie twarzy na zdjęciach oraz nagraniach a przyszłości może pomóc w stworzeniu takich technologii jak samojezdne samochody. Sam termin został spopularyzowany przez informatyka Arhura Samuela w roku 1959, był on autorem pierwszego działającego systemu tego typu, jego program automatycznie grał w warcaby i uczył się na podstawie poprzednich potyczek.

2.1. Rodzaje



Rys. 2.1: Rodzaje uczenia maszynowego Źródło: Własne

Algorytmy uczenia maszynowego można podzielić na trzy główne rodzaje w zależności od problemów, które mają one rozwiązywać są to:

1. **Uczenie nadzorowane** Jest to najczęściej wykorzystywany rodzaj uczenia maszynowego polega on na tym że maszyna uczy się na podstawie przykładów zawartych w danych treningowych uczenie nadzorowane można porównać do nauczyciela i ucznia gdzie dane

pełnią rolę nauczyciela a program ucznia. Algorytmy tego typu potrafią znaleźć odpowiednie zależności na podstawie etykiet przypisanym danym, które następnie wykorzystują w celu predykcji wcześniej nie analizowanych danych. Ważnym zagadnieniem w przypadku uczenia nadzorowanego jest tak zwany Overfitting polegający na przeuczeniu programu jednym zestawem treningowym przez co traci on umiejętność generalizacji problemu i nie jest w stanie poprawnie podejmować predykcji danych niewystarczająco podobnych do treningowych.
Przykładowe zastosowanie:

- Klasyfikacja - przewidywanie kategorii
  - rozpoznawanie elementów na zdjęciu
  - filtrowanie spamu w skrzynce mailowej
- Regresja - przewidywanie liczb
  - przewidywanie trendów finansowych lub ekonomicznych,
  - prognozowanie pogody

2. **Uczenie nienadzorowane** W przeciwieństwie do uczenia nadzorowanego uczenie nienadzorowane opiera się na braku nauczyciela a zadaniem maszyny jest znalezienie wzorców i zależności między analizowanymi obiektami samodzielnie. Dwie metody które są najczęściej wykorzystywane w uczeniu nienadzorowanym to:
   1. **Analiza składowych głownych** - Polega na zmniejszaniu wymiarowości danych poprzez odnajdywanie a następnie odrzucanie cech, które niosą ze sobą najmniejszą ilość informacji,
   2. **Analiza skupień (Klasteryzacja)** - Pozwala na identyfikację oraz grupowanie danych podobnych, które nie są w żaden sposób oznaczone. Może także pomóc w odnalezieniu anomalii nie pasujących do żadnej z wydzielonych grup.

   Wykorzystanie tego typu algorytmów pozwala na badanie danych nieoznaczonych, które są znacznie częściej spotykane niż dane oznaczone.
   Przykładowe zastosowanie:

   - Redukcja wymiarów
     - Wizualizacja danych “big data”
     - Kompresja danych
   - Klasteryzacja
     - Spersonalizowane reklamy
     - Systemy rekomendacyjne

3. **Uczenie ze wzmocnieniem** Uczenie ze wzmocnieniem polega na wykorzystaniu metody prób i błędów w taki sposób by maszyna została “nagrodzona” za wykonywanie czynności pożądanych oraz “karana” za popełnianie błędów. Sukces takiego systemu oparty jest na odpowiedniej implementacji systemu nagród, który może mieć całkowicie inne działanie w zależności od rozwiązywanego problemu. Ponieważ algorytmy uczenia ze wzmocnieniem dążą do zebrania jak największej ilości “nagrody” nie zawsze odnajdują one optymalne rozwiązanie.
   Przykładowe zastosowanie:

