ZARZĄDZANIE MEDIAMI
Tom 9(4)2021, s. 681–693
doi: 10.4467/23540214ZM.21.037.14580
www.ejournals.eu/ZM

Jędrzej Wieczorkowski orcid.org/0000-0002-1252-8975

Szkoła Główna Handlowa w Warszawie
Kolegium Analiz Ekonomicznych
jedrzej.wieczorkowski@sgh.waw.pl

Aleksandra Suwińska

MOWA NIENAWIŚCI W MEDIACH SPOŁECZNOŚCIOWYCH – MOŻLIWOŚCI AUTOMATYCZNEJ DETEKCJI I ELIMINACJI

Abstract

HATE SPEECH ON SOCIAL MEDIA – THE POSSIBILITY OF AUTOMATIC DETECTION AND ELIMINATION

The article deals with the issues of hate speech and other forms of verbal aggression on the Internet as well as the possibility of their automatic detection. The paper discusses the studies confirming the partial effectiveness of text mining methods in the automatic detection of hate speech on social media. Hate speech is related to verbal aggression resulting from belonging to a group (national, racial, religious, etc.) and has become a significant problem in the social and economic context. Automatic detection significantly support the management of online news websites and social media due to the moderation of the received content. Moreover, eliminating online hate speech reduces its negative social and economic effects. The linguistic and cultural specificity of the hate speech are the problem, and the gap so far is solving the problem in Polish conditions. The study used the Tweeter database. Then, methods such as artificial neural networks, naïve Bayes classifier and support vector machine were used. The obtained results confirm the thesis about the possibility of using text mining methods in the process of reducing hate speech, but at the moment the described methods do not allow for full automation of the elimination of such content. The issue was presented in the article primarily in the context of the significance and scale of the problem and the possibility of solving it, and less from the point of view of the technical details.

Keywords: hate speech, cyberbullying, social media, text mining, Twitter

JEL: L86, L82, C40, O30

Wprowadzenie

Popularność mediów społecznościowych wpływa na zmianę wagi niektórych problemów w relacjach międzyludzkich. Choć mowa nienawiści istniała w praktyce od zawsze, to nowe możliwości komunikacji wynikające z powszechności wykorzystywania internetu, a w szczególności mediów społecznościowych, nakazują zdecydowanie zmienić przyjęty punkt widzenia. Dostępność metod i narzędzi umożliwiających szerzenie swoich poglądów i docieranie z nimi do znaczącego kręgu odbiorców, a także powszechność dostępu do internetu dla osób w każdym wieku, skutkują wyraźnym rozszerzeniem skali problemów wynikających z agresji słownej, w tym mowy nienawiści i cyberprzemocy. Ignorowanie takich problemów skutkuje konsekwencjami społecznymi oraz ekonomicznymi. W rezultacie zmienia się spojrzenie z jednej strony na wolność wypowiedzi, z drugiej zaś na ewentualną cenzurę i ograniczenia udostępnianych treści.

W przypadku dopuszczenia kontroli treści zamieszczanych w sieci technicznym i organizacyjnym problemem stają się metody takiej kontroli. Wynikają one ze skali działalności serwisów społecznościowych oraz z dynamiki pojawiania się w nich nowych treści. Zobowiązanie administratorów serwisów internetowych do monitorowania treści skutkuje wysokimi kosztami procesu moderacji oraz niepożądanymi opóźnieniami w przepływie informacji. Z oczywistych względów niemożliwe jest więc stosowanie tradycyjnych metod cenzury, lecz rozwiązaniem mogą się okazać metody automatycznej detekcji i blokowania niepożądanych treści, w szczególności mowy nienawiści. Wymaga to jednak stosowania zaawansowanych metod analitycznych klasyfikujących teksty pod kątem ich potencjalnej szkodliwości. Mogą być one użyteczne choćby przy moderacji sekcji komentarzy lub forów internetowych. Daje to serwisom korzyści wizerunkowe, ale przede wszystkim może pozwolić na spełnienie prawnych wymogów stawianych przed firmami funkcjonującymi w wirtualnym świecie internetu, istotnie wspomagając zarządzanie mediami społecznościowymi.

Artykuł ma na celu scharakteryzowanie rosnącego problemu szerzenia mowy nienawiści w internecie, w szczególności w mediach społecznościowych, oraz zaproponowanie metod analitycznych, które ten problem są w stanie rozwiązać lub przynajmniej zredukować. Szczegółowym celem jest więc potwierdzenie tezy o skuteczności metod automatycznego wykrywania mowy nienawiści. Autorzy przeprowadzili badanie oparte na dokonywanych w języku polskim wpisach pochodzących z serwisu Twitter. W badaniu porównywano skuteczność różnych metod eksploracji tekstu. Nadmienić należy, że w dotychczasowych badaniach wykorzystywano głównie teksty w języku angielskim. Jednakże rozwiązania zaadaptowane z innych języków mogą nie być tak dokładne, jak te stworzone specjalnie dla jednego z nich. Potwierdzenie skuteczności metod automatycznej detekcji mowy nienawiści dla wszystkich języków, w tym dla języka polskiego, pozwoliłoby uznać rozwiązania za

bardziej uniwersalne. Takie badania były już prowadzone, lecz ich niewielka skala wskazuje na istnienie tutaj luki badawczej.

