

**Ewolucja metod badawczych i wykorzystywanych danych w badaniach z
dyscypliny ekonomia i finanse w obliczu rozwoju nowych technologii i
pojawiających się zjawisk w gospodarce światowej**

Aneta Hryckiewicz-Gontarczyk¹
Akademia Leona Koźmińskiego

Malwina Rzepka²
Akademia Leona Koźmińskiego

Streszczenie:

W ciągu ostatnich kilku dekad pojawiły się nowe sposoby wytwarzania, przechowywania i analizowania danych, które w połączeniu z nowymi technikami obliczeniowymi i algorytmicznymi powinny się przyczyniać do większej precyzji modelowania. Jednakże wyniki ostatnich badań z dziedziny ekonomii i finansów wskazują, że o ile wykorzystanie nowych technologii przyczynia się do większej precyzji estymacji w porównaniu do tradycyjnych modeli ekonometrycznych, to pojawia się coraz więcej przypadków, kiedy modele algorytmiczne generowały wyniki obarczone wysokim błędem statystycznym. Celem artykułu jest przegląd zastosowania nowoczesnych metodologii w badaniach z dziedziny ekonomii i finansów, jak również ocena wiarygodności zastosowanych modeli.

Słowa kluczowe: big data, algorytmika, finanse, ekonomia, badania

¹ Ekonomiczny Zakład Analiz Empirycznych, Akademia Leona Koźmińskiego, ul. Jagiellońska 57-59, 03-301 Warszawa, ahryckiewicz@alk.edu.pl

² Ekonomiczny Zakład Analiz Empirycznych, Akademia Leona Koźmińskiego, ul. Jagiellońska 57-59, 03-301 Warszawa

Abstract:

The last few decades have witnessed the creation of novel ways to produce, store, and analyse data, culminating in the emergence of the field of data science, which brings together computational, algorithmic, statistical and mathematical techniques towards extrapolating knowledge from big data. However, the recent research results document that while new models generate a more precise results as compared to the traditional econometrical models, there are also cases where the algorithmic techniques generate biased results. The aim of this article is to review the new methodologies which have been recently employed in the research in the field of economics and finance and assessment of their reliability.

Wstęp

Rewolucję przemysłową można zdefiniować jako szybką, gwałtowną i wielką zmianę w ekonomii napędzaną przekształceniem w metodzie i rodzaju podjętej pracy. Podczas pierwszej rewolucji przemysłowej nastąpiło zjawisko urbanizacji. Przejście od życia wiejskiego do społeczeństw miejskich, uprzemysłowionych gospodarstw. Moc parowa i postęp w rozwoju maszyn przekształcił zarówno rolnictwo jak i cały przemysł rzemieślniczy. Na skutek tego nastąpiła zasadnicza zmiana podstaw ówczesnej gospodarki. Druga rewolucja przemysłowa napędzana węglem, parą, żelazem pobudziła gwałtowny rozwój w obszarach energetycznych, naftowych i stalowych. Był to okres rozkwitu wielkich korporacji i globalizacji. Trzecia rewolucja jest powszechnie uznawana za rozpoczętą pod koniec dwudziestego wieku w 1969 roku³. Stanowiła podstawę przejścia od technologii mechanicznej do analogowej. Rozwój usprawnień w komunikacji, komputerach i wprowadzenie internetu stanowią kluczowe determinanty zmiany w tym okresie. Każda rewolucja wpłynęła na społeczeństwo oraz na badania naukowe w swój własny sposób, kierując się własną naturą postępu. Obecnie jesteśmy świadkami i czynnymi uczestnikami kolejnej rewolucji przemysłowej, która została przewidziana i określona przez Klausa Schwaba. Czwarta rewolucja przemysłowa, często określana też jako przemysł 4.0 została po raz pierwszy przedstawiona w artykule opublikowanym w 2015 roku, a później opisana w książce Schwaba (2015). Jak sam autor napisał „Stoimy u progu rewolucji technologicznej, która fundamentalnie zmieni sposób jaki żyjemy, pracujemy i odnosimy się do siebie nawzajem. W swojej skali, zakresie i złożoności, transformacja ta nie będzie podobna do niczego, czego ludzkość doświadczyła wcześniej...”⁴. Istnieje kilka powodów, dla których dzisiejsze przemiany nie są tylko przedłużeniem poprzedniej rewolucji, a raczej nadejściem kolejnej ery. Ery różniące się pod względem szybkości, skali i wpływu systemów. Prędkość z jaką doświadczamy zmiany nie ma historycznego przełożenia. Czwarta rewolucja rozwija się raczej wykładniczo niż liniowo. Rozmiar tych zmian zwiastuje transformację całych systemów na poziomie produkcji, zarządzania i analityki. Nadchodzą przełomy technologiczne w dziedzinach takich jak sztuczna inteligencja, robotyka, Internet rzeczy, pojazdy autonomiczne, druk 3D, nanotechnologia, obliczenia kwantowe, Big Data i wiele

³ Brian H. Roberts, (2015), The third Industrial Revolution: Implications for Planning Cities and Regions, ResearchGate

⁴ K. Schwab (2015), The Fourth Industrial Revolution What It Means and How to Respond. (<https://www.weforum.org/agenda/2016/01/the-fourth-industrial-revolution-what-it-means-and-how-to-respond>, pobrano 08.11.2022)

innych. Wszystkie te technologie leżące u podstaw rewolucji digitalizacji skorzystały z trzech potężnych trendów w komunikacji, procesowaniu informacji i przetwarzaniu danych⁵.

To właśnie ostatnia rewolucja stwarza ogromne możliwości dla naukowców, że ich modele będą dużo bardziej dokładne a dostęp do danych pozwoli na dużo wcześniejsze przewidywania negatywnych zdarzeń w różnych sektorach gospodarki. W ciągu ostatnich kilku dekad pojawiły się nowe sposoby wytwarzania, przechowywania i analizowania danych, które w połączeniu z nowymi technikami obliczeniowymi i algorytmicznymi pozwalają na dużo większą ekstrapolację wiedzy. Jednocześnie ruch otwartych danych — wyłaniający się z trendów politycznych, takich jak dążenie do otwartej nauki — zachęca do udostępniania i łączenia heterogenicznych danych badawczych za pośrednictwem dużych infrastruktur cyfrowych. Dostępność ogromnych ilości danych w formatach do odczytu maszynowego stanowi zachętę do tworzenia wydajnych procedur gromadzenia, organizowania, wizualizacji i modelowania tych danych. Infrastruktury te z kolei służą jako platformy do rozwoju sztucznej inteligencji, mającej na celu zwiększenie niezawodności, szybkości i przejrzystości procesów tworzenia wiedzy. Badacze ze wszystkich dyscyplin dążą do wykorzystania nowej rewolucji do łączenia i porównywania danych z różnych źródeł jako poprawę dokładności i mocy predykcyjnej odkryć naukowych oraz pomoc w określaniu przyszłych kierunków badań, co ostatecznie zapewnia nowy punkt wyjścia dla badań empirycznych.

