Marek Troszyński
Collegium Civitas
ORCID 0000-0002-3653-4018

MOWA NIENAWIŚCI W INTERNECIE. CZY ALGORYTMY MOGĄ KSZTAŁTOWAĆ SFERĘ PUBLICZNĄ?

Internet od momentu powstania traktowany był przez twórców i użytkowników jako przestrzeń przyjazna wolności wypowiedzi, „oddolna" odpowiedź na tradycyjny system medialny, w którym prawo głosu było przywilejem nielicznych. Kiedy w latach 70. powstają w USA pierwsze społeczności internautów, są one związane z ruchami kontrkulturowymi, co w szczególności dotyczy mieszkanców dużych miast Kalifornii [Castells 2010]. Oddolne organizacje non profit, jak Electronic Frontier Foundation stworzona przez Johna Barlowa, od początku swego funkcjonowania brały sobie za cel obronę wolności słowa (por. https://www.eff.org/issues/free-speech). Warto jednak pamiętać, że w tych społecznościach dużą rolę odgrywały wewnętrzne regulaminy i kodeksy etyczne, które skutecznie porządkowały komunikację w nowo postałym medium.

Tradycyjne media przez dekady kultywowały dziedzictwo liberalnej filozofii, wyraźnie oddzielającej przestrzeń publiczną od przestrzeni prywatnej. W tej pierwszej dominował język oficjalny, odnoszący się do dyskursu racjonalnego, pozbawionego emocji i osobistych odniesień. Wypowiedzi wychodzące poza poprawność polityczną nie były dopuszczane do debaty publicznej.

Dopiero po komercjalizacji Internetu i upowszechnieniu mediów społecznościowych oraz tak zwanych „treści tworzonych przez użytkowników" (*user generated content*) pojawił się problem zmieszania dyskursu prywatnego i publicznego. Emocje i niekontrolowany język, charakterystyczny dla wypowiedzi prywatnych, trafił do mediów. W 1996 roku Bill Clinton podpisuje Telecommunication Act, w którym jeden z zapisów zwalnia pośredników internetowych z odpowiedzialności za publikowane teksty. Odpowiedzialność redakcji za publikacje była fundamentalną zasadą w okresie dominacji mediów masowych. Zniesienie odpowiedzialności zostało powtórzone w wielu innych państwach, m.in. w Polsce, takie rozwiązanie przyjęto w uchwalonej w 2002 roku Ustawie o świadczeniu usług drogą elektroniczną (Dz.U. 2002 nr 144 poz. 1204). Efektem tego jest obecny stan prawny, pozwalający każdemu użytkownikowi Sieci na napisanie, co mu w duszy gra. Nie ma bowiem podmiotu, który jest odpowiedzialny za publikowane przez internautów teksty. Jednocześnie pośrednicy zarabiają na treściach umieszczanych na ich stronach. Celem poszczególnych podmiotów, które dostarczają treści, jest zogniskowanie uwagi widza na swojej

stronie przez jak najdłuższy czas. Dominującym przez wiele lat modelem rozliczeń na rynku reklamy internetowej jest *cost per mille* (CPM), czyli płatność za odsłony (wyświetlenia strony). Im więcej internautów odwiedzi daną stronę („wygeneruje odsłonę" w języku marketingu interaktywnego), tym więcej centów wpadnie do kieszeni właściciela strony. Połączenie tych dwóch czynników – zwolnienie pośredników z odpowiedzialności za teksty i zależność przychodów od liczby odwiedzających – sprawiło, że wydawcom (pośrednikom internetowym) zależy na jak największej liczbie internautów, którzy oglądają ich stronę. A najskuteczniejszym wabikiem dla czytających są osobiste, emocjonalne wypowiedzi innych osób. Buduje to przekonanie nadawców, że usunięcie jakichkolwiek treści uszczupli ich przychody, a zatem odbije się negatywnie na prowadzonym przez nich biznesie.