1. Zarys i skala problemu mowy nienawiści w mediach społecznościowych

Tematyka agresji słownej, w tym mowy nienawiści, poruszana jest w literaturze naukowej w ostatnich latach coraz częściej wraz z rozwojem popularności internetu. Szczególnie problem jest istotny w kontekście koncepcji Web 2.0, w której coraz istotniejszą rolę odgrywa treść generowana przez użytkowników różnorodnych sieciowych serwisów, w tym mediów społecznościowych.

Mowa nienawiści będąca specyficzną agresją słowną w kontekście niniejszych rozważań została zawężona do internetu. Anthony Cortese (2006) mowę nienawiści zdefiniował jako poniżanie ludzi na postawie ich rasy, pochodzenia etnicznego, religii, płci, wieku, stanu psychicznego, niepełnosprawności lub orientacji seksualnej. Karmen Erjavec i Melita Poler Kovačič (2012) definiują ją jako atak na tle rasowym, etnicznym, religijnym, politycznym, płci, orientacji seksualnej, statusu czy też jako atak skierowany do innych grup społecznych. Pojawienie się internetu nie wpłynęło znacząco na rozumienie pojęcia samej nienawiści, można ją więc traktować jako głęboką, trwałą, intensywną emocję wyrażającą gniew i wrogość wobec osoby, grupy lub obiektu. Przyjmuje się, że nienawiść charakteryzuje się chęcią wyrządzenia krzywdy lub spowodowania bólu obiektowi nienawiści oraz odczuwaniem przyjemności z powodu nieszczęścia tego obiektu (Reber, 1975). Jednakże takie starsze badania pokazują, że nienawiść – najczęściej traktowana jako skrajna niechęć, wrogość, życzenie komuś śmierci lub cierpienia, skrajne obrzydzenie lub złość – była zazwyczaj ukierunkowana na osoby wcześniej znane lub często nawet w przeszłości kochane, a uczucie nienawiści rzadko dotyczyło osób nieznanych (Aumer-Ryan, Hatfield, 2007). W tym aspekcie można zaobserwować wraz z rozwojem metod komunikacji elektronicznej wyraźne zmiany pośród adresatów uczucia nienawiści – w kierunku osób i grup anonimowych.

W kontekście niniejszych rozważań interesujące jest, jak mowa nienawiści rozumiana jest przez serwisy społecznościowe. Administratorzy serwisu Facebook (2019) przyjęli definicję mowy nienawiści jako wypowiedzi bezpośrednio atakującej inne osoby na podstawie tak zwanych cech chronionych, takich jak rasa, pochodzenie, narodowość, wyznanie, orientacja seksualna, kasta, płeć biologiczna, płeć kulturowa, tożsamość płciowa, poważna choroba lub niepełnosprawność. Ponadto stosowane są środki ochronne względem statusu imigracyjnego. Jako ataki rozumiane są treści namawiające do przemocy lub odczłowieczające inne osoby, sugerujące ich niższość lub wzywające do ich wykluczania bądź segregacji. Z kolei Twitter (2019) nie pozwala promować przemocy, atakować innych osób ani im

grozić z powodu ich rasy, przynależności etnicznej, pochodzenia, orientacji seksualnej, płci, tożsamości płciowej, religii, wieku, niepełnosprawności ani choroby.

Na gruncie polskiego prawa karnego – w art. 256 § 1 Kodeksu karnego – znajduje się między innymi zapis:

> Kto publicznie propaguje faszystowski lub inny totalitarny ustrój państwa lub nawołuje do nienawiści na tle różnic narodowościowych, etnicznych, rasowych, wyznaniowych albo ze względu na bezwyznaniowość, podlega grzywnie, karze ograniczenia wolności albo pozbawienia wolności do lat 2 (Kodeks karny, 1997, art. 256 § 1).

Należy odnotować, że definicja polskiego Kodeksu karnego jest zdecydowanie najwęższa, gdyż w przeciwieństwie do przedstawionych wcześniej, nie bierze pod uwagę na przykład dyskryminacji ze względu na orientację seksualną czy niepełnosprawność.

W konsekwencji zwraca się uwagę (*Polska*. Raport, 2018), że mimo iż polskie prawo zapewnia obywatelom prawo do równości i swobody wypowiedzi, to ich realizacja nie jest w pełni zgodna z międzynarodowymi standardami. Dotyczy to wspomnianego wyżej ograniczonego zakresu ochrony przed nawoływaniem do nienawiści znajdującego się w prawie karnym, a także trudności z rzeczywistym egzekwowaniem tych przepisów. Ofiary mowy nienawiści mogą dochodzić swoich praw nie tylko na gruncie prawa karnego, ale także postępowania cywilnego i administracyjnego, lecz niestety te przepisy wydają się nieskuteczne i są rzadko stosowane.