Czy jednak nowa rewolucja pozwoli nam „zrewolucjonizować” naukę i nie tylko dokonywać nowych odkryć, ale również poprawić modele, które do tej pory były obciążone wysokim błędem z racji braku dostępu do danych, niskiej jakości danych, czy też krótkich szeregów czasowych. Czy szeroko pojęte uczenie maszynowe pozwoli nam na poprawę precyzji estymacji i złagodzenie założeń, które często warunkowały prawidłowość modeli regresji? I w końcu w jakim stopniu różne dyscypliny, a przede wszystkim dyscyplina ekonomia i finanse mogą stać się beneficjentami nowej rewolucji. Odpowiedź na te pytania wciąż jest niejasna. Chociaż literatura akademicka dowodzi, że sztuczna inteligencja i techniki uczenia maszynowego mogą oferować wyższą precyzję w szacunkach, istnieje cały szereg badań, które pokazują, że nowe technologie są również obciążone pewnymi wadami, co skutkuje błędami we wnioskowaniu.

⁵ World Trade Organization (2018), The Future of World trade: How digital technologies are transforming global commerce, s. 24. (https://www.wto.org/english/res_e/publications_e/world_trade_report18_e.pdf pobrano: 08.11.2022)

Na przykład istnieją naukowe dowody wskazujące na to, że algorytmy sterujące asystentem wirtualnym firmy Amazon - Alexą lub Siri należącym do firmy Apple charakteryzują się dyskryminacją. Badania pokazują, że podczas gdy biały Amerykanin ma 92% wskaźnik dokładności, jeśli chodzi o zrozumienie przez asystenta głosowego, to już biała Amerykanka posiada ten wskaźnik na poziomie 79%. Natomiast Amerykanka rasy mieszanej ma tylko 69% szans na zrozumienie jej przez Alexę. Jednym z powodów różnic w rozpoznawaniu głosów jest dostarczenie stronniczych danych wsadowych, co przypisuje się lokalizacji siedziby tych firm (Amazon - Alexa, Google - Home i Apple - Siri - wszyscy mają siedzibę w Kalifornii) i dominującej w tym regionie grupie etnicznej i męskiej płci. Dane wskazują, że zaledwie 13,5% pracowników w dziedzinie uczenia maszynowego to kobiety, a zbiory danych wykorzystywane do budowania algorytmów są dużo bardziej skoncentrowane na białych mężczyznach.⁶

Podobne przypadki dyskryminacji płciowej zauważa się w komputerowych systemach wizyjnych, które zgłaszały wyższy poziom błędów w rozpoznawaniu kobiet, szczególnie tych o ciemniejszych odcieniach skóry aniżeli mężczyzn. Co raz większa część badań dokumentuje błędy w decyzjach algorytmicznych w odniesieniu do płci, koloru skóry, danych demograficznych, pochodzenia etnicznego lub narodowości. Wciąż jednak nie jest do końca wiadomo, jaka jest przyczyna takiego stanu. Co ważniejsze, większość badań pochodzi z sektora HR, marketingu, mediów społecznościowych, opieki zdrowotnej czy przestępczości, natomiast wciąż relatywnie mała część badań jest wciąż robiona w dziedzinie ekonomii i finansów. Wciąż nie do końca wiadomo, w jakich obszarach szerszy dostęp do danych i algorytmika może przysłużyć się lepszemu zrozumieniu mechanizmów sterujących rynkami finansowymi, wykrywaniu nadużyć, kryzysów finansowych, czy też recesji gospodarczych. A co najważniejsze, wciąż nie wiemy, jakimi błędami są obciążone modele stosujące „rewolucyjne” metody. Dlatego celem artykułu jest przegląd nowych metodologii badawczych stosowanych w naukach o finansach i ekonomii i zaprezentowanie ich dotychczasowego zastosowania, jak i ocena tych badań.

1. Dostęp do danych

⁶ <https://www.designweek.co.uk/issues/13-19-january-2020/unconscious-bias-ai-voice-assistants/> (pobrano: 08.11.2022)

1.1. Big Data

W wyniku przejścia od informacji analogowych do cyfrowych całkowita liczba danych cyfrowych gwałtownie wzrosła. Do 2016 r. ilość danych wygenerowanych w danym roku została oszacowana na 16,1 trylionów gigabajtów, a według prognoz ekspertów do 2025 ilość danych będzie dziesięciokrotnie wyższa niż w 2016 r. - około 163 tryliony gigabajtów⁷. Szerszy dostęp do danych stwarza cenne źródło informacji dla podmiotów prywatnych, jak i publicznych, jednakże o ich wartości świadczy dopiero odpowiednia analiza. Z tego względu rozpoczął się wyścig w poszukiwaniu rozwiązań do efektywnego zarządzania, przechwytywania i przetwarzania różnorodnych danych w celu uzyskania wartościowych informacji. Pomimo, że termin Big Data jest powszechnie znany i używany niestety, nie ma do tej pory uniwersalnej definicji użycia. Według firmy McKinsey⁸ termin „Big Data” odnosi się do zbiorów danych, których rozmiar przekracza możliwości typowych narzędzi w zakresie ich przechwytywania, przechowywania, zarządzania i analizowania. Gartner (2012) zaproponował popularną definicję Big Data jako „3V”⁹. Definicja ta związana jest z cechami Big Data, takimi jak wolumen, wysoka prędkość i wysoka wartość informacyjna, które wymagają efektywno-kosztowych, innowacyjnych form przetwarzania informacji. Współcześnie, pisząc o Big Data nie można pominąć kwestii związanych z ich cechami określonymi jako „4V”. Skrót ten odnosi się do początkowych liter czterech angielskich wyrazów: *Volume*, *Velocity*, *Variety* i *Veracity*.

Volume

Volume odnosi się do wolumenu zbioru danych, które są przechwytywane, przetwarzane i analizowane. Wielkość tych zbiorów danych jest ogromna i wymaga specjalnych technologii dedykowanych do ich obróbki.

