SGI

DISKUTUJ bez nenávisti

Stratégie na znižovanie nenávistných prejavov v online priestore

Lucia Kováčová
Jozef Miškolci
Edita Rigová

Diskutuj bez nenávisti:

Stratégie na znižovanie nenávistných prejavov v online priestore

Jozef Miškolci
Inštitút pre dobre spravovanú spoločnosť
Pedagogická fakulta, Univerzita Komenského v Bratislave

Lucia Kováčová
Inštitút pre dobre spravovanú spoločnosť

Edita Rigová
Inštitút pre dobre spravovanú spoločnosť

Bratislava 2017

Diskutuj bez nenávisti:
Stratégie na znižovanie nenávistných prejavov v online priestore

Autor a autorky: Jozef Miškolci, Lucia Kováčová, Edita Rigová
Preklad: Martina Kubánová
Jazyková korektúra: Susan Ryan
Recenzentka: Agnes Horváthová
Grafický dizajn: Tomáš Miško
Tlač: ADIN, s.r.o.

Slovak Governance Institute
Štúrova 3, 811 02, Bratislava, Slovakia
www.governance.sk

ISBN: 978-80- 972761-1- 9

S podporou:

SGI
NADÁCIA OTVORENEJ SPOLOČNOSTI
ROMANO KHER
RÓMSKY DOM

Podporené grantom z Foundation Open Society Institute v spolupráci s Think Tank Fund of the Open Society Foundations.

Obsah

Zoznam tabuliek

Zoznam grafov

Úvod

Sociálne médiá sú v súčasnosti zaplavené nenávistným obsahom. Vytvárajú tak prostredie, v ktorom sa rozličné sociálne skupiny (najmä etnické, rasové alebo náboženské skupiny, LGBTI, ľudia s postihnutiami, ženy, mládež a ďalší) stretávajú s urážkami, predsudkami a vyhrážkami v rôznej slovnej alebo grafickej podobe. Na príslušníkov a príslušníčky týchto skupín to má závažné dopady vo forme krátkodobých alebo dlhodobých zdravotných problémov, môže to viesť k ich vylúčeniu z diskusných fór (a tým aj z politického a občianskeho života) alebo dokonca až k diskriminačným praktikám alebo násiliu.

Rozširovanie nenávistného obsahu sa často mylne zamieňa a odôvodňuje právom na slobodu prejavu. Právo na slobodu prejavu však nie je absolútne a neumožňuje zneužívanie práv iných jednotlivcov, akým je napríklad právo na život bez diskriminácie, právo na slobodu myslenia, či právo na slobodu náboženského vyznania. Právo na slobodu prejavu v sebe preto nesie určitú zodpovednosť a užívatelia a užívateľky sociálnych médií by sa mali zapojiť do vytvárania bezpečného digitálneho prostredia, v ktorom sa všetci jednotlivci s rozličnými vlastnosťami môžu slobodne zapájať do diskusií a vyjadriť svoje názory bez vystavenia sa urážkam alebo vyhrážaniu.

Okrem obmedzenia nenávistných prejavov legislatívnymi opatreniami existuje aj súbor mäkkých nástrojov, ktoré je možné použiť na zmiernenie nenávistných prejavov v online priestore bez začatia stíhania, právneho procesu a následne uloženia sankcií. Nepriame nástroje konfrontácie nenávistných prejavov môžu použiť rozliční aktéri, ktorí by sa mali zapojiť do procesu zmierňovania nenávistných prejavov v online priestore, a to od bežných používateľov sociálnych médií, cez rodičovskú a učiteľskú verejnosť, poskytovateľov internetových služieb, až po vládne a mimovládne organizácie.

Hlavným cieľom tejto príručky je predstaviť nástroje a stratégie, ktoré môžu rozliční aktéri použiť pri konfrontácii s nenávistnými prejavmi v online priestore. Snahou príručky je motivovať týchto aktérov, aby využili rôzne nástroje a aktívne sa zapojili do vytvárania bezpečného digitálneho priestoru. Prvá kapitola obsahuje definíciu nenávistných prejavov v online priestore, ich rôzne ciele, formy, frekvenciu a dopady na cieľové skupiny a jednotlivcov. Druhá kapitola stručne opisuje vzdelávanie ako primárny nástroj na znižovanie nenávistných prejavov. Tretia kapitola obsahuje analýzu rozličných stratégií a nástrojov, ktoré môžu využiť najmä užívatelia a užívateľky internetu a poskytovatelia internetových služieb na znižovanie nenávistného obsahu v sociálnych médiách a online diskusiách. Štvrtá kapitola opisuje a analyzuje výsledky experimentálneho testovania zvolených stratégií protiargumentácie, konkrétne overovania faktov a osobnej skúsenosti, a to v kontexte zmierňovania nenávistných prejavov smerovaných voči členom a členkám rómskej komunity na sociálnych sieťach. Posledná kapitola je venovaná úlohe médií v boji proti nenávistným prejavom.

1 Čo sú nenávistné prejavy (v online priestore)?

Nenávistné prejavy môžeme definovať ako druh komunikácie alebo interakcie, kedy jednotlivec alebo celá spoločenská skupina zastrašovaná otvorenými alebo nepriamymi urážkami, vyhrážkami, politickými vyjadreniami, ktoré vyvolávajú nenávisť, pohoršenie alebo nesympatie, čo vedie k obťažovaniu, útlaku a diskriminácii [1]. Nenávistné prejavy sú často spojené so zovšeobecňovaním, pri ktorom sú urážky smerované na celú spoločenskú skupinu, napríklad „všetci Rómovia sú kriminálnici" alebo „všetci moslimovia sú teroristi"[2]. Nenávistné prejavy môžu tiež podnecovať násilie voči cieľovej skupine alebo jednotlivcom priamou formou (napríklad povzbudzovaním ostatných ľudí, aby zaútočili na cieľovú skupinu) alebo nepriamou formou (napríklad prirovnávaním cieľovej skupiny ku zvieratám alebo hmyzu, čo páchateľom uľahčuje nemať zľutovanie a páchať na tejto skupine násilie).[3]

„Nenávistné prejavy preto zahŕňajú napríklad nadávky založené na predsudkoch o identite, určité používanie hanobenia a prirovnaní, niektoré extrémistické politické a náboženské prejavy (napr. vyjadrenia smerujúce k tomu, že všetci moslimovia sú teroristi, alebo že homosexuáli sú ľudia druhej kategórie), ako aj určité zobrazenia ‚nenávistných symbolov' (napr. svastiky alebo horiace kríže). Takéto konanie označujeme ako nenávistné prejavy vtedy a do takej miery, ak vyjadrujú názor, že príslušnosť k určitej spoločenskej skupine oprávňuje posudzovať určitého človeka alebo pristupovať k nemu s pohŕdaním."

Robert Mark Simpson v publikácii *„Dignity, harm, and hate speech" („Dôstojnosť, ujma a nenávistné prejavy")* (2013, s. 701)

1.1 Formy nenávistných prejavov

Nenávistné prejavy v online priestore môžu mať nasledujúce formy[4]:

- **slovnú alebo písomnú formu**, ako napríklad vulgarizmy, vyhrážky alebo urážky vychádzajúce z rôznych charakteristík cieľovej skupiny, ale aj politické vyjadrenia o cieľovej skupine.
- **nenávistné symboly**, ako napríklad svastiky alebo určité čísla symbolizujúce nenávistné ideológie.
- **iné grafické materiály**, ako napríklad obrázky alebo mémy, ktoré zobrazujú cieľovú skupinu urážlivým spôsobom (napríklad ich zobrazovanie ako kriminálnikov alebo deviantov).

Nenávistné prejavy môžu byť šírené prostredníctvom verejne dostupných materiálov (napr. príspevky alebo obrázky zdieľané prostredníctvom sietí Facebook, Instagram alebo Twitter) alebo vo forme súkromných správ odosielaných do schránok jednotlivcov (napr. prostredníctvom aplikácií Facebook Messenger alebo WhatsApp).

Je dôležité uviesť, že nenávistné prejavy nemusia predstavovať len otvorené urážky, napríklad vulgarizmy alebo grafický materiál, ktorý priamo ponižuje cieľovú skupinu. Nenávistné prejavy môžu byť skryté a vyjadrené sofistikovanejším spôsobom. Nenávistné komentáre môžeme potom rozdeliť na[5]:

Otvorené odkazy – predstavujú otvorené a priame urážky namierené voči jednotlivcom a skupinám. Môžu sem patriť priame hanobenia a vulgarizmy alebo grafický materiál, ktorý otvorene šíri nenávisť a predsudky voči cieľovým skupinám alebo jednotlivcom.

Skryté odkazy – predstavujú rafinovanejší spôsob urážania jednotlivcov a skupín. Skryté nenávistné odkazy sú úzko spojené s takzvanými „maskovanými webovými stránkami", ktoré sa tvária ako zdroje poskytujúce neutrálne faktografické informácie o rozličných historických, spoločenských, politických a kultúrnych témach, ale v skutočnosti sú naplnené obsahom s predsudkami a nenávistnou propagandou. Príkladmi takýchto maskovaných webových lokalít je tlačový a online formát pseudovedeckého časopisu ZEM a VEK alebo rozličné stránky na sieti Facebook či webové lokality orientované na históriu, ktoré predstavujú „pravdivé" fakty o rozličných historických udalostiach a obsahujú napríklad popieranie existencie alebo rozsahu židovskej genocídy, čo vedie k vyvolávaniu antisemitizmu. V prípade propagandy zameranej proti LGBTI ľuďom prezentuje americká webová stránka Family Research Council (Rada pre výskum rodiny) nepravdivé výskumné dáta ako dôkaz škodlivosti homosexuality. Cieľom takýchto webových lokalít je pritiahnuť študentov a študentky, rodičov, učiteľov a učiteľky, alebo iných ľudí so záujmom o tieto témy, aby ich čítali a verili, že ich obsah je

možné považovať za vedecký, a preto oprávnený a rešpektovaný.

„Maskované webové stránky" charakterizujú nové stratégie komunikácie s verejnosťou alebo s určitými cieľovými skupinami. Skupiny a jednotlivci, ktorí spravujú tieto webové lokality, zmenili svoj jazyk a ich verejný prejav sa stal sofistikovanejší a slušnejší. Týmto spôsobom sa snažia stať akceptovateľnejšími pre širšie a rôznorodejšie publikum a vyhnúť sa tak postihom na základe legislatívy zameranej proti nenávistným prejavom[6].

Skrytý spôsob šírenia nenávisti nemá na cieľové skupiny menšie dopady ako otvorené nenávistné odkazy. S tým súvisia zistenia štúdie[7] z roku 1997, realizovanej na študentoch a študentkách univerzít ázijského a belošského pôvodu, ktoré ukázali, že cieľová sociálna skupina vnímala oveľa tvrdší dopad pri skrytých rasistických odkazoch než pri tých otvorených. Napriek tomu sú tvrdé a poburujúce nenávistné správanie a úmysly zriedkavejšie a je zložitejšie ich dokazovať. S podobným zistením prišla aj iná štúdia, kde autorský tím Mosher a Proenza (1968) uvádzajú, že príslušníci a príslušníčky cieľovej skupiny vystavenej rasovému hanobeniu nevnímali rozdielnu ujmu v prípade tvrdých vyjadrení nenávisti v porovnaní s miernymi[8].

1.2 Ciele nenávistných prejavov (v online priestore)

Môžeme identifikovať rozličné ciele nenávistných prejavov, z ktorých niektoré môžu byť úmyselné a systematické, a iné zase neúmyselné, kedy užívateľ sociálnych médií šíriaci nenávistný obsah si nie je vedomý dôsledkov svojho konania. V skutočnosti môže nenávisť šíriť každý z rôznych dôvodov, napríklad ak má niekto nízke sebavedomie alebo na sebe osobne zažíva nenávisť a šikanovanie. Ciele nenávistných prejavov môžeme kategorizovať nasledujúcim spôsobom[9]:

- Ventilácia vlastných obáv a frustrácií samotného diskutujúceho a ich nasmerovanie na cieľovú skupinu, ktorá je tak v pozícii obetného baránka
- Urážanie, obťažovanie a ponižovanie cieľovej skupiny
- Vyjadrenie podradnosti cieľovej skupiny
- Upevňovanie predsudkov a šírenie mýtov o cieľovej skupine
- Vyvolávanie nesympatií až nenávisti voči cieľovej skupine
- Vyvolávanie útlaku a násilia

1.3 Jednoduchosť šírenia nenávisti v digitálnom svete

Hoci sa nenávistné prejavy často objavujú na rozličných miestach, vrátane verejných priestorov, pracovísk, domácností alebo prostredníctvom médií, pre nenávistné prejavy v online priestore je typický jednoduchý spôsob ich šírenia prostredníctvom ktoréhokoľvek užívateľa alebo užívateľky sociálneho média, ktorý má prístup na internet. V skutočnosti môže jednotlivec ako užívateľ vytvárať a šíriť v online priestore nenávistný obsah z pohodlia svojho domova. A pretože sociálne médiá (ako napríklad Facebook, Twitter a rôzne online diskusné fóra) používa veľké množstvo užívateľov, takýto nenávistný obsah si môžu prezrieť (a ďalej zdieľať, lajkovať alebo komentovať) tisícky ďalších užívateľov a užívateliek.

„Každý môže byť zverejňovateľom, aj ten najprišernejší antisemita, rasista, bigot, homofób, sexista alebo ploditeľ nenávisti. Ľahkosť a rýchlosť, s akou je možné vytvoriť a v online priestore šíriť webové stránky, stránky sociálnych médií, videá a zvukové nahrávky, okamžité správy, vytvárajú nemožnosť sledovať, riadiť a bojovať s propagandou na internete."

Foxman & Wolf v publikácii *„Viral Hate: Containing its spread on the Internet" („Virálna nenávisť: Pohľad na jej rozšírenie na internete")*, (2013, s. 10)

Znamená to tiež, že užívateľ alebo organizovaná skupina užívateľov a užívateliek môžu ovplyvňovať verejný diskurz v online priestore, ktorý má vplyv na názory obrovského množstva ľudí. Týmto spôsobom môže byť veľmi jednoduché rozšíriť napríklad mýtus o nadmernej výške sociálnych dávok, ktoré poberajú Rómovia a Rómky v hmotnej núdzi na Slovensku. Mýtus sa dá rozšíriť rýchlo a môže byť zdieľaný a lajkovaný (čo zvyšuje viditeľnosť obsahu), a tak presvedčí bežnú verejnosť o privilegovanom postavení rómskych komunít. Vedie to k vyvolávaniu nenávisti a posilňovaniu už existujúcich predsudkov voči tejto etnickej menšine, čo vedie k ďalším tvrdým opatreniam.

Niektorí výskumníci a výskumníčky[10] naznačujú, že samotný dizajn diskusných fór môže prispievať k ľahkosti šírenia nenávistného obsahu v online priestore, pretože sociálne siete ponúkajú svojim používateľom obľúbený obsah alebo obsah, ktorý by sa im mohol potenciálne páčiť (napr. ak lajkovali podobný obsah), čím sa nenávistný materiál šíri ďalej. Takýto dizajn sociálnych médií pomáha ľuďom s podobnými nenávistnými názormi sa združovať, a následne radikalizovať. Na druhej strane môžu sociálne siete pomáhať združovať sa aj užívateľom a užívateľkám s otvoreným pohľadom na svet, ktorí sú proti nenávisti, a mobilizovať ich, aby realizovali aktivity proti nenávistným prejavom (pozri napr. časť 3.4 Komunitné „protestné" akcie v online priestore).

1.4 Frekvencia nenávistných prejavov v online priestore

Pretože každý môže byť vydavateľom a môže šíriť nenávistný obsah (vo forme príspevkov v sociálnych médiách alebo vo forme komentárov, blogov, súkromných správ atď.), nenávistné prejavy sú veľmi rozšíreným fenoménom. Výsledky online prieskumu o nenávistných prejavoch v digitálnom svete, ktorý realizovala Rada Európy v roku 2016, ukazujú, že 4 z 5 respondentov majú skúsenosť s nejakou formou nenávistných prejavov, pričom 2 z 5 vnímajú osobné ohrozenie z nenávistných prejavov[11].

Aj iné štúdie ukazujú vysoký výskyt nenávistného obsahu na internete. Podľa výskumu, ktorý realizoval Inštitút pre verejné otázky v roku 2016 Mladí ľudia v kyberpriestore – šance a riziká pre demokraciu[12], 69 % mladých ľudí na Slovensku má skúsenosť s nenávistnými prejavmi na sociálnych sieťach a iných diskusných internetových fórach. Iná štúdia[13] zo Spojených štátov amerických uvádza, že 61 % žien respondentiek uviedlo, že sú cieľom sexuálne zameraných nenávistných prejavov „každý deň" alebo „často", pričom 46 % účastníkov tmavej pleti uviedlo, že bolo cieľom rasovo zameraných nenávistných prejavov „každý deň" alebo „často".