Interesujące jest zróżnicowanie zachowań w zakresie szerzenia mowy nienawiści między światem rzeczywistym a wirtualnym. Do jego przyczyn zalicza się między innymi (np. Suler, 2004) tak zwaną dysocjacyjną anonimowość pozwalającą na oddzielenie akcji podejmowanych w sieci od tych, z którymi jesteśmy kojarzeni osobiście, a także tak zwaną dysocjacyjną wyobraźnię pozwalającą na rozdzielenie w umyśle osobowości świata rzeczywistego i internetowego. Innymi przyczynami są między innymi: asynchroniczność wypowiedzi w internecie z ewentualnością ucieczki po zamieszczeniu wpisu, fizyczna niewidzialność, a także minimalizacja autorytetu wynikająca z przeświadczenia, że w sieci wszyscy czują się sobie równi i nie pojawia się strach przed dezaprobatą osób o wyższym statusie, który mógłby hamować wyrażanie kontrowersyjnych opinii. W rezultacie obserwuje się (Tereszkiewicz, 2012) większą agresję słowną występującą w internetowych anonimowych grupach otwartych niż w wymagających logowania grupach zamkniętych. Ponadto agresja słowna wywołuje pewien dreszcz emocji, daje poczucie obrony swojego terytorium oraz stosowana jest jako narzędzie służące do pozbycia się ze swojego otoczenia osób i grup, które są traktowane jako złe lub gorsze. Inne badania (Jabłońska, 2017) wskazują, że agresja słowna jest nie tylko sposobem na wyładowanie negatywnych emocji, lecz stanowi również rozrywkę, poprawia samopoczucie, a także jest podejmowana jako działalność zarobkowa.

Powody pojawiania się mowy nienawiści w internecie są więc w praktyce zbliżone do przyczyn szerszego zjawiska – agresji słownej. Jednak, choćby ze względów

prawnych, należy wydzielić mowę nienawiści, która jest zazwyczaj karalna, od innych typów agresji słownej i mowy obraźliwej określanych także jako cyberprzemoc, które – choć również są etycznie naganne – przeważnie nie podlegają penalizacji. Z punktu widzenia zarządzania mediami społecznościowymi rozróżnienie takie może więc być kluczowe.

Skala mowy nienawiści w internecie jest trudna do oszacowania, gdyż bardzo niewielka część czynów formalnie karalnych skutkuje w Polsce wszczęciem postępowań sądowych. Według szacunków Rzecznika Praw Obywatelskich (2018) i Biura Instytucji Demokratycznych i Praw Człowieka, jedynie 5% przestępstw z nienawiści w Polsce wobec Ukraińców, migrantów z państw muzułmańskich i Afryki Subsaharyjskiej jest zgłaszane policji. Z kolei według badania przeprowadzonego w 2016 roku przez Centrum Badań nad Uprzedzeniami we współpracy z Fundacją Batorego (Winiewski i in., 2017) 80% młodzieży deklaruje, że spotkało się w internecie z wypowiedziami o charakterze islamofobicznym, 75% – antysemickim, a 71% – antyukraińskim.

Mowa nienawiści rani zarówno osoby indywidualne, jak i całe społeczeństwo. Krzywdy wyrządzane pojedynczym osobom to między innymi krzywdy psychiczne, zwłaszcza te wyrządzane dzieciom. Mowa nienawiści ogranicza perspektywy społeczne i finansowe, ponieważ ofiara często wycofuje się ze środowisk, w których doświadczyła dyskryminacji, i staje się bardziej ostrożna. Badanie Centrum Badań nad Uprzedzeniami (Winiewski i in., 2017) wskazuje, że osoby częściej spotykające się z mową nienawiści przejawiają większą niechęć do członków mniejszości, której ta mowa dotyczy. Zależność ta w polskich warunkach jest najsilniejsza w przypadku mniejszości takich, jak muzułmanie, Romowie, imigranci oraz osoby transseksualne. Autorzy badania pokazali również, że korelacja ta jest szczególnie widoczna wśród młodzieży – osoby młode doświadczające mowy nienawiści są mniej skłonne do akceptacji osób dyskryminowanych w swoim otoczeniu niż dorośli. Cortese (2006) twierdzi, że największe szkody są wyrządzane właśnie na poziomie społecznym. Mowa nienawiści wpływa również na gospodarkę, choć jej wpływ nie jest łatwy do oszacowania. Jednakże Lewis R. Gale i współautorzy (2002) udowodnili, że wyższy wskaźnik przestępstw z nienawiści wiąże się z wyższymi wskaźnikami nadużyć gospodarczych i bezrobocia.

W związku z opisanymi wyżej konsekwencjami naturalna wydaje się konieczność opracowania narzędzi służących do redukcji tych negatywnych zjawisk. Obecnie nacisk powinien być położony na przeciwdziałanie rozszerzaniu się mowy nienawiści w internecie, w szczególności w miejscach, gdzie zawartość tworzona jest przez samych użytkowników, czyli między innymi w mediach społecznościowych.

2. Opis zastosowanej metody badawczej

Eksploracja tekstu (ang. *text mining*) służy do odkrywania zależności występujących w tekstach oraz ich statystycznej analizy. Od pewnego czasu podejmuje się próby stosowania tych metod do automatycznej detekcji mowy nienawiści – jak wspomniano, dotychczas przede wszystkim w odniesieniu do języka angielskiego. Jednakże poszczególne języki posiadają swoją specyfikę utrudniającą przenoszenie doświadczeń między nimi.

