## 1. Wstęp

Obecnie wykrywanie emocji stało się bardzo częstym problemem informatycznym. W wielu dziedzinach okazuje się ono niezbędne. Pierwszą, najbardziej oczywistą branżą jest robotyka i rozwój sztucznej inteligencji. Wielu naukowców ciągle pracuje nad stworzeniem robota posiadającego ludzkie emocje i umiejącego je rozpoznać wchodząc w interakcje z człowiekiem. Ponadto wiedza na temat ludzkich emocji przydaje się w marketingu. Firmy chcą wiedzieć jakimi uczuciami kierują się ich klienci na przykład podczas dokonywania zakupów na stronie internetowej. Również w polityce, ta wiedza jest przydatna. Podczas ważnych wydarzeń wiele osób udziela się na forach internetowych. Dzięki możliwość automatycznego wykrywania emocji z tekstu można szybko dowiedzieć jaki wydźwięk ma dana sytuacja polityczna wśród obywateli. Ostatnią, najważniejszą rzeczą jest umiejętność rozpoznania emocji drugiego człowieka w kontaktach międzyludzkich.

Cielem jest wykrycie nacechowania emocjonalnego w krótkich formach wypowiedzi, czyli określenie, czy dana wypowiedź jest pozytywna czy negatywna. W poniższej pracy przedstawiono autorski algorytm oparty o polski WordNet Słowosieć oraz algorytm sztucznej inteligencji SVM. Obydwie metody pozwalają na automatyczne wykrywanie wydźwięku emocjonalnego podanego tekstu.

## Dane testowe

Poniższy rozdział przedstawia proces zbierania i przygotowania danych testowych. Pozyskano je z różnych źródeł internetowych przy pomocy autorskich skryptów.

### Zebranie danych

W celach przetestowania algorytmu przygotowano 5 zbiorów testowych. Zebrano je z 4 różnych portali społecznościowych zawierających wypowiedzi użytkowników. Przy wyborze portalu brano pod uwagę jakość komentarzy, występowanie oceny przypisanej do wypowiedzi, różnorodność tekstów i łatwość w pobraniu dużej ilości informacji ze strony. W celu uzyskania jak największego zbioru danych napisano skrypt w języku Python. Miał on za zadanie pobrać z każdego linku URL wypowiedzi użytkowników oraz ocenę. W kolejnym kroku przypisywano do zebranych wypowiedzi emocje według poniższego algorytmu:

Gdy skala ocen tekstu jest od 1-10

1. Wszystkie wypowiedzi, których oceny są poniżej 6, określa się jako negatywne.
2. Wypowiedzi z ocenami powyżej 5 określa się jako pozytywne.

Gdy skala ocen tekstu skala ocen jest o 1-5

1. Wszystkie wypowiedzi, których oceny są poniżej 4, określa się jako negatywne.
2. Wypowiedzi z ocenami powyżej 3 określa się jako pozytywne.

W końcowej fazie testowy zbiór składał się z krótkiej formy tekstowej, oceny użytkownika i przypisanej do oceny emocji. Został zapisany w formacie CSV.

|  |  |  |
| --- | --- | --- |
| **Komentarz** | **Emocja** | **Ocena** |
| Adam McKay zrealizował film swojego życia. Najambitniejszy, imponujący tempem, jednak bardzo przystępny i zrozumiały nawet dla takiego laika jak ja. | positive | 9.0 |
| Ten film tak mnie oczarował, że zapomniałem o irytującej klimatyzacji, zerkaniu na zegarek, a nawet zapomniałem o tym, że chce mi się pić. 130 minut seansu minęło zupełnie jak 30. Jednym słowem: miazga. | positive | 9.0 |
| Zręczny technicznie i aktorsko remake argentyńskiego arcydzieła, pozbawiony jego siły i polotu. | negative | 4.0 |
| Rzecz profesjonalnie zrobiona, ale pusta. Dramaty, przegrane uczucia, bolesne tragedie - wszystko to wypada w filmie bezdusznie. | negative | 4.0 |

##### Tabela prezentująca część pliku CSV z komentarzami ze strony MediaKrytyk

W poniższych podrozdziałach zostaną opisane opracowane zestawy danych.

### Zbiór ocen filmów z portalu FilmWeb[[1]](#footnote-0)

Jest to największy bank danych. Zebrano w nim 12821 komentarzy filmów. Posłużył on do wyuczenia się klasyfikatora SVM. Jednak w wielu przypadkach napisane przez użytkowników teksty były średniej jakości, zawierały dużo potocznych słów i niekoniecznie odnosiły się do wystawionej oceny filmu.

|  |  |
| --- | --- |
| **Komentarze** | **12821** |
| **Pozytywne** | 10423 |
| **Negatywne** | 2398 |

##### Tabela przedstawiająca stosunek negatywnych do pozytywnych komentarzy z Filmwebu

### Zbiór komentarzy z portalu MediaKrytyk

Dla porównania skuteczności algorytmów zebrano dane z drugiego portalu posiadającego komentarze filmów wraz z ocenami użytkowników. Strona MediaKrytyk zbiera wszystkie recenzje filmów. Wypowiedzi użytkowników są bardziej profesjonalne i zawierają dużo mniej słów potocznych. Ten zbiór testowy posiada 6073 komentarzy.

|  |  |
| --- | --- |
| **Komentarze** | **6073** |
| **Pozytywne** | 4124 |
| **Negatywne** | 1949 |

##### Tabela przedstawiająca stosunek negatywnych do pozytywnych komentarzy z MediaKrytyk

### Zbiór recenzji z portalu filmWeb

W celu sprawdzenia jak algorytm zachowuje się podczas wykrywania emocji w długich formach wypowiedzi zebrano 319 recenzji filmów, gdzie minimalna liczba znaków wynosiła 300.

|  |  |
| --- | --- |
| **Komentarze** | **319** |
| **Pozytywne** | 249 |
| **Negatywne** | 70 |

##### Tabela przedstawiająca stosunek negatywnych do pozytywnych recenzji z Filmwebu

### Zbiór recenzji produktów ze strony Opineo

Dla porównania różnych typów wypowiedzi zebrano recenzje produktów sprzętu AGD i RTF z portalu Opineo. Zbiór ten posiada 6311 komentarzy do produktów. Wypowiedzi zawierają informacje na temat jakości i użyteczności zakupionego sprzętu.

|  |  |
| --- | --- |
| **Komentarze** | **6311** |
| **Pozytywne** | 6067 |
| **Negatywne** | 244 |

##### Tabela przedstawiająca stosunek negatywnych do pozytywnych komentarzy z Opineo

### Zbiór recenzji restauracji z TripAdvisor

Również dla porównania typów wypowiedzi zebrano recenzje wrocławskich i warszawskich restauracji z portalu TripAdvisor. Zbiór ten składa się z 6919 komentarzy. Wypowiedzi dotyczą jakości obsługi, serwowanego jedzenia i wyglądu wnętrza restauracji czy baru.