1.5 Dopady nenávistných prejavov alebo prečo by sme im mali venovať pozornosť?

Nenávistné prejavy nie sú novým fenoménom. Nenávistné prejavy môžeme považovať za súčasť útlaku a diskriminácie skupín na základe ich rasy, etnicity, pohlavia, náboženského vyznania a viery, sexuálnej orientácie, postihnutia, veku alebo iných znakov. Väčšina genocíd a pogromov v dejinách začínala propagandou a zlomyseľnými tvrdeniami o cieľovej skupine. V dôsledku toho viedli nenávistné prejavy v dejinách k násiliu, útlaku a upevňovaniu rasových, etnických a iných predsudkov, a následne vylučovaniu určitých menšín a iných skupín z každého aspektu spoločnosti.

Jednotlivci, ktorí píšu nenávistné komentáre, si často neuvedomujú vážne dopady svojho konania. Na jednej strane môžu mať v momente svojho hnevu chuť napísať o určitej osobe alebo skupine niečo zlomyseľné, ale nevnímajú ničivé dôsledky, ktoré tým zapríčinia. Na druhej strane však niektoré nenávistné skupiny si sú plne vedomé svojho konania a úmyselne sa snažia šíriť nenávisť voči určitej skupine ľudí, posilňovať predsudky voči nim a demonštrovať tak svoju silu (ide najmä o organizované skupiny vyznávajúce rasistické ideológie).

1.5.1 Fyzické a psychologické dopady

Téma fyzických a psychologických dopadov nenávistných prejavov si získala značnú pozornosť akademickej obce najmä kvôli tomu, že oprávnenosť vytvárania legislatívnych obmedzení voči nenávistným prejavom (teda súdnym potrestaním ich šíriteľov) musí byť dobre odôvodnená a dokázaná, aby bola akceptovaná verejnosťou a súdmi. Fyzické a psychologické dopady môžu byť krátkodobé a dlhodobé. Krátkodobé fyzické a psychologické dopady nenávistných prejavov sú charakterizované súborom zdravotných problémov, ktoré výrazne znižujú kvalitu života dotknutých jednotlivcov. Výskum z roku 2004 ukazuje, že nenávistné prejavy môžu zapríčiniť bolesti hlavy, vysoký krvný tlak, vysokú tepovú frekvenciu alebo dokonca rizikové správanie, ako napríklad užívanie drog. Okrem toho nenávistné prejavy spôsobujú strach, obavy alebo smútok[14].

Dlhodobé dôsledky opakujúceho sa zneužívania prostredníctvom nenávistných prejavov môžu zahŕňať znižovanie sebavedomia alebo komplex menejcennosti, stanovovanie si menej ambicióznych životných cieľov, nočné mory, stiahnutie sa zo spoločenského života a depresie alebo mentálne ochorenia, ktoré dokazuje množstvo výskumných štúdií[15]. V tomto zmysle má miera

vlastnej identifikácie s menšinovou skupinou dopad na úroveň psychologických následkov alebo „na dôsledky v podobe sebaobviňovania, zvnútorňovania si negatívnych hodnotení a zlyhania pri hľadaní nápravy". Čím viac sa obete nenávistných prejavov identifikujú so svojou spoločenskou skupinou a nachádzajú v nej útočisko a útechu, tým menší je ničivý účinok[16]. Traumatizujúcim aspektom v nenávistných prejavoch je, že cieľové skupiny sú si vedomé extrémneho násilia, ktoré zažívajú alebo zažili v minulosti ich príslušníci a príslušníčky, a tak keď sú terčom nenávistných prejavov, pripomína im to, že aj oni sa jedného dňa môžu stať obeťou násilného útoku[17]. Sú pozbavení svojej dôstojnosti, reputácie a istoty vlastnej momentálnej bezpečnosti a bezpečnosti svojho sociálneho postavenia[18].

1.5.2 Spoločenské dopady

Opakovanosť a vysoký výskyt nenávistných prejavov spôsobuje, že sa nenávistné prejavy a predsudky napríklad o menejcennosti určitej etnickej menšiny stávajú spoločensky akceptovateľné. Spoločnosť napríklad začne prijímať názory, že určitá etnická menšina je menejcenná alebo je charakteristická negatívnymi vlastnosťami.

„…bežný výskyt takýchto nadávok znecitlivuje čitateľov a čitateľky, čím sa začnú nenávistné prejavy a znevažovanie menšín javiť ako normálne."

Foxman & Wolf v publikácii *„Viral Hate: Containing its spread on the Internet" („Virálna nenávisť: Pohľad na jej rozšírenie na internete")*, (2013, s. 31-32)

Aké sú širšie spoločenské dôsledky akceptácie nenávistných postojov voči menšinám alebo iných spoločenským skupinám? K dotknutej skupine, ktorá je vnímaná negatívnym spôsobom, sa nespravodlivo správajú rôzni členovia spoločnosti, vrátane policajtov a policajtiek, vyučujúcich, zamestnancov a zamestnankýň úradov práce, zamestnávateľov a iných relevantných aktérov. Títo prechovávajú predsudky voči danej skupine tak isto ako zvyšok spoločnosti, a tak je veľmi pravdepodobné, že i oni diskriminujú dotknuté skupiny v podobe odmietnutia rôznych služieb alebo poskytnutia dostatočne kvalitných služieb (napr. úrad práce neposkytne rekvalifikačné kurzy pre nezamestnaných alebo škola bude segregovať rómske deti). Toto vedie k nižšej kvalite života a chudobe. Dotknuté skupiny môžu byť vyčlenené z kultúrneho alebo politického života, keďže členom a členkám „majority" sa nepáči, keď práve oni majú politickú moc.

1.5.3 Nenávistné prejavy ako hrozba pre demokraciu

Nenávistné prejavy taktiež predstavujú ohrozenie demokracie, pretože umlčiavajú menšiny a odrádzajú ich od zapájania sa do verejných online diskusií. Pri snahe komunikovať cez rôzne kanály a zaradiť sa do spoločnosti, rôzni diskutujúci môžu dať pocítiť predstaviteľom menšín, že tu nie sú vítaní, čím ich umlčia, ich názory zosmiešnia lebo sú vnímané ako menejcenné. Vo svojom dôsledku môžu byť menšiny menej zastúpené v politických štruktúrach a mať menší vplyv na verejné politiky a celkovo politickú a verejnú sféru[19].

Toto sa týka všetkých verejných fór, vrátane internetu, ktorý môže byť tiež považovaný za politickú arénu či miesto výmeny a šírenia názorov, miesto, kde ľudia vytvárajú aliancie a získavajú dôležité informácie a ovplyvňujú názory ostatných (v niektorých prípadoch, v masovej miere pôsobiac na tisíce ľudí). Internet a obzvlášť online sociálne sieťovanie môže preto napĺňať rôzne demokratické funkcie. Môže byť priestorom pre občiansku angažovanosť a poskytovať členom určitej menšiny pocit plného občianstva. Autori Citron a Norton v ich štúdii[20] z roku 2011 to nazývajú „digitálne občianstvo". Podľa nich nenávistné prejavy predstavujú veľmi vážne ohrozenie tohto typu demokratickej angažovanosti, keďže to môže umlčiavať a odrádzať členov menšín od ich participácie.

1.5.4 Bezpečnostné dopady

Písanie nenávistných komentárov možno interpretovať, že autor alebo autorka komentáru „vypúšťa paru", a teda že zverejnením takéhoto obsahu sa znižuje pravdepodobnosť, že skutočne spácha aj trestný čin mimo online priestoru proti danej skupine. Napriek tomu, nenávistné prejavy nezabraňujú páchaniu nenávistných trestných činov. Nenávistné prejavy môžu dokonca viesť k vyvolaniu násilia a povzbudenia jednotlivcov k páchaniu trestných činov. Autori Foxman a Wolf[21] zmieňujú niekoľko prípadov prenasledovania, vrážd a samovrážd, ktoré boli vyvolané

nenávistnými prejavmi. Napríklad, v roku 1998 v USA bola na jednej rasistickej webovej stránke zverejnené informácie o matke (Bonnie Jouhari) dieťaťa zmiešaného „rasového" pôvodu. Následne bola vystavená nenávistnými prejavmi, obťažovaná telefonátmi a prenasledovaná aj doma. V inom prípade z roku 1999, David Copeland nastražil bombu s klincami, ktorá zabila troch ľudí a zranila vyše stovku ľudí. Neskôr vysvetlil svoje konanie slovami: „Zaútočil som na černochov, Pakistáncov a zvrhlíkov". Tyler Clementi spáchal samovraždu ako dôsledok kyberšikany od ľudí, ktorí ho videli na videu v romantickej situácii s človekom rovnakého pohlavia na Twitteri. Táto situácia bola nelegálne natočená tajnou kamerou jeho spolubývajúcim, ktorý video na Twitteri zverejnil.

Genocídy a pogromy sú tiež úzko prepojené s nenávistnými prejavmi, respektíve, s nenávistnou propagandou. Genocídy boli sprevádzané šírením nenávisti, predsudkov a dehumanizujúcich správ o dotknutých skupinách. Napríklad, genocíde v Rwande v roku 1994, resp. masovému zabíjaniu páchanému aj civilným obyvateľstvom, ktoré počas 100 dní viedlo k zavraždeniu približne 500 000 až 1 000 000 ľudí, predchádzala nenávistná propraganda. Členovia a členky etnickej skupiny Tutsiov, ktorá bola hlavným terčom masového zabíjania, boli nazývaní „švábmi" (inyenzi) a „hadmi" (inzoka). Tým však nemožno tvrdiť, že iba nenávistné prejavy samotné spôsobujú genocídu. Nenávistné prejavy predstavujú dôležitý nástroj na vytváranie spoločenskej klímy, verejnej mienky a hlavne postojov voči dotknutej skupine. V prípade genocídy v Rwande, pomenovania Tutsiov boli cielene použité za účelom dehumanizovať ich, teda vykresliť ich, že nie sú ľudskými bytosťami s cieľom ľahšie zapojiť väčšinové obyvateľstvo Hutuov do masového zabíjania. Konflikt v bývalej Juhoslávii v 90-tych rokoch 20. storočia bol taktiež vyvolaný a podnecovaný nenávistnými prejavmi v masmédiách voči rôznym etnickým a náboženským skupinám.

„Za účelom podnietiť jednotlivcov, aby páchali genocídu, podnecovanie v zmysle navádzania nie je dostačujúce; vyžaduje si to predošlé vytvorenie určitej klímy, v ktorej páchanie takýchto zločinov je umožnené. Nenávistná propaganda vedie k vytvoreniu takejto klímy"

W.K. Timmermann v publikácii *„The relationship between hate propaganda and incitement to genocide: A new trend in international law towards criminalization of hate propaganda" (Vzťah medzi nenávistnou propagandou a podnecovaním ku genocíde: Nové trendy v medzinárodnom práve ku kriminalizácii nenávistnej propagandy)* (2005, s. 257)

2 Vzdelávanie a kampane ako primárny spôsob odstraňovania predsudkov

Vzdelávanie ako spôsob odstraňovania nenávistných prejavov je založené na predpoklade, že v prvom rade je nutné adresovať korene nenávistných prejavov, ktorými sú najmä predsudky a mýty o rôznych spoločenských skupinách (napríklad etnických menšinách, ženách, LGBTI)[22]. Vzdelávanie je preto často odporúčané odborníkmi a odborníčkami z vedeckej obce či aktivistami a aktivistkami, ktorí poukazujú na rôzne výhody tohto prístupu, akými je okrem adresovania predsudkov aj podpora medzikultúrneho dialógu, zvyšovanie povedomia o tom, ako chrániť seba samých, ale aj ostatných pred nenávistnými prejavmi v online aj offline svete.

Keďže deti a mládež sú obzvlášť zraniteľné voči nenávistným prejavom, je potrebné sa zameriavať predovšetkým na túto skupinu. Súčasťou obsahu vzdelávacích aktivít môže byť celá rada tém, akým je napríklad zoznámenie sa s definíciou nenávistných prejavov, s ich škodlivým dopadom, ale aj všeobecne extrémizmom, demokraciou, či ľudskými právami, a to osobitne tých najzraniteľnejších skupín, akými sú napríklad menšiny. Vzdelávanie nemusí prebiehať len v školskom prostredí, ale tiež v domácnostiach, komunitách, prostredníctvom médií, občianskych združení, kultúrnych akcií či v samotnom online priestore.[23]

Vyučovanie v triede

- Vzdelávanie o nenávistných prejavoch, extrémizme či predsudkoch môže byť zahrnuté do školských osnov nielen na úrovni základných a stredných škôl, ale v primeranej forme aj na predškolskej úrovni. Osobitne dôležité je posilňovať kritické myslenie a čitateľskú gramotnosti. Okrem toho by mali byť

deti a mládež vedené k tomu ako sa chrániť v online priestore a zároveň neubližovať ostatným.[24]

Non-formálne a neformálne učenie

- Niektorí autori a autorky navrhujú inovatívne a interaktívne spôsoby učenia o nenávistných prejavoch, ako napríklad prostredníctvom videí, filmov či divadelných hier, ktoré vyvracajú predsudky a presadzujú začlenenie rôznych skupín do spoločnosti. Ďalšími nástrojmi v tomto smere môžu byť súťaže v písaní esejí o rôznych témach, akými sú napríklad výhody spoločenského začleňovania menšín, hudobné a umelecké festivaly, akým je napríklad festival Fusion v Bratislave o nových menšinách, vzdelávanie prostredníctvom médií (napríklad programy pre deti o témach rôznorodosti), ale tiež tzv. mainstreaming, ktorý môže byť realizovaný napríklad začlenením členov a členiek menšín do televíznych diskusií alebo filmov a to tak, že sú postupne braní ako súčasť spoločnosti.[25]

Vzdelávanie novinárov a novinárok

- Osobitne dôležité je vzdelávanie ľudí pracujúcich v médiách a študentov a študentiek žurnalistiky či masmediálnej komunikácie a to tak, aby si tieto skupiny osvojili citlivý jazyk a tým predchádzali ďalšiemu šíreniu predsudkov a mýtov o rôznych spoločenských skupinách.

Verejné kampane

Hlavným cieľom online a offline verejných kampaní je zvyšovať povedomie alebo záujem o istú tému, v tomto prípade nenávistných prejavov. Kampane však môžu slúžiť na monitorovanie nenávistného obsahu na internete alebo dokonca navrhujú riešenia ako s ním pracovať.

Verejné kampane môžu mať rôzne formy, akými sú napríklad videá, billboardy, mediálne články a blogy, a byť distribuované cez rozličné komunikačné platformy, akým je napríklad Facebook, bežné webstránky, YouTube alebo ďalšie masmédiá, akým je televízia, rádio a noviny.

Verejné kampane môžu byť rozdelené na:[26]

- Kampane určené na zvyšovanie povedomia – napríklad môžu zvyšovať povedomie o nenávistných prejavoch alebo diskriminácii
- Kampane na pozitívne zviditeľnenie členov a členiek menšín
- Kampane zamerané na zbieranie informácií o diskriminácii a zabraňovanie diskriminačných praktík

Kampane na zvyšovanie povedomia slúžia na vzdelávanie ľudí napríklad o tom, čo je nenávistný prejav, aké môže mať formy a negatívne dopady. Tieto kampane môžu byť zamerané tiež na to, ako sa majú užívatelia a užívateľky internetu chrániť pred nenávistnými prejavmi online, ako môžu reagovať, kde takéto prejavy nahlasovať alebo na koho sa majú obrátiť, pokiaľ sú terčom alebo svedkom nenávistného obsahu[27]. Jedným z príkladov takejto online kampane je European Ins@fe, ktorej cieľom je poskytovať informácie deťom, rodičom, pracovníkom a pracovníčkam v školstve a s mládežou o tom ako sa v digitálnom prostredí chrániť, ako používať rôzne online nástroje a za týmto účelom vytvárať rôzne stratégie. Ďalším príkladom kampaní na zvyšovanie povedomia je kampaň „Virtuálny rasizmus, skutočné dopady" (ang. „Virtual racism, real consequences") vedenou brazílskou organizáciou s názvom Criola. Obsahom kampane bolo umiestniť nenávistné prejavy na internete na billboardy v blízkosti bydliska tých, ktorí ich šírili (bez uvedenia ich identity) s cieľom poukázať na to, čo sú nenávistné prejavy a povzbudiť ľudí, aby proti nim vystúpili.

Cieľom kampaní na pozitívne zviditeľnenie členov a členiek menšín (alebo iných skupín, ktoré sú terčom nenávisti ako sú napríklad LGBTI alebo náboženské skupiny) je posilniť ich postavenie, a to prezentovaním ich v pozitívnom svetle[28]. Jedným z príkladov takejto kampane je „Syndróm Róm", ktorej cieľom bolo predstaviť úspešných Rómov a Rómky pracujúcich vo vede, v školstve, či na rôznych pozíciách v súkromnom sektore.