Velocity

Postępy w zarządzaniu danymi, w szczególności przetwarzaniu danych w chmurze, ułatwiają rozwój platform do bardziej efektywnego pozyskiwania, przechowywania i manipulowania dużymi ilościami danych. Dzięki tej charakterystyce dane są

⁷ D. Reinsel et al., (2017), Data Age 2025: The Evolution of Data to Life-Critical Don't Focus on Big Data; Focus on the Data That's Big, s.3 3 (<https://www.import.io/wp-content/uploads/2017/04/Seagate-WP-DataAge2025-March-2017.pdf> pobrano 08.11.2022)

⁸ <https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/big-data-the-next-frontier-for-innovation> (pobrane: 08.11.2022)

⁹ Gartner report, (2012), The Importance of „Big Data”: Definition, 1-9

gromadzone w czasie rzeczywistym i w rzeczywistym tempie. Efektywne usprawnianie szybkości pobierania danych doprowadzi do lepszego podejmowania decyzji na podstawie aktualnych danych wejściowych.

Variety

Ogólna różnorodność dużej ilości danych o różnych cechach, takich jak: strukturyzowane, niestrukturyzowane i częściowo strukturyzowane dane pochodzące z różnych źródeł to kolejny aspekt, który sprawia, że duża ich ilość jest wyzwaniem dla badaczy. Podczas gdy w przeszłości dane mogły być gromadzone jedynie w arkuszach kalkulacyjnych, czy bazach danych, dziś dane występują w wielu formach, takich jak: e-maile, pliki PDF, zdjęcia, filmy, nagrania audio, wskaźniki z telefonu komórkowego i wiele innych. Kwantyfikacja tych danych, ich analiza i porównywalność wciąż stwarza ogromne problemy dla badaczy.

Veracity

Czwartym wymiarem, na który należy zwrócić uwagę jest kwestia wiarygodności (ang. Veracity). Oznacza ona, że dane i ich analityka powinna być wolna od błędów i można na niej polegać. Dlatego „Veracity” odnosi się do jakości danych, które są wykorzystywane. Podejmowanie decyzji na podstawie mało wiarygodnych danych może doprowadzić do wnioskowania obciążonego wysokim błędem statystycznym.

Duże zbiory danych nie są czymś nowym. Ich zastosowanie od dawna można było znaleźć w takich dziedzinach, jak: astronomia, meteorologia, taksonomia i demografia. Podobnie badania biomedyczne - a zwłaszcza w takich poddziedzinach, jak epidemiologia, farmakologia i zdrowie publiczne - mają rozległą tradycję wykorzystywania danych o dużej objętości, szybkości, różnorodności i zmienności, a wiarygodność tych danych i wyników są regularnie dyskutowane i poddawane pod weryfikację przez pacjentów, rządy, firmy farmaceutyczne, ubezpieczycieli i instytucje publiczne (Bauer 2008). Poszerzenie dostępu do danych stwarza duże możliwości dla takich dziedzin, jak finanse i ekonomia. Z racji wrażliwości sektora finansowego na zachowanie podmiotów gospodarczych, wiele danych dotyczących kredytobiorców, wielkości finansowania, bankructw, czy też zachowań inwestorów było tajnych, a naukowcy musieli pracować na bardzo ograniczonych próbach, często mających bezpośrednio konsekwencje na wnioskowanie statystyczne. Otwarty dostęp do danych o większej ilości, częstotliwości, jak również różnorodności stwarza możliwości na ulepszenie obecnych modeli w finansach, a algorytmika na większą precyzję prognozowania.

1.2. Rozwój mediów społecznościowych

W ostatnim czasie na znaczeniu przybierają media społecznościowe, jako źródło dużej ilości danych. Rozwój mediów społecznościowych stworzył także platformę do szybkiego i czasami nieocenzurowanego ¹⁰wypowiadania się przez ludzi ze świata finansów, show-biznesu, czy też polityki. Osoby znane i popularne mogą publikować swoje przemyślenia na temat obecnej sytuacji ekonomicznej i finansowej, a ich opinie znajdują zawrotny zasięg. Co ciekawe, mogą mieć one znaczący wpływ na zachowanie się rynków akcji, czy też kryptowalut. Popularnym przykładem jest Elon Musk. Elon Musk był jedną z najpopularniejszych osób na Twitterze – posiadał ponad 69 milionów obserwatorów. Ante (2021) przeanalizował blisko 50 różnych zdarzeń, które skutkowały tweedami bądź zmianami na koncie Tweetera Elona Maska i ocenił ich wpływ na rynek kryptowalut. W wypadku każdego wydarzenia, występowała obserwowalna, znacząca korelacja pomiędzy tymi zdarzeniami a wzrostem wolumenu obrotów na rynku Bitcoina. Przykładowo: 30 minut po tym, jak Elon Musk zmienił swoje bio na Twitterze na „#bitcoin”, skumulowany ponadprzeciętny zysk z kapitału (ang. CAR - Cumulative Abnormal Return) wyniósł 6.31%; 13.2% 1 godzinę po tym wydarzeniu, a po 7 godzinach osiągnął aż 18.99%. Wydarzenia takie jak te pokazują, jak cennym źródłem informacji mogą być media społecznościowe i jak ważną rolę mogą pełnić w zachowaniu rynków finansowych. Portal Statista.com¹¹ raportuje, że w 2021 roku 4.6 miliardów ludzi korzystało z mediów społecznościowych; liczba ta od 2017 roku cały czas wzrasta. Portal prognozuje, że do 2027 roku liczba ta wyniesie 6 miliardów.

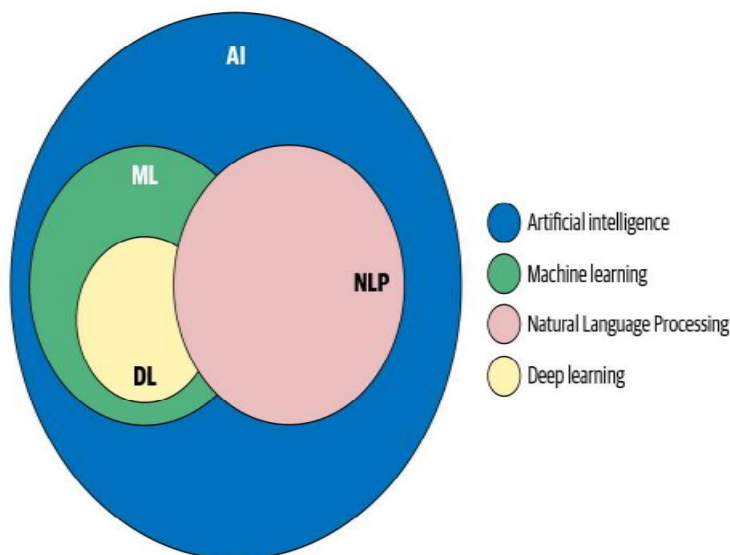
2. Nowoczesne metody badawcze wykorzystywane w badaniach w dziedzinie ekonomii i finansach