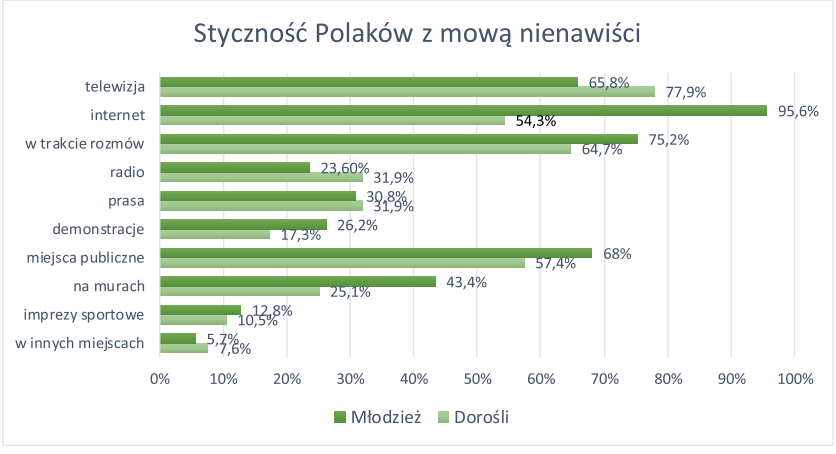
MOWA NIENAWIŚCI I JEJ PERCEPCJA

Z rozwojem mediów społecznościowych skorelowane jest upowszechnienie mowy nienawiści w przestrzeni publicznej. Zacznijmy od definicji. W polskim systemie prawnym brak jednoznacznego określenia tego zjawiska. W debacie najczęściej przywoływany jest artykuł 257 Kodeksu karnego: „Kto publicznie znieważa grupę ludności lub poszczególną osobę z powodu jej przynależności narodowej, etnicznej, rasowej, wyznaniowej albo z powodu jej bezwyznaniowości lub z takich powodów narusza nietykalność cielesną innej osoby, podlega karze pozbawienia wolności do lat 3".

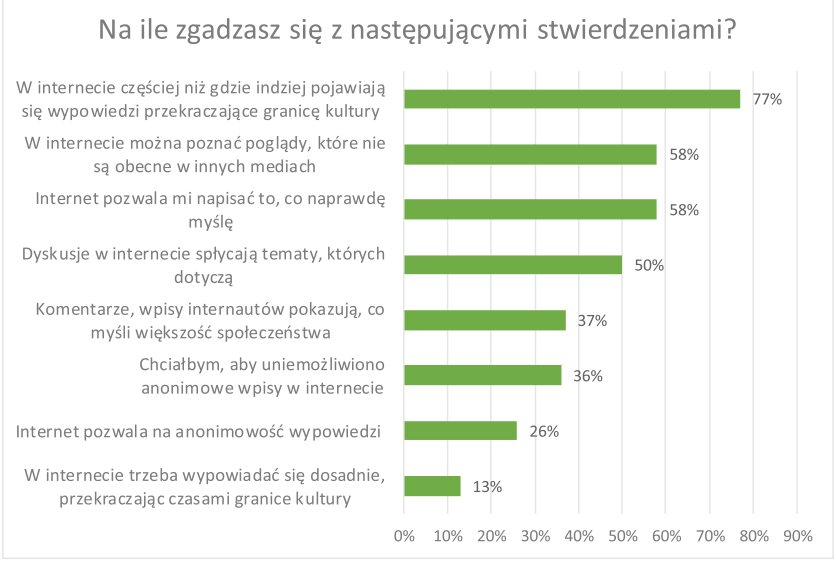
W naukach społecznych analiza mowy nienawiści ma ugruntowaną tradycję [por. Schauer 1992]. W polskiej literaturze socjologicznej temat ten podejmuje Lech Nijakowski, który tak określa to zjawisko: „mowa nienawiści polega na przypisywaniu szczególnie negatywnych cech i/lub wzywaniu do dyskryminujących działań, wymierzonych w pewną kategorię społeczną, przede wszystkim taką, do której przynależność jest postrzegana jako «naturalna» (przypisana), a nie z wyboru" [Nijakowski 2008, 132]. W tej definicji główny nacisk położony jest na przynależność do grupy oraz „przypisywanie negatywnych cech".

Przekraczanie granic wolności wypowiedzi w Internecie i występowanie mowy nienawiści jest przedmiotem wielu badań i analiz. W 2016 roku Centrum Badań nad Uprzedzeniami we współpracy z Fundacją Batorego przeprowadziło badanie przemocy werblanej wobec grup mniejszościowych. Badaniem objęto 1052 dorosłych oraz 682 młodych ludzi w wieku 16–18 lat. Badania te pokazują, że Internet okazał się miejscem, w którym młodzież najczęściej spotyka się z mową nienawiści skierowaną do grup mniejszościowych; co więcej, prawie nie ma młodych ludzi, którzy nie zetknęliby się w Internecie z przekraczeniem granic wolności wypowiedzi. Zdecydowanie mniejszy odsetek dorosłych, choć także znaczący (ponad 55%), przyznaje, że spotkał się w Internecie z mową nienawiści [Wiśniewski 2017].