|  |  |
| --- | --- |
| **Komentarze** | **6919** |
| **Pozytywne** | 1217 |
| **Negatywne** | 5702 |

##### Tabela przedstawiająca stosunek negatywnych do pozytywnych komentarzy z TripAdvisor

### Lematyzacja danych

Po utworzeniu zbiorów nastąpił proces ujednolicenia form wyrazów. W procesie tym biorą udział dwa ważne pojęcia: stemming oraz lematyzacja[[2]](#footnote-1). Stemming jest to proces doprowadzenia słowa do tak zwanego rdzenia pozbawionego wszelkich końcówek, będącego częścią wspólną dla danej części wyrazów. Lematyzacja według słownika języka polskiego jest to sprowadzanie formy fleksyjnej wyrazu do postaci słownikowej. Potrzebny jest do niej słownik lub rozbudowany zestaw reguł fleksyjnych dla danego języka. Dzięki tym zabiegom program jest wstanie rozpoznać pojedyncze słowa i ich znaczenie. Po ujednoliceniu form wyrazów dużo więcej słów ze zdania algorytm był wstanie odnaleźć w słowniku Słowosieci. Do lematyzacji słów oraz stemmingu użyto biblioteki PoliMorfologik[[3]](#footnote-2). Jest to projekt oparty o tworzenie słownika morfosyntaktycznego. Posiada około 215 tysięcy leksemów i ponad 3.5 miliona form wyrazowych. Jego zadaniem jest przedstawienie słowa w postaci znacznika morfosyntaktycznego. “Każdy znacznik morfosyntaktyczny jest ciągiem wartości rozdzielonych dwukropkami, np.: subst:sg:nom:m1 dla segmentu chłopiec. Pierwsza wartość, np. subst, określa klasę gramatyczą, następne zaś, np. sg, nom i m1 wartości odpowiednich dla tej klasy kategorii gramatycznych" [[4]](#footnote-3). Starano się sprowadzić wszystkie słowa do postaci rzeczownik, czasownika lub przymiotnika, ponieważ dla tych form było największe prawdopodobieństwo znalezienia w słowniku Słowosieci w dalszym etapie algorytmu.

|  |  |
| --- | --- |
| **Komentarz** | **Komentarz po lematyzacji** |
| Choć film ciężki i nudny, warto go zobaczyć. Zrozumieć procesy, które wpłynęły na nasze życia. | [warto, proces, nasze, ciężki, wpłynąć, i, na, życie, który, nudny, go, zobaczyć, film] |
| Wręcz genialny film, wykorzystujący inne ujęcie tematu kryzysu oraz przedstawiający go z zupełnie innej perspektywy. | [wykorzystywać, przedstawiać, oraz, ujęcie, inny, kryzys, inny, perspektywa, temat, genialny, go, zupełnie, z, film] |
| Film z ciekawym pomysłem i ze skopaną realizacją, gdzie powaga miesza się z groteską, a ciągłe puszczanie "oczka" do widza nuży już po drugim zdaniu Ryana Goslinga. | [do, zdanie, pomysł, z, widz, realizacja, ciągły, groteska, ciekawy, a, mieszać, i, powaga, drugi, już, gdzie, się, skopać, Ryan, z, puszczanin, po, nużyć] |
| Stanowi z pewnością przede wszystkim ciekawy pomysł na film, który niestety z kolejnymi minutami nieco się rozmywa. Mimo to, ze względu na humor i wykreowanie postaci i świata, ogląda się go z przyjemnością. | [wszystko, przed, z, się, świat, pomysł, oglądać, postać, wykreować, przyjemność, go, wzgląd, Stan, film, humor, i, pewność, na, ciekawy, to, niestety, który, kolejny, z, rozmywać, minuta, nieco] |

##### Przykłady komentarzy przed i po lematyzacji

### Znaki specjalne oraz wielkość liter

Ważnym elementem było również usunięcie znaków specjalnych ze zdań. Niektóre z nich, takie jak apostrof górny, powodowały przerwanie procesu lematyzacji. Inny przykładem był znak kropki, który uniemożliwiał w kolejnym etapie wyszukanie słowa w słowniku emocji.

|  |  |  |
| --- | --- | --- |
| **Modyfikacje** | **Komentarz** | **Komentarz po lematyzacji** |
| brak modyfikacji | Potężne rozczarowanie. Jeśli czerpaliście przyjemność z lektury książki Cline'a - w kinie będzie Was aż skręcać z zażenowania i niedowierzania, że z porządnej fabularnie pozycji wyciśnięto coś na tyle marnego i wypłyconego. Elementy, które składają się na głębię każdej historii: psychologia postaci, sensowne dialogi między nimi, osadzenie owych bohaterów w szerszym społeczeństwie | [czerpać, przyjemność, książka, z, lektura] |
| usunięcie znaku kropki i przecinka | Potężne rozczarowanie Jeśli czerpaliście przyjemność z lektury książki Cline'a - w kinie będzie Was aż skręcać z zażenowania i niedowierzania że z porządnej fabularnie pozycji wyciśnięto coś na tyle marnego i wypłyconego Elementy które składają się na głębię każdej historii: psychologia postaci sensowne dialogi między nimi osadzenie owych bohaterów w szerszym społeczeństwie | [rozczarowanie, czerpać, przyjemność, książka, z, lektura] |
| usunięcie średnika ze słowa “Cline'a” | Potężne rozczarowanie Jeśli czerpaliście przyjemność z lektury książki Clinea - w kinie będzie Was aż skręcać z zażenowania i niedowierzania że z porządnej fabularnie pozycji wyciśnięto coś na tyle marnego i wypłyconego Elementy które składają się na głębię każdej historii: psychologia postaci sensowne dialogi między nimi osadzenie owych bohaterów w szerszym społeczeństwie | [niedowierzanie, osadzić, być, bohater, że, sensowny, fabularnie, czerpać, psychologia, postać, aż, on, na, książka, lektura, każdy, między, porządny, dialog, który, wycisnąć, coś, składać, marne, historia, przyjemność, głębić, ow, kin, szeroki, społeczeństwo, i, rozczarowanie, tyle, pozycja, w, się, z, zażenowanie, skręcać] |

##### Przykład komentarza, w którym znaki specjalne miały wpływ na proces lematyzacji

Jak widać w powyższej tabeli niewielkie modyfikacje tekstu miały bardzo duży wpływ na wynik procesu lematyzacji.