¹⁰ <https://edition.cnn.com/2019/11/03/media/trump-twitter-typos/index.html>

¹¹ <https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/#:~:text=How%20many%20people%20use%20social,almost%20six%20billion%20in%202027>. (pobrano: 05.11.2022)

2.1. Metody uczenia maszynowego

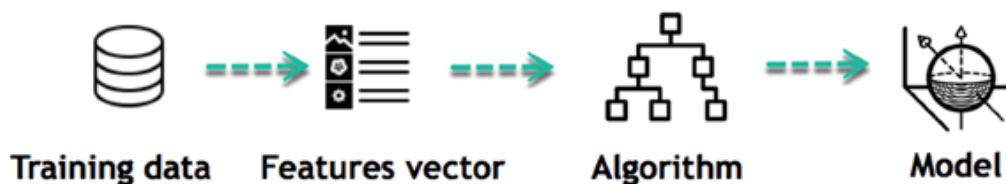
Uczenie maszynowe definiowane jest jako „algorytmy komputerowe, które są w stanie doskonalić się poprzez doświadczenie” (Tom M Mitchell, 1997, „Machine Learning”). Jest podzbiorem sztucznej inteligencji, które przyjmuje formę uczenia się poprzez doświadczenie lub obserwację przykładów. Następnie, algorytm szuka rozpoznawanych schematów korzystając z matematycznych modeli danych oraz statystyki w celu sporządzenia prognoz lub podejmowania decyzji bez bezpośrednich instrukcji. Komputer powinien uczyć się samodzielnie, bez ludzkiej ingerencji bądź pomocy i udoskonalać się poprzez dostarczanie mu dodatkowych informacji i danych.



Rycina 1: Elementy Sztucznej Inteligencji (AI)
Źródło: Vajjala et al 2020 - Practical NLP

Modele uczenia maszynowego zakładają, że użytkownik/naukowiec dobierze odpowiednie atrybuty (dane) wyjaśniające określony problem; wektor takich atrybutów nazywamy wektorem cech. Następnie, algorytm analizuje dostarczone dane. Model jest produktem końcowym – jest to algorytm, który został przeszkolony (ang. Pretrained) w wybranych cechach naszych danych.

Learning Phase



Rycina 2: Proces doskonalenia (uczenia się) modelu.

Źródło: <https://www.guru99.com/machine-learning-tutorial.html>

Przydatność tak zbudowanego modelu jest następnie sprawdzana. Modelowi podawane są nie widziane wcześniej dane – nazywane często danymi testowymi (ang. test data). Jest to wektor tych samych atrybutów, jednak jest to całkowicie inny zbiór danych.

Następnie, model dokonuje prognoz bądź klasyfikacji. Jeżeli używany algorytm należy do modeli uczenia nadzorowanego, możemy również dowiedzieć się, czy decyzje bądź przewidywania modelu są poprawne.

Inference from Model



Rycina 3: Proces inferencji z modelu.

Źródło: <https://www.guru99.com/machine-learning-tutorial.html>

2.2. Przetwarzanie języka naturalnego

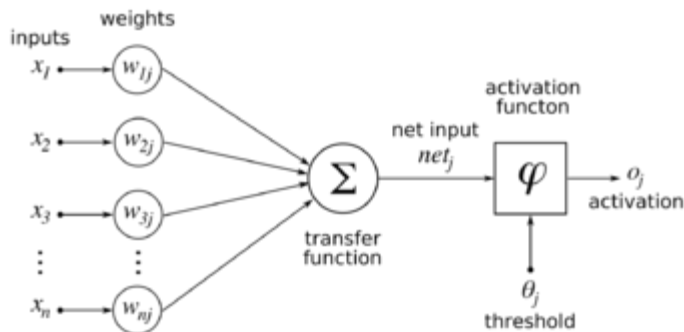
Przetwarzanie języka naturalnego (ang. Natural Language Processing – NLP) to podzbiór sztucznej inteligencji nastawiony na rozumienie, analizę oraz odczytywanie znaczenia języka naturalnego przez maszyny. Algorytmy NLP pozwalają na automatyczne podsumowywanie, tłumaczenia, rozpoznawanie różnych podmiotów, analizę sentymentu, rozpoznawanie mowy, czy też segmentację tematyczną. Przetwarzanie języka naturalnego uznawane jest powszechnie jako wymagający problem – język naturalny często bywa nieprecyzyjny. Zrozumienie języka naturalnego polega zarówno na zrozumieniu słów, jak i konceptów oraz połączeń pomiędzy nimi. Są one prostsze do nauczenia się dla człowieka niż dla maszyny – komputery napotykają problemy, gdy pojawia się ambiwalencja, nieunikniona w wypadku języka mówionego.

Przetwarzanie języka naturalnego wywodzi się z lingwistyki, a pierwsze modele opierały się na serii zdefiniowanych przez człowieka instrukcji. Wraz ze zwiększonymi możliwościami komputerów i upowszechnieniem uczenia maszynowego w latach 90 XX wieku potrzeba na długie, oparte na zasadach i wychwytywaniu wyjątków algorytmy zaczęły tracić na znaczeniu. Jednak przyczyniło się do rozwinięcia technologii sieci neuronowych, obecnie często stosowane do klasyfikowania tekstu, tłumaczenia języków, generacji tekstu bądź analizowania sentymentu.

2.3. Sieci neuronowe

Pierwsze modele uczenia głębokiego były modelami liniowymi, których motywacja była ugruntowana nauką o mózgu (Courville, Bengio, Goodfellow, 2015). Nasze neurony mogą być rozumiane jako funkcje matematyczne, które, w zależności od dostarczonych danych wejściowych zadziałają lub też nie. Dane, jakie przyjmie neuron, zależą od dendrytów (mogą być zatem dostarczone przez zmysły, ale również przez inny neuron). Soma jest miejscem, gdzie dane wejściowe są gromadzone i prowadzone są kalkulacje. Jeżeli wynik kalkulacji przekroczy określony próg, neuron aktywuje się (zadziała). Następnie, informacja o aktywacji lub jej braku przekazywana jest do aksonu, który poprzez swoje terminale przekazuje informację dalej do synaps, połączonych z dendrytami innych neuronów. Koncepcja sztucznego neuronu została zaproponowana przez Warrena McCullocka oraz Waltera Pittsa w 1943 roku. Naukowcy przedstawili prosty, matematyczny model neuronu, w którym wyjście

(output) zależało od ważonej sumy danych wejściowych (inputs), które dla danego progu przechodzą przez funkcję progową. Model ten dzieli dane wejściowe na dwa typy; podobnie, jak dane wyjściowe (outputs).



