UNIWERSYTET EKONOMICZNY W POZNANIU

Instytut Informatyki
i Ekonomii Ilościowej

Radosław Szymczyk

Analiza wydźwięku wypowiedzi o Głównym Urzędzie Statystycznym z wykorzystaniem sieci neuronowych

Sentiment analysis of statements about Statistics Poland using neural networks

Praca licencjacka

Promotor: dr Maciej Beręsewicz

Pracę przyjęto dnia:

Podpis promotora

Kierunek: Informatyka i ekonometria

Poznań 2021

Spis treści

Wstęp

W ostatnich latach rola internetu znacząco wzrosła w dziedzinie badań statystycznych. Dzięki wzrostowi ilości danych możliwych do analizy (internet rzeczy, działania użytkowników internetu) oraz rozbudowie rozwiązań chmurowych (ang*cloud computing*) badacze byli w stanie skonstruować nowe modele statystyczne oraz znacząco usprawnić te które były wykorzystywane dotychczasowo. Wynikiem tego był gwałtowny wzrost publikacji naukowych z dziedziny statystyki oraz ponowne zainteresowanie się modelami wcześniej ograniczonymi technicznie. Jednym z tego typu modeli były sztuczne sieci neuronowe. Ich koncepcja była już znana w wieku XX ale z powodu zbyt małej ilości mocy obliczeniowej oraz danych treningowych nie były one wykorzystywane powszechnie w zastosowaniach biznesowych. Natomiast wraz z wzrostem internetu zaczęto stosować je z sukcesami do bardzo skomplikowanych, wielowymiarowych problemów.

Jednym z tego typu problemów jest analiza wydźwięku wypowiedzi z plików tekstowych która stała się tematem przewodnim przedstawionej pracy licencjackiej. Postawiono sobie za cel analizę wydźwięku wypowiedzi użytkowników Twittera którzy w swoich wypowiedziach zawarli słowo kluczowe GUS, przy pomocy głębokich sieci neuronowych.

W pierwszym rozdziale zostanie przedstawiona krótka idea analiza wydźwięku wypowiedzi, źródła danych w badaniach statystycznych oraz możliwe wykorzystania tej metody w realnych problemach.

Drugi rozdział opisze szczegółowo tematykę sieci neuronowych. Przedstawi on krótką historię ich powstania oraz wszystkie zagadnienia techniczne które są wymagane do zrozumienia ich schematu działania.

Trzeci rozdział zawiera badanie empiryczne wykorzystujące pobrane dane z platformy Twitter. Przedstawia on różne podejścia do budowy modelu oraz wpływ niezbalansowanych danych na finalne wyniki predykcji.

Rozdział 1

Analiza wydźwięku w badaniach ekonomicznych

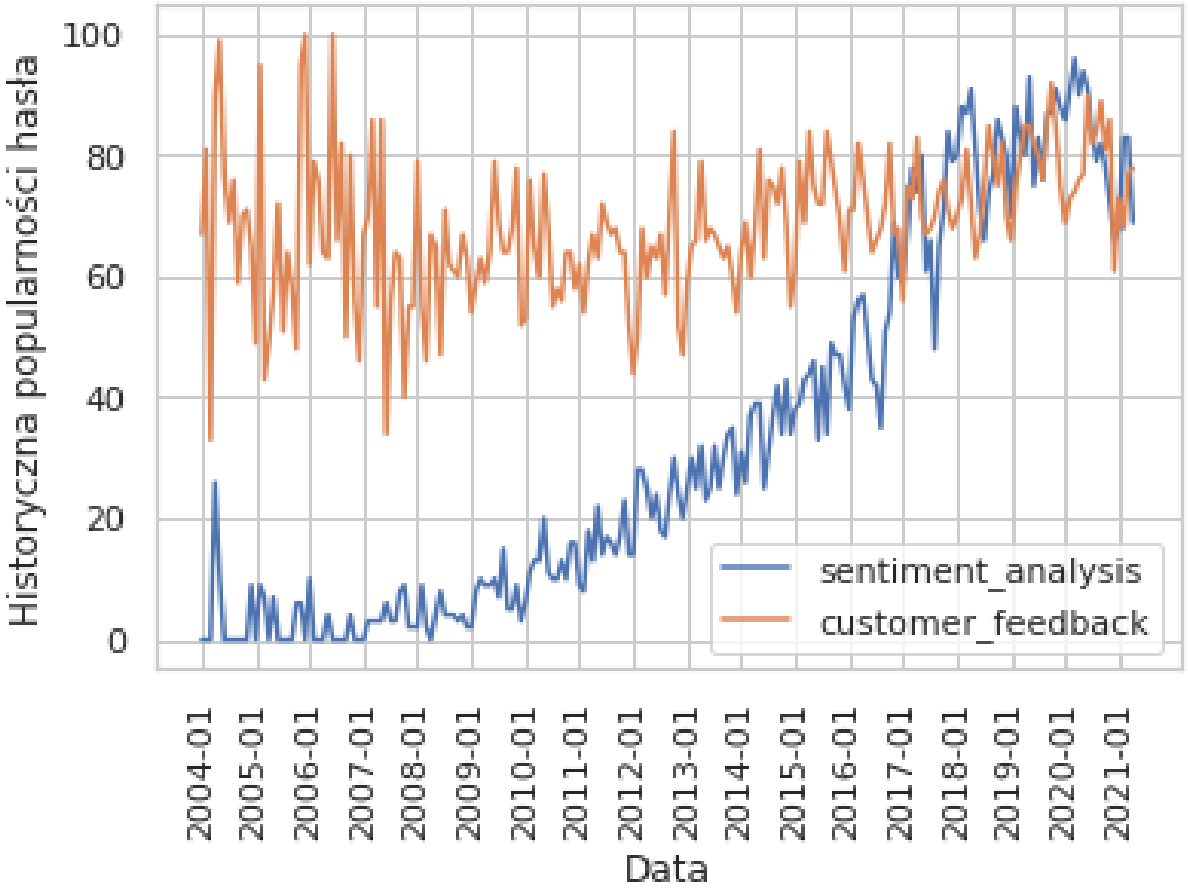
1.1 Idea analizy wydźwięku

Analiza wydźwięku jest zbiorem metod, technik i narzędzi wykorzystywanych do detekcji i wydobywania subiektywnych informacji takich jak opinie czy nastawienie autora z języka naturalnego (Liu 2009). W tradycyjnym podejściu analiza wydźwięku badała polarność opinii. Oznaczało to, że opinię wyrażoną w kierunku pewnego zjawiska lub przedmiotu klasyfikowało się do jednej z 3 grup (Dave i in. 2003):

- pozytywna,
- neutralna,
- negatywna.

Zainteresowanie opinią innych osób jest prawdopodobnie tak wiekowe jak sama werbalna komunikacja. Historycznie, liderzy byli zainteresowani opiniami swoich podwładnych aby przygotować się na sprzeciw wobec swoich działań lub zwiększeniu popularności w rządzonym przez nich regionie. Pierwszy próby pomiaru sprzeciwu wewnętrznego były podejmowane już w starożytnej Grecji (Richmond 1998). Głosowanie jako metoda do pomiaru opinii społeczeństwa na tematy polityczne ma swoje korzenie w 5 wieku p.n.e. w Atenach (Thorley 2004). Wysiłki mające na celu uchwycenie opinii publicznej poprzez kwantyfikację i pomiar na podstawie kwestionariuszy pojawiły się w pierwszych dekadach XX wieku (Droba 1931), natomiast w 1937 roku powstało czasopismo naukowe poświęcone pomiarze opinii publicznej (*oldartice* nodate).

W ostatnich latach zainteresowanie tematem analizy wydźwięku znacznie zyskało na popularności. Ponad 99 procent z 7000 publikacji poruszających zagadnienie analizy wydźwięku zostało wydanych po roku 2004 (Mäntylä i in. 2016).



Rysunek 1.1. Wykres pokazujący popularność fraz śentiment analysis oraz customer feedback w czasie

Źródło: Opracowanie własne na podstawie danych z Google Trends

Pierwsze akademickie badania poświęcone mierzeniu publicznej opinii były prowadzone podczas i po drugiej wojnie światowej i były skierowane głownie na tematy polityczne (Knutson 1945; Stagner 1940). Przełom w zainteresowaniu analizą sentymentów w środkowych latach 2000 wywołany gwałtownym rozwojem internetu. Analiza wydźwięku w tych pracach głownie skupiła się na recenzjach produktów zamieszczanych w internecie (Dave i in. 2003). Od tamtej pory zaczęto wykorzystywać analizę wydźwięku również do przewidywania rynków finansowych (Nassirtoussi i in. 2014) czy też mierzenia reakcji społeczeństwa po atakach terrorystycznych (Burnap, Williams i in. 2014). Ponadto badacze również zaczęli udoskonalać modele

zajmujące się analizą sentymentów z prostych kategorii odróżniających jedynie dwa skrajne stany emocjonalne(zdenerwowanie i zadowolenie) do bardziej wyspecjalizowanych odróżniających np. smutek,złości,przerażenie czy zdziwienia.

1.2 Źródła danych na potrzeby analizy wydźwięku

1.2.1 Tradycyjne metody pozyskiwania danych

Według Kaczmarczyk (2018) głównym kryterium podziału metod zbierania danych ze źródeł pierwotnych jest sposób zbierania, który jest określony przez rodzaj bodźca używanego podczas pomiaru cech jednostki próby. Kryterium te pozwala wyodrębnić dwie grupy zbierana danych. Grupę sondażową i grupę poza sondażową. Dodatkowo grupę sondażową można podzielić na dwie podgrupy:

- pośrednie metody sondażowe,
- bezpośrednie metody sondażowe.

1.2.2 Pośrednie metody sondażowe

Pośrednie metody sondażowe skupiają się na wysyłaniu ankiet oraz kwestionariuszy do respondentów. Podczas wypełniania ankiet respondent znajduje się w dużej odległości od prowadzonego badanie. Prowadzony w ten sposób pomiar jest w mniejszym stopniu kontrolowany przez badaczy i zapewnia dużą anonimowość respondenta.W porównaniu z bezpośrednimi metodami sondażowymi metody pośrednie są o wiele tańsze w zastosowaniu, nie wymagają fizycznej obecności respondentów oraz czasu na przeprowadzenie indywidualnego badania.Tego typu badania bardzo skorzystały na rozwoju internetowych mediów. W Moskowitz i Martin (2008) ukazano część zalet spowodowanych zamienieniem ankiet fizycznych na ankiety internetowe:

- niższy koszt i krótszy czas zbierania danych,
- łatwiejsze dotarcie do docelowych segmentów rynku,
- możliwości komunikacji globalnej oraz interaktywnej,
- brak wpływu ankietera na reakcje respondentów,
- możliwość natychmiastowej analizy danych i biedzące, wykorzystywanie danych do poprawy funkcjonowania modelu,

- większa elastyczność pomiaru,
- ułatwienie sprawdzania danych surowych w momencie ich otrzymywania.

Rodzaje metod	Metody zbierania danych	Wybrane techniki (odmiany) metod	Stosowane instrumenty pomiarowe
Pośrednie metody ankietowe	Ankieta: • pocztowa • internetowa • prasowa • ogólna (rozdawana) • opakowaniowa (towarowa) • faksowa • telefoniczna • radiowa • telewizyjna • komputerowa	• email survey • online survey • ATS (stacjonarna) • ATK (komórkowa)	• kwestionariusz ankietowy
Pośrednie metody heurystyczne	• metoda delficka • konkurs pomysłów • sesje wirtualne (*brainnetting*)	• klasyczna (pocztowa) • internetowa	• kwestionariusz delficki • szkicownik wizualny • pakiet symulacyjny
Wywiady pośrednie	• wywiad internetowy CAWI) • wywiad telefoniczny	• klasyczny • CATI	• kwestionariusz wywiadu
Panele konsumenckie	• panel pocztowy • panel internetowy • panel telefoniczny		• dziennik panelowy • kwestionariusz panelowy
Pośrednie wywiady grupowe	• telefoniczne (telekonferencje) • internetowe	• pisemne • ustne • łączone	• scenariusz wywiadu • komunikator głosowy • kamera komputerowa

Rysunek 1.2. Klasyfikacja pośrednich sondażowych metod zbierania danych ze źródeł pierwotnych

Źródło: S. Kaczmarczyk, 2014, rozdz. 7,8.