Kampane zamerané na zbieranie informácií o diskriminácii a zabraňovanie diskriminačných praktík slúžia najmä na monitorovanie webstránok, sociálnych médií či tých, čo nenávistné príspevky šíria[29]. Príkladmi takýchto kampaní, ktoré fungujú vo forme webstránok, je napríklad nemecká Hass-im netz.info alebo maďarský inštitút Athenea monitorujúce aktivity extrémistických skupín na internete. Takéto webstránky často poskytujú internetovým užívateľom a užívateľkám informácie nielen o stránkach s nenávistným obsahom, ale aj rady ako sa pred nenávistným obsahom na internete chrániť.

3 Stratégie a nástroje na znižovanie nenávistných prejavov v online priestore

Okrem legislatívnych opatrení a vzdelávania môžeme identifikovať súbor mäkkých nástrojov, ktoré môžu používať rozliční aktéri od ľudí pôsobiacich v oblasti vzdelávania cez jednotlivých užívateľov internetu, poskytovateľov internetových služieb (PIS), organizovaných skupín a komunít, ako napríklad mimovládne alebo verejné organizácie, a ďalších, ktorí sa snažia zmierňovať nenávistné prejavy v online priestore. Zmierňovanie nenávistných prejavov v online priestore si vyžaduje širšie zapojenie a spoluprácu medzi týmito aktérmi. Podstatné je, že čím viac sú takéto stratégie používané, tým výraznejšie sa podarí znížiť nenávistné prejavy v digitálnom svete. Účelom tejto časti je motivovať rozličných aktérov k činom a zapojeniu sa do zmierňovania nenávistného obsahu.

Nasledujúce nástroje je možné využiť v sociálnych médiách a na internetových diskusných fórach: protiargumentácia (overovanie faktov, osobná skúsenosť, „predstav si seba" a obzvlášť vytváranie „spojencov" pomocou odkazov na kľúčové slová (#, hashtag) ako doplnkový nástroj pri protiargumentácii), označovanie príspevkov (tzv. flagging) a nahlasovanie, pomenovanie a zahanbenie, zvýraznenie pozitívnych komentárov, používanie filtrovacieho softvéru.

3.1 Protiargumentácia

Najzákladnejšou stratégiou na zníženie nenávistných prejavov je ozvať sa proti nenávistným prejavom s vlastným prejavom. Inými slovami, každý jednotlivec sa môže ozvať a vyjadriť svoj nesúhlas s nenávistnými príspevkami v online diskusii.

„Snáď najdôležitejšie je jednoducho poskytnúť nevyvrátiteľný dôkaz, ktorým každému pripomenieme, že svet je plný ľudí dobrej vôle – ľudí, ktorí odmietajú nenávisť a držia sa hodnôt občianstva a rešpektu."

Foxman & Wolf v publikácii „Viral Hate: Containing its spread on the Internet (Virálna nenávisť: Pohľad na jej rozšírenie na internete)", (2013, s. 132)

Z pohľadu jednotlivca sa môže zdať, že protiargumentácia nemá žiaden dopad a že jednotlivec samotný nič zásadné nedokáže zmeniť. Napriek tomu má v praxi protiargumentácia často úlohu povzbudenia pre ostatných užívateľov a užívateľky sociálnych médií, ktorí nájdu motiváciu prispieť do diskusie, keď uvidia, že v diskusii sú aj konštruktívne komentáre[30]. V dôsledku toho môže vzrásť množstvo protiargumentov, ktoré nadobudnú prevahu nad ľuďmi s nenávistnými prejavmi, ktorí môžu byť v skutočnosti v menšine.

Hlavné princípy pri oponovaní voči nenávistným komentárom

Konfrontovať nenávistné prejavy v online diskusiách si vyžaduje určité pravidlá, aby mala celá táto aktivita šancu byť účinná. Maja Nenadović navrhuje vo svojom článku[31] z roku 2013 komplexnú stratégiu nazvanú PLEASE, určenú na efektívne reagovanie na ostatných diskutujúcich na online diskusných fórach. Odporúča aktívne sa zapájať do diskusií a postupovať podľa nasledujúcich krokov:

1. **Pause:** Dajte si pauzu a neberte diskusiu osobne, aby ste sa vyhli útokom a konfliktom.
2. **Listen:** Počúvajte pozorne argumenty, aby ste ich úplne pochopili.
3. **Express empathy:** Vyjadrite empatiu aj v prípade rasistických poznámok.
4. **Analyse:** Analyzujte komentáre a protiargumenty, ako aj predpoklady, na ktorých sú založené.
5. **Speak:** Hovorte a vyjadrite názory.
6. **Explain:** Trpezlivo vysvetľujte vlastné názory, pretože ľudia môžu mať odlišné vedomosti a životné skúsenosti.

Používatelia môžu využiť rozličné formy protiargumentov, konkrétne:

- Overovanie faktov – vyslovenie tvrdení na základe faktov
- Osobná skúsenosť
- „Predstav si seba" alebo stratégia rozprávania príbehov
- Humor a sarkazmus

3.1.1 Overovanie faktov

Pri zapojení sa do diskusie s cieľom konfrontovať nenávistné prejavy môže užívateľ využiť rozličné druhy argumentov. Jeden zo spôsobov, ako sa postaviť voči nenávistným prejavom, je **vyslovenie tvrdení na základe faktov (overovanie faktov)**, čo v praxi znamená vyvrátiť určité výroky, napr. rasistické stereotypy, pomocou dôveryhodných dát alebo vedeckých dôkazov[32]. Napríklad pri diskutujúcich na sieti Facebook, ktorí tvrdia, že príslušníci rómskej alebo inej etnickej menšiny poberajú špeciálne alebo nadmerné sociálne dávky len kvôli svojmu etnickému pôvodu, a preto majú privilégiá oproti príslušníkom „majority", je možné uviesť údaje o skutočnej výške sociálnych dávok podľa zdrojov Úradu práce, sociálnych vecí a rodiny a vyvrátiť takéto nepravdivé tvrdenia.

Vyslovenie tvrdení na základe faktov a overovanie faktov ako formu protiargumentácie môžu iniciovať nielen bežní používatelia internetu, ale aj školení profesionáli a profesionálky, ako napríklad moderátori a moderátorky internetových diskusií, novinári a novinárky (v prípade webových stránok spravodajských médií) alebo knihovníci a knihovníčky pracujúci s primárnymi zdrojmi, ktoré veľmi efektívne konfrontujú predsudky a nepravdivé informácie a môžu byť vnímaní ako autority, ktorých intervenciu je možné rešpektovať.

3.1.2 Osobná skúsenosť

Ďalšou stratégiou komunikovania s diskutujúcimi v online priestore je ponúkať argumenty **na základe osobnej skúsenosti.** Znamená to, že nepoužijeme napríklad štatistiky alebo iné druhy dát na vyvrátenie protiargumentov, ale použijeme vlastnú osobnú skúsenosť. Napríklad pri konfrontácii s tvrdením o neochote príslušníkov etnickej menšiny alebo prisťahovalcov pracovať, ktorá je príčinou vysokej nezamestnanosti týchto skupín ľudí, môžeme použiť vlastnú skúsenosť s tým, že veľká časť pracovníkov na stavbách v Bratislave je rómskeho pôvodu. Takéto príbehy by mali pochádzať zo skutočného života, aby boli autentické a pravdivé.

3.1.3 „Predstav si seba" alebo stratégia rozprávania príbehov

Niektorí výskumníci a výskumníčky[33] uvádzajú, že keď sú protiargumenty predstavené naratívnym spôsobom, teda vo forme príbehu, môžu ich oponenti ľahšie prijať. Rovnako pri výskume toho, kedy majú ľudia tendenciu súhlasiť s intervenciou v sociálnych médiách (pri boji proti kybernetickému šikanovaniu), výskumník van Laer vo svojej štúdii[34] z roku 2013 zistil, že keď sú fakty predložené vo forme príbehu s prvkom odkazujúcim na seba, ľudia majú tendenciu nechať sa ľahšie presvedčiť, že regulácia sociálnych médií je nutná. Napríklad používanie fráz typu „predstavte si sami seba" sa javí ako účinné pri presviedčaní ľudí o určitých faktoch.

3.1.4 Humor a sarkazmus

Ďalšou formou protiargumentácie je reagovať s humorom na komentáre obsahujúce nenávistný obsah, a ukázať tak absurdnosť nenávistných výrokov (buď formou písomného komentára alebo pripojením mémov ku komentárom). Napríklad je možné reagovať na nepravdivé výroky o vysokej miere kriminality príslušníkov a príslušníčok určitej etnickej skupiny komentárom o vysokej kriminalite členov a členiek „majority" v korupčných kauzách. Je však potrebné vedieť, že medzi sarkazmom a osobnými útokmi či ponižovaním oponentov v diskusii je len tenká hranica. Používanie tejto formy protiargumentácie by nemalo tiež viesť k ďalšiemu prehĺbeniu neporozumenia medzi diskutujúcimi a teda odklonu od zmysluplnej diskusie. Preto by táto forma protiargumentácie nemala prekročiť hranicu medzi robením si žartov z nenávistných príspevkov a útokmi na jednotlivých diskutujúcich.

3.2 Vytváranie „spojencov" pomocou odkazov na kľúčové slová (#, hashtag)

Konfrontácia nenávistných prejavov môže byť vyčerpávajúca a veľmi stresujúca, pretože užívateľ alebo užívateľka môže zažiť osobné útoky, napríklad vulgarizmy alebo urážky, na vlastnej koži. A to nielen ak patrí k určitej primárnej cieľovej skupine (napr. etnickej menšine), ale aj ako jej spojenec, ktorý chce do diskusie priniesť konštruktívne komentáre. Protiargumentácia môže

byť tiež navyše časovo náročná kvôli potrebe vyhľadávať fakty a dáta (napríklad pri vyvracaní mýtov a predsudkov). Preto je vhodné, aby intervenovali naraz viacerí diskutujúci, a aby sa navzájom zastali a povzbudili. Jedným z nástrojov, ktorý umožňuje zhromaždiť takýchto spojencov v jednej diskusii je použiť odkazy na kľúčové slová (hashtag), ktoré umožňujú používateľom nájsť určité miesto. V roku 2017 vznikla na Slovensku iniciatíva #somtu[35], ktorá pomáha zhromaždiť takýchto spojencov v diskusiách na sieti Facebook v prípade, že sa v istých diskusiách vyskytne nenávistný obsah. Takýto odkaz na kľúčové slová v praxi vytvára komunitu ľudí, ktorí môžu intervenovať voči tým diskutujúcim, ktorí sú často v prevahe a zapĺňajú online diskusie nenávistnými komentármi, nepravdivými tvrdeniami, predsudkami alebo stereotypmi, a v dôsledku toho odrádzajú príslušníkov zraniteľných skupín alebo kriticky mysliacich užívateľov a užívateľky od zapojenia sa do diskusií.

3.3 Označovanie príspevkov a nahlasovanie

Stratégiu označovania príspevkov a nahlasovania môžeme tiež považovať za doplnkovú k protiargumentácii, pretože funguje ako reakcia na nenávistné komentáre tým, že upozorňuje správcov diskusie na porušenie pravidiel diskusie alebo pomocou negatívneho hlasovania vyjadruje nesúhlas s komentármi. Obe stratégie prenášajú na komunitu užívateľov a užívateľok zodpovednosť za ochranu jej členov pred ujmou spôsobenou nenávistnými prejavmi. Nahlasovanie príspevkov v online diskusiách sa používa v prípade porušenia pravidiel diskusie, kedy konkrétny komentár porušuje štandardy a pravidlá diskusnej komunity. Tieto štandardy a pravidlá by preto mali byť užívateľom a užívateľkám dostupné a jasne napísané, aby sa všetci užívatelia jednoducho oboznámili s definíciou nenávistných prejavov a tým, ktoré prejavy porušujú práva iných ľudí a ktoré nie.

Označovanie príspevkov slúži na vyjadrenie názoru negatívnym hlasovaním o komentári, ktoré môže viesť aj k skrytiu komentára v online diskusii. Napríklad niektoré platformy pre online diskusie najmä v rámci spravodajských médií (SME, Guardian, Independent atď.) umožňujú užívateľom a užívateľkám negatívne hlasovať o komentároch, ktoré síce neporušujú interné pravidlá diskusie, ale mnohé skupiny ľudí by ich mohli považovať za urážlivé. Navyše to posilňuje postavenie komunity v rozhodovaní o tejto otázke a definovaní, čo znamená neakceptovateľné správanie a neželaný prejav, a kedy by mal byť zakázaný. To isté platí o tlačidle „lajku" v diskusiách na sieti Facebook.

3.4 Komunitné „protestné" akcie v online priestore

Protiargumentáciu nemusí koordinovať len jednotlivec, ale môže mať aj podobu komunitnej alebo skupinovej akcie. Konkrétne napríklad používateľ siete Facebook môže vytvoriť udalosť alebo fanúšikovskú stránku, ktorá by reagovala na rôzne stereotypy alebo nenávistné skupiny, alebo by sa vysmievala z názorov určitých nenávistných skupín a zároveň by vznikla komunita ľudí s podobnými názormi diskutujúcimi na sieti Facebook. Na Slovensku boli nedávno viaceré príklady takýchto iniciatív, ako napríklad „Naše Slovensko", ktoré sa vysmieva z ultrapravicovej politickej strany Kotleba - Ľudová strana Naše Slovensko. Zverejňuje rozličné obrázky parodujúce členov a sympatizantov tejto politickej strany.

3.5 Pomenovanie a zahanbenie webových stránok s nenávistným obsahom

V prípade už zmienených „maskovaných web stránok", ktoré obsahujú nepravdivé historické, politické alebo spoločenské fakty, je účinným spôsobom reagovania na ne, monitorovať a identifikovať ich, aby bola verejnosť (bežní používatelia internetu, ktorí hľadajú informácie) dobre informovaná o ich dôveryhodnosti. Slovenský projekt konspiratori.sk pre bežných užívateľov internetu nielen monitoruje a uvádza zoznamy takýchto „maskovaných webových stránok", ale upozorňuje aj komerčné spoločnosti a odrádza ich od umiestňovania online reklamy na takéto webové stránok.

3.6 Zvýraznenie pozitívnych komentárov

Potrebná je však nielen protiargumentácia, nahlasovanie alebo negatívne hlasovanie o škodlivom obsahu, ale veľmi dôležitá je aj možnosť užívateľov zvýrazniť dobré príklady pozitívnych komentárov (ako napríklad komentáre, ktoré vyvrátia rasistické predsudky)[36]. V diskusiách na sieti Facebook môže ako takáto stratégia slúžiť tlačidlo

„lajku", pretože „lajkovaním" pozitívneho obsahu (komentáre, príspevky, videá, obrázky) podporujú užívatelia a užívateľky iných ľudí, ktorí sú autormi určitého pozitívneho materiálu alebo konštruktívnych názorov, a vyjadrujú súhlas s príslušným obsahom. Zvýraznenie pozitívneho materiálu v online priestore môže byť možnosťou vtedy, ak sa používateľ z rôznych dôvodov nechce aktívne zapojiť do online diskusie, napríklad ak sa obáva osobných útokov iných diskutujúcich alebo nemá čas písať protiargumenty.

Poskytovatelia internetových služieb[37] ako kľúčoví aktéri pri zmierňovaní nenávistných prejavov v online priestore

Poskytovatelia internetových služieb (PIS) môžu zohrávať významnú úlohu pri adresovaní nenávistných prejavov v online priestore, pretože môžu stanoviť interné pravidlá a postupy pre užívateľov a užívateľky, ktorými zakážu určitý obsah, napríklad taký, ktorý rasovo alebo etnicky urážlivý. Najbežnejšími príkladmi takýchto interných politík sú **Dohody o používaní služieb a Pravidlá komunity**, ktoré umožňujú poskytovateľom internetových služieb odstrániť všetok nenávistný obsah pomocou definovania nenávistných prejavov a iných foriem neakceptovateľného správania v online priestore. Preto je potrebné, aby PIS používali svoje interné politiky na rozhodnutie, či ide o nenávistné prejavy, a riešenie týchto prípadov.

Interné politiky majú tiež vzdelávaciu úlohu. Pravidlá komunity nielenže umožňujú PIS zakazovať a odstraňovať nenávistný obsah, ale zároveň aj **vzdelávajú svojich užívateľov a užívateľky o neakceptovateľných prejavoch v online priestore** a informujú o ich zodpovednosti chrániť ostatných užívateľov[38].

Pravidlá diskusie sú prínosom aj pre samotných PIS, pretože ak by nenávistné prejavy v online priestore a kybernetická kriminalita neboli dostatočne regulované, odradí to užívateľov od zapájania sa do verejných online diskusií alebo akýchkoľvek iných aktivít, pri ktorých nechcú byť vystavení predsudkom a nenávisti, ale chcú diskutovať o rozličných témach na určitej úrovni kvality. S tým súvisia aj ťažkosti ohľadom kvality komentárov, pretože konštruktívne komentáre, ktoré poskytujú nové pohľady a perspektívy, priťahujú konštruktívne mysliacich užívateľov, zatiaľ čo nenávistné komentáre spôsobujú opak[39].