Rycina 4: Uproszczony model neuronu, zaproponowany przez McCullocka i Pittsa.

Stosowanie sieci neuronowych nie wymaga wstępnej selekcji danych (wybrania wektora cech), a zatem pozwala pominąć wielokrotne testowanie różnych konfiguracji i kombinacji danych treningowych. Modele same analizują wszystkie dostępne dane (nie eliminuje to zatem całkowicie narażenia na ludzkie uprzedzenia – dane dalej dostarcza człowiek) i wybierają z nich te, które najlepiej przekładają się na jakość klasyfikacji bądź prognoz. Ma to jednak istotny minus – by dojść do stabilnych i poprawnych lub zgodnych z prawdą wniosków, wymaga to dostarczenia bardzo dużej ilości obserwacji danych treningowych. Mimo to, postęp technologiczny ponownie umożliwił rozpowszechnienie się również takich rozwiązań, a sieci neuronowe osiągnęły imponujące rezultaty w dziedzinach widzenia komputerowego lub rozpoznawania mowy. Znajdują one zastosowanie także w analizie sentymentu.

2.4. Analiza sentymentu

Rozwój Internetu przyczynił się do zmian w ludzkim zachowaniu i uczynił go głównym źródłem informacji. Wiele osób korzysta z platform umożliwiających przekazanie swoich myśli – np. w formie wpisu, bądź pozostawienia polubienia, komentarza, bądź podzielenie się opinią w formie recenzji produktu lub usługi. Coraz większa popularność sieci społecznościowych przyczynia się do możliwości badania sentymentu poprzez zastosowanie technik języka naturalnego. Analiza sentymentu ma na celu skwantyfikować ludzkie opinie, myśli, emocje na wybrane tematy lub o wybranych aspektach – w tym, oczywiście, staje się

ważnym źródłem informacji o nastrojach inwestorów na rynkach finansowych, konsumentów, jak i stanie gospodarki.

Główną wadą analizy sentymentu jest trudność kwantyfikacji informacji, dlatego wyniki mogą być przekłamane. Wynika to z tego, iż na co dzień ludzie nie używają liczb do wyrażania emocji. Jednak komputery są coraz lepsze w odczytywaniu tonu tekstu (np. Grammarly), a zatem domniemywania intencji oraz stanu emocjonalnego piszącego. Pomocne mogą być także emotikony, które mają postać wyrazów twarzy, mających na celu przekazywanie emocji towarzyszących danej wypowiedzi, tak, by poza treścią tekstu zakomunikować także mimikę/stan emocjonalny piszącego. Emotikony często podzielone są na trzy grupy – mogą one wyrażać pozytywne, neutralne lub negatywne emocje. Przykładem emocji kwalifikowanych jako negatywne to smutek i depresja, a pozytywnych to miłość, radość i szczęście. Inną przesłanką o intencjach piszącego mogą być znaki interpunkcyjne, takie jak znaki zapytania lub wykrzykniki oraz ich liczba – im większa, tym dobitniejsza zazwyczaj treść wypowiedzi. Przy analizie języka używanego w sieciach społecznościowych można się także spotkać ze slangiem, który często jest dobitny i ma wyraźny ładunek emocjonalny, jak na przykład „lol” (ang. laughing out loud) lub „rotfl” (ang. rolling on the floor laughing).

3. Zastosowanie nowych metod badawczych w dziedzinie ekonomii i finansów

3.1. Zastosowanie analizy sentymentu do prognozowania wahań na rynku akcji

Możliwość prognozowania przyszłych zachowań na rynku papierów wartościowych jest kluczowe z perspektywy finansów, zwłaszcza w takich dziedzinach, jak optymalizacja portfela oraz zarządzanie ryzykiem. Zainteresowanie naukowców wzbudza także poszukiwanie zmiennych, które wykazywałyby na korelację ze zmiennością rynku i pomagały ją przewidywać. Większość badań w tym obszarze koncentrowała się na wykorzystaniu historycznych danych finansowych do regresji liniowych albo regresji probabilistycznych dając raz lepsze, raz gorsze wyniki.

Wraz z osiągnięciami NLP, uczenia maszynowego oraz uczenia głębokiego, a także rozwojem sieci społecznościowych, które okazały się bogatym źródłem danych, możliwym

stało się badanie korelacji pomiędzy opiniami wygłaszanymi w sieci a wahaniami obserwowanymi na rynkach. Obecnie, metody przetwarzania języka naturalnego pozwalają na analizę wiadomości oraz komentarzy w mediach społecznościowych w celu przewidywania zachowań rynków finansowych. W ostatnim czasie Deveikyte et al. (2020) zbadali związek pomiędzy sentymentem z Twittera odnoszącym się do spółek z FTSE100 (do badania zostały wybrane tylko wpisy z odpowiednimi tagami) a wahaniami indeksu FTSE100 i udowodnili istnienie silnej, negatywnej korelacji. P-value wyniosło znacznie mniej niż 0.05, podczas gdy wskaźnik korelacji -0.7, sugerując, iż średnia z sentymentu z wpisów na Twitterze z danego dnia pomaga przewidzieć wahania na rynku akcji dnia następnego. Kolejną istotną implikacją badania jest wyraźna sugestia, że wraz ze wzrostem sentymentu (a zatem ilości pozytywnych wpisów), rynek staje się mniej zmienny.