   - Tworzenie programów grających w gry
   - Samojezdne samochody

2.2. Algorytmy klasyfikacji

Z problemem klasyfikacji można się spotkać wszędzie tam, gdzie wykorzystując zbiór zmiennych objaśniających należy wskazać wartość przyjmowaną przez zmienną modelowaną. W problemach klasyfikacji zmienna modelowana może przyjmować wartości binarne (**klasyfikacja dwuklasowa**) lub jedną z wielu etykiet (**klasyfikacja wieloklasowa**). Aby wybrać odpowiedni do rozwiązywanego problemu algorytm klasyfikacji należy wziąć pod uwagę 4 czynniki:

- Złożoność czasowa - jak długo trwa uczenie oraz predykcja nowych danych,
- Interpretowalność - jak łatwo można wytłumaczyć decyzję podjętą przez system,
- Skalowalność - jak dużą ilość zasobów zużywa dany algorytm,
- Czynnik ludzki - Aby poprawnie ustawić parametry algorytmu potrzebna jest wiedza na temat jego działania. Ponieważ poprawne ich ustawienie może mieć większe znaczenie dla efektywności niż dobór najlepszego algorytmu często lepiej jest wykorzystać algorytm znany a nie optymalny.

2.2.1. KNN

Algorytm K najbliższych sąsiadów (**K nearest neighbours**) jest algorytmem stosowanym zarówno w problemach regresyjnych jak i klasyfikacji. Kroki jego działania można opisać w następujący sposób:

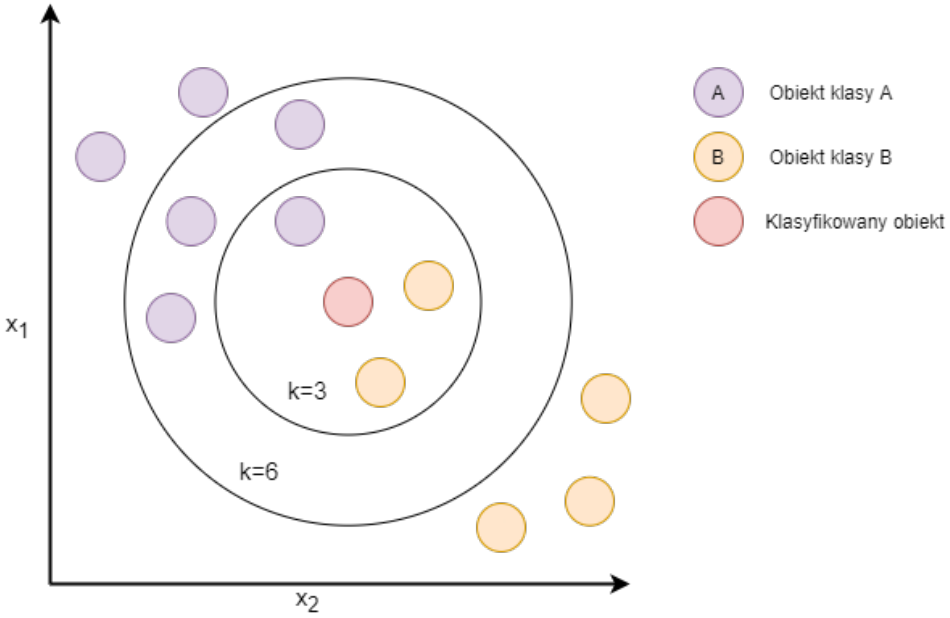
1. Umieszczenie wszystkich obiektów posiadających N cech jako punkty w N-wymiarowej przestrzeni
2. Obliczenie odległości między obiektem którego etykieta będzie przewidywana a każdym innym obiektem
3. Przypisanie do obiektu etykiety, którą posiada większość z K najbliższych obiektów.

Faza uczenia w KNN polega wyłącznie na wczytaniu danych treningowych do pamięci przez co jest ona bardzo szybka.

Najważniejszą kwestią w poprawnym implementowaniu algorytmu KNN jest odpowiednie wybranie liczby K, której optymalna wartość będzie się różnić w zależności od danych. W problemach klasyfikacji często wybiera się K o wartości nieparzystej aby uniknąć problemu remisu podczas zliczania sąsiadów. W przypadku problemów regresyjnych zamiast wykonywać głosowanie predykcję wykonuje się na podstawie średniej wartości K najbliższych sąsiadów.