K interným politikám sú komplementárne už spomenuté **systémy nahlasovania** (ako napríklad označovanie príspevkov, nahlasovacie tlačidlá, formuláre sťažností a horúca linka), ktoré umožňujú užívateľom nahlasovať škodlivý obsah, aby PIS mohol konať v zmysle odstránenia akéhokoľvek škodlivého obsahu, varovania užívateľov, ktorí porušujú interné pravidlá, alebo ich blokovanie v prípade závažného alebo opakovaného porušenia[40].

3.7 Filtrovací softvér

Ďalším technologickým nástrojom, ktorý by mohol plniť funkciu stratégiu na odstránenie nenávistného obsahu v online priestore (na webových stránkach, nie v sociálnych médiách), je filtrovací softvér, ktorý môže blokovať nenávistný materiál. Jedným z najstarších dobre známych filtrovacích softvérov je program Hatefilter, ktorý vydala organizácia Anti-Defamation League (Liga proti hanobeniu) v roku 1998 v USA, ktorý nielenže triedi a blokuje webové stránky obsahujúce nenávistný obsah, ale poskytuje jej používateľom a používateľkám aj informácie a vzdelávanie o nenávistných prejavoch. Filtrovacie nástroje používajú hlavne rodičia a ich hlavným cieľom je chrániť deti pred nenávistným materiálom na internete. Patria medzi ne rôzne nástroje rodičovskej kontroly, napríklad nástroje NetNanny, SurfWact, Parental Control alebo Webwatcher. Je dôležité uviesť, že takýto softvér filtruje a blokuje webové stránky, ktoré boli nahlásené monitorovacími organizáciami (napr. Anti-Defamation League, Gay and Lesbian Alliance against Defamation), ktoré sledujú lokality s nenávistným obsahom.

Filtrovacie nástroje môžeme rozdeliť do dvoch kategórií, a to na **nástroje používané klientom** a **nástroje používané serverom.** Filtrovacie nástroje na strane klienta iniciujú samotní užívatelia (napr. rodičia, školy, pracoviská), pričom nástroje na strane servera iniciujú a inštalujú PIS, pričom tieto môžu napríklad blokovať webové stránky, ktoré označili monitorovacie organizácie ako lokality s nenávistným obsahom.

Napriek tomu však filtrovací softvér čelí určitým obmedzeniam, najmä v súvislosti s diskusiami na sociálnych médiách (ako napríklad Facebook, Twitter), kde je obsah dynamický a neustále sa

mení[41]. Z tohto dôvodu sú diskusné platformy sietí Faceboook a Twitter silne kritizované za to, že nechránia svojich užívateľov a užívateľky od nenávistného obsahu.

3.8 Zhrnutie

Keďže nenávistný obsah výrazne zaplavuje kybernetický svet, rôzni aktéri od bežných internetových užívateľov a užívateliek, cez predstaviteľov médií, až po poskytovateľov internetových služieb by mali zasahovať a zapojiť sa do vytvárania bezpečného digitálneho sveta. Ako sme uviedli vyššie, tento cieľ je možné dosiahnuť pomocou rôznych nástrojov. V nasledujúcich kapitolách predstavíme experimentálne testovanie účinnosti dvoch foriem protiargumentácie, a to stratégiu overovania faktov a stratégiu osobnej skúsenosti, ako aj pohľady predstaviteľov médií v oblasti boja s nenávistnými prejavmi.

4 Experiment zameraný na zníženie nenávistných prejavov v sieti Facebook

4.1 Celkový zámer

V tomto experimente sme my, autorský kolektív tejto publikácie, chceli dosiahnuť dva hlavné ciele. Po prvé, chceli sme identifikovať, aké formy nenávistných prejavov voči rómskej komunite sa najčastejšie objavujú v diskusiách na sieti Facebook. Po druhé, chceli sme otestovať, ako ľudia reagujú, keď niekto vstúpi do diskusie na sieti Facebook s obhajobou rómskej menšiny, a ktoré stratégie sa javia ako najúčinnejšie v konfrontácii s nenávistnými prejavmi voči rómskej komunite.

4.2 Testované stratégie

Napriek našej vedomosti o tom, že existujú viaceré stratégie konfrontácie s nenávistnými prejavmi (pozri kapitolu 3), v snahe o jednoduchšiu realizáciu a presnejšie zacielenie experimentu sme sa rozhodli testovať len dve hlavné stratégie. V rámci prvej stratégie sme argumentovali pomocou osobnej skúsenosti s konkrétnymi ľuďmi alebo človekom z rómskej komunity, čo narušilo nenávistné alebo voči rómskej menšine zaujaté tvrdenie iného diskutujúceho. V rámci druhej stratégie sme argumentovali pomocou overovania faktov alebo výskumnými zisteniami, kvalitatívnymi alebo kvantitatívnymi zisteniami rôznych rešpektovaných osôb z výskumnej sféry, výskumných inštitúcií alebo pomocou oficiálnych údajov (napríklad Úradu práce, sociálnych vecí a rodiny), ktoré vyvracajú nenávistné tvrdenie iného diskutujúceho.

Príklad stratégie na základe osobnej skúsenosti:

„Každého istým spôsobom poburuje predstava toho, že živí niekoho, komu sa pracovať nechce. Chcem však podotknúť, že z mojej skúsenosti to nie je vždy tak, ako je to tu prezentované. Mám kamarátov, známych rómskeho pôvodu. Jeden z nich dostal správu z agentúry, cez ktorú sa chcel zamestnať, že Rómov neprijímajú. Chcem tým povedať, že sa môže ľahko stať, že niektorí Rómovia pracovať chcú alebo chceli, ale nebola im daná šanca zamestnať sa."

Príklad stratégie na základe overovania faktov:

„Nerozumiem však tomu, ako môžete veriť niečomu takému, že Rómovia dostávajú väčšie dávky ako ostatní ľudia, ktorí Rómovia nie sú. Nič také ako rómske dávky neexistuje. Ani stravu v škole nedostávajú len rómske deti. Dostávajú ju všetky deti bez rozdielu, ak pochádzajú zo sociálne znevýhodnených rodín. Tieto veci je možné ľahko si vygoogliť cez upsvar.sk. Mimochodom, dávky dostane jednotlivec: 61,60 €, jednotlivec s 1 – 4 deťmi: 117,20 €, dvojica s 1 – 4 deťmi: 160,40 € a dvojica s viac ako 4 deťmi v paušálnej sume 216,10 € s tým, že nezáleží na tom, či má 5 alebo 10 detí. Okrem toho, jednotlivci a rodiny môžu poberať príspevok na bývanie, čo zväčša Rómovia nie sú oprávnení poberať, lebo na to potrebujú vlastniť byt/dom alebo mať nájomnú zmluvu, čo teda v osadách nemajú. Odkaz http://www.upsvar.sk/socialne-veci-a-rodina/hmotna-nudza/pomoc-v-hmotnej-nudzi.html?page_id=363675"

V tomto experimente sme prostredníctvom stratégií uvedených vyššie testovali dve hypotézy, ktoré znejú nasledovne:

(1) Vstup do diskusií na sieti Facebook s prorómskymi argumentmi zvýši celkový podiel prorómsky zameraných komentárov (pričom do celkového

podielu nezapočítavame naše vlastné komentáre). Inými slovami, že obhajoba rómskej menšiny bude motivovať iných ľudí s prorómskymi postojmi tiež vstupovať do diskusií.

(2) Stratégia na základe osobnej skúsenosti bude účinnejšia pri odstraňovaní nenávistných prejavov voči rómskej komunite než stratégia na základe overovania faktov.

Na otestovanie týchto dvoch hypotéz sme sa rozhodli analyzovať spolu 60 diskusií, z ktorých každá sa nachádzala pod postom na sieti Facebook, ktorý sa týkal rómskej témy (napríklad spravodajský článok o vysokej nezamestnanosti Rómov zverejnený na sieti Facebook vo forme postu s priloženým linkom na článok). Všetky tieto posty na sieti Facebook boli zverejnené v období od apríla 2016 do januára 2017. Do 30 diskusií sme vstúpili s prvou alebo druhou stratégiou, prípadne s kombináciou oboch stratégií. Skupinu týchto 30 diskusií nazývame „intervenčná skupina". Do zvyšných 30 diskusií na sieti Facebook, ktoré sa týkali rómskej témy, sme nevstúpili a nazývame ich „kontrolná skupina". Týmto spôsobom môžeme porovnávať diskusie v intervenčnej skupine a kontrolnej skupine, aby sme mohli testovať prvú hypotézu. Na testovanie druhej hypotézy sme porovnávali naše vlastné komentáre len v rámci intervenčnej skupiny. V 10 týchto diskusiách sme použili výlučne stratégiu na základe osobnej skúsenosti. V 10 diskusiách sme použili výlučne stratégiu na základe overovania faktov. A vo zvyšných 10 diskusiách sme použili kombináciu oboch týchto stratégií.

V snahe testovať druhú hypotézu, teda či je jedna stratégia účinnejšia ako druhá, sme museli najprv definovať, čo považujeme za „účinnú" alebo „úspešnú" intervenciu. Za jednoznačne neúspešnú intervenciu v snahe odstrániť nenávistné prejavy sme považovali situáciu, kedy diskutujúci (autor pôvodného príspevku alebo iný diskutujúci v rámci rovnakého diskusného vlákna) reagoval na náš komentár ďalším komentárom obsahujúcim nenávistný prejav alebo predsudky. Za úspešnú alebo účinnú intervenciu sme považovali nasledujúce situácie:

- diskutujúci úplne uznal náš argument obhajujúci rómsku komunitu;
- diskutujúci čiastočne uznal náš argument obhajujúci rómsku komunitu;
- žiaden ďalší diskutujúci v rámci príslušného diskusného vlákna nereagoval na náš komentár, čo zastavilo nenávistné prejavy.

Poslednú uvedenú situáciu je najviac problematické považovať za úspech, pretože nepoznáme dôvody, pre ktoré diskutujúci nepokračovali v diskusii. Možným dôvodom je to, že ich naše argumenty presvedčili, ale aj to, že si nevšimli náš komentár alebo jednoducho nemali chuť pokračovať v diskusii napriek tomu, že sa ich nenávistný postoj voči rómskej komunite vôbec nezmenil. Jednako však bol aj v tejto situácii dosiahnutý celkový zámer odstrániť nenávistné prejavy v priestore diskusií na sieti Facebook, pretože komentáre s nenávistným obsahom nepokračovali. Preto sme aj túto situáciu hodnotili ako úspešnú.

4.3 Platforma pre experiment

Tohto experimentu sa zúčastnil štvorčlenný výskumný tím. Ako výskumný tím sme neustále vzájomne koordinovali naše postupy. Pôvodne sme sa snažili intervenovať z našich vlastných profilov na sieti Facebook, ale nakoniec sme sa s cieľom ochrany pred osobnými útokmi ostatných diskutujúcich rozhodli vytvoriť jeden falošný profil, ktorý sme všetci používali pri intervenovaní.

Intervenovali sme v diskusiách pod postami na sieti Facebook, ktoré sa týkali témy rómskej menšiny na Slovensku. Tieto posty boli zverejnené buď konkrétnymi poslancami a poslankyňami Národnej rady Slovenskej republiky alebo rôznymi slovenskými spravodajskými médiami, ako napríklad SME, aktuality.sk, RTVS a Topky, pretože patria medzi najpopulárnejšie spravodajské internetové stránky na Slovensku[42]Ani politici a političky, ani spravodajské média nezverejňujú často posty o rómskej komunite. Preto sme v podstate vstúpili do všetkých diskusií na sieti Facebook, ktoré sme počas výskumnej fázy našli a ktoré boli fyzicky dostupné pre našu intervenciu. Medzi týmito postami sme nerobili žiadnu selekciu. Kontrolnú skupinu diskusií tvorili všetky posty, ktoré sme prehliadli počas výskumnej fázy (pretože autorský kolektív pracoval aj na iných projektoch alebo počas dovoleniek nebol vždy dostupný pre intervenciu). V tabuľke 1 sú znázornené zdroje všetkých postov na sieti Facebook, ktoré sme zahrnuli do našej intervenčnej alebo kontrolnej skupiny diskusií. Rozličné

online spravodajské média mierne prevyšujú počet diskusií na sieti Facebook iniciovaných politickými predstaviteľmi a predstaviteľkami.

Priestor sociálnej siete Facebook je veľmi živým a neustále sa meniacim prostredím, ktoré prinieslo množstvo výziev pre samotný výskum a možnosť načrtnúť akékoľvek všeobecné výskumné závery. Uveďme charakteristické výzvy, ktoré ovplyvňujú validitu tohto výskumu:

- zdroje postov na sieti Facebook (Tabuľka 1) sa veľmi líšili skladbou ich publika (socio-ekonomické prostredie, dosiahnuté vzdelanie, voličské preferencie atď.);
- obsah postov sa výrazne líšil, niektoré posty už samotné obsahovali nenávistné prejavy s istými emocionálnymi konotáciami, napr. „Rómsky gang napadol turistu", zatiaľ čo iné posty používali relatívne neutrálne výrazy a obsah, napr. o politikách bývania;
- dĺžka diskusií a počet komentárov v skúmaných diskusiách sa výrazne líšil (najkratšia diskusia obsahovala 35 komentárov, pričom najdlhšia obsahovala 523 komentárov), rovnako sa líšila aj viditeľnosť týchto diskusií prostredníctvom zdieľania a lajkovania na sieti Facebook.

4.4 Výsledky

Intervencia zvyšuje podiel prorómskych komentárov (Hypotéza 1)

V 60 skúmaných diskusiách na sieti Facebook bolo analyzovaných 7 586 komentárov (4 027 v intervenčnej skupine a 3 559 v kontrolnej skupine). Každý komentár sme označili nasledujúcimi kódmi: (1) prorómsky komentár; (2) protirómsky komentár; (3) zmiešaný komentár (obsahoval aj prorómske, aj protirómske názory) alebo (4) nepodstatný komentár (netýkal sa témy rómskej menšiny).

Ako je znázornené v tabuľke 2, najviac zastúpené v intervenčnej aj kontrolnej skupine boli nepodstatné komentáre. V intervenčnej skupine sme prispeli 362 prorómskymi komentármi z celkového počtu 4 027 komentárov v tejto skupine. Hoci celkový podiel prorómskych komentárov bol

Tabuľka 1: Zdroje skúmaných diskusií na sieti Facebook

Zdroj	Kontrolná skupina (počet príspevkov)	Intervenčná skupina (počet príspevkov)
Aktuality (online spravodajské médium)	9	15
Topky (online spravodajské médium)	8	2
Boris Kollár (poslanec NR SR za hnutie Sme rodina – Boris Kollár)	5	5
Milan Krajniak (poslanec NR SR za hnutie Sme rodina – Boris Kollár)	3	2
Lucia Nicholsonová (poslankyňa NR SR za stranu Sloboda a Solidarita)	2	0
Alojz Hlina (bývalý poslanec NR SR, predseda Kresťanskodemokratického hnutia)	2	0
Milan Uhrík (poslanec NR SR za stranu Kotleba – Ľudová strana Naše Slovensko)	1	0
SME (tlačové a online spravodajské médium)	0	3
RTVS (televízne, rozhlasové a online médium)	0	2
Lucia Žitňanská (ministerka spravodlivosti za stranu Most-Híd)	0	1
Spolu	**30**	**30**

Tabuľka 2: Podiel prorómskych a protirómskych komentárov

Intervenčná skupina					
naše komentáre	prorómske (bez našich)	protirómske	zmiešané	nepodstatné	**spolu**
362	293	1203	78	2091	**4027**
9 %	7,3 %	29,9 %	1,95 %	51,9 %	**100%**

Kontrolná skupina				
prorómske	protirómske	zmiešané	nepodstatné	**spolu**
160	1581	99	1719	**3559**
4,5%	44,4 %	2,8 %	48,3 %	**100%**

v intervenčnej aj kontrolnej skupine skôr nízky, po vylúčení našich komentárov bol podiel prorómskych komentárov vyšší v intervenčnej skupine (7,3 %) v porovnaní s kontrolnou skupinou (4,5 %). Tento rozdiel je štatisticky významný na hladine pravdepodobnosti 0,0117. Na základe tohto výsledku môžeme usudzovať, že naša prvá hypotéza bola potvrdená. Preto môžeme tvrdiť, že prispievanie prorómskymi argumentami do diskusií na sieti Facebook motivuje iných ľudí vstúpiť do diskusie a tiež sa zastať rómskej komunity. Inými slovami, oplatí sa brániť diskriminovanú spoločenskú skupinu na online platforme, pretože to nielen zvyšuje celkovú viditeľnosť inkluzívnych postojov, ale motivuje aj iných prispieť k tejto viditeľnosti.