Kolejnym etapem procesu przygotowania danych było sprowadzenie wszystkich liter do małej wielkości. Był to znaczący etap, ponieważ bez niego biblioteka PoliMorfologik traktowała słowo zaczynające się z wielkiej litery jako własne, przez to nie sprowadzała go do formy fleksyjnej i tym samym uniemożliwiała znalezienie go w słowniku Słowosieci.

|  |  |
| --- | --- |
| **Komentarz** | **Komentarz po lematyzacji** |
| Marna - choć w obsadzie błyszcząca od gwiazd - wersja argentyńskiego thrillera. | [choć, błyszczący, thriller, od, Marna, wersja, w, argentyński, gwiazda, obsada] |

##### Przykład komentarza, w którym wielkość liter miała wpływ na proces lematyzacji. Słowo “Marna” program potraktował jako nazwę własną i nie zamienił jej na słowo “marny”

Dla podsumowania powyższych działań mających na celu poprawienie skuteczności działania algorytmu zmierzono jego skuteczność przed i po formatowaniu danych. Wyniki zamieszczono w tabeli poniżej.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Komentarze z portalu MediaKrytyk | Komentarze z portalu Opineo | Komentarze z portalu filmWeb | Komentarze z portalu TripAdvisor |
| Skuteczność algorytmu przed lematyzacją danych | 46% | 31% | 40% | 46% |
| Skuteczność algorytmu po lematyzacji danych | 59% | 57% | 48% | 73% |
| różnica punktów procentowych | 13 | 26 | 8 | 27 |

Na podstawie danych można stwierdzić, że średnio skuteczność algorytmu wzrosła o 18,5 punktów procentowych, dzięki dokonaniu lematyzacji słów oraz odpowiednim sformatowaniu badanych wypowiedzi.

Podczas analizy zdań natrafiono również na słowa, które biblioteka “PoliMorfologik” przekształca poprawnie, ale nie na części mowy odpowiednie dla słownika Słowosieci. Na przykład w zdaniu “*Mimo wielu wad, film Krzysztofa Langa ogląda się naprawdę dobrze.*” po lematyzacji wypisano słowo “dobrze”, gdyż jest to przysłówek. Niestety w Słowosieci nie występuje w takiej formie lecz w postaci przymiotnika - “dobry”.

## 8. Autorski algorytm oparty o Słowosieć

Nowatorska metoda została opracowana na podstawie polskiego WordNetu Słowosieć, który został stworzony na Politechnice Wrocławskiej. WordNet jest to leksykalna baza danych dla języka angielskiego[[5]](#footnote-4). Dla każdego zdania posiada zbiór synonimów, definicji, przykładów użycia i relacje pomiędzy nimi. Pierwszy WordNet powstał pod koniec lat 90’ na uniwersytecie w Princeton. Jest wykorzystywany w wielu aplikacjach związanych z analizą tekstów. Istnieje bardzo dużo odpowiedników WordNetu w innych językach, jednak z uwagi na bezpośrednie tłumaczenie z języka angielskiego, nie są one tak przydatne, ponieważ materiał leksykalny i sieć relacji silnie zależą od rozwiązań specyficznych dla języka angielskiego[[6]](#footnote-5).

Polski WordNet zaprojektowano i zaimplementowano w oparciu o polską semantykę leksykalną. Jego struktura odzwierciedla relacje semantyczno-leksykalne między polskim znaczeniami leksykalnymi. Jest on całkowicie wolny od bezpośredniego tłumaczenia z języka angielskiego, dlatego też jest wiernym odzwierciedleniem języka polskiego. Posiada on 80 tysięcy słów z określeniem emocjonalnym[[7]](#footnote-6).

Ważnym elementem Słowosieci jest adnotacja emocjonalna dla 30 tysięcy słów. Dla każdej pozycji przypisana jest **emocja podstawowa**: radość, zaufanie, cieszenie się na coś oczekiwanego, smutek, złość, strach, wstręt, zaskoczenie czymś nieprzewidywalnym, **wartość uniwersalna**: użyteczność, dobro drugiego człowieka, prawda, wiedza, piękno, szczęście, nieużyteczność, krzywda, niewiedza, błąd, brzydota, nieszczęście oraz **nastawienie emocjonalne**: - m, - s, 0, +s, +m (gdzie: - m oznacza mocno negatywne, - s słabo negatywne, 0 - neutralne, + s - słabo pozytywne, + m - mocno pozytywne). Na potrzeby zaproponowanego w ramach niniejszej pracy algorytmu pobrano informacje o nastawieniu emocjonalnym każdego słowa. Dzięki niemu możliwe jest rozpoznanie, czy dane słowo jest pozytywne, czy negatywne i tym samym określeniu, jak cała wypowiedź jest nacechowana.

|  |  |  |
| --- | --- | --- |
| **Adnotacja w słowniku emocji** | **Adnotacja w Słowosieci** | **Znaczenie** |
| -2 | -m | mocno negatywne |
| -1 | -s | słabo negatywne |
| 0 | 0 | neutralne |
| 1 | +s | słabo pozytywne |
| 2 | +m | mocno pozytywne |

##### Tabela prezentująca zapis w słowniku Emocji na podstawie Słowosieci

Do implementacji metody wybrano język Python, ponieważ jest bardzo wydajny pod względem przetwarzania dużej ilości danych. W celu przeszukiwania i tworzenia nowych tablic użyto biblioteki Pandas. Dzięki niej z łatwością można było przetworzyć cały słownik Słowosieci, który jest zapisany w postaci XML. Wybrano jedynie słowa z przypisanym nastawieniem emocjonalnym, czyli takie, które w pliku posiadały znacznik “lexical-unit”. Przetworzenie wszystkich pozycji ze Słowosieci i zapis wybranych do nowego pliku CSV zajęło 45 sekund. Powstały słownik składa się z dwóch elementów - słowa i wartości liczbowej, do której przypisana jest emocja. Wartości liczbowe słownika odpowiadają symbolom przypisanym do słów w Słowosieci.

Po stworzeniu słownika emocji zaczęto przetwarzać zgromadzone zbiory krótkich form wypowiedzi (informacje na temat przygotowania danych znajdują się w rozdziale 5), tak aby można było odnaleźć pojedyncze wyrazy.

|  |  |
| --- | --- |
| **Uczucie** | **Słowo** |
| **-1.0** | abstrakcyjny |
| **-1.0** | abstraktowy |
| **-2.0** | absurd |
| **-2.0** | absurdalność |
| **-2.0** | absurdalny |
| **-1.0** | achajski |
| **1.0** | adekwatność |
| **1.0** | adekwatny |

##### Tabela prezentująca przykładowe słownik emocji zawierający słowa z adnotacją emocjonalną

Algorytm w pierwszym etapie dokonuje procesu lematyzacji tekstu, dzięki któremu zwraca podany tekst w formie tablicy ze słowami w formie fleksyjnej. Następnie każde słowo jest wyszukiwane w słowniku uczuć i pobierana jest jego wartość (liczba w słowniku emocji). Jeżeli słowo nie zostało znalezione przypisywana jest mu wartość zero. Po wyszukaniu wszystkich słów zliczana jest ich wartość. Gdy jest ona wyższa od zera wypowiedź zostaje uznana za pozytywną, gdy zaś wartość jest ujemna tekst określany jest jako negatywny.