Zalety	Wady
Prosty do zrozumienia i interpretacji	Przechowuje wszystkie dane treningowe w pamięci co skutkuje wysoką złożonością pamięciową
Szybkość fazy uczenia	Wrażliwy na ilość danych i nieistotne cechy
Uniwersalność można go wykorzystać zarówno w problemach regresyjnych jak i klasyfikcyjnych	Kosztowny obliczeniowo
Nie wykonuje generalizacji danych przez co jest efektywny w przypadku danych nieliniowych	

Tab. 2.1: Wady i zalety KNN



Rys. 2.2: Graficzne przedstawienie algorytmu KNN Źródło: Własne

2.2.2. SVM

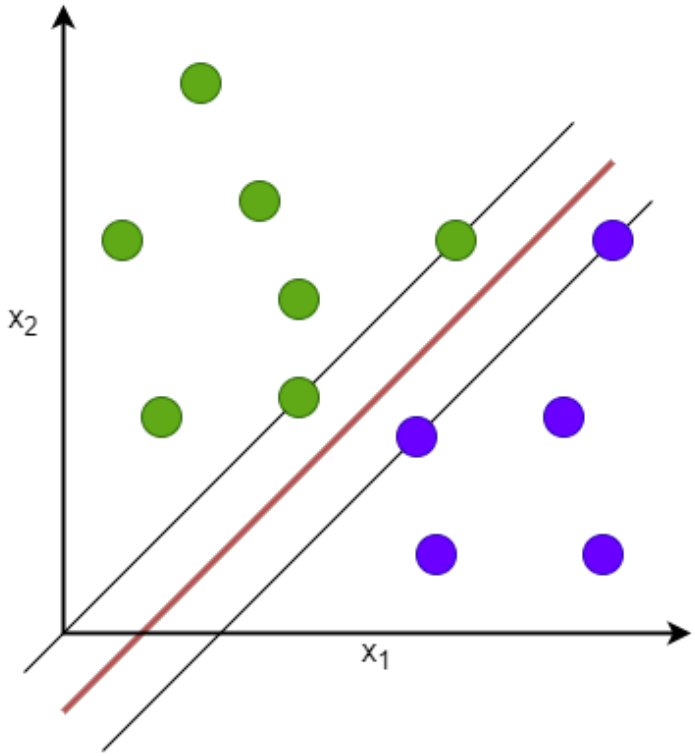
Klasyfikator SVM (**Support vector machines**) również można wykorzystywać zarówno w problemach regresyjnych i klasyfikcyjnych. Służą one do klasyfikacji binarnej, co oznacza że obiekty należy podzielić na dokładnie dwie klasy. SVM polega na znalezieniu takiej prostej lub płaszczyzny (w zależności od liczby cech), która w jak najlepszy sposób dzieli obiekty dwóch klas. Czasami idealny podział jest niemożliwy z czym klasyfikator ten radzi sobie w jeden z dwóch sposobów:

1. Ignorowanie punktów które uniemożliwiają podział,
2. Wykorzystanie tak zwanych kernel trick, które przekształcają dane do wyższego wymiaru w którym podział jest możliwy.

Aby odnaleźć optymalne rozwiązanie SVM skupiają się na punktach jak najbardziej skrajnych obu klas są one nazywane **wektorami nośnymi** zmiana ich położenia lub usunięcie całkowicie zmienia położenie płaszczyzny.