Nie je rozdiel v účinnosti dvoch testovaných stratégií (Hypotéza 2)

V 30 diskusiách v rámci intervenčnej skupiny sme použili pri jednotlivých komentároch buď stratégiu osobnej skúsenosti alebo stratégiu overovania faktov. Celkový počet komentárov so stratégiou overovania faktov bol výrazne vyšší ako so stratégiou osobnej skúsenosti (195 s overovaním faktov a 79 s osobnou skúsenosťou). Výskumný tím bol zručnejší v používaní faktov a dôkazov, ako vo vytváraní vymyslených osobných skúseností v prípade chýbajúcej vlastnej osobnej skúsenosti, ktorá by sa týkala diskutovanej témy. Táto výrazná nerovnováha v počte komentárov v rámci jednej a druhej stratégie nesie so sebou obmedzenia pri zovšeobecnení výskumných zistení.

Pri porovnávaní komentárov so stratégiou osobnej skúsenosti s komentármi so stratégiou overovania faktov sme zistili, že neexistuje štatisticky významný rozdiel v ich účinnosti. Komentáre so stratégiou osobnej skúsenosti zastavili nenávistné prejavy v 33,33 % prípadov, zatiaľ čo komentáre so stratégiou overovania faktov ich zastavili v 36,27 % prípadov. Diskutujúci čiastočne uznali argumenty komentárov so stratégiou osobnej skúsenosti v 2,56 % prípadov a pri komentároch so stratégiou overovania faktov v 4,15 % prípadov. Žiaden diskutujúci neuznal úplne náš prorómsky argument. Na základe všetkých týchto výsledkov a s vedomím metodologických obmedzení tohto výskumu môžeme usudzovať, že stratégia osobnej skúsenosti i stratégia overovania faktov majú podobný potenciál účinného odstraňovania nenávistných prejavov v online priestore, čiže ani jedna zo stratégií sa neukázala byť účinnejšou než tá druhá.

Diskusie zahltené protirómskymi predsudkami

V 60 diskusiách na sieti Facebook, ktoré sme zahrnuli do tohto výskumu, sme sledovali frekvenciu výskytu rôznych druhov protirómskych komentárov. Tieto komentáre môžeme rozdeliť do troch skupín:

1) Akí sú ľudia z rómskych komunít a ako sa správajú?
2) Aké politiky a opatrenia voči rómskej komunite sa v súčasnosti uplatňujú?
3) Aké politiky a opatrenia voči rómskej komunite by sa mali uplatňovať v budúcnosti?

Graf 1 znázorňuje rôzne druhy protirómskych komentárov, ktoré reagujú na niektorú z týchto otázok. Ako je zo zoznamu týchto druhov zrejmé, takmer žiaden diskutujúci nepoužíval výskumné zistenia alebo fakty na podloženie svojich argumentov. Týkalo sa to protirómskych aj prorómskych komentárov. Diskutujúci používali buď svoje osobné skúsenosti na podporu svojich argumentov alebo len použili všeobecné tvrdenia a prezentovali ich ako dobre známe a jednoznačné pravdy. Problémom je, že väčšina týchto „právd" boli len „mýty".

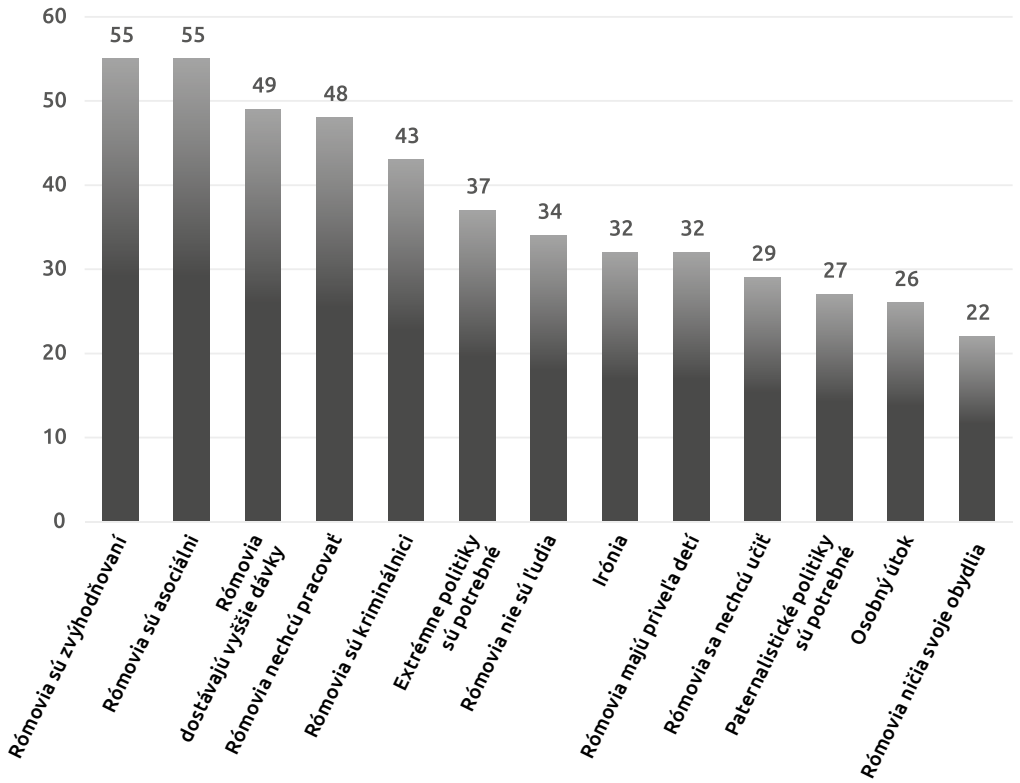
Akí sú ľudia z rómskych komunít a ako sa správajú?

Graf 1 znázorňuje, že ľuďom z rómskych komunít sú pripisované viaceré vlastnosti a druhy správania:

- asociálne správanie (spomenuté v 55 zo 60 diskusií);
- neochota pracovať (spomenuté v 48 zo 60 diskusií);
- kriminálnici (spomenuté v 43 zo 60 diskusií);
- prirovnávanie k zvieratám, hmyzu alebo veciam (spomenuté v 34 zo 60 diskusií);
- majú priveľa detí (spomenuté v 32 zo 60 diskusií);
- neochota vzdelávať sa (spomenuté v 29 zo 60 diskusií);
- ničenie vlastných príbytkov (spomenuté v 22 zo 60 diskusií).

Diskutujúci neuviedli, že veľká väčšina rómskych obyvateľov je integrovaná v spoločnosti, ale ich neustále zobrazovali ako „výnimky". Často tiež zobrazovali Rómov a Rómky ako trvale „neprispôsobivých" majorite, ako asociálnych kriminálnikov, že je to v ich DNA, že sa s tým narodili. Niekedy aj diskutujúci rómskeho pôvodu uvádzali niektoré nenávistné prejavy voči iným ľuďom rómskeho pôvodu.

Graf 1: Druhy protirómskych komentárov



Aké politiky a opatrenia voči rómskej komunite sa v súčasnosti uplatňujú?

Graf 1 znázorňuje, že diskutujúci sú prevažne presvedčení, že Rómovia a Rómky sú privilegovanou skupinou v tom zmysle, že majú lepší prístup k verejným službám (ako napríklad bezplatné vzdelávanie v materských školách, obedy zdarma v školách, nové byty zdarma, lieky zdarma), dostávajú vyššiu sumu sociálnych dávok alebo špeciálne sociálne dávky. Bolo to spomenuté v 55 zo 60 diskusií na sieti Facebook. V 49 prípadoch použili diskutujúci chybné informácie o výške sociálnych dávok. Výrazne nadhodnotili ich úroveň, v niektorých prípadoch až päťnásobne viac oproti skutočnosti.

Diskutujúci zvyčajne popierali existenciu diskriminácie voči Rómom a Rómkam a v prípade, keď ju pripustili, obhajovali ju ako oprávnenú, pretože, ako argumentovali, ľudia, ktorí ich diskriminovali, pravdepodobne konali na základe vlastných negatívnych skúseností s nimi. Pripisovali „kolektívnu vinu" všetkým ľuďom rómskeho pôvodu.

Diskutujúci často ignorovali najzákladnejšie sociologické poznatky o tom, že spoločenské fenomény sú veľmi komplexné, záleží pri nich na okolnostiach a sú ovplyvnené množstvom psychologických, sociálnych, kultúrnych a politických faktorov. Žiadne jediné politické opatrenie, ani na celoštátnej úrovni, nedokáže vyriešiť a radikálne zmeniť taký komplexný fenomén, ako je segregácia a diskriminácia Rómov a Rómok. Tento fenomén sa dá zmeniť len dlhodobým postupným procesom, ktorý si vyžaduje množstvo politík a aktivít na všetkých úrovniach spoločnosti (individuálnej, lokálnej, regionálnej, štátnej a medzinárodnej) a vo všetkých sférach života (zamestnanosť, vzdelávanie, bývanie, zdravie, kultúra atď.). Kvôli neschopnosti vnímať celkovú situáciu ľudí rómskeho pôvodu v jej komplexnosti a kvôli nepochopeniu toho, že žiadne jediné politické opatrenie alebo program, či aktivita nedokážu priniesť rýchlu a radikálnu spoločenskú zmenu, mali diskutujúci tendenciu obviňovať politikov a mimovládne organizácie z míňania prostriedkov použitých na inklúziu vylúčených rómskych komunít.

Aké politiky a opatrenia voči rómskej komunite by sa mali uplatňovať v budúcnosti?

Zjednodušené vnímanie spoločenských fenoménov viedlo niektorých diskutujúcich k navrhovaniu extrémistických opatrení na vyhladenie alebo vyhostenie ľudí rómskeho pôvodu. Bolo to spomenuté v 37 zo 60 diskusií na sieti Facebook. Zjednodušujúca perspektíva viedla mnohých diskutujúcich tiež k tomu, aby sa stavali do nadradenej pozície ľudí, ktorí vedia najlepšie, čo je pre Rómov dobré, a k presvedčeniu, že prísnejšie tresty pre nich vyriešia problém. Objavilo sa to v 27 zo 60 diskusií na sieti Facebook.

Zatiaľ čo niektorí diskutujúci uvádzali rasistické názory, ale dištancovali sa od toho, aby boli označovaní ako rasisti, začala sa objavovať skupina diskutujúcich, ktorí sa sami hrdo označovali ako rasisti. Keď sa diskutujúcim na oboch stranách minuli argumenty, často prešli do vulgárnych slovných útokov (v 26 zo 60 diskusií) alebo do irónie (v 32 zo 60 diskusií). Zdá sa, že sarkastické alebo ironické komentáre dokážu hýbať emóciami diskutujúcich na oboch stranách, ale javia sa ako neúčinné.

Reakcie na naše intervencie podporujúce ľudí rómskeho pôvodu

Veľa diskutujúcich reagovalo na uvedené výskumné zistenia s averziou a spochybňovali samotný výskum alebo to, že nereflektuje životné skúsenosti ľudí, ktorí bývajú v blízkosti rómskych osád. Často tiež spochybňovali dôveryhodnosť výskumníkov a výskumníčok ako takých, ktorí nemajú priamu skúsenosť s Rómami a Rómkami.

Ani používanie osobných skúseností s Rómami a Rómkami nebolo pre mnohých diskutujúcich presvedčivé. Uznávali, že je mnoho Rómov a Rómok, ktorí nezodpovedajú negatívnym stereotypom, ale tých označovali ako „výnimky".

V niekoľkých prípadoch, kedy diskutujúci čiastočne uznali naše argumenty narušujúce predsudky v určitej oblasti, napríklad ohľadom výšky sociálnych dávok alebo neochoty pracovať a študovať, tí istí diskutujúci uviedli následne niekoľko ďalších predsudkov. Ak by sme použili metaforu s cibuľou, tak hoci sa nám v niektorých prípadoch podarilo čiastočne vyvrátiť určitý konkrétny predsudok a odstrániť jednu vrstvu cibule, diskutujúci nezmenili svoj celkový postoj voči ľuďom rómskeho pôvodu

a uvádzali niekoľko ďalších predsudkov, takže cibuľa stále ostávala celá. Aby sme zmenili celkový postoj osoby, museli by sme pokračovať v diskusii a pokúšať sa odstrániť zvyšné vrstvy cibule.

Keď diskutujúci navrhli nejaké extrémistické a rozličné krátkozraké opatrenia, ako napríklad zrušenie všetkých sociálnych dávok, a keď sme podrobne vysvetlili dopady, ktoré by takéto opatrenie malo na rómskych obyvateľov a obyvateľky, ako aj na celkovú spoločnosť v krajine, diskutujúci často v diskusii už nepokračovali. Je možné interpretovať tento jav tak, že nakoniec ich naše argumenty a vysvetlenia čiastočne presvedčili.

4.5 Zhrnutie

Vstupovať do diskusií na sieti Facebook a prezentovať prorómske postoje sa oplatí bez ohľadu na to, či pritom používame stratégiu overovania faktov alebo osobnej skúsenosti, pretože to motivuje ďalších diskutujúcich s prorómskymi postojmi vyjadriť svoj názor. Diskutujúci s protirómskymi postojmi sú zväčša presvedčení, že ľudia rómskeho pôvodu sú privilegovaní a že sú vo svojej podstate asociálni. Vnímajú celkovú situáciu a dopady politík zjednodušene a ako priamočiaru záležitosť.

5 Regulácia online diskusií a odstraňovanie nenávistných prejavov v slovenských médiách

5.1 Respondenti a respondentky, s ktorými sa uskutočnili rozhovory

Keďže nenávistné prejavy v kybernetickom priestore sú relatívne novým fenoménom so závažnými dopadmi na cieľové skupiny, je dôležité preskúmať, akú úlohu pri znižovaní výskytu nenávistných komentárov by mali zohrávať médiá v online diskusiách, ktoré spravujú. Počas februára 2017 sa uskutočnili pološtruktúrované rozhovory s predstaviteľmi a predstaviteľkami médií, ktorí zodpovedajú za reguláciu online diskusií v rámci médií SME, Aktuality.sk, Trend, Denník N a Pravda. Predstavitelia média Topky vyplnili dotazník s otvorenými otázkami. Cieľom bolo preskúmať, ako médiá postupujú pri eliminácii nenávistných prejavov na svojich webstránkach a aké stratégie používajú pri regulácii online diskusií (vrátane diskusií prostredníctvom siete Facebook). Vo vzorke spravodajských médií sú zastúpené najpopulárnejšie spravodajské média na Slovensku[43].

5.2 Výsledky

Hlavné zistenia ukazujú, že spravodajské médiá monitorujú nenávistné prejavy v online diskusiách a využívajú pritom rozličné nástroje a stratégie. Príspevky pravidelne posudzujú zamestnanci a zamestnankyne príslušných médií, ktorí zodpovedajú za reguláciu online diskusií, najčastejšie editori, hlavní editori, editori sociálnych sietí alebo iní správcovia. Hoci regulácia je výlučnou zodpovednosťou ľudí na týchto pozíciách, nie je to ich jediná úloha. Vzhľadom na čas, ktorý musia venovať ostatným úlohám, majú len obmedzený čas, ktorý môžu venovať regulácii diskusií a odstraňovaniu nenávistných prejavov. Napriek tomu, že regulovanie diskusií si vyžaduje viac pracovných hodín, regulátori nemajú možnosť venovať sa tejto činnosti na plný pracovný úväzok. Priemerný čas venovaný tejto činnosti je približne hodina denne.

Pri hodnotení obsahu diskusných príspevkov sa editori a editorky sociálnych sietí najčastejšie riadia písomnými pravidlami, ktoré sú zvyčajne dostupné na webových stránkach ich médií, pričom slúžia najmä diskutujúcim ako pravidlá vhodnej účasti v online diskusii. Ak nie sú pravidlá vytvorené, rozhodovanie o vhodnosti obsahu závisí od subjektívneho hodnotenia regulátora. Päť zo šiestich médií, s ktorých predstaviteľmi bol realizovaný rozhovor, má takéto pravidlá diskusie. Avšak žiadne z médií nemá vytvorený samostatný formálny dokument, ktorý by slúžil ako pokyny pre regulátorov diskusií. Regulátori monitorujú celú diskusiu vrátane príspevkov, ktoré nahlásili iní diskutujúci, s výnimkou média SME, v ktorom sa regulátor zameriava iba na príspevky nahlásené inými diskutujúcimi.