1.2.3 Bezpośrednie metody sondażowe

Bezpośrednie metody sondażowe polegają na zadawaniu pytań w bezpośrednim kontakcie pomiędzy ankieterem i respondentem. Zaletą tego typu metod jest znaczy wzrost stopnia kontroli udzielanych odpowiedzi, o wiele większa częstotliwość odpowiedzi na zadane pytania oraz skrócenie czasu przeprowadzania badania względem metod pośrednich. Wadą metod bezpośrednich jest ich większy koszt oraz brak możliwości udoskonalenia ich z użyciem rozwiązań internetowych.

Rodzaje metod	Metody zbierania danych	Wybrane techniki metod	Instrumenty pomiarowe
Bezpośrednie metody ankietowe	• ankieta audytoryjna • ankieta bezpośrednia	• zwrot natychmiastowy • zwrot odroczony	• kwestionariusz ankietowy
Wywiady bezpośrednie indywidualne	• wywiad osobisty • wywiad osobisty swobodny • rozmowa (anamneza)	• wywiad w domu respondenta • wywiad w pasażu handlowym • wywiad w biurze respondenta • wywiad na ulicy • wywiad audytoryjny (CLT) • CATI	• kwestionariusz wywiadu • palmtop, tablet • scenariusz wywiadu • scenariusz rozmowy
Bezpośrednie metody heurystyczne	• burza mózgów • metoda synektyczna • metoda myślenia lateralnego • metoda morfologiczna	• klasyczna (Osborna) • Gordona-Little'a • Philips 66 • technika 635	• arkusz kontrolny • kwestionariusz • tablica morfologiczna
Panel konsumencki	• panel bezpośredni	• wywiady osobiste • wywiady grupowe (panele wrażliwości)	• dziennik panelowy • kwestionariusz wywiadu • mikrofony i kamery
Bezpośrednie metody jakościowe	• wywiad grupowy • osobisty wywiad pogłębiony • metody projekcyjne	• metody skojarzeń słownych • metody uzupełnień • metody konstrukcji • metody wyobrażeń	• scenariusz, mikrofon, kamera • kwestionariusz wywiadu • testy

Rysunek 1.3. Klasyfikacja bezpośrednich sondażowych metod zbierania danych ze źródeł pierwotnych

Źródło: Kaczmarczyk, 2014, rozdz. 8.

1.2.4 Metody poza sondażowe

Metody poza sondażowe skupiają się na analizie niewerbalnej, pytania są wykorzystywane w nich jedynie jako narzędzia pomocnicze. Przez swój charakter niewerbalny metody te mogą analizować nie tylko ludzi ale również przedmioty, zjawiska i zdarzenia.

Metody poza sondażowe dzielą się na 3 główne grupy (Kaczmarczyk 2018):

- pomiary fizjologiczne i neuromarketingowe, w których przedmiotami pomiarów są głównie cechy osobowe,
- metody obserwacji, etnograficzne i monitoringu, w których przedmiotami pomiarów są zarówno cechy osobowe , jak i rzeczowe,

- metody rejestracji i spisu oraz sensoryczne, w których obiektami pomiarów są głownie cechy rzeczy , w tym zdarzeń i zjawisk.

Rodzaje metod	Metody zbierania danych	Wybrane techniki metod	Instrumenty pomiarowe
Metody obserwacji	• obserwacja uczestnicząca • tajemniczy klient • obserwacja internetowa • inne metody obserwacji	• indywidualna • biznesowa • ekspercka • telefoniczna	• dziennik (arkusz) • zmysły (głównie wzrok)
Metody rejestracji spisu oraz monitoringu	• panel sklepowy (detaliczny) • audyt detaliczny i hurtowy • monitorowanie i bazy danych • rejestracja przez GPS • rejestracja telemetryczna	• rejestracja skaningowa • rejestracja RFID	• czynnik kodów kreskowych • skaner (czytnik RFID) • podręczne komputery • telemetr (wizometr)
Pomiary fizjologiczne	• pomiar fal mózgowych • pomiar ruchu gałek ocznych • pomiar wrażliwości skóry • inne pomiary fizjologiczne		• EEG • okulograf (eyetracker – ET) • wariograf (poligraf)
Metody sensoryczne	• degustacja • próbne użytkowanie • oceny próbek towarowych		• zmysły
Pozostałe metody pozasondażowe	• metody neuromarketingowe • metody etnograficzne	• funkcjonalny rezonans magnetyczny(technika BOLD) • optyczna tomografia absorpcyjna • techniki klasyczne • techniki internetowe (netografia)	• EEG • czujniki laserowe • zwykły notes • aparat fotograficzny • kamera filmowa • mikrofon

Rysunek 1.4. Klasyfikacja poza sondażowych metod zbierania danych ze źródeł pierwotnych

Źródło: Kaczmarczyk, 2014, rozdz. 9.

1.2.5 Ograniczenia źródeł danych

1.2.5.1 Reprezentatywność

Pomimo zachęcających wyników badań wykorzystujących media społecznościowe jako źródło danych (por. Sekcja 1.4) nigdy nie wyeliminowano problemu obciążenia populacyjnego. Demografia użytkowników mediów społecznościowych znacząco różni się od ogólnej struktury społeczeństwa. Przeciętna osoba używająca media społecznościowe jest młodsza oraz posiada

wyższe wykształcenie od osoby, która ich nie używa (Mellon i Prosser 2017). Również istotnym czynnikiem przy tego typu badaniach jest aktywność użytkowników, kobiety są bardziej skłonne do czynnego korzystania z mediów społecznościowych (częściej piszą np. komentarze) (Joinson 2008).

1.2.5.2 Anonimowość przy wystawianiu opinii

Małyszko (2015) porusza problem anonimowości w Internecie w badaniu opinii konsumenckich. Opinie oraz recenzje zamieszczane w internecie rzadko są weryfikowane co do ich wiarygodności oraz rzetelności. Dodatkowo wiele stron internetowych nie wymaga podawania swoich prawdziwych danych osobowych przy wystawianiu recenzji. Oba te aspekty utworzyły możliwość do manipulowania recenzjami przez sprzedawców którzy mogą w anonimowy sposób dodawać recenzje chwalące ich produkty lub krytykujące ich bezpośrednią konkurencję. Z uwagi na skalę problemu oraz możliwych korzyści zaczęto wprowadzać rozwiązania mające na celu ograniczenie praktyki tworzenia zmanipulowanych recenzji. Przykładowo na serwisie Amazon.com wprowadzono specjalne kategorie recenzji, będące z założenia bardziej wiarygodnymi:

- **Amazon Verified Puchase Reviews** – ten typ recenzji może być publikowany jedynie przez zarejestrowanych użytkowników którzy zakupili już oceniany produkty,
- **Amazon Vine** – utworzenie grupy wyselekcjonowanych osób których opinie były historycznie najbardziej wiarygodne.

Przykład praktykowania zmanipulowanych recenzji można było zaobserwować podczas błędu w serwisie Amazon.com, który upublicznił prawdziwe tożsamości autorów recenzji. Okazało się, że niektórzy znani pisarze zamieszczali bardzo pochlebne recenzje własnych książek z największą możliwą oceną. Taka praktyka miała wesprzeć ich sprzedaż i przekonać niezdecydowanych użytkowników na finalny zakup.

1.2.5.3 Boty

Integralną częścią mediów społecznościowych są konta użytkowników. Służą one do reprezentowania użytkownika na platformie. Zazwyczaj strony internetowe nie wymagają potwierdzenia swojej tożsamości oraz nie nakładają blokady na ilość kont którą może posiadać jeden

użytkownik. Brak ograniczenia w zakładaniu kont prowadzi do zagrożenia występowania Botów czyli sztucznych bytów mających na celu wykonywanie pewnych czynności w zastępstwie za człowieka. Boty są niebezpieczne pod kątem analizy danych ponieważ mogą zniekształcić prawdziwą zsumowaną opinię użytkowników (Rauchfleisch i Kaiser 2020). Boty niosą za sobą również inne zagrożenia jak np sianie dezinformacji które może mieć na przykład wpływ na wyniki wyborów (Shao i in. 2017).

1.3 Twitter jako źródło danych

1.3.1 Czym jest Twitter?

Twitter jest serwisem społecznościowym udostępniającym usługę mikroblogowania za pomocą ograniczonych wiadomości zwanych Tweetami. Powstał 21 marca 2006 roku i przeznaczony jest dla grup znajomych, współpracowników oraz rodziny do szybkiego wymieniania informacji.

Tweet jest wiadomością publikowaną na serwisie Twitter która może zwierać zdjęcia, nagrania wideo, linki lub tekst. W odróżnieniu od innych wiadomości na popularnych stronach społecznościowych Tweet ma ograniczony rozmiar. wideo.[1]

W porównaniu z innymi serwisami społecznościowymi każdy opublikowany tweet jest publiczny i możliwy do pobrania za pomocą udostępnionego przez Twitter API. Przy jednoczesnym generowaniu 500mln tweetów dziennie [2] oraz wyspecjalizowanym zapytaniom obejmującym słowa kluczowe, język w jakim został napisany tweet czy przedział czasowy w jakim mógł zostać zamieszczony jest to jedno z największych publicznych źródeł danych do wszelkiego rodzaju analiz.

1.4 Media społecznościowe w wybranych badaniach społecznych

1.4.1 Ceny akcji

W artykule (Bollen i in. 2011) badacze chcieli zbadać czy ogólny nastrój społeczeństwa uzyskany poprzez analizę wydźwięku wypowiedzi użytkowników mediów społecznościowych jest skore-

[1] https://help.twitter.com/en/using-twitter/how-to-tweet

[2] https://blog.twitter.com/official/en_us/a/2014/the-2014-yearontwitter.html

Rysunek 1.5. Przykładowy tweet

Źródło: https://twitter.com/elonmusk

lowany z wskaźnikami ekonomicznymi oraz czy na jego podstawie można prognozować przyszłe wartości ekonomicznych mierników. W przeprowadzaniu badania Bollen, Mao oraz Zeng zdecydowali się wybrać indeks DIJA [3] oraz pozyskać dane z platformy Twitter. Do zbadania hipotezy, że nastroje społeczne, mierzone przez OpinionFinder'a i GPOMS, mogą zostać wykorzystane do przepowiadani zmian wartości zamknięcia DJIA badacze użyli analizy przyczynowości Grangera oraz Samo-organizujących się sieci neuronowo-rozmytych. Ich wyniki wykazały ,że precyzja w przewidywaniu DIJA (która w pracy wyniosła 86.7 procent) może zostać znacząco zwiększona poprzez użycie ogólnego nastroju użytkowników mediów społecznościowych oraz może ona zredukować MAPE o ponad 6 procent.

1.4.2 Przewidywanie sprzedaży

W artykule (Asur i Huberman 2010) Sitaram Asur i Bernardo A. Huberman przestawili jak przy pomocy wiadomości publikowanych na Twitterze można przewidzieć realne wyniki sprzedaży. W swojej pracy skupili się na wynikach kasowych 24 popularnych filmów wydanych w 3 mie-

[3] Indeks Dow Jones Industrial Average jest drugim najstarszym i najbardziej znanym indeksem giełdowym. Należący do Down Jones Company, mierzy dzienne ruchy cen 30 dużych amerykańskich firm na Nasdaq i Nowojorskiej giełdzie

sięcznym badanym okresie. W wyszukiwaniu tweetów opisujących dane filmy badacze zdecydowali się użyć słów kluczowych które oznaczały tytuły filmowe (np. Avatar). Przy wykorzystaniu modelu liniowej regresji uzyskano wyniki które posiadały o wiele większą trafność niż te z Hollywood Stock Exchange oraz wykazały silną korelację pomiędzy ilością reakcji dla danego tematu w mediach społecznościowych a jego późniejszym rankingiem.W pracy również podkreślono znaczący wpływ analizy wydźwięku na realną poprawę prognoz wyników kasowych filmów.