Z kolei badanie przeprowadzone przez Związek Pracodawców Branży Internetowej IAB Polska (Internetowa kultura obrażania? 2017, N=1121) pokazuje, że przekonanie o Internecie jako miejscu, gdzie indziej pojawiają się wypowiedzi przekraczające granice kultury, jest rozpowszechnione. Jednocześnie bardzo wielu badanych postrzega Internet jako przestrzeń wolności, w której mogą spotkać się z poglądami nieobecnymi w innych mediach oraz wyrażać to, co naprawdę myślą. Z analizy danych



Rysunek 1
Styczność Polaków z mową nienawiści [źródło: Wiśniewski 2017]



Rysunek 2
[źródło: Internetowa kultura obrażania? 2017]

zastanych w Internecie, przeprowadzonej w ramach tego samego projektu badawczego, wynika, że jedynie kilka procent internautów przekracza granice kultury, używa wulgaryzmów czy obraża innych. Największy odsetek tego typu wypowiedzi występował na forach i w komentarzach do artykułów (4,7%). Jak zauważają autorzy raportu, zdecydowana większość wpisów przekraczających granice wypowiedzi dotyczyła polityki.

METODA – AUTOMATYCZNE PRZETWARZANIE JĘZYKA NATURALNEGO

W jaki sposób możemy zatem monitorować skalę tego zjawiska? Z jednej strony mamy działania internautów (indywidualnych osób), z drugiej ogromne, komercyjne firmy wykorzystujące treści publikowane przez użytkowników do zarabiania pieniędzy. Internauci publikują każdego dnia setki tysięcy komentarzy, postów, odpowiedzi w wątkach. Jest zupełnie oczywiste, że człowiek nie ma możliwości przejrzenia tak ogromnego zbioru wypowiedzi.

Z pomocą przychodzą nam narzędzia do automatycznego przetwarzani języka naturalnego (*natural language processing* – NLP). Narzędzia komputerowej analizy tekstów wykorzystywane są w socjologii od kilku dziesięcioleci. Początki sięgają pionierskich programów komputerowych dla języka angielskiego, korzystających ze słowników [Stone i in. 1966].

Jak działają narzędzia NLP? Dla większości metod konieczne jest wstępne przetwarzanie tekstu w celu otrzymania odpowiedniej reprezentacji. Liczba kroków i poziom przetworzenia zależą od algorytmu rozpoznającego pożądane zjawisko. Kolejne działania to: podział na słowa (tokenizacja), rozpoznawanie części mowy oraz przetwarzanie składniowe, płytkie lub głębokie. Zestawy narzędzi tego typu zawiera pakiet Stanford NLP [Manning i in. 2014] oraz dla języka polskiego – Multiserwis, utrzymywany w Instytucie Podstaw Informatyki PAN [http://zil.ipipan.waw.pl/].

W tego typu metodach brana jest pod uwagę kolejność występowania słów w tekście oraz kontekst zdaniowy, są one z sukcesem używane w rozpoznawaniu dłuższych fragmentów tekstu określonego typu, jak przykładowo frazy zawierające wydźwięk (ang. *sentiment*). Istnieje możliwość zastosowania tych metod również do innych problemów związanych z jakościową analizą tekstu, takich, w których kodowane przez badaczy zjawiska i treści mają charakter wielowyrazowy, silnie zależny od kontekstu.

Do rozpoznawania zjawisk na poziomie zdań oraz większych fragmentów tekstu stosowane są metody klasyfikacji z nadzorem oparte na reprezentacji *bag-of-words*, czyli biorącej pod uwagę nie kolejność słów w tekście, a tylko sam fakt ich wystąpienia. Przykładowe zastosowania obejmują klasyfikację tematyczną tekstów (stwierdzenie, czy tekst dotyczy sportu czy może polityki), analizy stylometryczne (identyfikacja różnego typu cech psychologicznych lub demograficznych piszącej osoby).

Nauki społeczne w wielu obszarach wykorzystują komputerowe narzędzia przetwarzania języka naturalnego. Dla języka angielskiego, którego dotyczy większość prowadzonych badań, można mówić o takich generalnych kierunkach, jak przewidywanie atrybutów osób piszących teksty (przykładowo ich emocje, płeć, wiek, przekonania polityczne i system wartości) czy przewidywanie zjawisk społecznych na podstawie zebranych tekstów (wyniki wyborów, epidemie chorób, notowania giełdowe).

W procesie monitorowania mowy nienawiści w mediach społecznościowych można wykorzystać narzędzia NLP oparte na uczeniu maszynowym (*machine learning*, *ML*). W tym przypadku to zespół badaczy (anotatorów) wybiera z dużego korpusu tekstów te wypowiedzi, które w ich zgodnej opinii są przykładami mowy nienawiści. Przy odpowiednio dużej liczbie tekstów tworzymy zbiór uczący dla algorytmu, który

następnie odnajduje specyficzne cechy takich wypowiedzi. W kolejnym kroku algorytm oznacza tego typu wypowiedzi w nowym zbiorze tekstów. Wynik jest weryfikowany przez zespół badawczy, który ocenia skuteczność działań maszyny. Pozwala to na przetestowanie wielu różnych metod i narzędzi, tak by wybrać najbardziej skuteczny algorytm.