V rámci regulovania online diskusií používajú médiá v rozličnej miere niekoľko stratégií, ako odstrániť obsah s nenávistnými prejavmi. Médiá používajú aj niektoré stratégie, ktoré môžu využiť diskutujúci, a tak aktívne prispieť k odstráneniu nenávistných prejavov na sociálnych sieťach. V nasledujúcom texte uvádzame všetky

spomenuté stratégie, ktoré používajú skúmané slovenské médiá:

1) **Odstránenie príspevkov**
je stratégiou, ktorú používajú všetky oslovené médiá na boj proti nenávistným prejavom v online prostredí.
2) **Zrušenie diskusií**
sa využíva v prípade, ak spravodajské články prinášajú citlivú tému, napríklad o rómskej menšine, migrantoch a migrantkách atď., a/alebo v jednotlivých prípadoch, keď je diskusia zaplavená nenávistnými prejavmi. Zrušenie diskusií je možné použiť aj ako štandardné nastavenie pre všetky zverejnené spravodajské články. Napríklad Denník N otvára diskusie len v rámci niekoľkých zverejnených článkov denne. To môže zabrániť prieniku obsahu s nenávistnými prejavmi.
3) **Blokovanie prístupu určitých prispievateľov do diskusií**
používajú média, ako napríklad Aktuality.sk, keď diskutujúci poruší pravidlá diskusií. Blokovanie môže byť krátkodobé alebo dlhodobé.
4) **Sledovanie histórie diskutujúcich**
slúži ako nástroj na skontrolovanie histórie príspevkov určitých diskutujúcich. Regulátori kontrolujú históriu príspevkov najčastejšie vtedy, keď dochádza k porušeniu pravidiel diskusie. Regulátor môže overiť, či a ako často v minulosti diskutujúci porušil pravidlá. Zistenia zavážia pri rozhodovaní o blokovaní prístupu.
5) **Komunikácia s diskutujúcimi mimo diskusie**
sa vo väčšine médií nepoužíva pravidelne. Regulátori môžu napríklad poskytnúť diskutujúcim vysvetlenie ohľadom blokovania prístupu alebo odstránenia príspevku a môžu ich upozorniť na nevhodný obsah, ktorý v príspevku zverejnili.
6) **Nahlasovanie príspevkov**
je stratégia, ktorú využívajú diskutujúci. Čitatelia a čitateľky môžu nahlásiť príspevky s nevhodným obsahom prostredníctvom nahlasovacieho systému umiestneného na webstránke s diskusiami, napríklad Topky.sk. Nahlásené príspevky potom vyhodnotí regulátor a odstráni ich, ak porušovali pravidlá diskusie.
7) **Hodnotenie komentárov**
je dostupné napríklad na webstránke SME a využívajú ho diskutujúci. Prostredníctvom hodnotenia vyjadrujú čitatelia a čitateľky svoj názor na komentáre zverejnené inými diskutujúcimi. Hodnotiaci systém môže mať podobu znakov plus a mínus (kde plus znamená dobrý príspevok a mínus jeho opak) a sú umiestnené nad alebo pod zverejneným príspevkom diskutujúceho.
8) **Skrytie príspevkov**
využívajú regulátori niektorých médií, ako napríklad Trend, pri príspevkoch, ktoré sú negatívne hodnotené. V dôsledku toho je potrebné kliknúť na skrytý príspevok, aby sa zobrazil na čítanie.
9) **Vytváranie tzv. blacklistov**
je stratégiou, ktorú využíva väčšina oslovených médií, napr. denník Pravda. Čierne zoznamy obsahujú zakázané slová, ktoré sú zaznamenané do špeciálneho systému a je možné ich kedykoľvek aktualizovať. Systém zneprístupní pre ostatných diskutujúcich všetky príspevky, ktoré obsahujú slová z tohto blacklistu.
10) **Moderovanie diskusií**
je v praxi zriedkavo používanou stratégiou a znamená vedenie diskusie zodpovedným zamestnancom či zamestnankyňou média. Hoci všetci oslovení predstavitelia médií si uvedomujú význam tejto stratégie, moderovanie využíva Denník N, a to tiež len príležitostne. Novinári a novinárky tohto média sa snažia byť prítomní v diskusiách k vlastným článkom.
11) **Overovanie faktov**
je osobitným druhom moderovania diskusií, ktorý je tiež používaný zriedkavo. Overovanie faktov sa používa zvyčajne pri obhajobe obsahu (napríklad správnosti dát uvedených v článku) príspevkov/spravodajských článkov novinárov a novinárok alebo vtedy, keď sa nájde chyba v takýchto príspevkoch. Overovanie faktov využívajú len médiá SME a Trend.

5.3 Prekážky, s ktorými sa médiá stretávajú

Napriek tomu, že médiá vnímajú zodpovednosť za obsah v online diskusiách (na svojich webových stránkach aj platformách siete Facebook, ktoré používajú) a uvedomujú si potrebu regulovať diskusie efektívnejšie, nedarí sa im to v plnej miere. Médiá čelia viacerým prekážkam v snahe o systematickejšiu reguláciu online diskusií. Jednou z najbežnejších prekážok je zrejme nedostatok ľudských zdrojov na reguláciu online diskusií. Regulácia diskusií je časovo náročná činnosť a regulátori majú v tejto

oblasti obmedzenia. Rovnako zamestnať viac ľudí a/alebo vytvoriť pozície výlučne na reguláciu diskusií si vyžaduje vysoký finančný vstup zo strany médií, čo je obzvlášť problematické pre malé médiá.

5.4 Odporúčania na zlepšenie regulácie

Oslovení predstavitelia a predstaviteľky médií odporúčajú niekoľko opatrení na zlepšenie regulácie online diskusií a účinné odstránenie nenávistných prejavov v online priestore. Po prvé, odporúčajú zvýšiť počet zamestnaných ľudí, ktorí by sa tejto činnosti venovali. Po druhé, odporúčajú zlepšiť vymáhanie práva, a to konkrétne zaviesť podrobnejšie vyšetrovanie, ktoré by v oprávnených prípadoch viedlo aj k sankciám. Po tretie, uvádzajú návrh spoplatniť zverejňovanie určitého počtu komentárov/príspevkov v online diskusiách ako jedno z možných riešení obmedzenia prístupu k diskusiám a počtu príspevkov v nich na zjednodušenie ich regulácie. Po štvrté, zjednotenie pravidiel diskusií medzi jednotlivými médiami by pomohlo médiám zjednotiť proces regulácie nenávistných prejavov na internete. V neposlednom rade, predstavitelia a predstaviteľky médií poukázali na potrebu zlepšiť reguláciu online diskusií samotnou sociálnou sieťou Facebook. Sieť Facebook obsahuje množstvo nenávistných príspevkov, z ktorých mnohé ostávajú online aj po ich nahlásení. Zvyčajne sú posty a komentáre zo siete Facebook odstraňované po vyhodnotení miery pravdepodobnosti, že by vyhrážka obsiahnutá v príspevku/komentári viedla k fyzickému útoku alebo inej forme zlého zaobchádzania. Okrem toho regulátor nemusí porozumieť obsahu nenávistného komentáru kvôli jazykovej bariére. Tieto postupy uplatňované sieťou Facebook komplikujú účinnosť boja proti nenávistným prejavom na internete.

5.5 Zhrnutie

Slovenské spravodajské médiá používajú väčšinu dostupných stratégií na reguláciu online diskusií, vrátane diskusií na sieti Facebook. Napriek tomu sa však ukazuje, že odstránenie nenávistných prejavov v online diskusiách nie je ich prioritou z hľadiska financií ani ľudských zdrojov. Regulátori diskusií venujú tejto činnosti len veľmi obmedzený čas a nie sú na ňu dostatočne vyškolení.

Použitá literatúra

- Banks, J. (2010). *„Regulating hate speech online“*. International Review of Law, Computers & Technology, 24(3), 233-239.
- Boeckmann, R. J., & Turpin-Petrosino, C. (2002). *Understanding the harm of hate crime*. Journal of Social Issues, 58(2), 207-225.
- Citron, D. K. & Norton H. (2011). *„Intermediaries and hate speech: Fostering digital citizenship for our information age“*. Boston University Law Review, 91(4), 1435 – 1484.
- Cohen-Almagor, R. (2011). *„Fighting hate and bigotry on the Internet“*. In: *Policy & Internet*, Volume 3, Issue 3
- Delgado, R., & Stefancic, J. (2004). *Understanding words that wound*. Boulder: Westview Press.
- Diakopoulos, N. & Naaman, M. (2011). *„Towards quality discourse in online news comments“*, In: *Proceedings of the ACM 2011 conference on Computer supported cooperative work*, 2011
- Durrheim, K., Greener, R. and Whitehead, K.A. (2015). *„Race trouble: Attending to race and racism in online interaction“*. British Journal of Social Psychology, 54, 84-99.
- Foxman, A.H & Wolf, Ch. (2013) *Viral hate. Containing its spread on the Internet*. New York: Palgrave Macmillan.
- Hrdina, M. (2016). *Identity, activism and hatred: Hate speech against migrants on facebook in the Czech republic in 2015*. Naše společnost, 14 (1), 38-47
- Leets, L. (2002). *Experiencing hate speech: Perceptions and responses to anti-semitism and antigay speec*h. Journal of Social Issues, 58(2), 341-361.
- Leets, L., & Giles, H. (1997). *Words as weapons - When do they wound? Investigations of harmful speech*. Human Communication Research, 24(2), 260-301.
- Liao, Q. Vera, Wai-Tat Fu (2013). *Beyond the filter bubble: interactive effects of perceived threat and topic involvement on selective exposure to information. Proceedings of the SIGCHI Conferehence on Human Factors in Computing Systems.* Dostupné na: https://dl.acm.org/purchase.cfm?id=2481326&CFID=953875486&CFTOKEN=97626206
- Maitra, I. (2012). *Subordinating speech*. In I. Maitra & M. K. McGowan (Eds.), *Speech and harm: Controversies over free speech* (s. 94-120). Oxford: Oxford University Press.
- McGonagle, T. A (2012). *„Survey and critical analysis of Council of Europe strategies for countering „hate speech““*. In: Herz, M. & Molnar, P. (2012). The content and context of hate speech. Cambridge University Press.
- Mosher, D., & Proenza, L. (1968). *Intensity of attack, displacement and verbal aggression*. Psychonomic Science, 12, 359-360.

- Molnar, P. (2012) *Responding to „hate speech with art education and the imminent danger test"* In: Herz, M. & Molnar P. (2012). *The content and context of hate speech: Rethinking regulation and responses*. New York: Cambridge University Press.

- Nenadović, M. (2013). *Applied debate: A weapon of fighting discrimination and building understandin*g. 4th International conference on argumentation, rhetoric, debate, and the pedagogy of empowerment. Dostupné na: http://d2ivco2mxiw5i2.cloudfront.net/app/media/5207

- Nielsen, L. B. (2002). *Subtle, pervasive, harmful: Racist and sexist remarks in public as hate speech*. Journal of Social Issues, 58(2), 265-280.

- Simpson, R. M. (2013). *Dignity, harm, and hate speech*. Law and Philosophy, 32, 701-728.

- Sorial, S. (2015). *Hate speech and distorted communication: Rethinking the limits of incitement*. Law and Philosophy, 34, 299-324.

- Stevens, T. & Neumann, P. R. (2009). *Countering online radicalisation: A strategy for action*. London: The International Centre for the Study of Radicalisation and Political Violence and the Community Security Trust.

- Timmermann, W. K. (2005). *The relationship between hate propaganda and incitement to genocide: A new trend in international law towards criminalization of hate propaganda?* Leiden Journal of International Law, 18, 257-282.

- Titley, G., Keen, E. & Foldi, L. (2012). *Starting points for combating hate speech online: Three studies about online hate speech and ways to address it*. Council of Europe, British Institute of Human Rights.

- van Laer, T. 2013. *„The means to justify the end: Combating cyber harassment in social media"*, Journal of Business Ethics, Volume 123, Issue 1, 85 – 98.

- Velšic, M. (2016). *Mladí ľudia v kyberpriestore: Šance a riziká pre demokraciu.* Bratislava: Inštitút pre verejné otázky

- West, C. (2012). *Words that silence? Freedom of expression and racist hate speech*. In I. Maitra & M. K. McGowan (Eds.), *Speech and harm: Controversies over free speech* (222-248). Oxford: Oxford University Press

Poznámky

1 Pozri Simpson, R. M., *Dignity, harm, and hate speech*, Law and Philosophy 32, s. 701 (2013)

2 Tamtiež. s. 702

3 Pozri Banks, J., *Regulating hate speech online*, International Review of Law, Computers & Technology, 24(3), s. 2 (2010)

4 Pozri Simpson, R. M., *Dignity, harm, and hate speech*, Law and Philosophy 32, s. 701 (2013)

5 Pozri Foxman, A.H & Wolf, Ch., *Viral hate. Containing its spread on the Internet*. New York: Palgrave Macmillan, s. 16 (2013)

6 Pozri Sorial, S., *Hate speech and distorted communication: Rethinking the limits of incitement*, Law and Philosophy 34 (2015)

7 Pozri Leets, L., & Giles, H., *Words as weapons - When do they wound? Investigations of harmful speech*, Human Communication Research, 24(2) (1997)

8 Pozri Leets, L., *Experiencing hate speech: Perceptions and responses to anti-semitism and antigay speech*, Journal of Social Issues 58(2) (2002)

9 Pozri napríklad Hrdina, M., *Identity, activism and hatred: Hate speech against migrants on Facebook in the Czech Republic in 2015*, Naše společnost, 14 (1), s. 39 (2016) alebo Foxman, A.H & Wolf, Ch., Viral hate. Containing its spread on the Internet. New York: Palgrave Macmillan (2013)

10 Pozri Hrdina, M., *Identity, activism and hatred: Hate speech against migrants on Facebook in the Czech Republic in 2015*, Naše společnost, 14 (1) (2016) alebo Liao, Q. Vera, Wai-Tat Fu, *Beyond the filter bubble: interactive effects of perceived threat and topic involvement on selective exposure to information* (2013), dostupné na: https://dl.acm.org/purchase.cfm?id=2481326&CFID=953875486&CFTOKEN=97626206

11 Pozri Council of Europe, *No hate speech movement survey on cyber hate speech* (2016), dostupné na: http://www.nohatespeechmovement.org/survey

12 Pozri Velšic, M., *Mladí ľudia v kyberpriestore: Šance a riziká pre demokraciu.* Bratislava: Inštitút pre verejné otázky (2016)

13 Pozri Nielsen, L. B., *Subtle, pervasive, harmful: Racist and sexist remarks in public as hate speech*, Journal of Social Issues 58(2) (2002)

14 Pozri Delgado, R., & Stefancic, J., *Understanding words that wound*, Boulder: Westview Press (2004)

15 Pozri napríklad Delgado, R., & Stefancic, J., *Understanding words that wound*, Boulder: Westview Press (2004); Boeckmann, R. J., & Turpin-Petrosino, C., Understanding the harm of hate crime, Journal of Social Issues 58(2)(2002) alebo Maitra, I., *Subordinating speech*, In I. Maitra & M. K. McGowan (Eds.), *Speech and harm: Controversies over free speech*, Oxford: Oxford University Press (2012)

16 Pozri Boeckmann, R. J., & Turpin-Petrosino, C., *Understanding the harm of hate crime*, Journal of Social Issues 58(2), s. 222 (2002)

17 Tamtiež.

18 Pozri Simpson, R. M., *Dignity, harm, and hate speech*, Law and Philosophy 32, s. 718 (2013) alebo Sorial, S. *Hate speech and distorted communication: Rethinking the limits of incitement*, Law and Philosophy 34, s. 306 (2015)

19 Pozri West, C. (2012). *Words that silence? Freedom of expression and racist hate speech.* In I. Maitra & M. K. McGowan (Eds.), *Speech and harm: Controversies over free speech* (s. 222-248). Oxford: Oxford University Press (2012)

20 Pozri Citron, D. K. & Norton H., *Intermediaries and hate speech: Fostering digital citizenship for our information age*, Boston University Law Review 91(4) (2011)

21 Pozri Foxman, A.H & Wolf, Ch., *Viral hate. Containing its spread on the Internet*, New York: Palgrave Macmillan (2013)

22 Pozri Durrheim, K., Greener, R. and Whitehead, K.A., *Race trouble: Attending to race and racism in online interaction*, British Journal of Social Psychology 54 (2015).; Molnar, P., *Responding to hate speech with art education and the imminent danger tes*t In: Herz, M. & Molnar P., *The content and context of hate speech: Rethinking regulation and responses*, New York: Cambridge University Press (2012) alebo Foxman, A.H & Wolf, Ch., *Viral hate. Containing its spread on the Internet*, New York: Palgrave Macmillan (2013)

23 Pozri Foxman, A.H & Wolf, Ch., *Viral hate. Containing its spread on the Internet*, New York: Palgrave Macmillan (2013) alebo Titley, G., Keen, E. & Foldi, L., *Starting points for combating hate speech online: Three studies about online hate speech and ways to address it*, Council of Europe, British Institute of Human Rights (2012)

24 Pozri Titley, G., Keen, E. & Foldi, L., *Starting points for combating hate speech online: Three studies about online hate speech and ways to address it*, Council of Europe, British Institute of Human Rights (2012)

25 Pozri Foxman, A.H & Wolf, Ch., *Viral hate. Containing its spread on the Internet*, New York: Palgrave Macmillan (2013) alebo Titley, G., Keen, E. & Foldi, L., *Starting points for combating hate speech online: Three studies about online hate speech and ways to address it*, Council of Europe, British Institute of Human Rights (2012)

26 Pozri Titley, G., Keen, E. & Foldi, L., *Starting points for combating hate speech online: Three studies about online hate speech and ways to address it*, Council of Europe, British Institute of Human Rights, s. 63(2012),

27 Tamtiež. s. 64

28 Tamtiež. s. 69

29 Tamtiež. s. 74

30 Pozri Foxman, A.H & Wolf, Ch., *Viral hate. Containing its spread on the Internet*. New York: Palgrave Macmillan (2013)

31 Pozri Nenadović, M., *Applied debate: A weapon of fighting discrimination and building understanding. 4th International conference on argumentation, rhetoric, debate, and the pedagogy of empowerment* (2013) dostupné na: http://d2ivco2mxiw5i2.cloudfront.net/app/media/5207

32 Pozri Foxman, A.H & Wolf, Ch., *Viral hate. Containing its spread on the Internet*. New York: Palgrave Macmillan (2013)

33 Pozri napríklad Leets, L., *Experiencing hate speech: Perceptions and responses to anti-semitism and antigay speech*, Journal of Social Issues, 58(2) (2002) alebo van Laer, T. *The means to justify the end: Combating cyber harassment in social media*, Journal of Business Ethics, 123(1) (2013)

34 Pozri van Laer, T. *The means to justify the end: Combating cyber harassment in social media*, Journal of Business Ethics, 123(1) (2013)

35 #somtu vyjadruje prítomnosť a podporu pri čelení nenávistných komentárov v diskusiách na Facebooku. Slovenská iniciatíva #somtu je podobnou iniciatívou k švédskej #jagärhär (znamená „som tu" tiež), ktorá má rovnaký cieľ.