Zalety	Wady
Efektywny w wielowymiarowych przestrzeniach	Nie jest on przystosowany do dużej ilości danych treningowych
Działa nawet w przypadku gdzie liczba danych treningowych jest mniejsza od ilości ich cech	Nie radzi sobie dobrze jeżeli obiekty różnych klas nachodzą na siebie
Przechowuje w pamięci tylko niewielką ilość danych treningowych	Trudne do zinterpretowania wyniki predykcji

Tab. 2.2: Wady i zalety SVM



Rys. 2.3: Wizualizacja algorytmu SVM Źródło: Własne

2.2.3. MLP

Działanie algorytmu Perceptronu wielowarstwowego (**Multi layered perceptron**) opiera się na wykorzystaniu modelu sztucznego neuronu, który na podstawie określonej funkcji aktywacji oblicza na wyjściu pewną wartość na podstawie ważonych sum danych wejściowych.

Funkcje aktywacji dzieli się na:

- **Funkcje progowe** z wyjściem binarnym
- **Funkcje liniowe** z wyjściem ciągłym
- **Funkcje nieliniowe** z wyjściem ciągłym

W sieciach wielowarstwowych najczęściej wykorzystuje się funkcje nieliniowe poniważ wykazują one najwieksze zdolności do nauki.

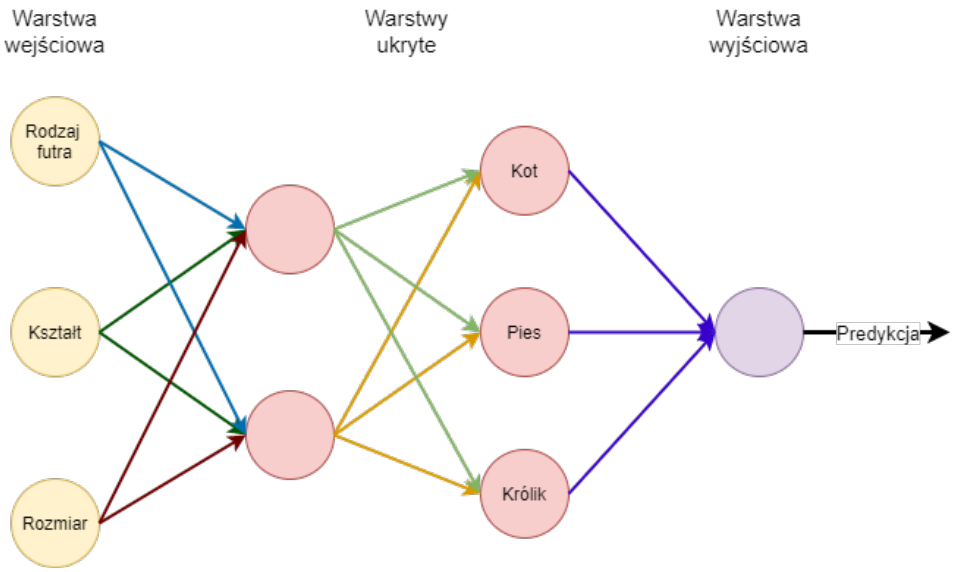
Perceptron wielowarstwowy składa się z trzech warstw sztucznych neuronów:

1. **Warstwy wejściowej** - są to neurony zapewniające całej sieci informacje czyli zazwyczaj dane treningowe nie wykonują one żadnych obliczeń a jedynie przekazują informacje do kolejnych warstw.
2. **Warstwy ukrytej** - wykonują one odpowiednie obliczenia zmieniając i przekazując informacje z warstwy wejściowej do wyjściowej. Sieć może składać się z dowolnej liczby warstw ukrytych.
3. **Warstwy wyjśćiowej** - wykonują one ostatnie obliczenia na danych a następnie zwracają wynik.

Uczenie takiej sieci polega na takiej modyfikacji wag na wejściu neuronów aby poprawić wyniki klasyfikacji. Podczas gdy perceptron jednowarstwowy czyli nieposiadający żadnej warstwy ukrytej może nauczyć się wyłącznie funkcji liniowych perceptron wielowarstwowy pozwala na nauczenie się skomplikowanych funkcji nieliniowych.