Przeprowadzone badania pokazały [Troszyński, Wawer 2017], że kluczowym czynnikiem nie jest jakość algorytmu, ale stopień zgodności zespołu kodującego.

Zgodność między anotatorami zależy od jakości instrukcji (im bardziej szczegółowe i precyzyjne, tym wyższa zgodność) oraz złożoności semantycznej i syntaktycznej tekstu. Jednak zazwyczaj instrukcje nie są wystarczające: w praktyce okazuje się, że istnieje wiele przypadków brzegowych, noszących cechy kilku kategorii, potencjalnie możliwych do zaklasyfikowania na różne sposoby. Rozwiązaniem tego problemu jest opracowanie (i wspólne szczegółowe omówienie) określonej liczby przypadków tego typu, zidentyfikowanych jako problematyczne, a także przyjęcie wzorcowych rozstrzygnięć, nawet arbitralnych [Troszyński, Wawer 2017].

Zatem aby rzetelnie oznaczać mowę nienawiści w tekstach z mediów społecznościowych, konieczne jest przeprowadzenie bardziej skutecznego treningu zespołu badawczego. Polega on na realizacji w obrębie zespołu kolejnych zadań oznaczania takich samych partii tekstu i porównywaniu wyników pracy poszczególnych analityków. Istotą treningu jest nie mierzenie zgodności, ale dyskusja i „tłumaczenie się" kodujących z podjętych decyzji. Dopiero tak wytrenowany zespół może przygotować materiał uczący dla algorytmów uczenia maszynowego. Takie działania w istotny sposób zwiększą również zgodność kodowania (*intercoder agreement*), co przyczynia się do poprawy jakościowej analizy tekstów.

Dlaczego zatem tego typu metody nie stały się powszechnym narzędziem eliminującym mowę nienawiści ze sfery publicznej? Po pierwsze, cześć firm pośredniczących w komunikacji internetowej stosuje zaawansowane algorytmy analizy, by eliminować treści niepożądane. Problem jest tu raczej po stronie regulaminów tych platform i rodzaju treści, które traktują one jako dopuszczalne. Mówimy o dyskursie publicznym, zatem każda decyzja ma charakter polityczny, stoi za tym określona wizja ładu społecznego. Narzędzia, które stosuje na przykład Facebook, wspierane regulaminem społeczności dość skutecznie eliminują funkcjonowanie mowy nienawiści na tej platformie. Upowszechnienie takich rozwiązań we wszystkie portalach jest możliwe, jednak doprowadziłoby do zmniejszenia liczby odsłon na poszczególnych stronach, a co za tym idzie, ograniczyłoby przychody pośredników.

Drugą kwestią jest gotowość do oddania władzy nad kształtem sfery publicznej algorytmom, nawet jeśli te są wytrenowane na wcześniejszych decyzjach ludzi. Decyzje o usunięciu nienawistnych postów, podejmowane w sposób zautomatyzowany, już dzisiaj budzą wiele napięć i kontrowersji, w szczególności w sytuacji, gdy decyzja podjęta przez algorytm wydaje się błędna. Globalność Internetu oznacza także wyprowadzenie tych decyzji poza władzę państw narodowych, co jest kolejnym czynnikiem wywołującym sprzeciw zainteresowanych podmiotów.

Przy wszystkich różnicach światopoglądowych i postawach politycznych reprezentowanych w tej dyskusji idea „władzy algorytmów", automatyzacji decyzji, pozbawienia człowieka wpływu na komunikację międzyludzką wywołuje powszechny, i chyba zrozumiały, sprzeciw.

LITERATURA

Castells M, 2010, Społeczeństwo sieci, Warszawa.

Internetowa kultura obrażania?, 2017, Warszawa. https://www.iab.org.pl/baza-wiedzy/raport-internetowa-kultura-obrazania/, (dostęp: 12.09.2019).

Manning C.D. i in., 2014, The Stanford CoreNLP Natural Language Processing Toolkit. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. The Association for Computational Linguistics. ACL, System Demonstrations.

Nijakowski L., 2008, Mowa nienawiści w świetle teorii dyskursu [w:] A. Horolets, red., Analiza dyskursu w socjologii i dla socjologii, Warszawa, s. 113–133.

Schauer F., 1992, The Sociology of the Hate Speech Debate, 37 Villanova Law Review 805.

Stone Philip J. i in., 1966, The General Inquirer: A Computer Approach to Content Analysis, Cambridge.

Troszyński M., Wawer A., 2017, Czy komputer rozpozna hejtera? Wykorzystanie uczenia maszynowego (ML) w jakościowej analizie danych, „Przegląd Socjologii Jakościowej", t. 13, nr 2, s. 62–80.

Wiśniewski M. i in., 2017, Mowa nienawiści, mowa pogardy. Raport z badania przemocy werbalnej wobec grup mniejszościowych, Warszawa.