36 Pozri Foxman, A.H & Wolf, Ch., *Viral hate. Containing its spread on the Internet*, New York: Palgrave Macmillan, s. 140 (2013)

37 Poskytovatelia internetových služieb alebo poskytovatelia online služieb predstavujú aktérov, primárne organizácie, ktoré poskytujú služby súvisiace s prístupom a používaním internetových služieb. Najčastejšie internetové služby sú elektronické poštové služby, novinové články a videá (ako napríklad noviny Guardian, Independent), sťahovateľné materiály a programy (hry, filmy), internetové predajne, zábavný obsah ako napríklad videé a obrázky (YouTube), alebo tiež sociálne sieťovanie vrátane diskusných stránok (napr. Facebook, Twitter), atď. Môžu byť komerčné, súkromné, štátne, neziskové atď.

38 Pozri Citron, D. K. & Norton H., *Intermediaries and hate speech: Fostering digital citizenship for our information age*, Boston University Law Review, 91(4), s. 1441 (2011)

39 Pozri Diakopoulos, N. & Naaman, M., *Towards quality discourse in online news comments*, In: *Proceedings of the ACM 2011 conference on Computer supported cooperative work*, (2011) alebo Citron, D. K. & Norton H., *Intermediaries and hate speech: Fostering digital citizenship for our information age*, Boston University Law Review, 91(4), s. 1441, (2011)

40 Pozri Banks, J., *Regulating hate speech online*, International Review of Law, Computers & Technology, 24(3), s. 233-239 (2010); McGonagle, T. *A Survey and critical analysis of Council of Europe strategies for countering „hate speech"*, In: Herz, M. & Molnar, P. , *The content and context of hate speech*. Cambridge University Press (2012); Titley, G., Keen, E. & Foldi, L., *Starting points for combating hate speech online: Three studies about online hate speech and ways to address it*, Council of Europe, British Institute of Human Rights (2012)

41 Pozri Cohen-Almagor, R., *Fighting hate and bigotry on the Internet, Policy & Internet*, 3(3) (2011) alebo Stevens, T. & Neumann, P.R., *Countering online radicalisation: A strategy for action*, London: The International Centre for the Study of Radicalisation and Political Violence and the Community Security Trust (2009)

42 Výskumná vzorka bola vytvorená na základe zistení prieskumu najpopulárnejších médií v treťom a štvrtom kvartály roku 2015, ktorá realizovala MEDIAN SK dostupné na: https://medialne.etrend.sk/internet-grafy-a-tabulky.htm

43 Tamtiež.

O autorkách a autorovi

Lucia Kováčová
je výskumníčka v Inštitúte pre dobre spravovanú spoločnosť SGI v Bratislave. Získala titul MA vo verejnej politike na Stredoeurópskej univerzite v Budapešti so špecializáciou na rovnosť a sociálnu spravodlivosť. Výskumne sa zaoberá pracovnou integráciou, sociálnou ekonomikou a inkluzívnym vzdelávaním znevýhodnených detí a mladých.

Jozef Miškolci
získal doktorandský titul vo vzdelávaní na Univerzite v Sydney v Austrálii v roku 2014. Jeho hlavnými výskumnými oblasťami sú inkluzívne vzdelávanie, vzdelávacie politiky a ľudské práva vo vzdelávaní. V súčasnosti pracuje ako výskumný pracovník na Pedagogickej fakulte Komenského univerzity v Bratislave a čiastočne v Inštitúte pre dobre spravovanú spoločnosť SGI.

Edita Rigová
je juniorná výskumníčka v Inštitúte pre dobre spravovanú spoločnosť SGI. Získala titul MPA vo verejnej administratíve na IDHEAP – Swiss Graduate School of Public Administration na Univerzite v Lausanne vo Švajčiarsku. Edita tiež absolvovala stáž v Agentúre Európskej únie pre základné práva vo Viedni. Výskumne sa zaoberá politikami na začleňovanie rómskej menšiny.

SGI

DISCUSS without hate

Strategies to Counter Hate Speech Online

Lucia Kováčová
Jozef Miškolci
Edita Rigová

Discuss Without Hate:

Strategies to Counter Hate Speech Online

Jozef Miškolci
Slovak Governance Institute
Faculty of Education, Comenius University in Bratislava

Lucia Kováčová
Slovak Governance Institute

Edita Rigová
Slovak Governance Institute

Bratislava 2017

Discuss Without Hate: Strategies to Counter Hate Speech Online

Authors: Jozef Miškolci, Lucia Kováčová, Edita Rigová
Translation: Martina Kubánová
Copy-editing and proofreading: Susan Ryan
Reviewer: Agnes Horváthová
Graphic design: Tomáš Miško
Printed by: ADIN, s.r.o.

Slovak Governance Institute
Štúrova 3, 811 02, Bratislava, Slovakia
www.governance.sk

ISBN: 978-80- 972761-1- 9

Subsidised by

SGI
OPEN SOCIETY FOUNDATION
ROMANO KHER
RÓMSKY DOM

Supported by a grant from the Foundation Open Society Institute in cooperation with the Think Tank Fund of the Open Society Foundations.

Contents

List of Tables

List of Figures

Introduction

Social media is currently greatly overwhelmed by hateful content, creating an environment where different social groups (mainly ethnic, racial or religious groups, LGBTI, people with disabilities, women, youth and others) encounter insults, prejudices, and threats in different verbal or graphic forms. This hostile online environment severely impacts members of these groups, particularly in terms of short or long-term health issues, and might contribute to their exclusion from discussion forums (and therefore from political and civic life) and the emergence of discriminatory practices or violence.

Spreading hateful content is often wrongfully confused with—and justified by—the right to freedom of expression. The right to freedom of expression is, however, not absolute and does not allow people to abuse the rights of other individuals, such as their right to be free of discrimination, the right to freedom of thought, and the right to freedom of religion. The right to freedom of expression carries certain responsibilities, and social media users should participate in creating a safe digital environment where all individuals with various characteristics are able to freely engage in discussions and express their opinions without being insulted and threatened.

Besides restricting hate speech by the legislative measures, there are a set of soft tools that could be used to mitigate hate speech online without starting prosecution or a lawsuit and subsequently imposing sanctions. The soft tools of countering hate speech can be applied by a variety of stakeholders from basic social media users, parents, teachers, and Internet service providers to governmental and non-governmental bodies which should all engage in the processes of mitigating hate speech online.

The main aim of this handbook is to present the tools and strategies that various stakeholders might use to deal with hateful content online. The handbook aspires to encourage different stakeholders to take action, utilise the tools, and actively participate in creating a safe online space. The first chapter deals with the definition of hate speech online, its various aims, forms, frequency and the consequences for targeted groups and individuals. The second chapter briefly describes educational tools as a primary way to decrease hate speech. In the third chapter, different strategies and tools are analysed that can be used primarily by the Internet users and Internet Service Providers to reduce hateful content in social media and online discussions. The fourth chapter describes and analyses the results of experimental testing of the selected counter-speech strategies, particularly fact-checking and personal experience in the context of addressing anti-Roma hate speech. The final chapter deals with the role of the media in countering hate speech.

1 What is hate speech (online)?

Hate speech can be defined as a kind of communication or interaction where the affected individual or social group is intimidated by explicit or implicit insults, threats, or political statements that are inciting hatred, disgust or antipathy and which results in harassment, oppression and discrimination. Hate speech is often related to generalisations when the insults are targeted at the entire social group such as "the Roma are all criminals" or "all Muslims are terrorists". Hate speech might even incite violence against the targeted group or individual in an explicit way (for example, by encouraging others to attack the targeted group) or in an implicit way (when the targeted group is compared to animals or insects which makes it easier for offenders to have no mercy and commit violence).[3]

"Hate speech thus includes things like identity-prejudicial abuse and harassment, certain uses of slurs and epithets, some extremist political and religious speech (e.g. statements to the effect that all Muslims are terrorists, or that gay people are second-class human beings), and certain displays of 'hate symbols' (e.g. swastikas or burning crosses). We classify such activities as hate speech if, and insofar as, they convey the idea that belonging to a particular social group warrants someone's being held in or treated with contempt"

Robert Mark Simpson in *"Dignity, harm, and hate speech"* (2013, p. 701)

1.1 Forms of hate speech

Hate speech online may take the following forms[4]:

- **Verbal or written forms** such as vulgarisms, threats or insults based on different characteristics of the targeted group, but also political statements about the targeted group.
- **Hate symbols** such as swastikas or certain numbers symbolising hate ideologies.
- **Other graphic materials** such as pictures or memes depicting targeted groups in an insulting way (for example, portraying them as criminals or deviants).

Hate speech might be spread by publicly available material (e.g., posts or pictures shared on Facebook, Instagram or Twitter) or in the form of private messages sent to individuals' inboxes (e.g., via Facebook Messenger or WhatsApp).

Importantly, hate speech does not have to present open insults, for example, vulgarisms or graphic material explicitly disgracing the targeted group. Hate speech might be well-hidden and expressed in a more sophisticated way. Hateful comments may be then divided into[5]:

Overt messages – present open and direct insults targeting individuals and groups. These may include direct slurs and vulgarisms or graphic material that openly spread hate and prejudice against the targeted group or individuals.

Covert messages – present subtle way of insulting individuals and groups. Covert hate messages are closely associated with so-called "cloaked websites" which masquerade as sources of neutral, factual information about various historical, social, political and cultural topics. However, these websites are virtually filled with extremely prejudiced content and hate propaganda. Examples of such cloaked websites are the online and offline pseudoscientific magazine ZEM a VEK or different history-oriented Facebook pages or websites presenting "true" facts about different historical events. These sources contain, for example, denial of the existence or the scope of the Jewish genocide and inciting language, e.g., anti-Semitism. In the case of anti-LGBTI propaganda, the American website of the Family Research Council presents fake research data in order to prove wrongness of homosexuality. The goal of such websites is to lure students, parents, teachers, or other interested individuals into reading and believing that their content is scientific, and therefore legitimate and respectable.

"Cloaked websites" are characterised by new strategies for communicating with the public or their targeted groups. Groups and individuals administering these websites have modified their language, and their public speeches have become more sophisticated, polite and civil. Through this, they aim to become more acceptable to a wider and more diverse audience, and evade being captured by anti-hate speech legislation[6].

Covert ways of spreading hate does not have a lesser impact on the affected than overt hateful

messages. In this respect, in the study[7] from 1997 conducted on Asian and Caucasian university students, it was demonstrated that the covert racist messages appeared to the targeted social group to have more severe an impact on them than the overt ones. Nonetheless, severe and outrageous hate conduct and intentions are rarer and more difficult to demonstrate. A similar finding was previously revealed in another study by Mosher and Proenza (1968) in which the group members targeted with racial slurs did not perceive a difference in harm between severe and mild expressions of hate[8].

1.2 Aims of hate speech (online)

Several different aims of hate speech can be identified. Some of them might be intentional and systemic, and some might be unintentional, such as when a social media user spreads hateful content but is not aware of consequences of his or her actions. Virtually anyone can spread hate, and it can be due to different reasons such as low self-esteem, or because he or she personally experienced hate or bullying. Aims of hate speech might be categorised into[9]:

- Ventilating and releasing discussant's own fears and frustrations, and projecting these onto the targeted group as a scapegoat.
- Insulting, harassing and disgracing the targeted group.
- Expressing the inferiority of the targeted group.
- Reinforcing prejudice and spreading myths about the targeted group.
- Inciting antipathy or even hatred against the targeted group.
- Inciting oppression and violence.

1.3 Easiness of spreading hate in the digital world

Although hate speech frequently occurs in different spaces—including public places, workplaces, households or through the media—hate speech online is characterised by its inherent virality, and it can be spread by any social media user with access to the Internet. A single user may create and spread online hateful content from the comfort of his or her own home. And since social media (such as Facebook, Twitter and different online discussion forums) is used by a large number of users, such hateful .

> *"Everyone can be a publisher, even the most vicious anti-Semite, racist, bigot, homophobe, sexist, or purveyor of hatred. The ease and rapidity with which websites, social media pages, videos and audio downloads, and instant messages can be created and disseminated online make Internet propaganda almost impossible to track, control, and combat."*
>
> Foxman & Wolf in *"Viral Hate: Containing its spread on the Internet"*, (2013, p. 10)

This also means that a single user or an organised group of users are allowed to shape the public discourse online, which in turn affects the opinions of masses. In this way, it can be very simple to spread, for example, the myth about the excessive amount of social allowances that the Roma in Slovakia receive. The myth might be spread quickly and shared and "liked" (which increases the visibility of the content), and as a result, it convinces the general public of the privileged position of the Roma. This leads to inciting hatred and reinforcing already existing prejudices against this ethnic minority, which results in further oppressive policies.

Additionally, several scholars[10] suggest that the design of the discussion platforms themselves may contribute to ease of spreading hate content online, because social networks suggest popular content or content they are likely to like to their users (for example, by "liking" similar content), and then further spread hateful material. In that way, this design helps to gather people with similar hateful opinions who may then get radicalised. In contrast, social networks may also assist open-minded users who are against hate to gather and mobilise them so they can conduct anti-hate speech activities (see section 3.1.4. Community "protest" action online).

1.4 Frequency of hate speech online

As everyone can be a publisher and spread hateful content (in the form of social media posts or comments, blog posts, private messages, etc.), hate speech is a very common phenomenon. Results of the online survey on hate speech in the digital world conducted by the Council of Europe in 2016 showed that 4 out of 5 respondents experienced

some form of hate speech, while 2 out of 5 felt personally threatened by hate speech[11].

Other studies also show the high incidence of hateful content on the Internet. According to the research conducted by the Institute for Public Affairs in 2016 *Young people in cyberspace – chances and risks for democracy*[12], 69% of young Slovaks experience hate speech on social media and other online discussion forums. Another study[13] from the United States revealed that 61% of female participants reported that they were targeted by sexually suggestive hate speech "every day" or "often", while 46% of participating people of colour were targeted by race-related hate speech "every day" or "often".

1.5 Impact of hate speech or why should we bother about hate speech online?

Hate speech is not a new phenomenon. Hate speech can be considered as a component of oppression and discrimination of groups based on race, ethnicity, gender, religion and belief, sexual orientation, disability, age, or others. Most genocides and pogroms throughout the history started with propaganda and vicious statements about the targeted group. As a result, hate speech has historically led to violence, oppression and reinforcing racial, ethnic or other prejudices and, therefore, excluding certain minorities and other groups from every aspect of society.

Individuals writing hateful comments are often not aware of the severe consequences of their actions. On the one hand, they might intend to write something nasty about a particular person or a group, frequently in a moment of anger, but do not see the damaging impact on the targeted group. On the other hand, some hateful groups are fully aware of their actions and intentionally attempt to spread hatred, reinforce prejudices, and display power over a certain group (mainly those organised groups based on racist ideologies).

1.5.1 Physical and psychological impact

The topic of physical and psychological impact of hate speech has gained significant academic attention, primarily because the legitimacy of creating legislative restrictions on hate speech has to be well-justified and proven in order to be accepted by the public and jurisprudence. More particularly, the physical and psychological impacts can be short-term and long-term. The short-term physical and psychological impact of hate speech is characterised by a set of health issues which significantly lower the quality of life of affected individuals. The research from 2004 shows that hate speech may cause headaches, high blood pressure, rapid pulse rate, or even risky behaviours such as drug-taking. Besides that, hate speech causes anxiety, fear and sadness[14].