detectingEmotionsInText(words\_list)

1. Pobierz słownik emocji
2. Pobierz liczbę elementów (wyrazów) z tablicy words\_list
3. result ← 0
4. emotions\_list ← list()
5. if liczba elementow tablicy words\_list > 0
6. then emotionsCounter ← 0
7. for each element in words\_list
8. if element == slowo ze slownika emocji
9. dodaj element do emotions\_list
10. emotionsCounter += wartosc znalezionego slowa w slowniku emocji
11. result = emotionsCounter/numberOfWords

Aby sprawdzić skuteczność algorytmu przebadano go na 5 różnych zbiorach danych. Poniżej przedstawiono wyniki badania.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Recenzje** | **FilmWeb** | **MediaKrytyk** | **Opineo** | **TripAdvisor** |
| Liczba komentarzy | 319 | 12821 | 6073 | 6311 | 6919 |
| liczba komentarzy z fałszywym określeniem emocji | 171 | 6622 | 2493 | 2697 | 1900 |
| liczba komentarzy z poprawnym określeniem emocji | 148 | 6199 | 3580 | 3614 | 5019 |
| Skuteczność (%) | 46 | 48 | 59 | 57 | 73 |

Jak widać w powyższej tabeli najsłabsze wyniki otrzymano w teście dla zbioru recenzji. Była to mała baza tekstów posiadających długie wypowiedzi. Najlepsze zaś osiągnięto dla komentarzy z TripAdvisor. Liczba komentarzy była porównywalna z dwoma innymi zbiorami z MediaKrytyk i Opineo. W celu sprawdzenia jakie parametry wpływają na skuteczność zaproponowanej metody przebadano następujące zależności:

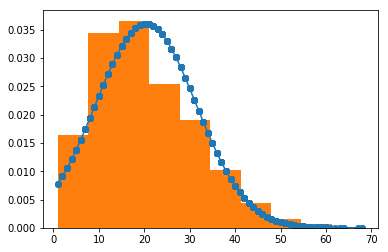
* długość wypowiedzi
* ocenę użytkownika
* liczbę znalezionych słów w słowniku Słowosieci
* typ wypowiedzi (pozytywna, negatywna)

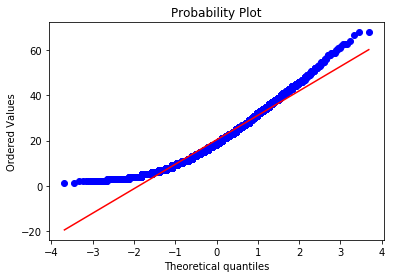
od poprawności określenia wydźwięku emocjonalnego wypowiedzi. W poniższych podrozdziałach zostaną opisane badania każdej z tych zależności dla poszczególnych zbiorów testowych.

### Długość wypowiedzi

W poniższym podrozdziale zostanie sprawdzona hipoteza mówiąca, że **długość wypowiedzi ma wpływ na poprawne wykrycie wydźwięku emocjonalnego danej wypowiedzi**. Przed wyborem odpowiedniej metody sprawdzającej zostanie wykonany test D'Agostino-Pearsona[[8]](#footnote-7). Do badania wybrano ten test, ponieważ nie jest znane odchylenie standardowe ani wartość średnia badanego zbioru. Ma on zbadać, czy dane są rozkładu normalnego. Przyjęto hipotezę zerową, że zbiór danych dotyczących długości wypowiedzi jest zbiorem o rozkładzie normalnym oraz poziom istotność wynosi 0.05. Wynik testu przedstawiono w poniżej tabeli:

|  |  |
| --- | --- |
| Statystyka K-kwadrat | 441.5382895181308 |
| wartość p | 0.0 |

Można stwierdzić, że rozkład nie jest rozkładem normalnym, ponieważ wartość współczynnika p jest mniejsza od poziomy istotności. Dlatego też odrzucono hipotezę zerową i przyjęto hipotezę alternatywną, że zbiór danych testowych nie ma rozkładu normalnego. Dla potwierdzenia zamieszczono wykres kwantyl-kwantyl oraz histogram, gdzie wyraźnie widać, że dane odbiegają od układu liniowego. 



Ponieważ dane nie są rozkładu normalnego sprawdzono korelację zmiennych przy pomocy współczynnika korelacji rangowej Spearmana[[9]](#footnote-8). Służy on do badania jak silna jest związek monotoniczny pomiędzy dwoma cechami. Wartość współczynnika mieści się w zakresie <-1,1>, gdzie dla wartości zbliżonych od -1 przyjmuje się silną ujemną zależność monotoniczną, a dla wartości zbliżonych do 1 przyjmuje się silną dodatnią zależność monotoniczna. W przypadku, gdy wartość współczynnika korelacji jest równa lub zbliżona do zera przyjmuje się, że nie istnieje monotoniczna zależność między zweryfikowanymi danymi. Jednak może istnieć związek niemonotoniczny. Poniżej przedstawiono wynik badania współczynnika korelacji rangowej Spearmana:

|  |  |
| --- | --- |
| Korelacja (rs) | -0.022930981099555894 |
| Wartość pvalue | 0.07395915683613187 |

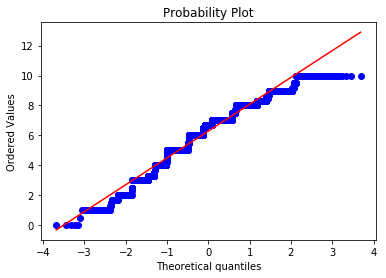
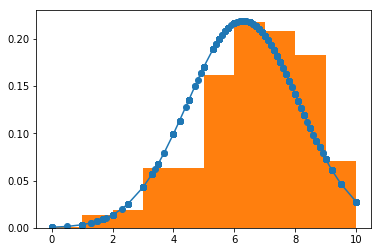
Jak widać wynik korelacji jest bliski zeru, dlatego też można przyjąć, że zmienne nie są skorelowane i przyjąć, że poprawność wykrywania emocji nie zależy od długość tekstu. Jednak na podstawie pvalue nie ma też podstaw do odrzucenia hipotezy zerowej.

### Ocena użytkownika

W poniższym podrozdziale sprawdzono hipotezę: **ocena użytkownika ma wpływ na poprawność wykrywania nacechowania emocjonalnego w wypowiedzi użytkownika**. Przed zbadaniem korelacji danych zdano, czy dane pochodzą z rozkładu normalnego. Zbadano jest również testem D'Agostino-Pearsona. Przyjęto hipotezę zerową, że zbiór danych dotyczących oceny użytkownika jest zbiorem o rozkładzie normalnym oraz poziom istotność wynosi 0.05. Wynik testu przedstawiono w poniżej tabeli:

|  |  |
| --- | --- |
| Statystyka K-kwadrat | 261.3714270755038 |
| wartość p | 0.0 |

Według powyższych danych wynika, że wartość pvalue jest niższa od poziomu istotności. Można odrzucić hipotezę zerową i przyjąć hipotezę alternatywną, że zbiór nie ma rozkładu normalnego. Dla potwierdzenia wyników badań przedstawiono wykres kwantyl-kwantyl oraz histogram.