1.4.3 Wskaźnik ufności konsumenckiej

W artykule (Daas i Puts 2014) badacze porównali dane pozyskane od firmy Coosto[4] oraz miesięczne raporty dotyczące wskaźnika zaufania konsumentów sporządzone przez Statistics Netherlands. Autorzy wykazali wysoką korelację badanych danych (r=0.9). Pomimo siedmiodniowego opóźnienia w przedstawieniu zaufania konsumentów poprzez media społecznościowe oraz obciążania które objawia się znacznie większym pozytywizmem użytkowników mediów społecznościowych , autorzy proponują połączenie tradycyjnych ankiet telefonicznych oraz analizy wydźwięku wiadomości zamieszczanych na mediach społecznościowych aby zwiększyć precyzję oraz częstotliwość mierzenia wskaźnika zaufania konsumentów.

1.4.4 Przewidywanie wyników wyborów

W pracy (Burnap, Gibson i in. 2016) badacze skupili się nad zagadnieniem przewidywania wyników wyborów. Użyli oni analizę wydźwięku oraz zbadali wcześniejsze poparcie partii dla poszczególnych kandydatów aby utworzyć najbardziej zbliżoną prognozę przydziału miejsc w parlamencie podczas generalnych wyborów w UK w 2015r. Pomimo zachęcających wyników nadal w sferze przewidywania wyników kluczowym jest pozbycie się obciążenia populacyjnego, ponieważ demografia użytkowników mediów społecznościowych znacząco różni się od ogólnej demografii społeczeństwa. Z artykułu (Mellon i Prosser 2017) można dowiedzieć się ,że przeciętny użytkownik mediów społecznościowych jest młodszy oraz bardziej wyedukowany od przeciętnej osoby która nie używa mediów społecznościowych. Autorzy artykułu zaznaczają również istotność uwzględnienia geo-lokacji która w znaczący sposób podnieść trafność następnych badań skupiających się na prognozowaniu wyborów.

[4]firma która gromadzi wiadomości w języku holenderskim z wszystkich największych mediów społecznościowych takich jak Twitter czy Facebook ale także z publicznych forów oraz blogów.

1.4.5 Wskaźnik oglądalności telewizji

W badaniu przeprowadzonym w Crisci i in. (2018) badacze zdecydowali się na próbę przewidzenia publiczności popularnych programów telewizyjnych za pomocą tweetów oraz ich sentymentów. W badanych godzinach analizowali oni częstotliwość cytowanych programów telewizyjnych oraz ich sentymenty.

1.4.6 Wykorzystanie mediów społecznościowych opartych na lokalizacji w metodach badania podróży

W artykule (Abbasi i in. 2015) badacze skupiają się w jaki sposób dane z mediów społecznościowych mogą posłużyć w analizowaniu aktywności mieszkańców danego regionu oraz odwiedzającego go turystów. Przy pomocy data miningu skupionemu na technikach językowych byli w stanie określić typ użytkownika (turysta/rezydent) oraz utworzyli dwie mapy. Pierwsza przedstawiała najbardziej odwiedzane miejsca w Sydney przez turystów a druga dystrybucję tweetów na podstawie wyodrębnionych celów podróży(zakupy, rozrywka, jedzenie, praca itp.).

1.5 Podsumowanie

W pierwszym rozdziale zostały przedstawione podstawowe pojęcia związane z tematem pracy. Przedstawiono wpływ internetu na badania statystyczne oraz idące za tym wady oraz zalety.

W następnym rozdziale zostanie przedstawione uczenie maszynowe wraz z założeniami, następnie zostanie przedstawiony schemat działania i budowa sieci neuronowych oraz biblioteka TensorFlow która dzięki interfejsowi `keras` implementuje sieci neuronowe w języku `Python`.

Rozdział 2

Sieci neuronowe

2.1 Historia powstania sieci neuronowych

Sztuczne sieci neuronowe zostały pierwszy raz zaprezentowane w 1943 roku przez Warrena McCulloch'a oraz Waltera Pitts'a (McCulloch i Pitts 1943). W swojej pracy badacze zaprezentowali uproszczony model obliczeniowy przedstawiający jak neurony biologiczne mogą współpracować w mózgach zwierząt aby rozwiązywać skomplikowane obliczenia wykorzystując rachunek zdań (ang. *propositional logic*). Była to pierwsza skonstruowana architektura sztucznych sieci neuronowych.

Wczesne sukcesy pobudziły wyobraźnię wielu zainteresowanych. Uważano, że odkryta technologia jest o krok od utworzenia w pełni inteligentnych maszyn. Kiedy stało się jasne w latach sześćdziesiątych, że jeszcze przez długi czas obietnica nie może zostać spełniona postanowiono przekierować fundusze badawcze na bardziej obiecujące badania. Na początku lat osiemdziesiątych powstały nowe architektury oraz lepsze techniki nauki modelów, które wywołały kolejny wzrost zainteresowania koneksjonizmem (badaniem sieci neuronowych). Na początku lat dziewięćdziesiątych powstało wiele konkurencyjnych technik uczenia maszynowego które posiadały lepsze rezultaty oraz podstawy teoretyczne (np. Maszyny wektorów nośnych). W połączeniu z dość wolnym rozwojem sieci neuronowych po raz kolejny zaprzestano ich dalszych badań (Géron 2017). Obecnie obserwuje się kolejną falę zainteresowania sztucznymi sieciami neuronowymi która może być najbardziej przełomową ponieważ:

- W świecie internetu rzeczy oraz Big Data istnieje bardzo dużo danych możliwych do analizy przy pomocy sieci neuronowych;

- Sztuczne sieci neuronowe radzą sobie lepiej (niż inne techniki uczenia maszynowego) w dużych i skomplikowanych problemach (np. widzenie maszynowe);
- Wzrost mocy obliczeniowej ukazany przez prawo Moore'a pozwala na trenowanie sieci na o wiele większych zbiorach danych oraz w rozsądnych okresach czasu;
- Ciągłe sukcesy sztucznych sieci neuronowych powodują wzrost zainteresowania i przeznaczanych funduszy na ich rozwój;
- Dodano małe poprawki do algorytmów które istniał w latach dziewięćdziesiątych które spowodowały duży pozytywny wpływ na ich działanie.

Obecnie wykorzystuje się sieci neuronowe w bardzo skomplikowanych wielowymiarowych problemach takich jak:

- Rozpoznawanie i klasyfikowanie obrazów (Zdjęcia Google, autonomiczne samochody);
- Usługi rozpoznawania mowy (Apple Siri);
- System polecenia (Youtubie, Spotify);
- Uczenie komputera grania w gry (DeepMind AlphaGo lub AlphaStar);
- Przetwarzanie języka naturalnego (wykrywanie wzorców/analiza wydźwięku).

2.2 Sztuczne sieci neuronowe

2.2.1 Reprezentacja danych

Każdy obiekt przed wprowadzeniem do sztucznej sieci neuronowej przekształca się w któryś z obiektów algebry liniowej (Goodfellow i in. 2016):

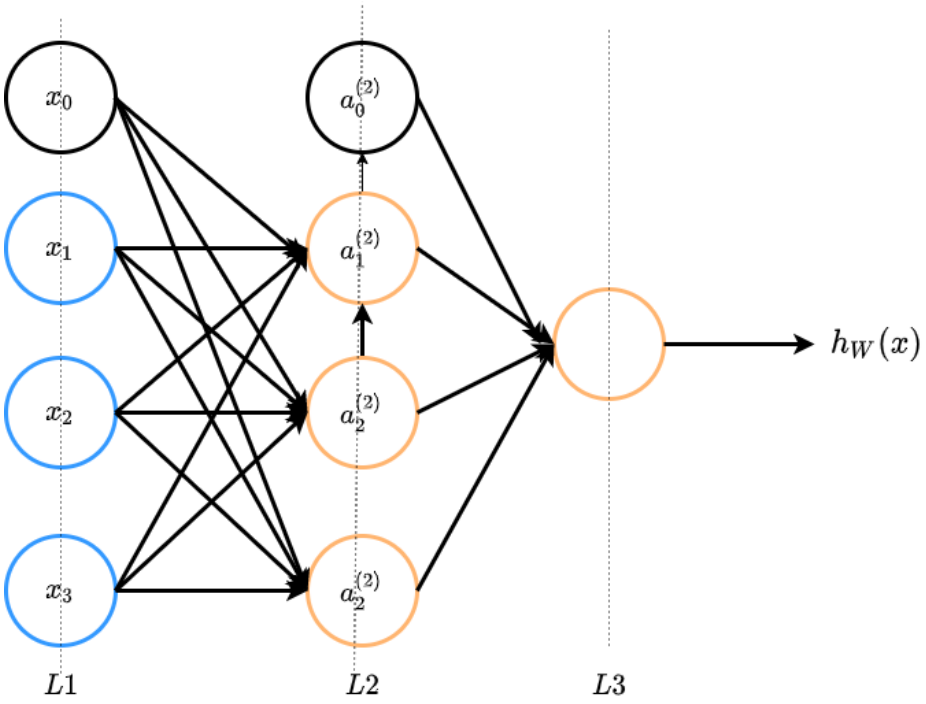
- Skalar to jedna liczba, co odróżnia go od innych obiektów algebry liniowej, które są zwykle tablicami złożonymi z wielu liczb;
- Wektor to tablica liczb. Liczby ustawione są po kolei oraz posiadają określony typ (Naturalne, Rzeczywiste itp.). Traktuje się wektory jako identyfikację punktów; w przestrzeni, przy czym każdy podaje współrzędną na innej osi. Elementy wektora reprezentuje się w postaci kolumny ujętej w nawiasy kwadratowe;
- Macierz to dwuwymiarowa tablica liczb;
- Tensory są tablicami o większej liczbie wymiarów. W ogólnym przypadku liczby tablicy tworzą regularna siatkę ze zmienną liczba osi, którą nazywamy tensorem.

2.2.2 Reprezentacja sieci neuronowej

Sieć neuronową można określić jako funkcję która dla określonych **x** czyli danych wejściowych ma za zadanie dopasować określone Y czyli dane wyjściowe. Każda sieć neuronowa składa się z 3 typów warstw neuronów:

- **Warstwa wejściowa** – pobiera ona dane wejściowe. Ilość neuronów w niej zawarta zależy od wielkości danych wejściowych. Dla przykładu, warstwa wejściowa prostej sieci neuronowej mającej na celu klasyfikację liczb zdefiniowanych na 28x28 pixelowym obrazie będzie posiadała 784 neurony (po jednym neuronie dla każdej wartości poszczególnego pixela).
- **Warstwy ukryte** – można przedstawić je w postaci czarnej skrzynki. Użytkownik nie jest w stanie sprawdzić działania tych warstw. Ilość warstw oraz zawartych w nich neuronów nie jest wprost zależna od danych wejściowych oraz wyjściowych.
- **Warstwa wyjściowa** – jej wartości wynikają z wyjścia poprzedzającej ją warstwy ukrytej. Ilość neuronów zależna jest od rozwiązywanego problemu. W przypadku kategoryzacji ilość neuronów uzależniona jest od możliwych kategorii. Wyjątkiem jest klasyfikacja binarna która wykorzystuje jeden neuron do przedstawienia wyjścia.

Warstwy połączone są ze sobą za pomocą poszczególnych neuronów. Każdy neuron połączony jest z wszystkimi neuronami warstwy poprzedniej. Połączenia przekazują wartości z funkcji aktywacji poprzedniego neuronu przemnożone przez odpowiadające im wagi oraz dodane obciążenie. Wagi oraz obciążenie nazywane są również parametrami modelu. Są one inicjalizowane losowo a następnie podczas trenowania są one zmieniane tak aby uzyskać jak najmniejszą funkcję kosztu.