3.2. Zastosowanie sztucznej sieci neuronowej do prognozowania bankructwa firmy

Badania na temat bankructwa, jego przyczyn, i próby jego prognozowania są obszerne i umotywowane potrzebą ich zapobiegania poprzez wcześniejsze ostrzeżenie. Nieograniczone tylko do bankructw, próby ujęcia ilościowego problemów przedsiębiorstwa zostały zapoczątkowane w latach 30 XX wieku, tuż po Wielkim Kryzysie. Fitzpartrick (1932) próbował zdiagnozować różnice w porównywanych wskaźnikach finansowych firm, które przetrwały kryzys i tych, które zbankrutowały. Winakor i Smith (1935) oraz Merwin (1942) podążyli jego śladem. Jednak pierwszy znaczący postęp miał miejsce w 1966, kiedy Beaver (1966) po raz pierwszy zastosował narzędzia znane ze statystyki. Beaver (1966) próbował zastosować model statystyczny do analizy porównawczej sprawozdań finansowych firm, które dalej wciąż działały na rynku i tych, które zbankrutowały. W tym celu zbadał 79 przedsiębiorstw, które zbankrutowały i pochodziły z 38 różnych sektorów. Autor badania zestawiał je z 79 firmami, które dalej działały na rynku. Obie grupy nie różniły się znacząco wielkością aktywów, jak i przynależnością do sektorów. Beaver (1966) udowodnił, że wskaźniki finansowe są bardziej odpowiednie do przewidywania nie samej upadłości firm, ale przyszłych wydarzeń, które mogą mieć wpływ na kondycję finansową firmy.

Do 1980 roku za najlepszą metodę prognozowania ryzyka bankructwa firm w badaniach naukowych była technika MDA (ang. Multiple Discriminant Analysis – Analiza dyskryminacyjna wielu czynników) Jednakże metoda ta cechowała się wieloma wadami, takimi, jak: liniowość, normalność rozkładu oraz założenie niezależność między przyjętymi

do analizy zmiennymi. W 1980, Ohlson (1980) zastosował regresję logistyczną (ang. Logistic Regression - LR) do prognozowania ryzyka bankructwa firm. Metoda ta również opierała się na zastosowaniu wskaźników finansowych. Autor pokazał, że prognozy modelu na jeden rok w przód były w 96.12% poprawne, 95.55% dla dwóch lat w przód i 92.84% dla trzech lat w przód.

Tradycyjne modele przewidywania bankructwa polegają jedynie na migawkowej ocenie kondycji finansowej przedsiębiorstwa mierzonej w okresie t w celu przewidzenia, czy dana firma zbankrutuje w okresie $t + 1$ (Shumway, 2001; Taffler, 1983). Du Jardin i Severin (2011) podkreślają, że tradycyjne modele w większości nie przewidują możliwości uczenia się firm na własnych błędach i zmiany swojego zachowania w przyszłości, tym samym nie dopuszczając możliwości by przedsiębiorstwo poprawiło swoją sytuację. Z tego powodu dotychczasowa precyzyjność modeli ekonometrycznych jest znacząco mniejsza dla prognoz o horyzoncie dalszym niż jeden rok. Du Jardin and Séverin (2011) udowadniają, że wielkość błędu może nawet sięgać w granicach 15%. Wspomniany model Altmana (1968) wykazywał się 95% skutecznością w klasyfikacji bankructwa firmy na 1 rok przed, ale wskazywał już również tylko 48% precyzyjności na 3 lata przed bankructwem. W 1983 poprawiona wersja modelu osiągnęła 97.1% poprawności na 1 rok przed bankructwem i 69.7% na 3 lata przed bankructwem. Bez względu na metodę (liniowa czy nieliniowa, regresja czy klasyfikacja) główną wadą obecnych modeli jest ich krótki horyzont czasowy (Du Jardin i Séverin, 2011).

Początek lat 90 XX wieku przyniósł rozwój modeli sztucznej sieci neuronowej (ang. artificial neural network – ANN) zaproponowany przez Odoma i Sharda (1990). Odom i Sharda (1990) zbadali 128 niewypłacalnych amerykańskich firm produkcyjnych wykorzystując modele sieci neuronowych wykazując dużo wyższą precyzję w prognozowaniu bankructw firm niż wcześniejsze tradycyjne modele. Zhang et al. (1999) porównał sztuczne sieci neuronowe z regresją logistyczną, włączając do modeli dodatkowe wskaźniki finansowe. Zastosowana w konstrukcji modelu 5-krotna technika walidacji krzyżowej przewyższyła poprawność prognoz tworzonych przez model LR (77%) o 11 punktów procentowych i wyniosła 88%. Wiele badań pokazało, że sztuczne sieci neuronowe przewyższają klasyczne metody w swojej precyzyjności prognoz (Tam i Kiang, 1992; Medsker, Trippi i Turban, 1993; Coats and Fant, 1993; Zahedi, 1996; Adya and Collopy, 1998).

3.3. **Uczenie głębokie oraz prognozowanie zmian na rynku akcji**

Obecnie uczenie głębokie zyskuje coraz większy rozgłos w badaniach nad prognozowaniem zmian na rynku akcji, zwłaszcza, że sprawdza się do tego lepiej niż niektóre metody uczenia maszynowego (Nikou, Mansourfar and Bagherzadeh, 2019). Zhang, Shi, Zhang, i Shi (2006) pokazują, że sieci neuronowe były skuteczniejsze niż uczenie maszynowe w wypadku prognozowania zdarzeń wykorzystujących długie szeregi czasowe. W szczególności rekurencyjne sieci neuronowe (ang. Recurrent neural networks – RNN), a dokładniej long short-term memory (Długa pamięć krótkotrwała, w skrócie LSTM) uważane za jeden z wariantów RNN są bardzo użyteczne do modelowania danych sekwencyjnych, takich jak szeregi czasowe. Mechanizm długiej pamięci krótkoterminowej umożliwia uwzględnienie kontekstu danych – ich historii – w procesie modelowania. Hiransha et al. (2018) zaprezentowali zestawienie czterech różnych modeli uczenia głębokiego: perceptronu wielowarstwowego (MLP), sieci neuronowej rekurencyjnej, sieci długiej pamięci krótkotrwałej i sieci neuronowej splotowej (CNN). Celem tego badania było przewidzenie cen dwóch akcji na każdym z dwóch różnych wiodących rynków na świecie: Nowojorskiej Giełdzie Papierów Wartościowych i Narodowej Giełdzie Papierów Wartościowych w Indiach. Modelem wykorzystanym jako punkt odniesienia dla wyników prognoz modeli uczenia głębokiego był model ARIMA. W wyniku tych badań stwierdzono, że zestawione modele uczenia głębokiego są w stanie zidentyfikować wzorce, które istnieją w danych dotyczących cen akcji. Modele uczenia głębokiego przewyższają zatem model ARIMA pod względem możliwości przewidywania. W innym badaniu Baughman, Chard, Foster, Haas i Wolski (2018) próbowali przewidzieć ceny akcji firmy Amazona za pomocą modelu LSTM, jak i modelu ARIMA. Użyto trzech warstw ukrytych w LSTM, a wyniki wykazały, że nastąpiła średnia redukcja błędu średniokwadratowego o 60%, a w niektórych przypadkach nawet o 95% w porównaniu z punktem odniesienia – modelem ARIMA. Gao, Zhang i Yanga (2020) porównywali wyniki prognoz zachowania rynku akcji dla czterech modeli – perceptronu wielowarstwowego (MLP), sieci długiej pamięci krótkotrwałej, splotowej sieci neuronowej i sieci neuronowej opartej na uwadze. Wyniki dowiodły, że sztuczne sieci neuronowe dobrze sprawdzają się w wyjaśnianiu nieliniowych zależności w długich