Zalety	Wady
Szybkie wykonywanie predykcji po nauczeniu modelu	Brak możliwości interpretacji wyniku predykcji
Efektywny dla nieliniowych danych posiadających wiele cech takich jak zdjęcia	Uczenie bardzo złożone obliczeniowo co skutkuje długim czasem uczenia
Działają najlepiej dla dużej ilości danych treningowych	Użytkownik ma niewielki wpływ na działanie sieci

Tab. 2.3: Wady i zalety MLP



Rys. 2.4: Graficzne przedstawienie perceptronu wielowarstwowego Źródło: Własne

2.2.4. Decision trees

Algorytm Decision trees polega na stworzeniu drzewa decyzyjnego w którym korzeń i węzły są poszczególnymi cechami zbioru danych a liście odpowiadają klasom, które tym danym należy przypisać, ostatnim elementem są krawędzie którymi reprezentuje się wartości cech. Kroki algorytmu są następujące:

1. Wybranie najlepszej cechy
2. Dodanie gałęzi odpowiadającym poszeczególnym wartościom wybranej cechy
3. Podział zbioru danych na podzbiory zgodnie z wartościami cechy
4. Jeżeli wszystkie elementy pozdzbiorów należą do tej samej klasy zakończenie gałęzi liściem. W przeciwnym wypadku powtórzenie kroków od 1 do 4 dla każdego podzbioru

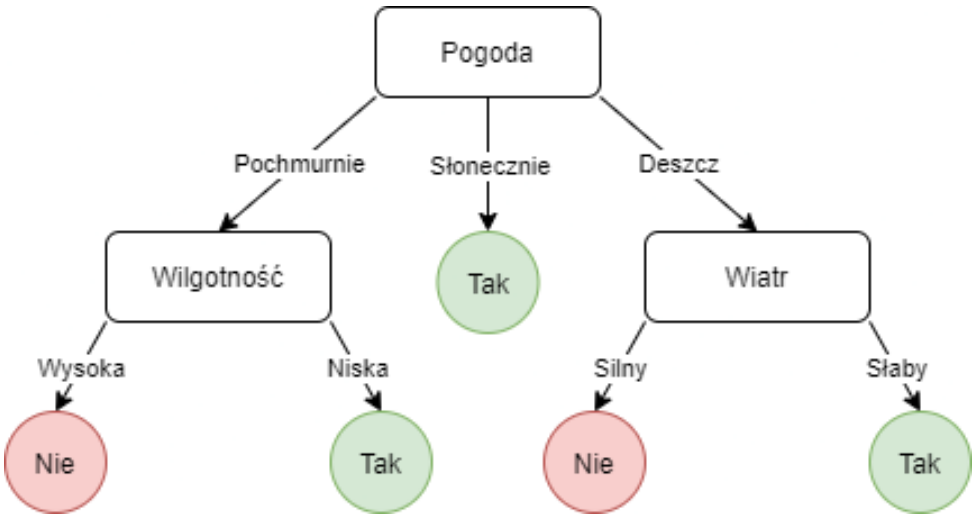
Aby dokonać optymalnego podziału algorytm musi wybrać cechy powodujące jak najlepsze podzielenie różnych klas a więc niosące ze sobą największą ilośc informacji, wykorzystuje się w tym celu entropię, której wartość pozwala określić średnią ilość informacji niesioną przez poszczególne cechy.

$$H(x) = -\sum_{i=1}^{n} p(x_i) \log_2 p(x_i). \tag{2.1}$$

Wzór na entropię 2.1, gdzie $p(x_i)$ jest prawdopodobieństwem wybrania klasy x_i

Zalety	Wady
Łatwe do zinterpretowania wyniki bardzo wygodne do zwizualizowania	Wrażliwy na przeuczenie
Automatyczna selekcja ważnych cech. Obecność cech zbędnych nie ma wpływu na efektywność	Potrzeba stworzenia nowego drzewa podczas dodawania nowych danych
Szybkie wykonywanie predykcji	Małe zmiany w danych treningowych maja duży wpływ na predykcję

Tab. 2.4: Wady i zalety Drzew decyzyjnych



Rys. 2.5: Przykładowe drzewo decyzyjne podejmujące decyzję czy wyjść na zewnątrz Źródło: Własne