Rysunek 2.1. Graficzna reprezentacja prostej sieci neuronowej

Źródło: Opracowanie własne

gdzie:

- x_1, x_2, x_3 oznaczają dane wejściowe
- x_0 oraz $a_0^{(2)}$ oznaczają neurony obciążenia (ang. *bias neurons*), neurony obciążenia zawsze równe są 1 więc $x_0 = a_0^{(2)} = 1$
- $L1$ oznacza pierwszą warstwę w sieci neuronowej która jest warstwą wejściową (ang. *input layer*)
- $L2$ oznacza drugą warstwę w sieci neuronowej która jest warstwą ukrytą (ang. *hidden layer*)
- $L3$ oznacza trzecią warstwę w sieci neuronowej która jest warstwą wyjścia (ang. *out layer*)
- $a_i^{(j)}$ oznacza aktywacja neuronu i w warstwie j (2.2.3)
- $h_W(x)$ wartość hipotezy h od x która jest parametryzowana za pomocą macierzy wag W. W przypadku użycia innej macierzy wag W otrzymamy inną hipotezę $h_W(x)$ lub inną funkcję która mapuje pewne X na Y

2.2.3 Funkcja aktywacji

Funkcją aktywacji nazywamy funkcję która przetwarza(w sposób zależny od wyboru rodzaju funkcji) zsumowane, przemnożone przez macierz wag $W^{(j)}$ dane wejściowe $x_1, x_2, x_3..., x_n$ .Uzyskany przez nią wynik nazywamy aktywacją neuronu.

Wartość poszczególnych funkcji aktywacji można zapisać przy pomocy następującego wzoru:

$$a_i^{(j)} = g(W_{in}^{(j-1)}x_1 + W_{in}^{(j-1)}x_2 + ... + W_{in}^{(j-1)}x_n) \tag{2.1}$$

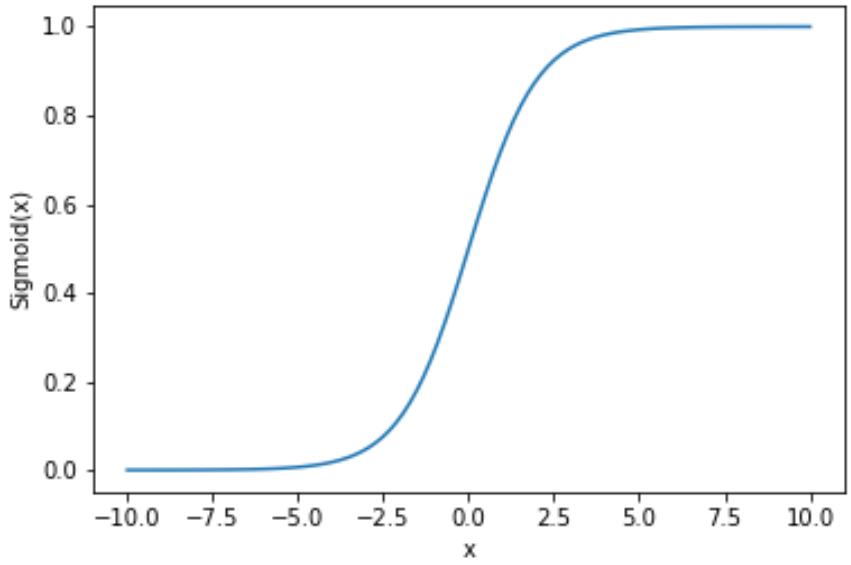
gdzie:

- $a_i^{(j)}$ oznacza aktywacja neuronu i w warstwie j
- $x_1, x_2, x_3..., x_n$ oznacza poszczególne dane wejściowe
- $g(x)$ - oznacza przyjętą funkcję do obliczenia aktywacji neuronu (np.Sigmoid)
- $W_{in}^{(j-1)}$ jest to pewna wartość odpowiadająca wartości z kolumny i i wiersza n macierzy $W^{(j-1)}$. Gdzie wartość i opisuje wybrany neuron a wartość n wybrane wejście x_n
- macierz $W^{(j)}$ nazywamy macierzą wag (parametrów) odpowiadającą połączeniom z warstwy j tej do warwy $j+1$. Jeżeli sieć posiada s_j neuronów w warstwie j oraz s_{j+1} neuronów w warstwie $j+1$ to macierz $W^{(j)}$ będzie posiadała wymiary s_{j+1} na $s_j + 1$ gdzie 1 oznacza dodatkowy neuron obciążeń (ang. *bias neuron* który zapisujemy jako x_0.

Finalny wynik sieci neuronowej również jest obliczany przy pomocy funkcji aktywacji. W takim przypadku do jego obliczenia używa się wyników z otatniej warstw ukrytej (ang. *hidden layer*) która równocześnie znajduje się przed warstwa wyjściową (ang. *output layer*)

Przykłady poszczególnych funkcji aktywacji $g(x)$:

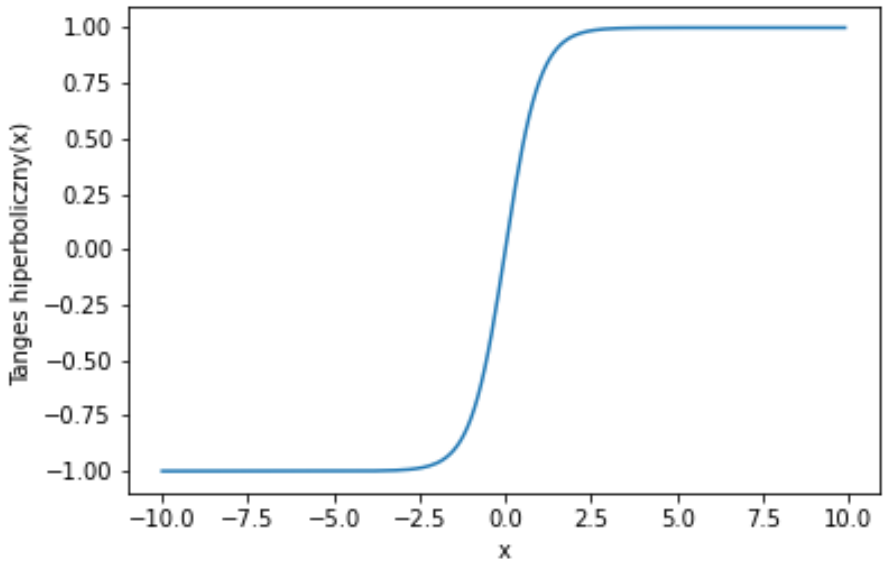
1. Funkcja Sigmoid którą określa wzór: $\frac{1}{1+e^{-x}}$ reprezentowana przez wykres:



Rysunek 2.2. Funkcja sigmoid

Źródło: Opracowanie własne przy pomocy bibliotekii matplotlib

2. Funkcja tangensa hiperbolicznego którą określa wzór: $\frac{e^x - e^{-x}}{e^x + e^{-x}}$ reprezentowana przez wykres:



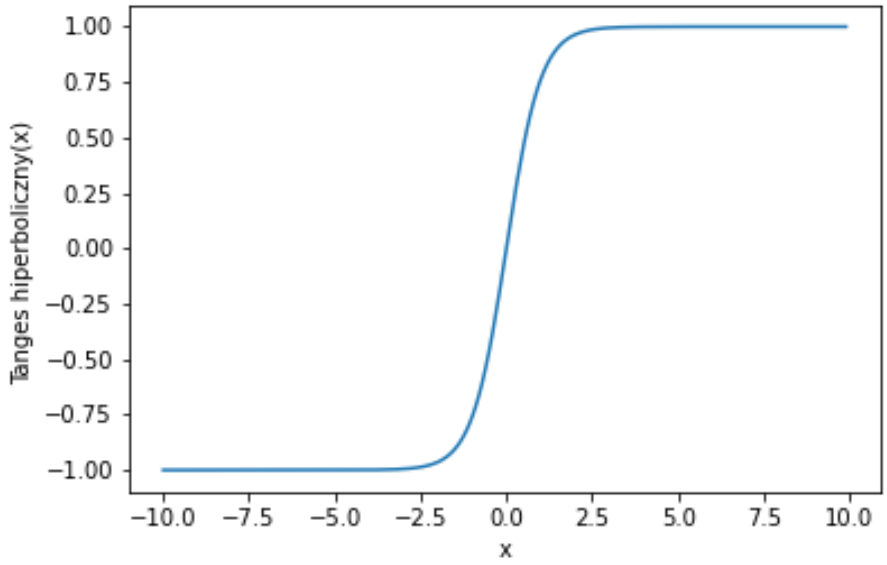
Rysunek 2.3. Funkcja tangesa hiperbolicznego

Źródło: Opracowanie własne przy pomocy bibliotekii matplotlib

3. Funkcja Relu (ang. *Rectified linear unit*) którą określa wzór:

$$max(0, x) = \begin{cases} 0 & \text{gdy } x \leq 0 \\ x & \text{gdy } x > 0 \end{cases} \tag{2.2}$$

reprezentowana przez wykres:



Rysunek 2.4. Funkcja tangesa hiperbolicznego

Źródło: Opracowanie własne przy pomocy bibliotekii matplotlib

4. Funkcja Softmax którą określa wzór: $\frac{e^{x_i}}{\sum_{j-1}^{J} e^{x_i}}$

2.2.4 Funkcja kosztu

Funkcja kosztu określa jak dobrze została zrealizowana dana predykcja dla określonych danych wejściowych.Im jest ona większa tym mniej dokładna okazała się predykcja. W drodze treningu sieci neuronowej wartość funkcji kosztu dąży do minimum.W przypadku sieci neuronowych funkcja kosztu jest zmodyfikowaną funkcją kosztu występującą w regresji liniowej postaci :

$$J(W) = \frac{1}{m} \sum_{i=1}^{m} Cost(h_W(x^{(i)}), y^{(i)}) \tag{2.3}$$

gdzie:

$$Cost(h_W(x), y) = \begin{cases} -log(h_W(x)) & \text{gdy } y = 1 \\ -log(1 - h_W(x)) & \text{gdy } y = 0 \end{cases}$$

więc:

$$J(W) = -\frac{1}{m}[\sum_{i=1}^{m} y^{(i)} \log h_W(x^{(i)}) + (1 - y^{(i)}) \log\big(1 - h_W(x^{(i)})\big)] \tag{2.4}$$

W przypadku funkcji kosztu sieci neuronowej zamiast pojedynczej wartości wyjścia posiadamy K wyjść. K zależy od typu badanego problemu klasyfikacji. W przypadku klasyfikacji binarnej K zawsze jest równe 1(0 jeżeli wystąpiła klasa A i 1 jeżeli wystąpiła klasa B. W klasyfikacji która posiada więcej niż 3 klasy liczba K równa jest liczbie klas. Jest to spowodowane przypisaniem dla każdej klasy indywidualnego neuronu. Zatem funkcja kosztu sieci neuronowej posiada postać:

$$J(W) = -\frac{1}{m}[\sum_{i=1}^{m}\sum_{k=1}^{K} y_k^{(i)} \log\big(h_W(x^{(i)})\big)_k + (1 - y_k^{(i)}) log((1 - h_W(x^{(i)}))_k)] \tag{2.5}$$

2.2.5 Propagacja wsteczna

W 1986 roku David Rumelhart, Geoffrey Hinton i Ronald Williams opublikowali przełomowy artykuł (Rumelhart i in. 1986) w którym wprowadzili koncepcję algorytmu propagacji wstecznej. Algorytm propagacji wstecznej jest to algorytm gradientu prostego (ang. *Gradient Descent* przy użyciu skutecznej techniki automatycznego obliczania gradientu:algorytm propagacji w dwóch przebiegach (po jednym w przód i w tył (ang. *one forward ,one backward* jest w stanie obliczyć gradient błędu sieci w odniesieniu do każdego parametru modelu. Może zatem określić w jaki sposób należy zmodyfikować parametry (wagi połączeń i neurony obciążeń (ang. *bias neuron*) tak aby zmniejszyć wartość błędu. Po uzyskaniu tych gradientów realizowany jest klasyczny algorytm gradientu prostego a cały cykl jest powtarzany do momentu osiągnięcia zbieżności z rozwiązaniem (Géron 2017) Zatem aby obliczyć $minJ(W)$ należy obliczyć funkcję kosztu $J(W)$ oraz pochodną $\frac{\partial}{\partial W_{ij}^{(l)}} J(W)$. Do obliczenia tej pochodnej wykorzystuje się następujący wzór:

$$\frac{\partial}{\partial W_{ij}^{(l)}} J(W) = a_j^{(l)} \delta_i^{(l+1)} \tag{2.6}$$

gdzie:

- $a_j^{(l)}$ to funkcja aktywacji w neuronie j w warstwie l

- $\delta_j^{(l)}$ to błąd w neuronie j w warstwie l który oblicza się na dwa sposoby:
  1. $\delta_j^{(L)} = a_j^{(L)} - y_j$ dla warstwy wyjścia gdzie L oznacza łączną liczbę warstw
  2. $\delta_j^{(l)} = (W^{(l)})^T \delta^{(l+1)} . * g'(W^{(l-1)} a^{(l-1)})$ dla warstw ukrytych gdzie l oznacza numer warstwy ukrytej

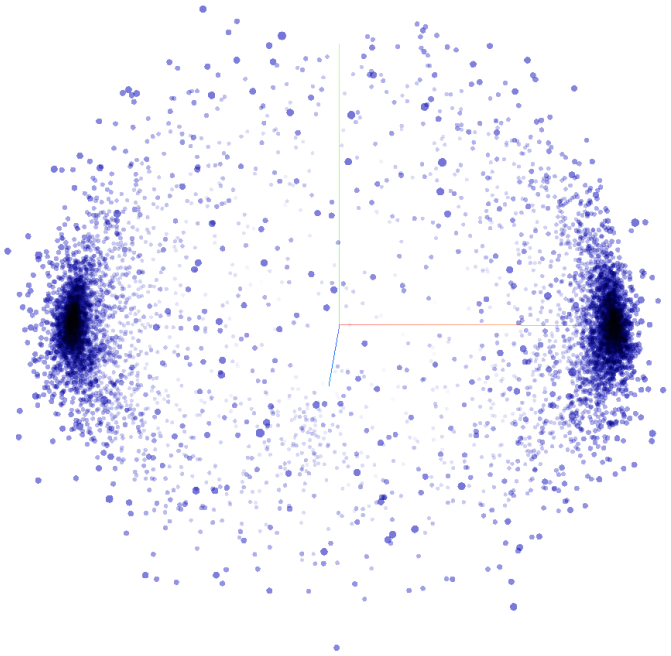
2.2.6 Wektory właściwościowe

Wektorem właściwościowym (osadzeń) (ang. *Embeddings*) nazywamy modyfikowalny wektor gęsty reprezentujący kategorię (Géron 2017). Domyślnie ich wartość jest inicjalizowana losowo. Wektory właściwościowe są modyfikowalne i wraz z procesem nauki modelu zbliżają się do wartości optymalnych. W przypadku gdy dwa wektory reprezentują podobną kategorię np. tweety które charakteryzują się negatywnym wydźwiękiem wypowiedzi, algorytm gradientu prostego będzie zbliżał je do siebie tworząc skupienia na wielowymiarowym wykresie[1]. Dzięki wektorom właściwościowym oraz ich reprezentacji danych, sieci neuronowe są w stanie uzyskiwać dokładniejsze prognozy dlatego proces uczenia dąży przeważnie do uczynienia z wektorów właściwościowych przydatnych reprezentacji kategorii. Jest o tak zwane uczenie reprezentacji (Géron 2017).

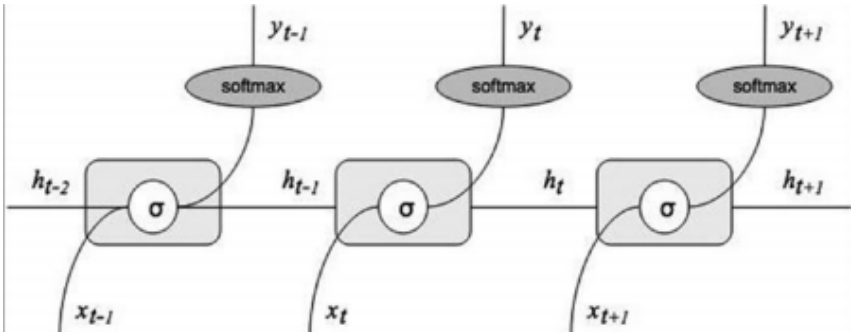
2.2.7 Sieci Rekurencyjne

Rysunek 2.6 przedstawia budowę rekurencyjnej sieci neuronowej, każdy węzeł x_t reprezentuje pewien wektor wejścia jednostki t, h_t reprezentuje pewny wektor wyjścia jednostki t przekazywany do kolejnej jednostki $t+1$ a y_t reprezentuje wektor wyjścia jednostki t przekształcony przez funkcję aktywacji softmax . Każdy węzeł oprócz pierwszego w całej sieci neuronowej otrzymuje wyjście z poprzedzającego go węzła $t-1$ jako dodatkowe wejście h_{t-1}. Na wyjście każdego węzła t oznaczane przez y_t nie tylko wpływają jego dane wejściowe ale również wnioskowanie dokonane przez poprzednie węzły. Dzięki tej funkcji sieci rekurencyjne są idealnym typem do rozpoznawania wzorców w sekwencjach danych.Podstawową budowę modeli RNN charakteryzuje jednak wysoka podatność na problem znikających i eksplodujących gradientów (Sarang 2021).

[1]Zazwyczaj wykorzystuje się wykresy 16 bądź 32 wymiarowe.



Rysunek 2.5. Przykład analizy skupień wektorów właściwościowych na podstawie danych z imdb



Rysunek 2.6. Rekurencyjna sieć neuronowa

Źródło: Artificial Neural Networks with Tensorflow 2 , Poornachandra Sarang, Rozdział 7, str 294.(Sarang 2021)

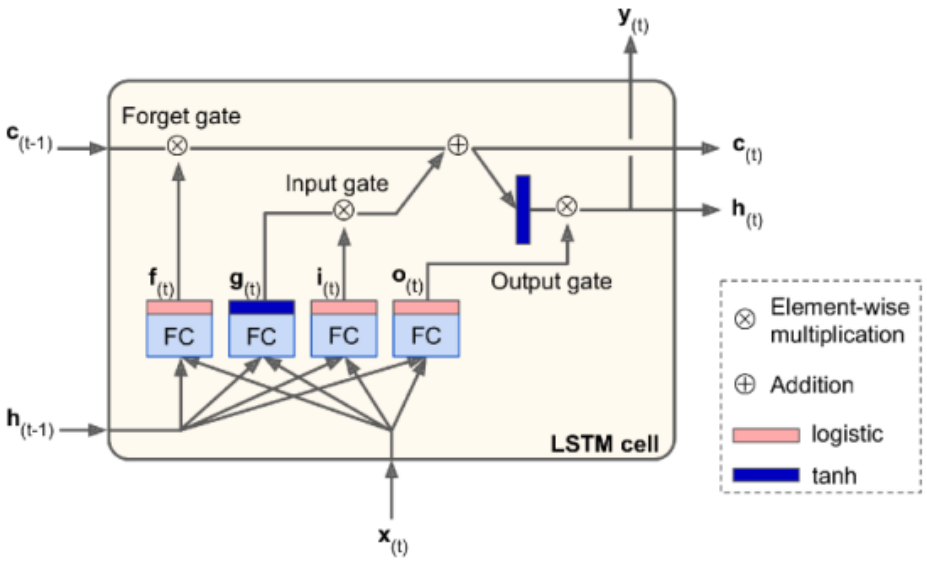
2.2.8 Zanikające i eksplodujące gradienty

Znikające i eksplodujące gradienty (ang. *The Vanishing and Exploding Gradient*) to nazwy na bardzo często występujące problemy sieci RNN. Podczas nauki modelu sieci neuronowych gradienty poddawane są wstecznej propagacji aż do warstwy początkowej. Wartości gradientów maleją wraz z przebiegiem algorytmu do warstw niższych .W przypadku głębszych sieci,

ze względu na długie łańcuchy, rozszerzenie gradienty kurczą się wykładniczo, zbliżając się do wartości znacznie poniżej 1 i czasami zanikają powodując zatrzymanie dalszej nauki modelu. Nazywa się ten problem znikającymi gradientami. Z drugiej strony jeżeli posiadają one o wiele większe wartości powyżej 1, stają się one większe wraz z postępem łańcucha i mogą spowodować eksplodowanie modelu. Ten problem nazywa się problemem eksplodujących gradientów (Sarang 2021).

2.2.9 Sieci wykorzystujące komórki długiej pamięci krótkotrwałej

Sieci wykorzystujące komórki długiej pamięci krótkotrwałej zwane w skrócie LSTM (ang. *long short-term memory*) są specyficzną architekturą sieci RNN. Zostały zaprezentowane pierwszy raz w 1997 roku przez Hochreiter'a oraz Schmidhuber'a (Hochreiter i Schmidhuber 1997). Główną zaletą sieci LSTM jest ich zdolność do długoterminowego zapamiętywania informacji przy jednoczesnym wyeliminowaniu problemu zanikających i eksplodujących gradientów. LSTM'y tak samo jak sieci rekurencyjne posiadają formę łańcucha powtarzających się modułów ale ich struktura jest o wiele bardziej skomplikowana. (Sarang 2021)



Rysunek 2.7. Struktura komórki LSTM

Źródło: Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 2nd Edition, rozdział 15, rysunek 15.9 (Géron 2017)

Komórka LSTM rozdzielona jest pomiędzy dwa wektory h i c. Wektor h interpretuje się jako stan krótkotrwały a c jako długotrwały. Podstawową koncepcją jest to ,że sieć uczy się co ma zapamiętywać w stanie długotrwałym, co odrzucać oraz co z niego odczytywać. Stan długotrwały porusza się w prawą stronę. Najpierw przechodzi przez bramkę zapominającą (ang. *forget gate*) mającą na celu porzucenie części wspomnień zapisanych wcześniej, następnie dodaje nowe wspomnienia za pomocą operacji sumowania(dodawane są dane wyselekcjonowane przez bramkę wejściową (ang. *input gate*)). Wynikowy stan c jest przesyłany dalej bez żadnych modyfikacji. Każdy cykl usuwa pewne wspomnienia oraz dodaje nowe. Po operacji sumowania stan jest kopiowany i przeprowadzony przez bramkę tangensa hiperbolicznego po czym otrzymany rezultat jest filtrowany przez bramkę wyjściową (ang. *output gate*). Uzyskuje się w ten sposób pewien stan wyjściowy h który jest również stanem wyjściowym y. (Géron 2017)

Nowe wspomnienia generowane są przez prosty algorytm. Najpierw do czterech różnych w pełni połączonych warstw (na rysunku oznaczone jako FC) zostaje wprowadzony wektor x oraz poprzedni wektor h. Każda z tych warstw posiada oddzielne zadanie do spełnienia. Warstwa główna generuje wartości wektora g. Jej zadaniem jest analizowanie bieżących x i h. W komórce podstawowej występuje tylko ta warstwa, a jej wynik jest przekazywany do wektorów y i h. Z kolei w komórce LSTM wynik uzyskiwany w tej warstwie nie jest wypuszczany dalej, lecz jego na istotniejsze elementy są przechowywane w stanie długotrwałym. Trzy pozostałe warstwy są kontrolerami bramek. Korzystają one z logicznej funkcji aktywacji, dlatego ich wyniki są w zakresie od 0 do 1.Ich wyniki przechodzą przez operację iloczynu elementowego, więc w przypadku wartości 0 bramka zostaje zamknięta a wartość 1 ją otwiera (Géron 2017) .Poszczególne bramki przeznaczone są do konkretnych funkcji:

- Bramka zapominająca (sterowana przez wektor f) definiuje które składniki pamięci długotrwałej powinny być zapomniane;
- Bramka wejściowa(sterowana przez wektor i) określa które elementy wektora g powinny być dodawane do pamięci długotrwałej;
- Bramka wyjściowa (sterowana przez wektor o służy do określenia które składowe stanu długotrwałego powinny być odczytywane i wysyłane na wyjście w danym okresie(do wektora h i y);

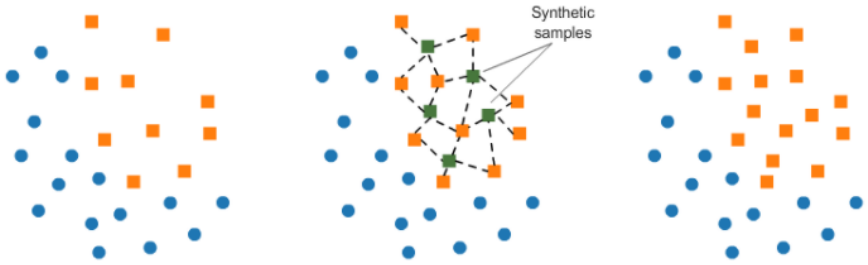
Podsumowując komórka LSTM jest w stanie rozpoznawać bardziej istotne dane i przechowywać je w dłuższym okresie czasu, tak długo jak są potrzebne. Potrafi ona odpowiednio od-

czytywać potrzebne informacje oraz zastępować informacje nieprzydatne informacjami o większym (w danym momencie) znaczeniu.