Podstawową metodą jest wyodrębnianie słownictwa (wyrazów) traktowanego jako ofensywne. Powstaje wtedy problem rozróżnienia treści rozumianych jako mowa nienawiści od pozostałej mowy obraźliwej (Kwok, Wang, 2013). Ponadto wyrazy uznawane zazwyczaj za ofensywne często używane są też w innym kontekście (Davidson i in., 2017). Innym problemem jest celowa błędna pisownia słów uważanych za ofensywne lub na przykład rozdzielanie liter znakami interpunkcyjnymi, aby utrudnić działanie filtrów.

William Warner i Julia Hirschberg (2012) zauważyli, że nienawiść w kierunku konkretnych grup charakteryzuje się wykorzystaniem mało licznego zbioru stereotypowych słów. Zaobserwowali oni też, że mowa nienawiści często wykorzystuje dobrze znane stereotypy w celu zdyskredytowania jednostki lub grupy, co może ułatwić odróżnienie jednej formy mowy nienawiści od innej poprzez identyfikację w tekście takiego stereotypu. Każdy z nich charakteryzuje się specyficznymi epitetami, frazami, pojęciami, metaforami i zestawieniami, odpowiadając nienawistnym intencjom. Jako przykład dla warunków amerykańskich podano mowę antylatynoską, która często odnosi się do przekraczania granicy, antyafrykańską mówiącą o bezrobociu lub wychowaniu przez samotnych rodziców czy język antysemicki odnoszący się do pieniędzy, bankowości i mediów. Znajomość poszczególnych stereotypów, dzięki opracowaniu modelu oddzielnie dla każdego z nich, pozwoliłaby zbudować kompleksowy model dla całego problemu mowy nienawiści. Rozwiązanie takie jest jednak trudno wykonalne oraz mało uniwersalne, więc autorzy niniejszego artykułu nie zdecydowali się na takie ujęcie.

Z punktu widzenia zastosowania konkretnych metod eksploracji danych podejmowano dotąd próby wykorzystywania co najmniej regresji logistycznej, naiwnego klasyfikatora Bayesa, drzew decyzyjnych, lasów losowych i liniowej maszyny wektorów nośnych. Ujęcie czysto leksykalne uznane zostało za niewystarczające między innymi ze względu na trudność rozróżniania mowy nienawiści od innych treści uznawanych za obraźliwe. Według Thomasa Davidsona i współautorów (2017) regresja logistyczna i liniowa maszyna wektorów nośnych przyniosły znacząco lepsze rezultaty niż pozostałe wymienione metody. Inne badania (Warner, Hirschberg, 2012; Kwok, Wang, 2013) wskazują także na wysoką skuteczność zastosowania naiwnego klasyfikatora Bayesa. Z kolei rzadko wykorzystuje się w omawianym zagadnieniu sztuczne sieci neuronowe, mimo że model ten jest coraz częściej wykorzystywany do skomplikowanych zagadnień klasyfikacji.

W konsekwencji autorzy niniejszego artykułu zdecydowali się na potrzeby przeprowadzonego badania zastosować dwie z wyżej wymienionych najlepiej dotąd sprawdzonych metod: naiwny klasyfikator Bayesa (NKB) i maszynę wektorów nośnych (MWN), lecz dodatkowo także sztuczne sieci neuronowe (SSN) z jedną warstwą ukrytą.

Jako badany zbiór danych wykorzystano polskojęzyczne wpisy z serwisu Twitter wstępnie podzielone według stopnia ich szkodliwości na trzy kategorie (klasy): tekst nieszkodliwy (0), cyberprzemoc (1), mowa nienawiści (2)[1]. Rozróżnienie pomiędzy typami treści szkodliwych (1 oraz 2) nastąpiło na podstawie stwierdzenia, czy szkodliwe działanie wymierzone jest w kierunku osoby prywatnej (cyberprzemoc – 1), czy osoby publicznej lub grupy osób (mowa nienawiści – 2). Teksty brane pod uwagę w badaniu zawierają szeroki przekrój tematów – od prywatnych, neutralnych przemyśleń przez dyskusje kibiców sportowych aż do opinii politycznych i publicznej krytyki wybranych osób.

Dane źródłowe zawierały 10 047 obserwacji w zbiorze treningowym oraz 1000 obserwacji w zbiorze testowym, jednak dla niektórych rekordów pełna treść wpisu nie była dostępna. W rezultacie w niniejszym badaniu wykorzystano 9224 obserwacji, z tego 8368 w zbiorze treningowym i 856 w testowym. Zdecydowaną większość zebranych tekstów stanowią treści nieszkodliwe (7806, czyli 93,3%). Za cyberprzemoc uznano 169 rekordów (2,02%), a za mowę nienawiści – 393 (4,7%). Na potrzeby modelowania ze zbioru treningowego wydzielono zbiór walidacyjny o wielkości 10% zbioru treningowego, więc końcowe proporcje prezentowały się następująco: zbiór treningowy 81,6% (7531 rekordów), zbiór walidacyjny: 9,1% (837 rekordów), zbiór testowy: 9,3% (856 rekordów).