2.2.5. Naive Bayes

Naiwny klasyfikator Bayesowski (**Naive Bayes**) jest probabilistycznym klasyfikatorem, który zakłada wzajemną niezależność wszystkich cech stąd naiwność algorytmu. Stworzony przez niego model opiera się na obliczeniu prawdopodobieństw wystąpienia poszczególnych cech pod warunkiem że występują one w danej klasie. Pozwala to przy użyciu twierdzenia Bayesa 2.2 obliczyć prawdopodobieństwa warunkowe określające z jak duża pewnością występująca kombinacja cech prowadzi do wystąpienia róznych klas na podstawie czego algorytm wykonuje predykcję.

$$P(A|B) = \frac{P(A|B)P(A)}{P(B)} \tag{2.2}$$

gdzie P(A) i P(B) prawdopodobieństwa wystąpienia zdarzeń A i B, P(A|B) prawdopodobieństwo wystąpienia zdarzenia A pod warunkiem wystąpienia zdarzenia B.

Algorytm ten jest często wykorzystywany w systemach czasu rzeczywistego, ponieważ działa bardzo szybko jednak jego efektywność w porównaniu do bardziej skomplikowanych algorytmów jest mniejsza.

Zalety	Wady
- Bardzo szybki nawet w przypadku przewidywania wieloklasowego	Jeżeli pewna cecha nie pojawi się w danych treningowych a istnieje w danych testowych algorytm nie będzie w stanie wykonać jego predykcji
Działa nawet dla małej ilości danych	Przyjmuje niezależność cech co w rzeczywistości jest prawie niespotykane
Dobra efektywność w przypadku klasyfikacji danych wielowymiarowych jak na przykład w przypadku klasyfikacji tekstu	Mniejsza efektywność w porównaniu z bardziej skomplikowanymi algorytmami

Tab. 2.5: Wady i zalety Naive Bayes

2.3. Zagrożenia

Ogromny rozwój uczenia maszynowego i sztucznej inteligencji spowodował że znalazły one wiele zastosowań w różnych dziedzinach życia codziennego. Powoduje to, że wiele osób skupia się na pozytywach związanych z tymi technologiami, jednakże niektórzy dostrzegają ogromne zagrożenia płynące z niewłaściwego wykorzystania ich i katastrofalne skutki, do których mogą doprowadzić. Wśród takich osób znajdują się naukowcy tacy jak jak Stephen Hawking, Elon Musk, Steve Wozniak i Bill Gates. Według Elona Muska sztuczna inteligencja niesie ze sobą większe zagrożenie niż bomby atomowe.

Dwa głowne sposoby w jaki sztuczna inteligencja może powodować zagrożenie to:

1. **AI zaprogramowane do czynienia krzywdy** czyli na przykład wykorzystanie uczenia maszynowego w takich narzędziach jak broń, która w niewłaściwych rękach stanowiłaby zagrożenie dla całej ludzkości.
2. **AI zaprogramowane w dobrych celach, jednak odnajdujące krzywdzący innych optymalny sposób osiągnięcia sukcesu** może to być związane ze złym przekazaniem celu jaki maszyna ma osiągnąć przez programistę. Przykładowo mogłby to być samojezdny samochód, który podczas jazdy zwraca uwagę tylko i wyłącznie na to by jak najszybciej dojechać do celu podróży. Samochód taki nie przestrzegał by istniejących praw drogowych ponieważ ograniczałyby one jego prędkość.

Przykłady te pokazują że zagrożenie nie płynie z sytuacji gdzie maszyny odwracają się przeciwko ludzkości, a z sytuacji w której maszyny wykonują poświęcone im zadania za dobrze i nie w sposób przewidziany przez stwórcę systemu.