2.3 Operacje na danych

2.3.1 metoda SMOTE

Metoda SMOTE to skrót od angielskiego *Synthetic Minority Oversampling Technique*. Jest to technika służąca do radzenia sobie z problemem niezbalansowanych klas w zbiorze danych. Polega ona na wygenerowaniu dodatkowych, syntetycznych danych (ang. *oversampling*) tak aby wszystkie klasy posiadały równą liczbę wystąpień w zbiorze danych. Metoda ta korzysta z już poznanych danych z grup mniejszościowych (ang. *minority class*) i tworzy z nich nowe dane o tej samej etykiecie. W przypadku danych które zawierają jedynie dwie zmienne (x,y) nowe dane mogą być reprezentowane jako pewne losowo wygenerowane punkty pomiędzy istniejącymi punktami klasy mniejszościowej co reprezentuje rysunek 2.8.

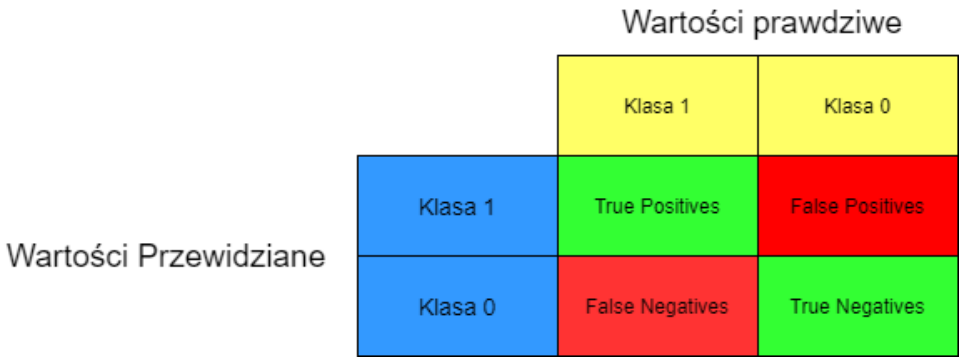


Rysunek 2.8. Przedstawienie działania metody SMOTE na przykładzie dwóch zmiennych

Źródło: https://www.kdnuggets.com/2019/09/5-sampling-algorithms.html

2.3.2 Ocena modelu

Podczas trenowania modelu ważne jest aby na bieżąco mierzyć jakość jego predykcji. Aby tego dokonać tworzy się macierz błędu zaprezentowaną na rysunku poniżej:



Rysunek 2.9. Macierz błędu dla problemu klasyfikacji binarnej

Źródło: Opracowanie własne

gdzie :

- (True positive - prawdziwie pozytywne) czyli model przewidział przykład jako pozytywny i faktycznie był on pozytywny w danych prawdziwych
- (True negative - prawdziwie negatywne) czyli model przewidział ,że przykład jest negatywny oraz rzeczywiście był negatywny w danych prawdziwych
- (False positive - fałszywie negatywne) czyli model przewidział ,że przykład jest pozytywny ale był on przykładem negatywnym w danych prawdziwych
- (False negative - prawdziwie negatywne) czyli model przewidział ,że przykład jest negatywny ale był on przykładem pozytywnym w danych prawdziwych

Przy wykorzystaniu wartości zaprezentowanych w macierzy błędu można następnie obliczyć metryki które ukażą poszczególne właściwości modelu: Przykładami tych metryk są:

1. Trafność (ang. *Accuracy*) - jest to stosunek ilości poprawnie przewidzianych wartości do wszystkich wartości. Parametr używany dla danych gdzie częstotliwość występowania poszczególnych klas jest równa. W przypadku dużych dysproporcji zawsze wskaże na grupę bardziej liczną. Oblicza się go za pomocą wzoru:

$$\frac{\text{True positive} + \text{True negative}}{\text{True positive} + \text{True negative} + \text{False positive} + \text{False negative}} \tag{2.7}$$

2. Wrażliwość (ang.*Recall lub sensitivity* - mierzy ile z faktycznie pozytywnych przypadków udało się poprawnie zakwalifikować jako pozytywne.Używa się go w problemach gdzie

istotna jest wysoka wykrywalność.Oblicza się go za pomocą wzoru:

$$\frac{\text{True positive}}{\text{True positive} + \text{False negative}} \tag{2.8}$$

3. Precyzja (ang. *Precision* - pozwala ocenić ile z przypadków przewidzianych jako pozytywne faktycznie jest pozytywnych. Oblicza się go za pomocą wzoru :

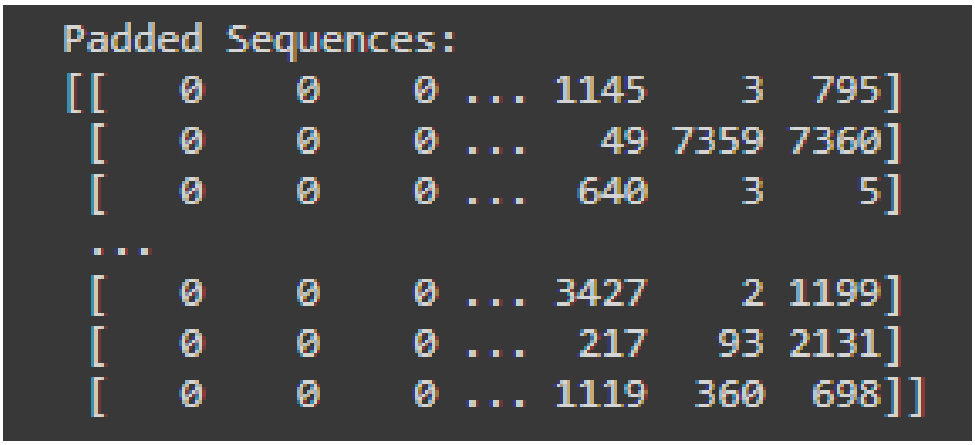
$$\frac{\text{True positive}}{\text{True positive} + \text{False positive}} \tag{2.9}$$

2.3.3 Przekształcanie danych tekstowych – Tokenizator

Tokenizator to narzędzie które przekształca dane przedstawione w postaci ciągów znaków na tensory liczbowe.Taka operacja jest wymagana, ponieważ modele uczenia maszynowego nie są w stanie uczyć się z danych tekstowych.Proces przetwarzania danych można podzielić na 3 główne etapy:

1. Pierwszym etapem jest sporządzenie słownika słów przy wykorzystaniu danych treningowych.Przypisuje on każdemu unikalnemu słowu daną wartość liczbową (np.kot:1).
2. Drugim etapem jest utworzenie wektorów (ang. *sequences*) odpowiadających poszczególnym zdaniom.Wektory te składają się z liczb utworzonych w etapie pierwszym.Ich długość zależna jest od ilości słów zawartych w zdaniu.
3. Ostatnim etap odpowiada za ujednolicenie długości wektorów utworzonych w etapie drugim.Model sieci neuronowych wymaga ujednoliconych danych wejściowych aby działał poprawnie. Ujednolicenie polega na zgromadzeniu wszystkich zdań w jednej macierzy(ang.*padded*). Tak skonstruowana macierz posiada liczbę wierszy równą ilości zdań w zbiorze danych oraz liczbę kolumn odpowiadającą liczbie najdłuższego badanego zdania. W przypadku gdy zdanie było krótsze od zdania najdłuższego różnicę w ich długości wypełnia się zerami. W zależności od przyjętej metody zera mogą wypełnić wektor przed zdaniem (pre) lub po zdaniu(post)

W przypadku przygotowywania danych testowych do późniejszej walidacji modelu wykorzystuje się jedynie etap drugi oraz etap trzeci. Wykorzystują one słownik utworzony przy pomocy modelu treningowego który został wzbogacony o jedną dodatkową zmienną oznaczaną jako OOV token (ang. *Out of vocabulary*) która oznacza słowo które nie zostało ujęte w zbiorze treningowym a pojawiło się w zbiorze testowym.



Rysunek 2.10. Przykład macierzy padded z wartością wypełniania pre

Źródło: Opracowanie własne

2.4 Podsumowanie

W ostatnich latach sieci neuronowe drastycznie zyskują na popularności, dzieje się tak, ponieważ są w stanie poradzić sobie z bardzo skomplikowanymi i wielowymiarowymi problemami które dotychczas były poza zasięgiem systemów uczenia się maszyn.Jednym z takich zastosowań jest analiza tekstu a jego szczególną podgrupą jest analiza wydźwięku wypowiedzi. Przy wykorzystaniu komórek LSTM oraz sieci neuronowych w następnym rozdziale zostało przeprowadzone badanie mające na celu skonstruowanie modelu który poradziłby sobie z problemem klasyfikacji wypowiedzi użytkowników Twittera na temat Globalnego Urzędu Statystycznego.

Rozdział 3

Empiryczna ocena wydźwięku wypowiedzi o Głównym Urzędzie Statystycznym

3.1 Pozyskiwanie i przetwarzanie danych

3.1.1 Źródło danych

Do przeprowadzenia analizy sentymentu wykorzystano dane pozyskane z strony internetowej Twitter. Pobrano trzydzieści sześć tysięcy tweetów które zawierały jedno z słów kluczowych. Wszystkie słowa kluczowe były kombinacją słowa GUS z dodatkowymi znakami bądź wyrazami np. #GUS, gus, GUS polska itp. Do pozyskania danych została wykorzystana biblioteka Twint. Twint przy pomocy metody web scrappingu pobrał dane z rozszerzeniem json .Każdy tweet posiadał 36 zmiennych w tym id, datę , nazwę użytkownika, język i treść wiadomości.

3.1.2 Przetwarzanie danych

Proces przetwarzania danych polegał na kilku etapach:

1. Pierwszy etap polegał na wyselekcjonowaniu tweetów jedynie napisanych w języku polskim. Aby usunąć wszystkie tweety obcojęzyczne zastosowano potrójne przeskanowanie danych. Pierwsze skanowanie zostało zastosowane podczas pobierania danych w parametrze `lang="pl"` biblioteki Twint. Pierwszy etap nie okazał się skuteczny w pełni więc zastosowano kolejne dwa aby zminimalizować prawdopodobieństwo tweeta w innym języku już w zbiorze testowym. Drugim etapem było usunięcie wszystkich tweetów które

nie miały wartości ”pl”w jednej z 36 oryginalnych kolumn pobranych z Twittera. Trzecim etapem było zastosowanie biblioteki `langdetect` i utworzenie zbioru z Tweetów które biblioteka bezpośrednio zakwalifikowała jako tweety w języku polskim.

2. Drugim etapem było usunięcie wszystkich znaków które nie niosły ze sobą żadnego znaczenia semantycznego. Były to wszystkie emotikony oraz znaki specjalne (np.#) lub linki do stron.Do wyselekcjonowania ich z ciągów znaków oraz następnego usunięcia użyto biblioteki że”która polega na zastosowaniu wyrażeń regularnych.
3. Trzecim etapem było zastosowanie Tokenizatora z biblioteki `Tensorflow` tak aby z danych tekstowych uzyskać wektory na których mógłby nauczy się model. Jego szczegółowe działanie zostało opisane w 2.3.3.