Do przygotowania modeli zbiór poddano przepróbkowaniu (ang. *oversampling*) ze względu na silnie niezbalansowanie próby pomiędzy poszczególnymi kategoriami, następnie dane wyczyszczono ze znaków specjalnych i interpunkcyjnych, poddano je lematyzacji (tj. każde słowo zostało sprowadzone do swojej formy podstawowej). Następnie dla każdej z trzech badanych kategorii zidentyfikowano po pięć najczęściej występujących słów. Zaobserwowano, że wśród tekstów nieszkodliwych najczęściej pojawiają się wyrazy często występujące w języku polskim (np. mieć, móc, wiedzieć), czyli słowa, których można się spodziewać w analizie typowego tekstu. W przypadku cyberprzemocy oraz mowy nienawiści proporcje się zmieniają – wśród najczęściej występujących słów pojawiają się między innymi nazwy profili osób publicznych, co znaczy, że słowa te były kierowane pod konkretny adres.

[1] Zbiór danych był stworzony i udostępniony w ramach konkursu PolEval 2019 na potrzeby zadania dotyczącego klasyfikacji tekstów umieszczonych na portalu Twitter na treści szkodliwe i nieszkodliwe. Nadzór nad zadaniem, z którego dane zostały wykorzystane w tej pracy, sprawował Michał Ptaszyński (Associate Professor z Kitami Institute of Technology w Japonii). Treści analizowanych tekstów oraz przypisane klasy są dostępne pod adresem: https://github.com/ptaszynski/cyberbullying-Polish (dostęp: 17.12.2019).

Wszystkie eksperymenty zostały przeprowadzone przez autorów z wykorzystaniem języka Python oraz głównie trzech pakietów – numpy, nltk oraz sklearn.

3. Wyniki badania

Wszystkie trzy algorytmy osiągnęły dokładność mierzoną jako stosunek liczby prawidłowo sklasyfikowanych tekstów niezależnie od kategorii do wszystkich treści w zbiorze, na poziomie powyżej 80%. Specyficzność została policzona ze względu na treści nieszkodliwe (kategoria 0), czyli jako stosunek liczby prawidłowo sklasyfikowanych treści szkodliwych (kategorie 1 i 2) do rzeczywistej liczby treści tych kategorii. W tabeli 1 zaprezentowano porównanie takich wskaźników, jak specyficzność, dokładność oraz czas uczenia dla wszystkich trzech badanych algorytmów.

Tabela 1. Porównanie wyników zastosowanych metod

	Specyficzność	Dokładność	Czas uczenia
Sieci neuronowe	52,17%	82,83%	28 min
Naiwny klasyfikator Bayesa	41,30%	86,68%	< 1 min
Maszyna wektorów nośnych	39,13%	90,19%	129 min

Źródło: opracowanie własne.

Najlepiej z klasyfikacją treści szkodliwych poradziły sobie sztuczne sieci neuronowe, stało się to jednak kosztem dokładności. Warto również zwrócić uwagę na czas uczenia algorytmu, który był z kolei zdecydowanie najkrótszy dla naiwnego klasyfikatora Bayesa. W tabeli 2 przedstawiono porównanie wyników klasyfikatorów w postaci tablicy pomyłek.

Tabela 2. Tablica pomyłek

		Wartość rzeczywista								
		0			1			2		
Metoda		SSN	NKB	MWN	SSN	NKB	MWN	SSN	NKB	MWN
Wartość przewidywana	0	661	704	736	8	7	11	24	34	39
	1	23	18	2	1	2	1	9	10	6
	2	80	42	26	3	3	0	47	36	35

Źródło: opracowanie własne.

Metoda SSN sklasyfikowała poprawnie 47 tekstów zawierających mowę nienawiści (kategoria 2), co daje wynik lepszy od algorytmu NKB o 11 rekordów, a od MWN – o 12 rekordów, czyli odpowiednio o 13,75% oraz 15,00% wszystkich rekordów w tej grupie. Żaden z algorytmów nie sklasyfikował natomiast poprawnie więcej niż 2 z 12 rekordów zawierających cyberprzemoc (kategoria 1), najczęściej były one oznaczane błędnie jako treści nieszkodliwe lub jako mowa nienawiści. Trudności w klasyfikacji tekstów tego typu mogą wynikać z mało reprezentatywnej grupy tekstów uczących – ta klasa była najmniej liczna. Wzrost specyficzności modelu SSN odbył się kosztem dokładności, co jest szczególnie widoczne w przypadku klasy 0. Treści nieszkodliwe zostały błędnie uznane za mowę nienawiści w 80 z 764 przypadków, co daje wynik odpowiednio o 4,97% i 7,07% rzeczywistej liczby treści nieszkodliwych wyższy niż metody NKB i MWN.

Podsumowując, przy zachowaniu zadowalającego poziomu dokładności klasyfikacji (powyżej 80%) najlepsze wyniki osiągnął model sztucznej sieci neuronowej z jedną warstwą ukrytą. Model tego typu wymagał najwięcej konfiguracji parametrów i osiągnął niższą dokładność niż algorytmy dwóch pozostałych typów, ale jako jedyny pozwolił na poprawne sklasyfikowanie ponad połowy szkodliwych treści, w tym 59% tekstów oznaczonych jako mowa nienawiści.