szeregach czasowych. Ponadto, w swojej analizie Fischera i Kraussa (2017) wykorzystał model LSTM do prognozowania rynku akcji. Badawcze wykazali, że precyzja wyników modelu była dużo wyższa aniżeli w przypadku regresji logistycznej. Warto dodać, że w tym wypadku prognozowanie zostało zdefiniowane jako problem klasyfikacyjny – algorytm miał przewidzieć, czy cena akcji wzrośnie czy zmaleje. Innym ciekawym badaniem okazało się badanie Anghel, Cioara i Zanc (2019), którzy wprowadzili i przetestowali skumulowaną sieć LSTM na różnych kryptowalutach i danych dotyczących kursów wymiany walut. Wyniki pokazały wysoką dokładność przewidywanych wartości kursów walut.

4. Zakończenie

Rozwój nowych technologii i dostępu do danych stwarza ogromne możliwości dla nauki i rozwoju wielu dziedzin, w tym finansów i ekonomii. Wyniki badań wskazują, że w wielu przypadkach większy dostęp do danych i metody algorytmiczne są bardzo przydatne do prognozowania zachowania na rynkach finansowych. Modele wykorzystujące techniki algorytmiczne wykazują się większą precyzyjnością aniżeli tradycyjne modele ekonometryczne, a dodatkowe źródła danych, tak jak tweety z social mediów potrafią mieć wysoką wartość prognostyczną. Jednak zachowanie modeli jest w dużym stopniu uzależnione od danych wejściowych. O ile modelowanie zachowań na rynkach finansowych jest oparte na twardych danych historycznych, o tyle modele algorytmiczne dot. decyzji kredytowych, czy też bankructw przedsiębiorstw zależne są od atrybutów danych podanych przez samego badacza, jak również dobranej próby. Powoduje to, że tego typu modele mogą również generować błędne wyniki. Dlatego ważny jest dalszy rozwój badań wskazujący na szanse, ale również na zagrożenia wykorzystywania big data i technik algorytmicznych w dziedzinie finansów i ekonomii.

Bibliografia:

Opracowania naukowe oraz książki:

Adya, M., Collopy, F., 1998. How effective are neural networks at forecasting and prediction? A review and evaluation. *J. Forecast.* 17, 481–495. [https://doi.org/10.1002/\(SICI\)1099-131X\(199809\)17:5/6<481::AID-FOR709>3.0.CO;2-Q](https://doi.org/10.1002/(SICI)1099-131X(199809)17:5/6<481::AID-FOR709>3.0.CO;2-Q)

Alghawi, I.A., Yan, J., Wei, C., 2014. Professional or interactive: CEOs' image strategies in the microblogging context. *Computers in Human Behavior* 41, 184–189. <https://doi.org/10.1016/j.chb.2014.09.027>

Altman, E.I., 1968. FINANCIAL RATIOS, DISCRIMINANT ANALYSIS AND THE PREDICTION OF CORPORATE BANKRUPTCY. *The Journal of Finance* 23, 589–609. <https://doi.org/10.1111/j.1540-6261.1968.tb00843.x>

Ante, L., 2021. How Elon Musk's Twitter Activity Moves Cryptocurrency Markets. *SSRN Journal*. <https://doi.org/10.2139/ssrn.3778844>

Atiya, A.F., 2001. Bankruptcy prediction for credit risk using neural networks: A survey and new results. *IEEE Trans. Neural Netw.* 12, 929–935. <https://doi.org/10.1109/72.935101>

Bauer, S., 2008. Mining data, gathering variables and recombining information: the flexible architecture of epidemiological studies. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* 39, 415–428. <https://doi.org/10.1016/j.shpsc.2008.09.008>

Baughman, M., Haas, C., Wolski, R., Foster, I., Chard, K., 2018. Predicting Amazon Spot Prices with LSTM Networks, in: *Proceedings of the 9th Workshop on Scientific Cloud Computing*. Presented at the HPDC '18: The 27th International Symposium on High-Performance Parallel and Distributed Computing, ACM, Tempe AZ USA, pp. 1–7. <https://doi.org/10.1145/3217880.3217881>

Beaver, W.H., 1966. Financial Ratios As Predictors of Failure. *Journal of Accounting Research* 4, 71. <https://doi.org/10.2307/2490171>

Coats, P.K., Fant, L.F., 1993. Recognizing Financial Distress Patterns Using a Neural Network Tool. *Financial Management* 22, 142. <https://doi.org/10.2307/3665934>

Deveikyte, J., Geman, H., Piccari, C., Proveti, A., 2020. A Sentiment Analysis Approach to the Prediction of Market Volatility. <https://doi.org/10.48550/ARXIV.2012.05906>

du Jardin, P., Séverin, E., 2011. Predicting corporate bankruptcy using a self-organizing map: An empirical study to improve the forecasting horizon of a financial failure model.