3.1.3 Etykietowanie

Etykietowanie zostało wykonane za pomocą subiektywnej oceny autora pracy licencjackiej.Ręcznie zostało oznaczonych 5 tysięcy wiadomości tekstowych, które zostały przyporządkowane do jednej z trzech poniższych grup:

- -1 - kategoria reprezentująca tweety o negatywnym wydźwięku wypowiedzi,
- 0 - kategoria reprezentująca tweety o neutralnym wydźwięku wypowiedzi,
- 1 - kategoria reprezentująca tweety o pozytywnym wydźwięku wypowiedzi.

Sam proces oznaczania zabrał duże pokłady czasu. Wymagał on przeczytania każdej wiadomości oraz zastanowienia się nad jej wydźwiękiem. W procesie oznaczania kierowano się przypuszczaną intencją autora wiadomości. Jeżeli wiadomość miała na celu obrazić lub upokorzyć inną osobę bądź instytucję została ona zaliczana jako negatywna.Wiadomość pozytywna charakteryzowała się grzecznym wyrazem zadowolenia lub podziękowania wobec jej adresata. Natomiast wiadomość neutralna była to wiadomość która z założenia autora nie niosła wraz z swoim znaczeniem dodatkowych emocji. Takie wiadomości najczęściej reprezentowały pewne fakty statystyczne przygotowane przez instytucje GUS.

W przypadku zbiorów danych które są nieoznaczone zazwyczaj należy oznaczyć ręcznie przynajmniej pewną część tych danych. W przypadku oszustw finansowych z użyciem kart kredytowych czy też wykrywaniu wczesnych etapów raka wykonują tą operację specjaliści w tych dziedzinach. Zbiór badanych danych w pracy nie zawierał natomiast bardzo technicznych cech więc do jego etykietowania autor zastosował swoją dążącą do obiektywizmu opinię. W przy-

padku dalszych prac w tej dziedzinie autor sugeruje zwiększenie liczby osób odpowiedzialnych za proces etykietowania tak aby uzyskać bardziej optymalną ocenę (w skali populacji) na temat poszczególnych wypowiedzi

3.1.4 Problem niezbalansowanych danych

Problem niezbalansowanych danych występuje gdy jedna lub więcej klas posiada znacząco większą liczbę obserwacji. W przypadku nauki modelu opartego na sieciach neuronowych taki typ danych może wpłynąć na wiarygodność mierników wydajności (w szczególności precyzji). W danych przygotowanych do nauki modelu występują następujące liczebności poszczególnych klas:

- 4307 tweetów o neutralnym wydźwięku wypowiedzi,
- 432 tweety o negatywnym wydźwięku wypowiedzi,
- oraz 219 tweetów o pozytywnym wydźwięku wypowiedzi.

W tego typu zbiorze danych można spodziewać się precyzji na poziomie ilorazu najbardziej licznej grupy (w tym przypadku grupy neutralnej o liczebności 4307) oraz łącznej liczby tweetów. Aby zapobiec tego typu generalizacją które może wykorzystywać model przy procesie nauki stosuje się metody over lub under samplingu. Polegają one na dodaniu (over) lub usunięciu (under) danych tak aby wyrównać liczebności w poszczególnych zbiorach. W pracy wykorzystano metodę over samplingu aby nie tracić potencjalnych informacji z odrzuconych danych. Wykorzystaną metodą była metoda SMOTE która została opisana w 2.3.1

Innymi przykładami w których można zastosować metody dodania lub usunięcia danych mogą być problemy przedstawione w sekcji 3.1.3

3.2 Wyniki

3.2.1 Wyniki modelu podstawowego

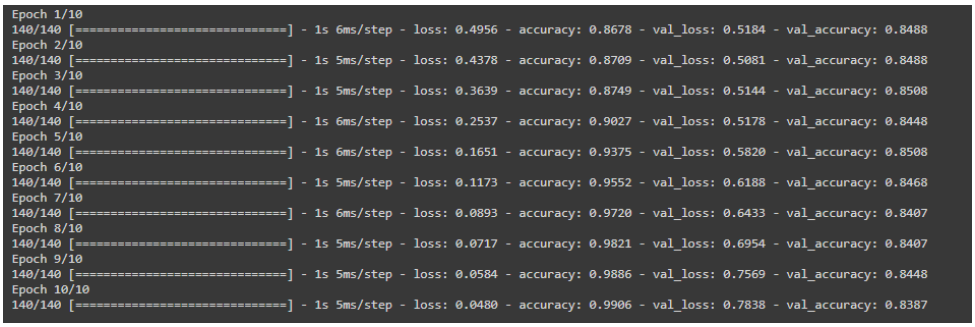
Model podstawowy zbudowany został za pomocą bibliotekii Tensorflow. Zawierał on 4 warstwy:

1. Warstwę wektorów właściwościowych (Embeddings) - reprezentującą słowa w przestrzeni;

2. Warstwę Flatten która spłaszcza wejście do jednego wektora;
3. Pierwszą warstwę neuronową która zawierała 5 jednostek obliczeniowych oraz posiadała funkcję aktywacji Relu;
4. Drugą warstwę neuronową która posiadała 3 jednostki obliczeniowe oraz funkcję aktywacji softmax.

Model wykorzystywał sparse categorical crossentropy jako funckję straty oraz funkcję adam jako optymalizator.

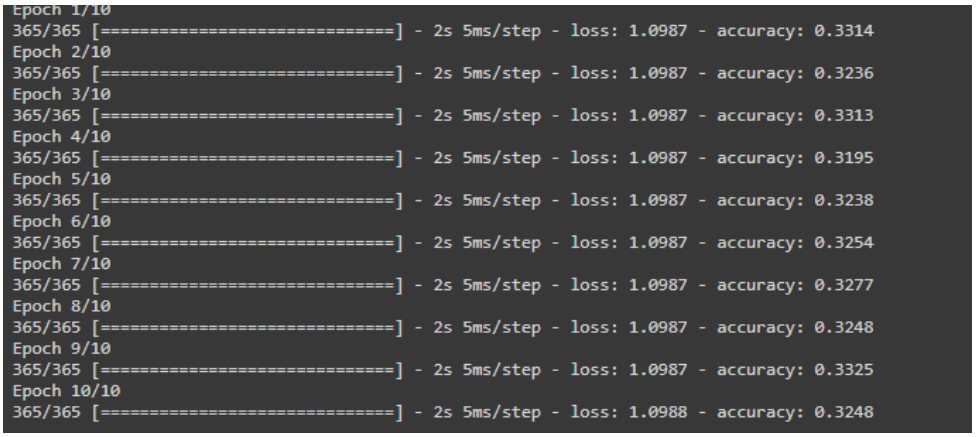
Model został wytrenowany dwukrotnie, raz przy pomocy danych po etapie przetworzenia 3.1.2 oraz drugi raz po zastosowaniu metody SMOTE 2.3.1. Podczas pierwszej próby model już w 4 epoce osiągnął poziom precyzji na poziomie ponad 90 procent, zakończył proces uczenia na poziomie ponad 99 procent a w zbiorze testowym oscylowała ona wokół wartości 85 procent. Takie wyniki zostały uzyskane ,ponieważ model posiadał zbyt dużą ilość danych jednej kategorii i zaczął oznaczać wszystkie tweety etykietą grupy najliczniejszej (jako tweet neutralny).Problem ten został opisany w 3.1.4



Rysunek 3.1. Reprezentacja procesu uczenia się modelu nie przekształconego podstawowego

Źródło: Opracowanie własne

Podczas drugiej próby wyniki trafności (ang.*accuracy* wynosiły około 33 procent. Takie wyniki zostały uzyskane , ponieważ model próbował ślepo trafić w jedną z trzech grup. Model nie był w stanie wyuczyć się poprawnie wzorców jakie definiowały przynależność tweetów do danych klas dlatego w pracy zdecydowano się na zastosowanie modelu bazującego na sieciach LSTM.



Rysunek 3.2. Reprezentacja procesu uczenia się modelu podstawowego używajacego metody SMOTE

Źródło: Opracowanie własne

3.2.2 Wyniki modelu LSTM

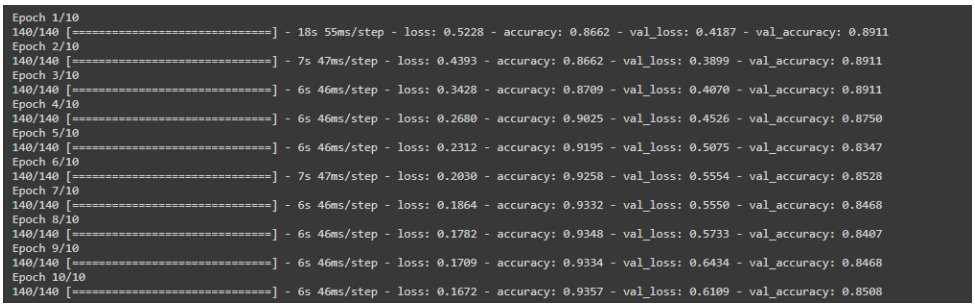
Sieci LSTM powszechnie wykorzystuje się w pracach poświęconych sekwencjom danych. Kluczową ich funkcjonalnością jest przechowywanie części informacji z poprzednich komórek do następnych dzięki czemu są w stanie odkryć wzorce które definiują np. przynależność do danej klasy. W pracy zdecydowano się na model który zawierał 5 warstw:

1. Warstwę wektorów właściwościowych (Embeddings) - reprezentującą słowa w przestrzeni;
2. Warstwę Flatten która spłaszcza wejście do jednego wektora;
3. Pierwszą dwukierunkową warstwę LSTM która zawierała 64 jednostki obliczeniowe;
4. Drugą dwukierunkową warstwę LSTm która zawierała 32 jednostki obliczeniowe;
5. Pierwszą warstwę neuronową która zawierała 5 jednostek obliczeniowych oraz posiadała funkcję aktywacji relu;
6. Drugą warstwę neuronową która posiadała 3 jednostki obliczeniowe oraz funkcję aktywacji softmax.

Wykorzystano również tak jak w przypadku modelu podstawowego sprase categorical crossentropy jako funkcję straty oraz adam jako optymalizator.

Wytrenowano model dwukrotnie w tych samych warunkach jak w przypadku modelu podstawowego: raz przy wykorzystaniu danych przetworzonych oraz raz przy wykorzystaniu metody SMOTE.

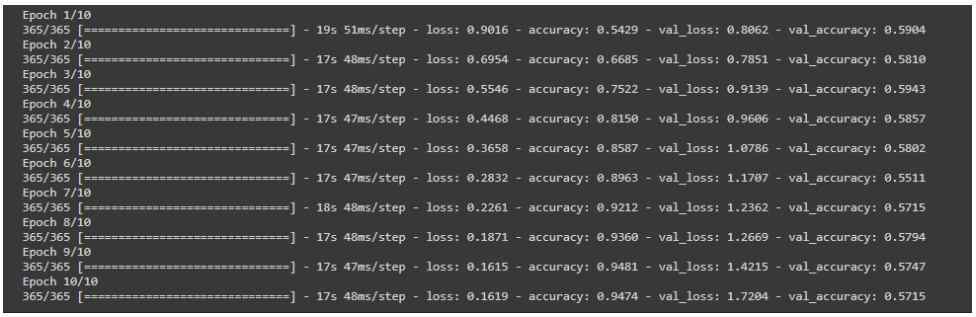
Wyniki pierwszej próby które zawierały dane przetworzone nie różniły się znacząco od wyników uzyskanych w pierwszej próbie w modelu podstawowym.Rezultaty były głownie uwarunkowane przez duży brak równowagi między liczebnościami poszczególnych klas. Precyzja była odwzorowaniem prawdopodobieństwa uzyskania prawidłowego wyniku przy zaznaczaniu wszystkich obserwacji etykietą najczęściej występującą (w tym przypadku etykietą definiującą wypowiedź neutralną).Pomimo braku widocznych różnic pomiędzy wynikami tych dwóch modeli (które nie wykorzystują techniki SMOTE) można zauważyć, że problem niezbalansowanej liczby przykładów znacząco wpływa na otrzymany wynik (niezależnie od struktury sieci) i wymaga on zastosowania dodatkowej metody (np. SMOTE)



Rysunek 3.3. Reprezentacja procesu uczenia się modelu nie przekształconego LSTM

Źródło: Opracowanie własne

Wyniki drugiej próby modelu LSTM znacznie różniły się od wyników drugiej próby modelu podstawowego. W przypadku modelu wykorzystującego metodę SMOTE oraz sieci LSTM w zbiorze treningowym można było zaobserwować płyny wzrost trafności (ang.*accuracy*) wraz z kolejnymi epokami oraz łagodny spadek wskaźnika straty. W przypadku zbioru testowego natomiast trafność oscylowała w okolicach około 58 procent podczas całego procesu uczenia się.