4. Dyskusja i znaczenie osiągniętych wyników

Pomimo niedoskonałości otrzymanych wyników najważniejszym wnioskiem z badania jest potwierdzenie faktu, że metody eksploracji danych, podobnie jak w przypadku języka angielskiego, sprawdzają się również w językach należących do innych grup, w szczególności w języku polskim. Potwierdza to wnioski ze stosunkowo nielicznych wcześniej przeprowadzonych badań. Przykładowo Flor-Miriam Plaza Del-Arco i współautorzy (2020) użyli metody maszyny wektorów nośnych, regresji logistycznej i drzewa decyzyjnego, aby znaleźć teksty ksenofobiczne lub mizoginistyczne napisane w języku hiszpańskim. Z kolei Valentino Santucci z zespołem (2018) wykorzystali między innymi model maszyny wektorów nośnych do wykrywania mowy nienawiści w języku włoskim. W kontekście języków słowiańskich Bohdan Andrusyak i współautorzy (2018) próbowali wykrywać mowę nienawiści w języku ukraińskim i rosyjskim w warunkach ograniczonej dostępności oznaczonych danych, wykorzystując w konsekwencji jako dane wejściowe słownik terminów obraźliwych. Nieliczne artykuły koncentrowały się na języku polskim: Renard Korzeniowski i współpracownicy (2019) używali wstępnie wytrenowanych modeli, a Maciej Biesek (2019) oceniał między innymi model maszyny wektorów nośnych, i to ten model osiągał najlepsze wyniki.

Zróżnicowanie wyników dotychczasowych badań, a także wyniki osiągnięte przez autorów niniejszego artykułu, pokazują, jak ważne jest niezależne badanie każdego języka w regionalnych uwarunkowaniach kulturowych, i jak wiele jest

jeszcze do zrobienia, aby opracować efektywne metody detekcji mowy nienawiści, w tym te dostosowane do polskiej specyfiki.

Charakterystyczne dla przeprowadzonego przez autorów badania są krótkie, zazwyczaj jednozdaniowe teksty, gdyż są one typowe dla serwisu Twitter. Należy mieć na względzie fakt, że dobór metod eksploracji danych jest specyficzny dla analizowanego zbioru danych. Wykorzystywane w badaniu metody będą więc prawdopodobnie osiągać najlepsze wyniki na tekstach o podobnej długości, w których słowa występują zazwyczaj jednokrotnie. Wnioski dotyczące skuteczności algorytmów, poza analizą wpisów w serwisie Twitter, można więc także rozciągnąć na przykład na wsparcie moderowania sekcji komentarzy w serwisach internetowych, zwłaszcza tam, gdzie istnieje limit wprowadzonych znaków.

Takie wsparcie dla serwisów internetowych jest niezbędne z kilku powodów. Przede wszystkim administrowanie jakimkolwiek portalem mającym dużą liczbą aktywnych użytkowników wymaga ogromnego wkładu pracy. Na przykład redakcja „The Guardian" przeprowadziła analizę 70 mln komentarzy zamieszczonych pod artykułami w okresie 2006–2016. Jak się okazało, 2% z nich, to jest 1,4 mln, naruszało przyjęte standardy społeczności, między innymi zawierały mowę nienawiści (Gardiner i in., 2016). Liczba komentarzy umieszczanych przez internautów może sięgać kilkuset w ciągu godziny. Dane „The Guardian" dają średnią 810 komentarzy na godzinę. Z kolei szacunek dokonany przez autorów niniejszego artykułu na podstawie polskich internetowych portali informacyjnych wskazał nawet 200 komentarzy zamieszczonych w ciągu godziny od opublikowania artykułu[2].

Omawiany problem nie jest tylko kwestią wizerunkową lub komfortu czytelników, ale również zobowiązaniem prawnym. W polskich uwarunkowaniach za treść opublikowanego wpisu odpowiada przede wszystkim jego autor, jednak administrator portalu internetowego może ponosić taką samą odpowiedzialność, jeżeli nie udowodni, że o szkodliwych treściach nie wiedział (Wernik, 2018).

W konsekwencji media społecznościowe walczą z mową nienawiści i doskonalą swoje algorytmy w celu szybszego i trafniejszego blokowania niechcianych treści. Na przykład Facebook raportuje, że dzięki sztucznej inteligencji w pierwszym kwartale 2020 roku udało się mu usunąć 88,8% mowy nienawiści, zanim została ona zaraportowana przez użytkowników. Do moderowania treści w samych Stanach Zjednoczonych zatrudnionych jest 15 tys. osób (Thomas, 2020). Część portali informacyjnych decyduje się z kolei na wyłączenie funkcji komentarzy na rzecz angażowania społeczności w innych miejscach, na przykład właśnie na Twitterze i Facebooku. Dzięki temu liczba anonimowych komentatorów jest minimalizowana, a obowiązek moderacji treści przechodzi na inny podmiot. Tą drogą poszły takie portale, jak Recode, Reuters, The Week (Ellis, 2015), a w Polsce – Onet.