Po przyjęciu hipotezy alternatywnej można zbadać korelację zmiennych posługując się współczynnikiem korelacji rangowej Spearmana. Badając zależność zmiennych przyjęto hipotezę zerową, mówiącą o tym, że ocena użytkownika zależy od skuteczności działania algorytmu na poziomie istotności 0.05. Wyniki badania przedstawiono w poniższej tabeli:

|  |  |
| --- | --- |
| Korelacja (rs) | -0.07070419863571814 |
| Wartość pvalue | 0.0 |

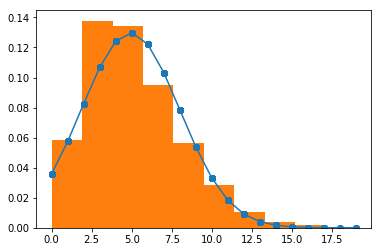
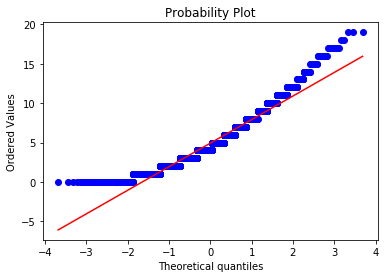
Wynika z nich, że wartość pvalue jest mniejsza od poziomu istotności oraz korelacja danych jest zbliżona do zera. Dlatego też można odrzucić hipotezę zerową i przyjąć hipotezę alternatywną, że ocena użytkownika nie zależy od wyniku metody.

### Liczba znalezionych słów w słowniku Słowosieci

W poniższym podrozdziale zweryfikowano hipotezę, że liczba znalezionych słów w słowniku Słowosieci ma wpływ na poprawność wykrycia nacechowania emocjonalnego w tekście. Przed sprawdzeniem korelacji zmiennych sprawdzono, czy dane są rozkładu normalnego. Tak jak we wcześniejszych podrozdziałach, posłużono się testem D'Agostino-Pearsona. Przyjęto hipotezę zerową mówiącą, że rozkład liczby znalezionych słów w słowniku jest rozkładem normalnym na poziomie istotności 0.05. Wyniki testu przedstawiono w poniższej tabeli.

|  |  |
| --- | --- |
| Statystyka K-kwadrat | 628.3854562034508 |
| wartość p | 0.0 |

Według badań wartość pvalue jest mniejsza od poziomu istotności, więc można odrzucić hipotezę zerową i przyjąć hipotezę alternatywną - rozkład danych nie jest rozkładem normalnym. Dla potwierdzenia zamieszczono poniżej wykres kwantyl-kwantyl oraz histogram.

Po stwierdzeniu, że badana korelacja dotyczy zmiennych z rozkładu, który nie jest rozkładem normalnym, posłużono się po raz kolejny współczynnikiem korelacji rangowej Spearmana. Przyjęto poziom istotności 0.05. Wyniki przedstawiono w poniżej.

|  |  |
| --- | --- |
| Korelacja (rs) | 0.03757292701421932 |
| Wartość pvalue | 0.0034063580813208723 |

Według powyższej tabeli wynik korelacji danych jest bliski zeru, a wartość pvalue jest mniejsza od poziomu istotności. Odrzucono więc hipotezę mówiącą, że liczba słów wykrytych w Słowosieci zależy od wyniku algorytmu.

### Typ wypowiedzi

W poniższym podrozdziale postawiono hipotezę, że typ wypowiedzi (pozytywny lub negatywny) ma wpływ na poprawność rozpoznania nacechowania emocjonalnego tekstu. Dzięki temu stwierdzeniu będzie można określić, czy algorytm osiąga lepsze wyniki dla wypowiedzi nacechowanych pozytywnie czy negatywnie. W tym celu podzielono zbiór danych na dwa podzbiory posiadające tylko komentarze pozytywne lub negatywne. Dla nich sprawdzono skuteczność metody. Wynik badania przedstawiono w poniższych tabelach.

|  |  |
| --- | --- |
| Komentarze pozytywne | |
| Liczba komentarzy z fałszywie wykrytym wydźwiękiem emocjonalnym | 1846 |
| Liczba komentarzy z pozytywnie wykrytym wydźwiękiem emocjonalnym | 2278 |
| Skuteczność (%) | 55 |

|  |  |
| --- | --- |
| Komentarze negatywne | |
| Liczba komentarzy z fałszywie wykrytym wydźwiękiem emocjonalnym | 647 |
| Liczba komentarzy z pozytywnie wykrytym wydźwiękiem emocjonalnym | 1302 |
| Skuteczność (%) | 67 |

Dla komentarzy negatywnych skuteczność algorytmu jest większa o 12 punktów procentowych. Według powyższego badania można przyjąć, że postawiono hipoteza jest prawdziwa i typ komentarza ma wpływ na poprawność wykrycia nacechowania emocjonalnego wypowiedzi.

## Algorytm SVM

Klasyfikator SVM jest metodą z dziedziny Machine Learning. Oznacza to, że algorytm sam uczy się na podstawie wcześniejszych doświadczeń. SVM polega na klasyfikacji binarnej, czyli podzieleniu zbioru na dwie klasy[[10]](#footnote-9). W przypadku problemu wykrywania emocji dzieli on zestaw danych na negatywne i pozytywne. Klasyfikuje dane na podstawie zbioru uczącego się, który wcześniej został mu podany.

W pracy zaimplementowano algorytm SVM przy użyciu biblioteki Scikit-learn w Pythonie[[11]](#footnote-10). Scikit-learn oferuje szereg narzędzi do uczenia maszynowego oraz analizy danych. W celach badawczych wykorzystano funkcję StratifiedKFold, TfidfVectorizer i LinearSVC.

### Funkcja StratifiedKFold

Funkcja implementuję metodę K-składowej walidacji krzyżowej Polega ona na podziale zbioru na K mniejszych podzbiorów o takiej samej bądź bardzo zbliżonej do siebie liczebności danych oraz powtórzeniu K-razy optymalizacji modelu.

### Funkcja TfidfVectorizer

Dzięki tej funkcji dokonano normalizacji danych tekstowych, tak aby mogły zostać użyte do wyuczenia klasyfikatora. Przy jej pomocy powstaje macierz, w której wartości komórek odpowiadają iloczynowi dwóch niezależnych od siebie czynników – częstości tematu TF i odwrotnej częstości dokumentu IDF.[[12]](#footnote-11) TF skaluje każdy wektor, aby suma jego składowych zawsze wynosiła 1. IDF jest miarą występowania danego tematu w korpusie tekstu.

