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

Spis tabel

Streszczenie

Wstęp

0.1. Motywacja

Rozdział 1

Informacje nieprawdziwe w dobie Internetu

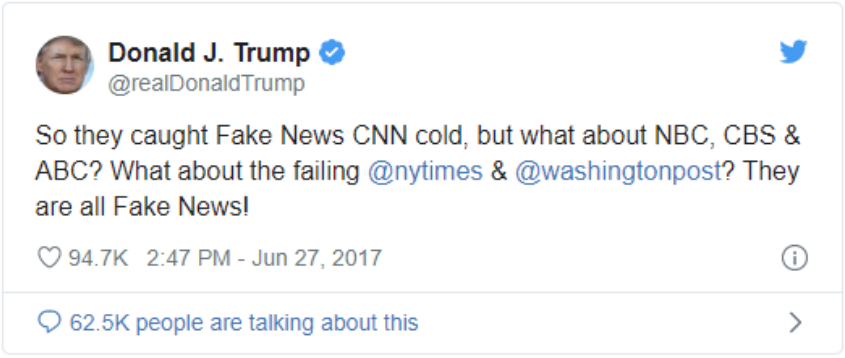
Pojęcie “Fake news” odnosi się do informacji, które pomimo że nie posiadają pokrycia z rzeczywistością są przedstawiane jako prawdziwe w mediach takich jak np.: wiadomości, artykuły, portale społecznościowe itd.. Zwrot ten jest neologizmem i w języku angielskim oznacza dosłownie “Fałszywe wiadomości”. 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.

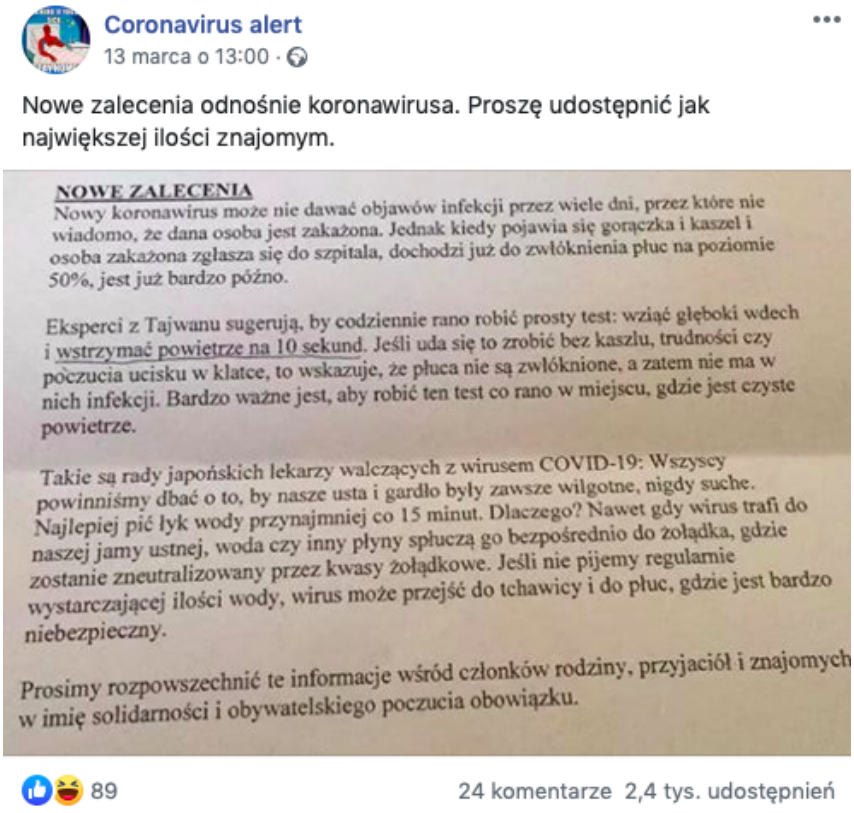
Dzięki pojawieniu się takich portali społecznościowych jak Facebook, Twitter oraz Instagram każdy użytkownik Internetu może opowiedzieć o swoich przemyśleniach, wydarzeniach z życia każdej osobie zainteresowanej. Portale te pozwoliły nie tylko na przekazywanie informacji o sobie ale także opowiadanie 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.