[2] Analiza własna artykułów opublikowanych w dniach od 31 lipca do 1 sierpnia 2020 r. na portalach: wp.pl, wyborcza.pl oraz gazeta.pl (dostęp: 1.08.2020).

Wsparcie moderacji treści zamieszczanych w internecie przy wykorzystaniu zaprezentowanych rozwiązań, w szczególności modelu opartego na sztucznych sieciach neuronowych, może skutkować bardziej efektywnym i szybszym klasyfikowaniem zamieszczanych przez użytkowników treści, przynosząc korzyści zarówno finansowe (przede wszystkim w zakresie automatyzacji pracy), jak i społeczne. Należy jednak wziąć pod uwagę, że przedstawione rozwiązania i otrzymane wyniki dotyczą krótkich treści napisanych w języku polskim. Ponadto przyjęta w badaniu definicja mowy nienawiści, to jest fakt, że szkodliwe treści kierowane są bezpośrednio do grupy osób lub osoby publicznej, może odbiegać od standardów przyjętych w danym portalu. Szczegóły rozwiązań muszą więc każdorazowo zostać dopasowane do konkretnych potrzeb. Zaproponowane metody mogą się jednak okazać ważną wartością dodaną w działaniach dotyczących ograniczenia mowy nienawiści w internecie.

Wnioski i podsumowanie

Problem automatycznej detekcji w internetowych treściach mowy nienawiści i, szerzej, innych form przemocy słownej jest złożony, gdyż wpływają na niego między innymi zagadnienia wciąż doskonalonych metod eksploracji tekstu, specyfika wykorzystywanych języków i uwarunkowania kulturowe. Niemniej przemiany technologiczne z jednej strony wpłynęły na metody komunikacji i w konsekwencji na szerzenie się różnego typu agresji słownej, z drugiej – dają narzędzia do przeciwdziałania temu zjawisku. Przeprowadzone badanie pozwoliło osiągnąć postawiony na początku cel, to jest potwierdziło możliwości przynajmniej częściowego zautomatyzowania procesu wykrywania szkodliwych treści w internecie.

W artykule przedstawiono skuteczność trzech metod analitycznych: sztucznych sieci neuronowych, naiwnego klasyfikatora Bayesa i maszyny wektorów nośnych. Ich efektywność można uznać za wystarczającą do wspierania podejmowania decyzji lub przyspieszenia manualnej klasyfikacji tekstów. Uzyskane wyniki nie pozwalają jednak na pełną automatyzację procesu, ponieważ nawet najlepiej klasyfikujący model, czyli model sztucznej sieci neuronowej, prawidłowo wykrywa niewiele powyżej połowy szkodliwych treści, dokładność zaś wyniosła niewiele powyżej 80%. Oznacza to jednak, że wykorzystując tę relatywnie prostą i mało złożoną obliczeniowo metodę, jaką jest sieć neuronowa z jedną warstwą ukrytą, można znacząco przyspieszyć i częściowo zautomatyzować moderację zamieszczanych treści.

Wykorzystanie zaprezentowanego modelu lub innych rozwiązań opartych na metodach eksploracji tekstu może w praktyce przynieść zarządzającemu serwisem internetowym szereg korzyści finansowych i wizerunkowych. Jednakże stosowanie gotowych rozwiązań wymaga ich weryfikacji, w szczególności przykładów użytych do uczenia algorytmu, i sprawdzenia, w jakim stopniu przypisane klasy są zgodne z definicją mowy nienawiści ustaloną przez wykorzystujący je podmiot. Tego typu

rozwiązania mogą być także dostosowywane do wykrywania innych treści o agresywnym charakterze.

Zaprezentowane metody charakteryzują się krótkim czasem uczenia i brakiem konieczności wykorzystywania znacznych mocy obliczeniowych, co stanowi ich niewątpliwą zaletę, ale w konsekwencji nie wykorzystują one w pełni możliwości współczesnych metod eksploracji danych. Biorąc pod uwagę uwzględnienie tej oceny, lecz także w kontekście nieustannego postępu w zakresie technologii i metod analizy danych, naturalnym polem dalszych badań będzie analiza efektywności coraz bardziej zaawansowanych rozwiązań, takich jak na przykład głębokie sieci neuronowe.

Bibliografia

Andrusyak B., Rimel M., Kern R. (2018). *Detection of Abusive Speech for Mixed Sociolects of Russian and Ukrainian Languages*. Proceedings of Twelfth Workshop RASLAN, s. 77–84.

Aumer-Ryan K., Hatfield E. (2007). *The Design of Everyday Hate: A Qualitative and Quantitative Analysis*. „Interpersona: An International Journal on Personal Relationships", 1(2), s. 143–172. DOI: 10.5964/ijpr.v1i2.11.

Biesek M. (2019). *Comparison of Traditional Machine Learning Approach and Deep Learning Models in Automatic Cyberbullying Detection for Polish Language*. Proceedings of the PolEval 2019 Workshop, s. 121–126.

Cortese A. (2006). *Opposing Hate Speech*. Westport: Praeger Publishers.

Davidson T., Warmsley D., Macy M., Weber I. (2017). *Automated Hate Speech Detection and the Problem of Offensive Language*. Proceedings of the Eleventh International AAAI Conference on Web and Social Media ICWSM, s. 512–515.