### Funkcja linearSVC

Jest to funkcja implementująca klasyfikator SVM z liniową funkcją jądra. Wykorzystuje się ją między innymi do badań danych tekstowych.

W celu zbadania skuteczności działania klasyfikatora przeprowadzono badania na każdym z 5 zbiorów testowych. W pierwszym przypadku wyuczono algorytm zbiorem komentarzy z FilmWebu, a następnie podano mu 4 inne zbiory testowe. W drugim przypadku z każdego zbioru wydzielono część treningową i testową. 80% dany posłużyło za dane treningowe, zaś 20% za dane testowe. Wyniki obu badań przedstawiono poniżej.

### Komentarze ze strony mediaKrytyk

**Zbiór danych treningowych:** 12821 komentarzy z portalu filmWeb

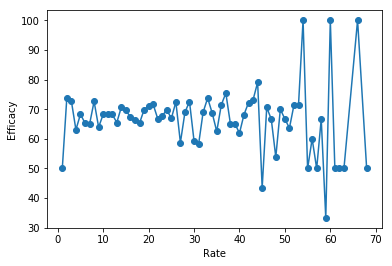
**Zbiór danych testowych:** 6073 komentarzy z mediaKrytyk

|  |  |  |
| --- | --- | --- |
|  | Komentarze przed lematyzacją | Komentarze po lematyzacji |
| liczba komentarzy z fałszywym określeniem emocji | 1964 | 1929 |
| liczba komentarzy z poprawnym określeniem emocji | 4109 | 4144 |
| Skuteczność (%) | 68 | 68 |

Analizując powyższą tabelę widać, że w przypadku klasyfikatora SVM lematyzacja słów nie wpływa na skuteczność wykrywania wydźwięku emocjonalnego. Dodatkowo sprawdzono zależność długości tekstu od skuteczności metody dla zbioru po lematyzacji. Postawiono hipotezę zerową mówiącą o tym, że **długość tekstu wpływa na poprawność określenia nacechowania emocjonalnego wypowiedzi** na poziomie skuteczności 0.05. W celu potwierdzenia hipotezy wyliczono korelację danych przy pomocy współczynnikiem korelacji rangowej Spearmana, ponieważ już wcześniej stwierdzono (rozdział <numer>), że dane długość tekstu nie są rozkładu normalnego. Wynik korelacji przedstawiono poniżej.

|  |  |
| --- | --- |
| Korelacja (rs) | -0.0033961812096073274 |
| Wartość pvalue | 0.7913108138727895 |

Wynik korelacji danych jest bliski zeru, dlatego też nie można powiedzieć, że długość badanego tekstu ma wpływ na wynik algorytmu. Jednak wartość pvalue jest większa od poziomu istotności, co nie pozwala na odrzucenie hipotezy zerowej. Może istnieć związek niemonotoniczny. Poniżej przedstawiono wykres rozrzutu (Scatterplot) reprezentujący zależność pomiędzy danymi.



Następnie postawiono hipotezę, że **istnieje zależność pomiędzy oceną użytkownika a wynikiem końcowym algorytmu**. W przypadku występowania tej korelacji można by stwierdzić, że metoda uzyskuje lepsze wyniki dla ocen skrajnych (1 i 10) lub wyłącznie dla ocen pozytywnych bądź negatywnych. Poprawność hipotezy sprawdzono współczynnikiem korelacji rangowej Spearmana, gdyż ocena użytkownika nie jest zbiorem o rozkładzie normalnym (sprawdzenie rozkładu danych w rozdziale <numer>). Wynik korelacji przedstawiono w poniższej tabeli:

|  |  |
| --- | --- |
| Korelacja (rs) | 0.6118957548382578 |
| Wartość pvalue | 0.0 |

Wynik współczynnika korelacji interpretuje się według poniższej skali[[13]](#footnote-12):

rxy = 0 zmienne nie są skorelowane

0 < rxy < 0,1 korelacja nikła

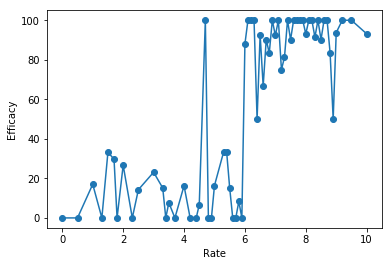
0,1 rxy < 0,3 korelacja słaba

0,3 rxy < 0,5korelacja przeciętna

0,5 rxy < 0,7 korelacja wysoka

0,7 rxy < 0,9 korelacja bardzo wysoka

0,9 rxy < 1 korelacja prawie pełna.

Wartość korelacji jest nieco większa niż 0.5, dlatego można przyjąć, że istnieje silna zależność pomiędzy oceną użytkownika a wynikiem końcowym. Dla potwierdzenie hipotezy przedstawiono również wykres badanej zależności.   
Widać, że dla ocen powyżej 6 skuteczność algorytmu silnie wzrasta, zaś dla ocen mniejszych od 6 skuteczność metody nie przekracza 40%. Potwierdzono to również badając skuteczność klasyfikatora SVM wyłącznie dla komentarzy pozytywnych lub negatywnych. Wynik badania przedstawiono poniżej.

|  |  |
| --- | --- |
| Komentarze pozytywne | |
| Liczba komentarzy z fałszywie wykrytym wydźwiękiem emocjonalnym | 353 |
| Liczba komentarzy z pozytywnie wykrytym wydźwiękiem emocjonalnym | 3771 |
| Skuteczność (%) | 91 |

|  |  |
| --- | --- |
| Komentarze negatywne | |
| Liczba komentarzy z fałszywie wykrytym wydźwiękiem emocjonalnym | 1606 |
| Liczba komentarzy z pozytywnie wykrytym wydźwiękiem emocjonalnym | 343 |
| Skuteczność (%) | 17 |

Wyraźnie widać, że klasyfikator SVM dużo lepiej wykrywa pozytywne nacechowanie emocjonalne w krótkich formach wypowiedzi zebranych ze strony mediaKrytyk.

### Komentarze ze strony Opineo

**Zbiór danych treningowych:** 12821 komentarzy z portalu filmWeb

**Zbiór danych testowych:** 6311 komentarzy dotyczących produktów ze strony Opineo

|  |  |  |
| --- | --- | --- |
|  | Komentarze przed lematyzacją | Komentarze po lematyzacji |
| liczba komentarzy z fałszywym określeniem emocji | 585 | 492 |
| liczba komentarzy z poprawnym określeniem emocji | 5726 | 5819 |
| Skuteczność (%) | 90 | 92 |

Z poniższej tabeli można wyciągnąć wniosek, że lematyzacja ma niewielki wpływ na skuteczność klasyfikatora dla badanego zbioru. Ponadto sprawdzono, czy długość wypowiedzi ma wpływ na wynik algorytmu. Postawiono hipotezę, że **istnieje zależność pomiędzy długością komentarza a poprawnością wykrycia wydźwięku emocjonalnego wypowiedzi** na poziomie skuteczności 0.05. Zweryfikowano ją przy pomocy współczynnika korelacji rangowej Spearmana i wynik przedstawiono poniżej.