Istotnym problemem jest także bezrobocie będące skutkiem zastępowania ludzi maszynami, ponieważ są one znacznie tańsze w utrzymaniu niż pracownicy a z wykorzystaniem uczenia maszynowego mogą one opanować niektóre zadania lepiej od ludzi. Rozwiązaniem na takie problemy mogłyby być obowiązkowo przestrzegane zasady bezpiecznego korzystania ze sztucznej inteligencji, których złamanie wiązałoby się z odpowiedzialnością karną.

Rozdział 3

Przetwarzanie języków naturalnych

Systemy przetwarzania języków naturalnych (**Natural language processing**) nazywane w skrócie NLP, oznaczają systemy mogące zrozumieć mowę i pismo ludzkie w takiej formie jaką ludzie posługują się na codzień. Programy takie mogą wykonywać zadania od zliczania częstotliwości występowania danego słowa w tekście do automatycznego pisania artykułów.

Głownym problemem w tworzeniu tego typu systemów jest to że do komunikacji z komputerem zazwyczaj potrzebne jest posługiwanie się precyzyjnymi komendami danego języka programowania, mowa ludzka jednak nie zawsze jest precyzyjna i jej znaczenie może różnić się w zależności od kontekstu czy różnego rodzaju regionalnych dialektów. Systemy NLP są często wykorzystywane w takich celach jak:

- Asystenci głosowi na przykład (*Siri, Alexa, Cortana*) - są to urządzenia, które wykonują komendy wypowiedziane w ich kierunku przez użytkownika.
- Wyodrębnianie ważnych informacji z tekstu w celu późniejszej analizy.
- Analiza sentymentu czyli wnioskowanie na podstawie tekstu opinii użytkowników na dany temat.
- Sprawdzanie błędów ortograficznych
- Tłumaczenie tekstu na inne języki
- Chatboty wykorzystywane przez wiele firm w celu posiadania całodobowej zautomatyzowanej obsługi klienta.

Rozwój NLP ma bardzo duże znaczenie dla osób niepełnosprawnych, które często tylko dzięki ich pomocy są w stanie nawiązać interakcję z technologią pozwalającą im na znaczne podniesienie jakości życia.

3.1. Przygotowywanie danych tekstowych

Aby ułatwić analizę ogromnej ilości danych tekstowych potrzebnych do poprawnego nauczenia systemu NLP, wykonuje się na nich różnego rodzaju operacje. Operacje te powodują że tekst nie traci swoich najważniejszych cech natomiast znacznie zmniejsza się moc obliczeniowa potrzebna do nauki algorytmów uczących wykonywanych na nim. Najczęściej wykorzystywanymi tego typu operacjami są:

- Zamiana wszystkich dużych liter na małe
- Usunięcie znaków specjalnych
- Usunięcie tak zwanych “Stop words” - są to bardzo często występujące w danym języku słowa, które zazwyczaj nie wnoszą istotnych dla analizy informacji

- Tokenizacja - polega na podziale tekstu na mniejsze częsci zwane tokenami. W przypadku dużych bloków tekstu może to być podział na zdania a w przypadku zdań podział na słowa itd..
- Lematyzacja - oznacza ona sprowadzenie grupy wyrazów stanowiących odmianę danego zwrotu do wspólnej postaci, pozwala to na traktowanie ich jako to samo słowo.
- Stemming - jest to proces usunięcia końcówki fleksyjnej pozostawiając tylko tematu czyli nośnika znaczenia wyrazu.

Wykonanie wybranych operacji na tekscie daje na wyjściu skrócone dane tekstowe, które można następnie przeanalizować lub wykonać na nich wektoryzację co pozwala na wykorzystanie ich w różnych algorytmach uczenia maszynowego.

3.2. Wektoryzacja

Ponieważ algorytmy sztucznej inteligencji potrafią uczyć się tylko z danych przedstawionych w formie numerycznej, aby móc wykorzystać je w kombinacji z NLP potrzebny jest pewien sposób zamiany formy tekstowej na liczbową, tak by straciły ona jak najmniej swoich najważniejszych cech a zarazem były czytelne dla maszyny.