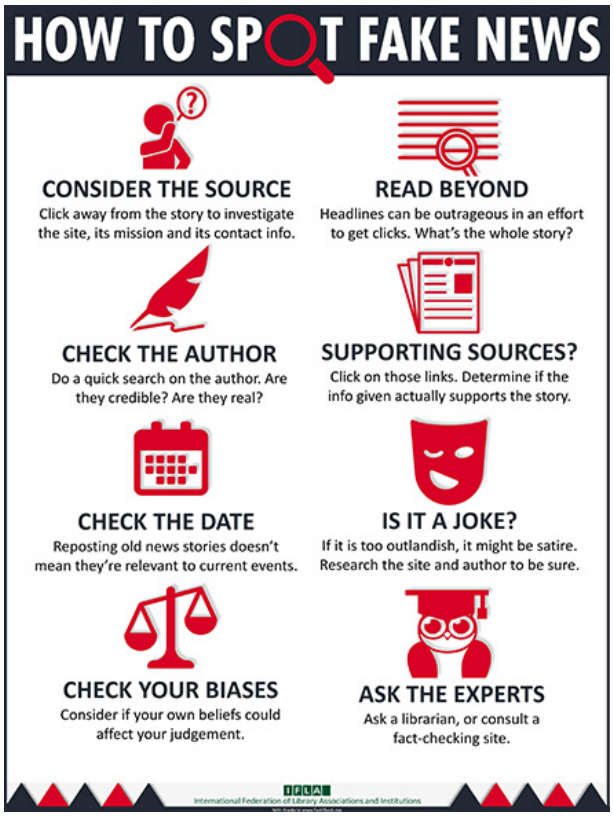
Rys. 1.3: Przykład fałszywej informacji na portalu facebook Źródło: https://www.facebook.com/

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]

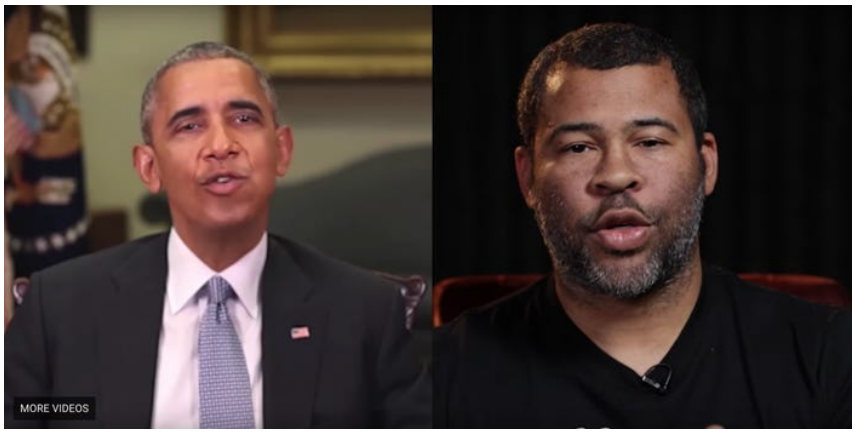
Pomimo istnienia wielu sposobów ochrony przed fałszywymi informacjami ich ilość powoduje, że bardzo łatwo zostać oszukanym. Z tego powodu bardzo pomocne byłoby stworzenie oprogramowania pozwalającego na automatyczne rozpoznanie czy informacja jest prawdziwa. Implementacja takiego systemu w mediach społecznościowych 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

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. Algorytmy tego typu można podzielić na dwa rodzaje:
   - 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. Wykorzystanie tego typu algorytmów pozwala na badanie danych nieoznaczonych, które są znacznie częściej spotykane niż dane oznaczone.
   - l
3. **Uczenie ze wzmocnieniem** Uczenie ze wzmocnieniem polega na wykorzystaniu metody prób i błędów w sposób gdzie maszyna zostaje "nagrodzona" za wykonywanie czynności pożądanych lub "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 poszukują dążą do zebrania jak największej ilości "nagrody" nie zawsze znajdują one optymalne rozwiązanie.
   - l

2.2. Algorytmy klasyfikacji

2.2.1. KNN

2.2.2. SVC

2.2.3. MLP

2.2.4. Binary trees

2.2.5. Naive Bayes

2.3. Wykorzystanie

2.4. Zagrożenia

Rozdział 3

Przetwarzanie języków naturalnych

3.1. Normalizacja danych tekstowych

3.2. Wektoryzacja

3.2.1. Bag of words

3.2.2. TfIDF

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.