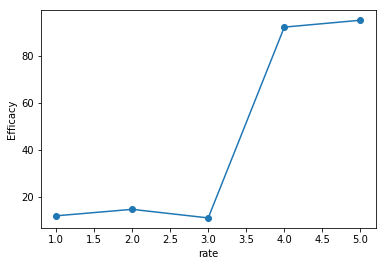
|  |  |
| --- | --- |
| Korelacja (rs) | -0.10287844873473703 |
| Wartość pvalue | 0.0 |

Widać, że nie występuje monotoniczna zależność pomiędzy długością tekstu a skutecznością algorytmu. Wartość pvalue jest mniejsza od poziomu skuteczności, dlatego też odrzucamy hipotezę zerową i przyjmujemy hipotezę alternatywną, że **nie istnieje monotoniczna zależność między ilością wyrazów w wypowiedzi a poprawnością działania klasyfikatora SVM dla zbioru komentarzy z Opineo**.

Badając wypowiedzi przyjęto również hipotezę, że **ocena użytkownika ma wpływ na skuteczność wykrycia nacechowania emocjonalnego komentarza** na poziomie skuteczności 0.05. W celu udowodnienia hipotezy również posłużono się współczynnikiem korelacji rangowej Spearmana.

|  |  |
| --- | --- |
| Korelacja (rs) | 0.25295067304989016 |
| Wartość pvalue | 0.0 |

Według powyższej tabeli można powiedzieć, że występuje słaba korelacja pomiędzy oceną użytkownika a wynikiem działania klasyfikatora. Jednak wartość pvalue jest równa zero, co zmusza nas do odrzucenia hipotezy zerowej i przyjęcie hipotezy alternatywnej mówiącej o tym, że **ocena użytkownika nie ma wpływu na skuteczność wykrycia nacechowania emocjonalnego**.

Mimo przeprowadzonego testu z wykresu rozrzutu wynika, że poprawność działania algorytmu wzrasta wraz ze wzrostem oceny użytkownika.

Dlatego też postanowiono zbadać skuteczność klasyfikatora osobno dla komentarzy pozytywnych oraz dla komentarzy negatywnych. Wynik badania przedstawiono w poniższej tabeli.

|  |  |
| --- | --- |
| Komentarze pozytywne | |
| Liczba komentarzy z fałszywie wykrytym wydźwiękiem emocjonalnym | 5696 |
| Liczba komentarzy z pozytywnie wykrytym wydźwiękiem emocjonalnym | 371 |
| Skuteczność (%) | 93 |

|  |  |
| --- | --- |
| Komentarze negatywne | |
| Liczba komentarzy z fałszywie wykrytym wydźwiękiem emocjonalnym | 30 |
| Liczba komentarzy z pozytywnie wykrytym wydźwiękiem emocjonalnym | 214 |
| Skuteczność (%) | 12 |

Po sprawdzeniu skuteczności dla obu grup, można stwierdzić, że dla komentarzy produktów klasyfikator SVM wykrywana nacechowanie emocjonalne bardzo dobrze dla pozytywnych wypowiedzi użytkowników.

### Komentarze ze strony TripAdvisor

**Zbiór danych treningowych:** 12821 komentarzy z portalu filmWeb

**Zbiór danych testowych:** 6919 komentarzy dotyczących restauracji ze strony TripAdvisor

|  |  |  |
| --- | --- | --- |
|  | Komentarze przed lematyzacją | Komentarze po lematyzacji |
| liczba komentarzy z fałszywym określeniem emocji | 1339 | 1312 |
| liczba komentarzy z poprawnym określeniem emocji | 5580 | 5607 |
| Skuteczność (%) | 80 | 81 |

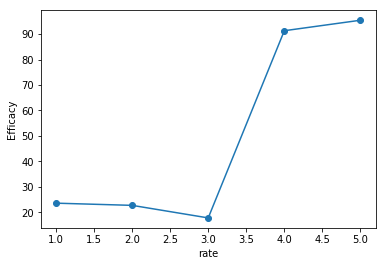
Analizując powyższą tabele można stwierdzić, że lematyzacja nie wpływa na skuteczność klasyfikatora. Aby lepiej przyjrzeć się bazie komentarzy recenzji restauracji przebadano zależność długości wypowiedzi od poprawności wykrycia nacechowania emocjonalnego tekstu. Postawiono hipotezę, że **długość wypowiedzi ma wpływ na poprawność metody klasyfikatora SVM** na poziomie istotności 0.05. Chcąc potwierdzić powyższe zdanie wyliczono współczynnika korelacji rangowej Spearmana, wiedząc wcześniej, że długość wypowiedzi nie należy do rozkładu normalnego. Wynik badania przedstawiono poniżej.

|  |  |
| --- | --- |
| Korelacja (rs) | -0.1606279284998240 |
| Wartość pvalue | 0.0 |

Wynika, że nie występuje korelacja pomiędzy długością wypowiedzi a wynikiem klasyfikatora. Wartość pvalue jest mniejsza od poziomu istotności, dlatego też można odrzucić hipotezę zerową, że poprawność metody jest zależna od długość wypowiedzi użytkownika i przyjąć hipotezę alternatywną informującą, że **poprawność działania algorytmu nie jest monotonicznie zależna od długości badanego komentarza**.

Również jak dla wcześniejszych algorytmów sprawdzono poprawność stwierdzenie, że **ocena użytkownika ma wpływ na jakość działania algorytmu sztucznej inteligencji** na poziomie istotności 0.05. Aby móc udowodnić powyższe zdanie zmierzono współczynnik korelacji rangowej Spearmana.

|  |  |
| --- | --- |
| Korelacja (rs) | 0.5570890875211718 |
| Wartość pvalue | 0.0 |

Z powyższej tabeli wynika, że istnieje silna korelacja pomiędzy oceną użytkownika a poprawnością działania metody wykrywania wydźwięku emocjonalnego wypowiedzi. Mimo to wartość pvalue jest mniejsza od przyjętego poziomu istotności, dlatego też należy odrzucić hipotezę zerową i przyjąć hipotezę alternatywną, która mówi, że **ocena użytkownika nie zależy monotonicznie od jakości działania algorytmu sztucznej inteligencji**. Badanie jednak nie odrzuca tezy, że może istnieć niemonotoniczna zależność pomiędzy danymi. Poniżej zamieszczono wykres ilustrujący tą zależność.