- Decision Support Systems 51, 701–711. <https://doi.org/10.1016/j.dss.2011.04.001>
- Fischer, T., Krauss, C., 2018. Deep learning with long short-term memory networks for financial market predictions. *European Journal of Operational Research* 270, 654–669. <https://doi.org/10.1016/j.ejor.2017.11.054>
- Fitzpatrick, P. (1932). A comparison of the ratios of successful industrial enterprises with those of failed companies, *Certified Public Accountant*, 6, 727-731
- Gao, P., Zhang, R., Yang, X., 2020. The Application of Stock Index Price Prediction with Neural Network. *MCA* 25, 53. <https://doi.org/10.3390/mca25030053>
- Goodfellow, I., Bengio, Y., Courville, A., 2016. *Deep learning, Adaptive computation and machine learning*. The MIT Press, Cambridge, Massachusetts.
- Gordini, N., 2014. A genetic algorithm approach for SMEs bankruptcy prediction: Empirical evidence from Italy. *Expert Systems with Applications*, 41, 6433–6445. <https://doi.org/10.1016/j.eswa.2014.04.026>
- Hiransha M, Gopalakrishnan E.A., Vijay Krishna Menon, Soman K.P., 2018. NSE Stock Market Prediction Using Deep-Learning Models. *Procedia Computer Science*, 132. <https://doi.org/10.1016/j.procs.2018.05.050>
- Java, A., Song, X., Finin, T., Tseng, B., 2007. Why we twitter: understanding microblogging usage and communities, in: *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis - WebKDD/SNA-KDD '07*. Presented at the the 9th WebKDD and 1st SNA-KDD 2007 workshop, ACM Press, San Jose, California, pp. 56–65. <https://doi.org/10.1145/1348549.1348556>
- Lee, J., Cho, J., 2005. Consumers' use of information intermediaries and the impact on their information search behavior in the financial market. *Journal of Consumer Affairs* 39, 95–120. <https://doi.org/10.1111/j.1745-6606.2005.00005.x>
- M, H., E.A., G., Menon, V.K., K.P., S., 2018. NSE Stock Market Prediction Using Deep-Learning Models. *Procedia Computer Science* 132, 1351–1362. <https://doi.org/10.1016/j.procs.2018.05.050>
- Merwin, C.L., 1942. *Financing small corporations in five manufacturing industries, 1926–36*. National Bureau of Economic Research, Inc.
- Medsker, L., Turban, E., Trippi, R.R., 1993. *Neural Network Fundamentals for Financial Analysts*. *JOI* 2, 59–68. <https://doi.org/10.3905/joi.2.1.59>
- Miller, V., 2008. *New Media, Networking and Phatic Culture*. *Convergence* 14, 387–400. <https://doi.org/10.1177/1354856508094659>
- Mitchell, T.M., 1997. *Machine Learning*, McGraw-Hill series in computer science. McGraw-Hill, New York.
- Nikou, M., Mansourfar, G., Bagherzadeh, J., 2019. Stock price prediction using DEEP

- learning algorithm and its comparison with machine learning algorithms. *Intell Sys Acc Fin Mgmt* 26, 164–174. <https://doi.org/10.1002/isaf.1459>
- Odom, M.D., Sharda, R., 1990. A neural network model for bankruptcy prediction, in: 1990 IJCNN International Joint Conference on Neural Networks. Presented at the 1990 IJCNN International Joint Conference on Neural Networks, IEEE, San Diego, CA, USA, pp. 163–168 vol.2. <https://doi.org/10.1109/IJCNN.1990.137710>
- Ohlson, J.A., 1980. Financial Ratios and the Probabilistic Prediction of Bankruptcy. *Journal of Accounting Research* 18, 109. <https://doi.org/10.2307/2490395>
- Roberts, B., 2015. The third industrial revolution: Implications for planning cities and regions. Working Paper Urban Frontiers 1.
- Shumway, T., 2001. Forecasting Bankruptcy More Accurately: A Simple Hazard Model. *J BUS* 74, 101–124. <https://doi.org/10.1086/209665>
- Singh, Gurmeet and Singla, Ravi, 2019. Corporate Bankruptcy Prediction Using Altman' Z-Score Model: the Effect of Time and Methodology on Accuracy of the Model. *Ceeol Com*, 11, pp.58-72.
- Taffler, R.J., 1983. The Assessment of Company Solvency and Performance Using a Statistical Model. *Accounting and Business Research* 13, 295–308. <https://doi.org/10.1080/00014788.1983.9729767>
- Tam, K.Y., Kiang, M.Y., 1992. Managerial Applications of Neural Networks: The Case of Bank Failure Predictions. *Management Science* 38, 926–947. <https://doi.org/10.1287/mnsc.38.7.926>
- Smith, R.F. and Winakor, A.H., 1935. Changes in the financial structure of unsuccessful industrial corporations. University of Illinois.
- Zahedi, F., A Meta-Analysis of Financial Applications of Neural Networks, *International Journal of Computational Intelligence and Organizations*, Vol. 1, No. 3, 1996, pp. 164-178.
- Zanc, R., Cioara, T., Anghel, I., 2019. Forecasting Financial Markets using Deep Learning, in: 2019 IEEE 15th International Conference on Intelligent Computer Communication and Processing (ICCP). Presented at the 2019 IEEE 15th International Conference on Intelligent Computer Communication and Processing (ICCP), IEEE, Cluj-Napoca, Romania, pp. 459–466. <https://doi.org/10.1109/ICCP48234.2019.8959715>
- Zavyalova, A., Pfarrer, M.D., Reger, R.K., Shapiro, D.L., 2012. Managing the Message: The Effects of Firm Actions and Industry Spillovers on Media Coverage Following Wrongdoing. *AMJ* 55, 1079–1101. <https://doi.org/10.5465/amj.2010.0608>
- Zhang, Z., Shi, C., Zhang, S., Shi, Z., 2006. Stock Time Series Forecasting Using Support Vector Machines Employing Analyst Recommendations, in: Wang, J., Yi, Z., Zurada, J.M., Lu, B.-L., Yin, H. (Eds.), *Advances in Neural Networks - ISNN 2006*, Lecture Notes in Computer Science. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 452–457. https://doi.org/10.1007/11760191_66

Strony www:

1. <https://www.weforum.org/agenda/2016/01/the-fourth-industrial-revolution-what-it-means-and-how-to-respond>, pobrano 08.11.2022)
2. (https://www.wto.org/english/res_e/publications_e/world_trade_report18_e.pdf
pobrano: 08.11.2022)
3. <https://www.designweek.co.uk/issues/13-19-january-2020/unconscious-bias-ai-voice-assistants/> (pobrano: 08.11.2022)
4. <https://www.import.io/wp-content/uploads/2017/04/Seagate-WP-DataAge2025-March-2017.pdf> (pobrano 08.11.2022)
5. <https://edition.cnn.com/2019/11/03/media/trump-twitter-typos/index.html>
(pobrano 05.11.2022)
6. <https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/#:~:text=How%20many%20people%20use%20social,almost%20six%20billion%20in%202027.> (pobrano: 05.11.2022)
7. <https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/big-data-the-next-frontier-for-innovation> (pobrane: 08.11.2022)

Spis ilustracji:

Rycina 1. Elementy Sztucznej Inteligencji (AI);

Źródło: Vajjala et al. 2020, Practical NLP

Rycina 2. Proces doskonalenia (uczenia się) modelu.

Źródło: <https://www.guru99.com/machine-learning-tutorial.html>

Rycina 3. Proces inferencji z modelu.

Źródło: <https://www.guru99.com/machine-learning-tutorial.html>

Rycina 4. Uproszczony model neuronu, zaproponowany przez McCullocka i Pittsa.