Operacje takie nazywa się wektoryzacją tekstu, ponieważ reprezentacja liczbowa będąca wynikiem ich to najczęściej wektory. Wybór poprawnego sposobu zamiany tekstu może mieć ogromne znaczenie dla efektywności nauczonego modelu.

Jedne z najpopularniejszych metod wektoryzacji to:

- **Bag of words** jest to metoda wektoryzacji, w której każdemu unikalnemu symbolowi z tekstu przypisuje się liczbę odpowiadającej jego ilości w analizowanym tekście. Symbolami mogą być całe zdania, słowa bądź NGramy. Metoda ta w żaden sposób nie zachowuje informacji o porządku ani kontekście występujących symboli a jedynie o częstotliwości ich występowania. Wektor wynikowy metody "Bag of words" ma wymiar równy liczbie unikalnych symboli w tekście, przez co przy analizie dużej ilości dokumentów ma dużą złożoność pamięciową aby naprawić ten problem najpierw na danych wykonuje się metody przygotowywania danych tekstowych.
- TFIDF (ang. **Term frequency inverse document frequency**) jest to metoda przypisująca, każdemu symbolowi jego wagę w kontekście analizowanych dokumentów. Aby obliczyć tę wagę potrzebuje ona dwa elementy:
  1. Częstość symboli - którą uzyskuje się dzieląc liczbę wystąpień symbolu przez liczbę symboli w całym dokumencie
  2. Odwrotną częstość w dokumentach

  Po obliczeniu obu tych wartości oblicza się tzw. TFIDF score, której wartość określa jak ważny jest dany symbol dla dokumentu w kontekście wykonywanej analizy. Wynik ten oblicza się na podstawie wzoru 3.1, gdzie
  - x - symbol
  - y - dokument
  - TF - częstość symbolu
  - N - ilość wszystkich dokumentów
  - df - ilość dokumentów w których pojawił się symbol x

$$TFIDF_{x,y} = TF_{x,y} * \log \frac{N}{df_x} \tag{3.1}$$

Rozdział 4

Projekt i implementacja systemu

4.1. Wykorzystane technologie

4.2. Wymagania funkcjonalne

4.3. Implementacja

Rozdział 5

Ocena eksperymentalna

5.1. Cel Badań

5.2. Warunki przeprowadzonego eksperymentu

5.3. Wyniki

5.4. Analiza wyników wraz z oceną statystyczną

5.5. Wnioski z badań

Rozdział 6

Podsumowanie

Literatura

[1] Fake news. it's complicated. https://firstdraftnews.org/latest/fake-news-complicated/. Dostęp dnia: 15-06-2020.

[2] Krótki przewodnik po fake newsach o koronawirusie. https://www.cyberdefence24.pl/krotki-przewodnik-po-aktualnych-fake-newsach-o-koronawirusie. Dostęp dnia: 15-06-2020.

[3] Obrona przestrzeni informacyjnej na przykładzie litwy, Łotwy i estonii. https://warsawinstitute.org/pl/obrona-przestrzeni-informacyjnej-na-przykladzie-litwy-lotwy-estonii/. Dostęp dnia: 15-06-2020.

[4] Słownik języka polskiego. https://sjp.pwn.pl/. Dostęp dnia: 15-06-2020.

[5] Wstrzymaniem oddechu nie sprawdzisz, czy masz koronawirusa. https://demagog.org.pl/analizy_i_raporty/wstrzymaniem-oddechu-nie-sprawdzisz-czy-masz-koronawirusa/. Dostęp dnia: 15-06-2020.

[6] M. Dice. *The True Story of Fake News: How Mainstream Media Manipulates Millions*. The Resistance Manifesto, 2017.

[7] F. L. Mott. *American Journalism*. Macmillan, 1941.

Dodatek A

Opis załączonej płyty CD/DVD

Na załączonej płycie znajduje się niniejsza praca w formacie PDF oraz pliki z kodem źródłowym aplikacji wykorzystanej do wykonania badań.