Dodatkowo, aby potwierdzić tezę zmierzono skuteczność algorytmu dla dwóch osobnych zbiorów komentarzy i jej wyniki zamieszczono w poniższych tabelach.

|  |  |
| --- | --- |
| Komentarze pozytywne | |
| Liczba komentarzy z fałszywie wykrytym wydźwiękiem emocjonalnym | 5365 |
| Liczba komentarzy z pozytywnie wykrytym wydźwiękiem emocjonalnym | 337 |
| Skuteczność (%) | 94 |

|  |  |
| --- | --- |
| Komentarze negatywne | |
| Liczba komentarzy z fałszywie wykrytym wydźwiękiem emocjonalnym | 975 |
| Liczba komentarzy z pozytywnie wykrytym wydźwiękiem emocjonalnym | 242 |
| Skuteczność (%) | 19 |

Z przedstawionych danych wynika, że istnieje bardzo wysoka skuteczność dla komentarzy mających wydźwięk pozytywny a bardzo niska dla komentarzy mających wydźwięk negatywny.

### Recenzje ze strony FilmWeb

**Zbiór danych treningowych:** 12821 komentarzy z portalu filmWeb

**Zbiór danych testowych:** 319 recenzji filmów ze strony filmWeb

|  |  |  |
| --- | --- | --- |
|  | Komentarze przed lematyzacją | Komentarze po lematyzacji |
| liczba komentarzy z fałszywym określeniem emocji | 77 | 94 |
| liczba komentarzy z poprawnym określeniem emocji | 242 | 225 |
| Skuteczność (%) | 75 | 75 |

Według powyższej tabeli można ponownie stwierdzić, że lematyzacja tekstu nie wpływa na wynik klasyfikatora. W kolejnym kroku zbadano poprawność hipotezy, mówiącej, że **długość tekstu zależy od poprawności wykrywania nacechowania emocjonalnego wypowiedzi** na poziomie skuteczności 0.05. Do przeprowadzenia badania wybrano współczynnik korelacji rangowej Spearmana, wiedząc z wcześniejszych testów, że dane nie należą do rozkładu normalnego. Wyniki przedstawiono w poniższej tabeli:

|  |  |
| --- | --- |
| Korelacja (rs) | 0.01123520632483853 |
| Wartość pvalue | 0.8415699152679084 |

Według powyższego współczynnika wynika, że istnieje niewielka zależność pomiędzy długością testu a skutecznością algorytmu. Wartość pvalue jest wyższa od poziomu skuteczność, dlatego też nie podstaw do odrzucenia hipotezy zerowej i można przyjąć, że **długość tekstu ma wpływ na poprawność wykrywania nacechowania emocjonalnego wypowiedzi**.

Drugą badaną zależnością jest ocena użytkownika. Postawiono hipotezę, że o**cena użytkownika zależy skuteczności działania klasyfikatora SVM**. Aby potwierdzić hipotezę wyznaczono współczynnik korelacji rangowej Spearmana, który przedstawiono w poniższej tabeli.

|  |  |
| --- | --- |
| Korelacja (rs) | 0.4456226408609416 |
| Wartość pvalue | 0.0 |

Analizując wynik można powiedzieć, że występuje przeciętna koleracja pomiędzy oceną użytkownika a poprawnym działaniem metody sztucznej inteligencji. Zaś wynik pvalue informuje o tym, że należy odrzucić hipotezę zerową i przyjąć hipotezę alternatywną, że **ocena użytkownika nie zależy monotonicznie od skuteczności wykrywania nacechowania emocjonalnego w recenzjach filmów**. By sprawdzić, czy istnieje niemonotoniczna zależność między zmiennymi przebadano poprawność działania algorytmu dla osobnych grup komentarzy. Wyniki przedstawiono w poniższych tabelach.

|  |  |
| --- | --- |
| Komentarze pozytywne | |
| Liczba komentarzy z fałszywie wykrytym wydźwiękiem emocjonalnym | 207 |
| Liczba komentarzy z pozytywnie wykrytym wydźwiękiem emocjonalnym | 42 |
| Skuteczność (%) | 83 |

|  |  |
| --- | --- |
| Komentarze negatywne | |
| Liczba komentarzy z fałszywie wykrytym wydźwiękiem emocjonalnym | 52 |
| Liczba komentarzy z pozytywnie wykrytym wydźwiękiem emocjonalnym | 18 |
| Skuteczność (%) | 25 |

Z pomiarów wynika, że rodzaj recenzji (pozytywny bądź negatywny) ma wpływ na końcowy wynik klasyfikatora SVM.

## Porównanie algorytmu sztucznej inteligencji z autorskim algorytmem opartym o Słowosieć

W poniższym podrozdziale

* algorytm Słowosieci jest niezależny od danych treningowych
* algorytm Słowosieci lepiej sprawdza się dla negatywnych wypowiedzi zaś klasyfikator SVM dla pozytywnych
* w obu algorytmach długość badanej wypowiedzi nie ma wpływu na wynik algorytmu
* w obu algorytmach ocena użytkownika nie ma monotonicznego wpływu na wynik

1. Portal społecznościowy *http://www.filmweb.pl* [↑](#footnote-ref-0)
2. Speech and Language Processing, Daniel Jurafsky, James H. Martin, ©2009 |Pearson [↑](#footnote-ref-1)
3. PoliMorfologik<https://github.com/morfologik/polimorfologik> [↑](#footnote-ref-2)
4. fragment publikacji *Korpus IPI PAN. Wersja wstępna, Adam Przepiórkowski, Aleksander Buczyński, Jakub Wilk* [↑](#footnote-ref-3)
5. G. A. Miller, R. Beckwith, C. D. Fellbaum, D. Gross, K. Miller. 1990. | WordNet: An online lexical database. [↑](#footnote-ref-4)
6. Marek MaziarzA, Maciej PiaseckiA, Ewa RudnickaA, Stan SzpakowiczB, Paweł KędziaA | plWordNet 3.0 – a Comprehensive Lexical-Semantic Resource [↑](#footnote-ref-5)
7. Strona Słowosieci (http://plwordnet.pwr.wroc.pl/wordnet/) [↑](#footnote-ref-6)
8. D'Agostino R.B. and Pearson E.S. (1973), Tests of departure from normality. Empirical results for the distribution of b2 and sqrt(b1) [↑](#footnote-ref-7)
9. Kendall M.G. (1938), A new measure of rank correlation. Biometrika, 30, 81-93 [↑](#footnote-ref-8)
10. Cichosz P., *Systemy uczące się*, WNT Warszawa, 2000 [↑](#footnote-ref-9)
11. http://scikit-learn.org/stable/index.html [↑](#footnote-ref-10)
12. ANALIZA SKUPIEŃ I REDUKCJA WYMIAROWOŚCI **W HIERARCHICZNYM MODEL**U KORPUSOWYM JĘZYKA Jan WICIJOWSKI, Bartosz ZIÓŁKO Akademia Górniczo-Hutnicza, Katedra Elektroniki (http://www.dsp.agh.edu.pl/\_media/pl:maszynopis.pdf) [↑](#footnote-ref-11)
13. *Michalski T., (2008).* Statystyka*, WSiP, Warszawa, s. 152* [↑](#footnote-ref-12)