Ellis J. (2015). *What happened after 7 news sites got rid of reader comments*. NiemanLab, https://www.niemanlab.org/2015/09/what-happened-after-7-news-sites-got-rid-of-reader-comments (dostęp: 30.08.2020).

Erjavec K., Kovačič M.P. (2012). *You Don't Understand, This is a New War! Analysis of Hate Speech in News Web Sites' Comments*. „Mass Communication and Society", 15, s. 899–920. DOI: 10.1080/15205436.2011.619679.

Facebook (2019). *Standardy społeczności. Propagowanie nienawiści*, https://www.facebook.com/communitystandards/hate_speech (dostęp: 8.11.2019).

Gale L.R., Health W.C., Ressler R. (2002). *An Economic Analysis of Hate Crime: Eastern*. „Economic Journal", 28(2), s. 203–216.

Gardiner B., Mansfield M., Anderson I., Holder J., Louter D., Ulmanu M. (2016). *The Dark Side of Guardian Comments*. „The Guardian", 12.04.2016, https://www.theguardian.com/technology/2016/apr/12/the-dark-side-of-guardian-comments (dostęp: 22.07.2020).

Jabłońska M. (2017). *Doświadczanie agresji słownej w cyberprzestrzeni wśród cyfrowych tubylców*. „Ekonomiczne Problemy Usług", 126(2), s. 175–183.

Kodeks karny, 1997. Ustawa z dnia 6 czerwca 1997 r. – Kodeks karny. Dz.U. 1997 nr 88 poz. 553 ze zm.

Korzeniowski R., Rołczyński R., Sadownik P., Korbak T., Możejko M. (2019). *Exploiting Unsupervised Pre-Training and Automated Feature Engineering for Low-Resource Hate Speech Detection in Polish*. Proceedings of the PolEval 2019 Workshop, s. 141–148.

Kwok I., Wang Y. (2013). *Locate the Hate: Detecting Tweets against Blacks*. Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence, AAAI Press, s. 1621–1622.

Plaza-Del-Arco F.M., Molina-González M.D., Ureña-López L.A., Martín-Valdivia M.T. (2020). *Detecting Misogyny and Xenophobia in Spanish Tweets Using Language Technologies*. „ACM Transactions on Internet Technology (TOIT)", 20(2), Article 12, s. 1–19. DOI: 10.1145/3369869.

Polska. Reagowanie na mowę nienawiści. Raport (2018). London: ARTICLE 19.

Ptaszynski M., Pieciukiewicz A., Dybala P. (2019). *Dataset for Automatic Cyberbullying Detection in Polish Language*, https://github.com/ptaszynski/cyberbullying-Polish#readme (dostęp: 17.12.2019).

Reber A.S. (1975). *The Penguin Dictionary of Psychology*. New York: Penguin Press.

Rzecznik Praw Obywatelskich (2018). *Jedynie 5% przestępstw motywowanych nienawiścią jest zgłaszanych na policję – badania RPO i ODIHR/OBWE*, https://www.rpo.gov.pl/pl/content/jedynie-5-przestepstw-motywowanych-nienawiscia-jest-zglaszanych-na-policje-badania-rpo-i-odihrobwe (dostęp: 8.11.2019).

Santucci V., Spina S., Milani A., Biondi G., Di Bari G. (2018). *Detecting Hate Speech for Italian Language in Social Media*. Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA).

Suler J. (2004). *The Online Disinhibition Effect*. „Cyberpsychology & Behavior", 7(3), s. 321–326. DOI: 10.1089/1094931041291295.

Tereszkiewicz A. (2012). *Do Poles Flame? Aggressiveness on Polish Discussion Groups and Social Networking Sites*. W: L. Laineste, D. Brzozowska, W. Chłopicki (red.). *Estonia and Poland: Creativity and Tradition in Cultural Communication*, 1, s. 221–236. Tartu: ELM Scholarly Press.

Thomas Z. (2020). *Facebook Content Moderators Paid to Work from Home*. BBC News, https://www.bbc.com/news/technology-51954968 (dostęp: 30.08.2021).

Twitter (2019). *Zasady dotyczące zachowań przepełnionych nienawiścią*, https://help.twitter.com/pl/rules-and-policies/hateful-conduct-policy (dostęp: 8.11.2019).

Warner W., Hirschberg J. (2012). *Detecting Hate Speech on the World Wide Web*. Proceedings of the Second Workshop on Language in Social Media. Montreal: Association for Computational Linguistic, s. 19–26.

Wernik A. (2018). *Odpowiedzialność administratora strony za nienawistne wpisy*. Infor, https://www.infor.pl/prawo/prawa-konsumenta/konsument-w-sieci/2800913,2,Odpowiedzialnoscadministratora-strony-za-nienawistne-wpisy.html (dostęp: 22.07. 2020).

Winiewski M., Hansen K., Bilewicz M., Soral W., Świderska A., Bulska D. (2017). *Mowa nienawiści, mowa pogardy. Raport z badania przemocy werbalnej wobec grup mniejszościowych*. Warszawa: Fundacja im. Stefana Batorego.