Rysunek 3.4. Reprezentacja procesu uczenia się modelu wykorzystującego SMOTE oraz LSTM

Źródło: Opracowanie własne

3.3 Analiza wyników klasyfikacji

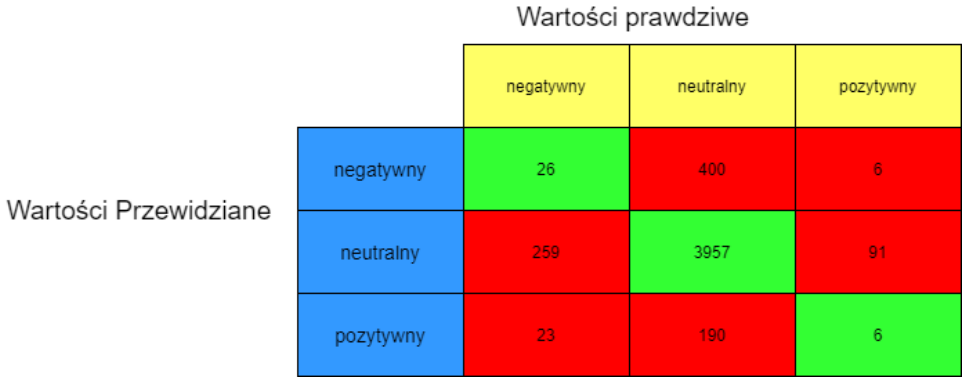
3.3.1 Problemy modelu

Model nie sprostał oczekiwaniom i w obecnej postaci nie jest dobrym klasyfikatorem wydźwięku wypowiedzi.Według autora pracy powodem tego mogą być:

1. Zbyt mała ilość danych – sieci neuronowe wymagają większych zbiorów danych aby prawidłowo rozpoznawać wzorce oraz przypisać poszczególne wiadomości do określonych klas.
2. Zbyt prosty model który zawierał zbyt małą ilość parametrów do dobrego opisania badanych wzorców. Rozwój modelu oraz dodanie kolejny warstw mogłoby spowodować wzrost ogólnej trafności modelu ale jedynie w połączeniu z większym zbiorem treningowym.Większa ilość parametrów wymaga większej ilości danych do prawidłowej konfiguracji.
3. Zbyt mała wariancja techniki SMOTE. Wygenerowane przykłady nie odpowiadały przyszłym przykładom które wystąpiły w zbiorze treningowym.

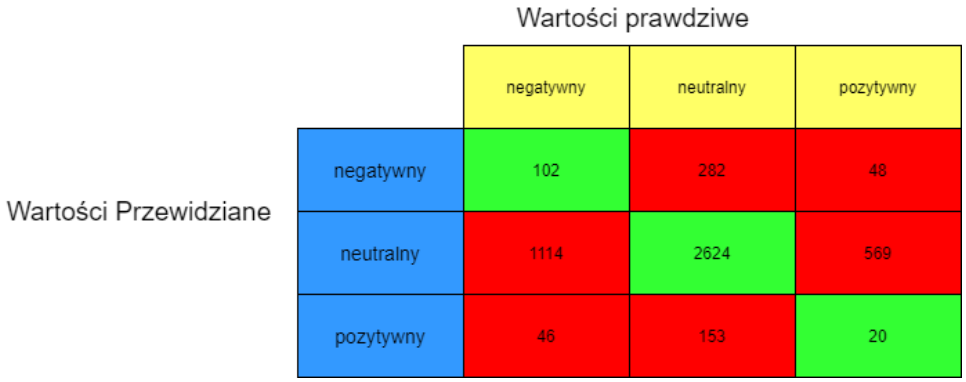
3.3.2 Ocena modelu przy pomocy macierzy błędu

Metoda SMOTE poprawiła ogólny wynik prawdziwie pozytywnych dla klasy negatywnej oraz pozytywnej ale znacząco obniżyła ilość poprawnie zakwalifikowanych jako neutralne na rzecz błędnie zakwalifikowanych jako negatywne.



Rysunek 3.5. Macierz błędu wykorzystująca LSTM ale nie wykorzystująca metody SMOTE

Źródło: Opracowanie własne



Rysunek 3.6. Macierz błędu modelu wykorzystującego metodę SMOTE oraz LSTM

Źródło: Opracowanie własne

3.3.3 Przykład błędnej klasyfikacji

Dobrym przykładem jest tweet: *eksport do Chin w pierwszych 11 miesiącach 2020 roku według danych GUS wyniósł 3,06 miliarda dolarów, a wartość importu z Chin do Polski do listopada 2020 roku wzrosła o 10,8 procent do kwoty 33,62 miliarda dolarów. Deficyt 1 do 10, ale było jeszcze gorzej 1 do 12*.

Oznaczony został on jako tweet neutralny, ponieważ zawiera jedynie informacje która reprezentuje wielkość eksportu do Chin. Model natomiast oznaczył tweet jako negatywny, mogło to być spowodowane dość negatywnym zakończeniem wiadomości które wynikało z użycia mocno nacechowanym słów. Fraza *"ale było jeszcze gorzej"*mogła być wyznacznikiem dla modelu aby skategoryzować prezentowaną wiadomość do klasy negatywnej.

Podsumowanie

Celem niniejszej pracy było skonstruowanie modelu, który byłby w stanie z dobrą dokładnością zakwalifikować wypowiedzi użytkowników Twittera na temat GUS do trzech klas (pozytywnej, neutralnej oraz negatywnej).

Zbiór danych treningowych zawierał około 5 tysięcy tweetów i został pobrany przy pomocy biblioteki `Twint`. Przed utworzeniem modeli skupiono się na prawidłowym przekształceniu oraz wyczyszczeniu danych. Zrobiono to za pomocą bibliotek: `langdetect`, `re` oraz modułu `Tokenizer` z biblioteki `Tensorflow`. Następnym krokiem było utworzenie 4 odrębnych modeli tak aby móc porównać wyniki przy zastosowaniu różnych metod.

Stworzono modele które wykorzystywały metodę SMOTE oraz komórki LSTM. Najlepiej wypadł model który wykorzystywał komórki LSTM oraz metodę SMOTE. Jego trafność oscylowała wokół 95 procent na zbiorze testowym. Natomiast w przypadku zbioru testowego posiadał jedynie 58 procent trafności. Powodem takich wyników jest zdecydowanie zbyt mała ilość danych uczących oraz zbyt mała złożoność modelu.

Podsumowując sieci neuronowe wykorzystujące komórki LSTM oraz metodę SMOTE mogą być dobrą metodą do analizy sentymentu ale potrzebują bardzo dużych ilości danych aby prawidłowo rozpoznać wzorce które definiują poszczególne klasy.

Bibliografia

Abbasi, A., Rashidi, T., Maghrebi, M. & Waller, S. (2015). Utilising Location Based Social Media in Travel Survey Methods: bringing Twitter data into the play. https://doi.org/10.1145/2830657.2830660

Asur, S. & Huberman, B. A. (2010). Predicting the future with social media. *2010 IEEE/WIC/ACM international conference on web intelligence and intelligent agent technology*, *1*, 492–499.

Bollen, J., Mao, H. & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of computational science*, *2*(1), 1–8.

Burnap, P., Gibson, R., Sloan, L., Southern, R. & Williams, M. (2016). 140 characters to victory?: Using Twitter to predict the UK 2015 General Election. *Electoral Studies*, *41*, 230–233.

Burnap, P., Williams, M. L., Sloan, L., Rana, O., Housley, W., Edwards, A., Knight, V., Procter, R. & Voss, A. (2014). Tweeting the terror: modelling the social media reaction to the Woolwich terrorist attack. *Social Network Analysis and Mining*, *4*(1), 206.

Crisci, A., Grasso, V., Nesi, P., Pantaleo, G., Paoli, I. & Zaza, I. (2018). Predicting TV programme audience by using twitter based metrics. *Multimedia Tools and Applications*, *77*(10), 12203–12232.

Daas, P. J. & Puts, M. J. (2014). *Social media sentiment and consumer confidence* (tech. rep.). ECB Statistics Paper.

Dave, K., Lawrence, S. & Pennock, D. M. (2003). Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. *Proceedings of the 12th international conference on World Wide Web*, 519–528.

Droba, D. (1931). Methods used for measuring public opinion. *American Journal of Sociology*, *37*(3), 410–423.

Géron, A. (2017). *Hands-on machine learning with Scikit-Learn and TensorFlow : concepts, tools, and techniques to build intelligent systems*. O'Reilly Media.

Goodfellow, I. J., Bengio, Y. & Courville, A. (2016). *Deep Learning* [http://www.deeplearningbook.org]. MIT Press.

Hochreiter, S. & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735–1780.

Joinson, A. N. (2008). Looking at, looking up or keeping up with people? Motives and use of Facebook. *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, 1027–1036.

Kaczmarczyk, S. (2018). Zalety i wady metod zbierania danych przez Internet w badaniach marketingowych. *Zeszyty Naukowe. Organizacja i Zarządzanie/Politechnika Śląska*.

Knutson, A. L. (1945). Japanese opinion surveys: the special need and the special difficulties. *Public Opinion Quarterly*, *9*(3), 313–319.

Liu, B. (2009). Handbook chapter: sentiment analysis and subjectivity. Handbook of natural language processing. *Handbook of Natural Language Processing. Marcel Dekker, Inc. New York, NY, USA*.

Małyszko, J. (2015). Automatyczne przetwarzanie recenzji konsumenckich dla oceny użyteczności produktów i usług. *Poznań: Uniwersytet Ekonomiczny w Poznaniu. Pobrane z: http://www. wbc. poznan. pl/Content/354211/Malyszko_Jacek_rozprawa_doktorska. pdf*.

Mäntylä, M., Graziotin, D. & Kuutila, M. (2016). The Evolution of Sentiment Analysis – A Review of Research Topics, Venues, and Top Cited Papers. *Computer Science Review*, *27*. https://doi.org/10.1016/j.cosrev.2017.10.002

McCulloch, W. S. & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, *5*(4), 115–133.

Mellon, J. & Prosser, C. (2017). Twitter and Facebook are not representative of the general population: Political attitudes and demographics of British social media users. *Research & Politics*, *4*(3), 2053168017720008.

Moskowitz, H. R. & Martin, B. (2008). Optimising the language of email survey invitations. *International Journal of Market Research*, *50*(4), 491–510.

Nassirtoussi, A. K., Aghabozorgi, S., Wah, T. Y. & Ngo, D. C. L. (2014). Text mining for market prediction: A systematic review. *Expert Systems with Applications*, *41*(16), 7653–7670.

oldartice [Accessed: 2021-04-23]. (nodate).

Rauchfleisch, A. & Kaiser, J. (2020). The False positive problem of automatic bot detection in social science research. *PloS one*, *15*(10), e0241045.

Richmond, J. (1998). Spies in ancient Greece. *Greece & Rome*, *45*(1), 1–18.

Rumelhart, D. E., Hinton, G. E. & Williams, R. J. (1986). Learning representations by back-propagating errors. *nature*, *323*(6088), 533–536.

Sarang, P. (2021). *Artificial Neural Networks with TensorFlow 2*. Springer.

Shao, C., Ciampaglia, G. L., Varol, O., Flammini, A. & Menczer, F. (2017). The spread of fake news by social bots. *arXiv preprint arXiv:1707.07592*, *96*, 104.

Stagner, R. (1940). The cross-out technique as a method in public opinion analysis. *The Journal of Social Psychology*, *11*(1), 79–90.

Thorley, J. (2004). *Athenian democracy*. Psychology Press.

Spis rysunków