The long-term impact of repeated hate speech abuse may include damaged self-esteem or feelings of inferiority, lower aspiration level, nightmares, withdrawal from society, and depression or mental illness, which are proved by a variety of research studies[15]. In this respect, the level of identification with the minority groups seems to have an impact on the level of psychological damage or "engaging in self-blame, internalising negative evaluations, and failing to seek redress." The more the hate speech victims identify with their social group and find refuge and consolation in them, the lower the damaging impact[16]. The traumatising component in hate speech seems to be that the targeted groups are conscious of the extreme violence their members suffer or have suffered historically, and when they are targeted by hate speech, they are reminded of these and warned that a lethal attack on them might ensue one day as well[17]. They are deprived of their dignity, reputation and assurance regarding their immediate security and the security of their social status[18].

1.5.2 Social impact

Repetitiveness and the high occurrence of hate speech causes hate speech and prejudices about, for example, ethnic minorities to become socially acceptable. For example, the society accepts views that the certain ethnic minority is inferior or is characterised by negative attributes.

"…the common appearance of such epithets desensitizes readers, making hate speech and the denigration of minorities appear normal."

Foxman & Wolf in *"Viral Hate: Containing its spread on the Internet"*, (2013, p. 31-32)

What is the broader social consequence of normalising hate speech and accepting hateful attitudes towards minorities or other social groups? The affected group, when viewed in a negative way, is treated unfairly by members of society, including police officers, school teachers, employees of Labour Offices, employers, and other important stakeholders. They hold prejudice against the targeted group just like the rest of the society, and are likely to discriminate against the targeted group in the form of denying services, providing with low quality services (for example, not offering re-qualification courses in the Labour Offices or school segregation). This leads them to a lower quality of life and poverty. Affected groups might be excluded in cultural or political life as they are not welcomed by the members of the "majority" holding political power.

1.5.3 Hate speech as a threat to democracy

Hate speech also poses a threat to democracy because it silences minorities and discourages them from participating in the public discussions online. When attempting to communicate through different channels and getting engaged with the public, minority members might be made to feel unwelcome by other discussants, their voices might be silenced and opinions belittled because they are viewed as inferior. As a result, the minorities might be less represented in the political structures and have less influence on public policies and the overall political and public sphere.[19]

This applies to all public forums, including the Internet, which might be considered as a political arena, a place for exchanging and spreading opinions, a place where people make alliances or acquire important information and influence others' opinions (in some cases, in a massive way by targeting thousands of people). The Internet and online social networking in particular may, therefore, fulfil some democratic functions. It may be a platform for civic engagement and provide members of particular minority groups with a sense of full citizenship. Authors Citron and Norton, in their studyi[20] from 2011, call it a "digital citizenship". According to them, hate speech poses a serious threat to this type of democratic engagement since it may silence and discourage minority members from participation.

1.5.4 Security impact

One would consider that writing hateful comments provides the writers with the chance to "blow off steam" meaning that after writing such content, they are less likely to commit an "offline" crime against the targeted group. Hate speech, however, does not prevent the committing of hate crimes. Virtually, hate speech may incite violence and encourage individuals to commit different criminal acts. Authors Foxman and Wolf[21] report on the number of cases of hate-related stalking, murders and suicides. For instance, in 1998 in the USA, a white supremacist website featured a mother (Bonnie Jouhari) of a biracial child. As a result, she was seriously attacked by hate speech, harassed by phone calls and stalked at home. In 1999, David Copeland planted nail bomb killing three and injuring more than a hundred, while explaining his actions with the words "I bombed the blacks, Paki's, [and] Degenerates". Tyler Clementi committed suicide after being cyberbullied by strangers who watched him in a same-sex romantic moment on Twitter. The moment was illegally captured on spycam by his roommate who tweeted the video.

Genocides and pogroms are also closely associated with hate speech and, more particularly, with hate propaganda. Genocides were accompanied by the spreading of hatred, prejudice, and dehumanising messages about the targeted group. For example, in the Rwandan genocide in 1994, where an estimated 500 000 – 1 000 000 of people were killed within 100 days, the mass killing (committed by civilians as well) was preceded by hateful propaganda in which Tutsi (the ethnic group primarily targeted in the mass slaughter) were called "cockroaches" (inyenzi) and "snakes" (inzoka). That is, however, not to say that hate speech alone causes genocide. Hate speech presents an important tool for shaping social climate, public opinion and, more particularly, the attitudes towards the targeted group. In the case of the Rwandan genocide, the names the Tutsis were called was used with the aim to dehumanise them, and to portray them as not being humans in order to make it easier to engage the Hutu civilians in the slaughter. The conflict in the Former Yugoslavia in 1990s was also ethnically-based and stirred and sustained by hate speech in mass media.

"In order to incite individuals to commit genocide, incitement in the sense of instigation is insufficient; it requires the prior creation of a certain climate in which the commission of such crimes is possible. Hate propaganda leads to the creation of such climate"

W.K. Timmermann in *"The relationship between hate propaganda and incitement to genocide: A new trend in international law towards criminalization of hate propaganda"* (2005, p.257)

2 Education and campaigning as a primary way of targeting prejudice

Education as a strategy to tackle hate speech is based on the fact that it is necessary to **target the roots of hate speech**, which are often prejudices and myths about certain societal groups (e.g., ethnic minorities, women, LGBTI), to diminish hate speech both online and offline[22]. Therefore, education and training are widely recommended by scholars and activists who highlight their different benefits, such as addressing prejudices and fostering intercultural dialogue, raising awareness and responsibility for the protection of the offline or online community from hate speech.

Young people and children are especially vulnerable to hate speech, and therefore, it is necessary to primarily target this age group. A full range of topics might be a part of educational activities, such as the definition of hate speech, harmful impact of hate speech on people and, more generally, cultural and political diversity, history, extremism, democracy, and human rights, especially of minorities and vulnerable groups, etc. Importantly, education on hate speech can take place not only in schools, but also in households, communities, the media, civic organisations, cultural events, or in the online space.[23]

Classroom teaching

- Education on hate speech, extremism and prejudice may be incorporated in the school curriculum of pre-primary, primary and secondary schools in an age-appropriate way. It is especially important to foster critical thinking and literacy, and children should be taught to learn how to protect themselves in the cyber world and not to harm others[24]

Non-formal and informal learning

- Several authors suggest innovative and interactive ways of learning about hate speech, such as promoting movies, speeches and playlets that counter against prejudices and advocate for the inclusion of different groups in society; essay contests targeted at children and youth about various topics, such as the benefits of inclusion; music or arts festivals celebrating diversity, such as the festival Fusion in Bratislava celebrating diversity in the society; education through media (programmes for children on diversity issues) but also so-called mainstreaming in terms of, for example, involving minority members in TV discussions and movies so they are perceived as an integral part of society.[25]

Training of journalists and students of journalism

- Especially important is the training of journalists and students of journalism, so they can acquire sensitive language to avoid triggering prejudices and myths about certain groups in their media outputs.

Public campaigns

The primary aim of offline or online public campaigns is to raise awareness and attract attention towards a particular issue, but at the same time, they may also serve, for example, as a way to monitor hate websites. They educate about the problem, provide information, or even propose solutions, which is significantly important for democratic and active civil society.

Public campaigns may have different forms (e.g., videos, billboards, blogs, articles) and run in different communication platforms such as Facebook, websites, YouTube, or in the mass media such as TV, radio or newspapers.

Public campaigns may be categorised into:[26]

- **Awareness campaigns** – raise awareness on hate speech or discrimination in the public;
- **Affirmative campaigns** – present members of minorities in a positive way;
- **Obstructive campaigns** – aims to collect information about discrimination and restrict discriminatory actions

Awareness campaigns' role is to educate people about what an expression of hate speech means, what forms it may take, or what the impact can be. They may also teach people how to protect themselves against the impact of hate speech, how to react, where to report it, or whom to engage in attempts of tackling hate speech[27]. One example of such an online campaign is the European Ins@fe, which aims to provide children, parents, teachers and youth workers with information and advice about how to stay safe in the digital environment, how to use online tools safely, and for this purpose develop different strategies. Another example is the campaign "Virtual racism, real consequences" led by a Brazilian organisation called Criola, where hateful Facebook messages were posted on the billboards near hate speakers' homes (without their name being released) with the purpose of demonstrating what hate speech is and encouraging people to speak up against racism.

The aim of the **affirmative campaign** is characterised by empowering vulnerable groups, such as ethnic minorities, LGBTQ, or religious groups by presenting them in a positive light[28]. One of the examples was a campaign called "Syndrome Roma" ("Syndróm Róm") which presented various successful Roma, such as economists, scientists, civil workers, and teachers.

Obstructive campaigns serve to collect information about the discriminatory practices or hate speech, mainly sites, violators or targeted groups[29]. The websites mapping and gathering information about the hate groups, such as extreme right ones—Hass-im netz.info (German) or Athenea Institute (Hungarian)—can be seen as examples of obstructive campaigns. Frequently, these sites not only gather information about different hate groups or sites, but also provide the Internet users with information on how to deal with it and protect themselves from the harm of hate speech.

3 Strategies and tools to reduce hate speech online

Besides legislative measures and education, a set of *soft* tools have been identified that might be utilised by a variety of actors from educational actors, individual Internet users, Internet service providers (ISPs), organised groups such as non-governmental and governmental organisations, and others, to mitigate hate speech online. Mitigation of hate speech online requires broader involvement and cooperation between these stakeholders. Importantly, the more such strategies are employed, the more thoroughly the cyber hate will be diminished. This part serves to encourage different stakeholders to take action and participate in alleviating hateful content.

The following tools might be utilised in social media and Internet discussion forums: counter-speech (fact-checking, personal experience, "imagine yourself" and, particularly, creating allies with hashtags as a complementary tool to counter-speech), flagging and reporting, naming and shaming, highlighting positive comments, and filtering software.

3.1 Counter-speech

The most basic strategy to reduce hate speech is to raise the voice and counter hate speech with more speech. In other words, any individual user can speak up and express his or her disagreement with hateful comments in an online discussion.

"Perhaps the most important is to simply provide irrefutable evidence to remind everyone that the world is full of people of good will – people who reject hatred and embrace the values of civility and respect."

Foxman & Wolf in *"Viral Hate: Containing its spread on the Internet"*, (2013, p. 132)

For a single user, it may also seem that counter-speech does not have an impact and they might have an impression that he or she alone will not make a difference in any significant way. Nevertheless, in practice, the counter-speech often serves a role of encouragement for other users who are motivated to contribute to the discussion when they see that there are also constructive comments in the discussion[30]. As a result, the number of counter-arguments may rise and diminish hate speakers who are in the minority.

Main principles when opposing hateful comments

Countering hate speech in online discussion requires certain rules so the whole process of opposing hate comments is likely to be efficient. Maja Nenadović in her paper[31] from 2013 proposes the complex strategy named PLEASE to effectively cope with other discussants in the online discussion forums. She advises to actively engage in discussions and to:

1. **Pause** and not take the discussion **personally in order to prevent attacks and conflicts;**
2. **Listen** carefully to arguments to competently understand them;
3. **Express empathy** even in cases of racist remarks;
4. **Analyse** comments and counter-arguments and the assumptions they are built on;
5. **Speak** and express opinions; and
6. Patiently **explain** own viewpoints as people often have different knowledge and life experiences.

Users may utilise different forms of counter-speech, namely:

- **Fact-checking - making factual claims**
- **Personal experience**
- **"Image yourself" or the strategy of storytelling**
- **Humour and sarcasm**

3.1.1 Fact-checking

When engaging in discussion with the aim to counter hate speech, the user may use different kinds of arguments. One way to stand against hate speech is **making factual arguments (fact-checking)**, which, in practice, means to refute certain claims, e.g., racial stereotypes with reliable data or scientific proof[32]. For example, in the case of Facebook discussants claiming that the Roma or other ethnic minority members receive extra or excessive amount of social allowances just because of their ethnic status, and thus are privileged over "majority" members, one can bring the data about the real amount of social allowances from the Central Labour Office and refute such false claims.

Making factual claims and fact-checking as a form of counter-speech may be initiated not only by common Internet users but also by trained professionals, such as Internet discussion moderators, journalists (in the case of news media websites) or by librarians who work with primary resources, which confront prejudices and misinformation very effectively and may be viewed as authorities, and thus their intervention might be respected.

3.1.2 Personal experience

Another strategy to argue with discussants online is to provide arguments based on **personal experience.** It means that the discussant does not use, for example, statistics or other kinds of data to refute the counter-arguments, but his or her own personal experience. For example, to counter the claim about the unwillingness of ethnic minority members or people with an immigrant background to work as a cause of high unemployment rate of such groups, the discussant may use his or her personal experience about seeing the Roma working in various construction sites all around Bratislava.. Such stories should come from a real-life experience so they are authentic and true.

3.1.3 "Imagine yourself" or the strategy of storytelling

Several scholars[33] argue that when the counter-arguments are presented in the narrative form, thus in the form of a story, they might be accepted by opponents more easily. Similarly, when examining when people tend to agree with an intervention in social media (to fight cyber bullying), the scholar van Laer in his study[34] from 2013 explored that when the facts are presented in the narrative format with the self-referencing element, people are prone to be convinced more easily. For example, using phrases like "imagine yourself" seems to be efficient to persuade people about certain facts.

3.1.4 Humour and sarcasm

Further forms of counter-speech are to make fun of comments containing hateful content in order to show the absurdity of hateful claims (either in the form of written comments or posting memes under comments). For example, one may address

false claims of the high criminality rate of members of ethnic group with comments on the high criminality of white men engaged in corruption cases. However, it is necessary to be aware that there is a thin line between sarcasm, personal attacks and the degradation of counter-discussants. Also, using this form of counter-hate speech does not have to lead to further deepening the misunderstanding between discussants and thus meaningful discussion. Therefore, this form of counter-speech should not cross the line between making fun of hateful comments and attacking individual discussants.

3.2 Searching for allies with hash tags

Countering hate speech might be exhaustive and very stressful since the user may experience personal attacks such as vulgarisms and insults, not only as a member of the primary targeted group (e.g., ethnic minority member) but also as an ally who wants to present constructive comments in the discussion. Moreover, counter-speech might be time-consuming due to searching for facts and data (for example, when refuting myths and prejudices). Therefore, it is desirable to have multiple discussants intervene at the same time to support and encourage each other. One of the tools to gather such allies in one discussion is through hashtags which enables users to identify places. In 2017, the initiative #somtu[35], aspires to bring together such allies in the Facebook discussion filled with hateful comments. Such a hashtag practically creates a community of people so they can intervene against the discussants that are often in majority, burden online discussions with hateful comments, false claims, prejudices or stereotypes and, as a result, discourage targeted groups and/or critically thinking users from engaging in the discussion.

3.3 Flagging and reporting

The strategy of flagging and reporting might also be considered as complementary to counter-speech as they act as a reaction to hateful comments by notifying discussion administrators about the violation of community standards or down-voting, and thus expressing the wrongfulness of the comments. Both strategies make the user community responsible for protecting its own members from harm of hate speech. **Reporting** comments in online discussions is used in case of violating community standards when a particular comment breaches community standards and guidelines. In this respect, community standards and guidelines should be accessible and clear so every user can easily get familiar with the definition of hate speech and which speech acts violate other people's rights and which do not.

Flagging serves to express an opinion by down-voting a comment, which may also result in hiding the comment in the online discussion. For example, certain online discussion platforms, mainly of newspapers (SME, Guardian, Independent, etc.), enable users to vote down comments that do not violate internal rules but that many groups may consider offensive. In addition, it empowers the user community to decide on this issue, define what the unacceptable behaviour and undesirable speech means, and when it should be banned. The same criteria may apply to the "like" button in the Facebook discussions.

3.4 Community "protest" action online

Counter-speech can be coordinated not only by a single user but may also take a form of a community or a group action. More specifically, for example, a Facebook user can create an event or fan-page that would challenge different stereotypes or hate groups, or even make fun of the opinions of particular hate groups and, at the same time, it would create a community of like-minded Facebook discussants. In Slovakia, there have been several recent examples of such initiatives, such as "Naše Slovensko"("Our Slovakia") which makes fun of the far right political party Ľudová strana Naše Slovensko (People´s Party Our Slovakia). It posts different pictures parodying the members and sympathisers of this political party.

3.5 Naming and shaming hateful websites

In the case of the already mentioned "cloaked websites" containing false historical, political or social facts, the efficient way to deal with them is to monitor and identify them, so the public (basic Internet users seeking data) are well-informed about their credibility. The Slovak project *konspiratori.sk* not only monitors and lists such "cloaked websites" for the sake of basic Internet users, but also

notifies commercial companies and discourages them from placing their online advertisements on such websites.

3.6 Highlighting positive comments

Very importantly, it is not only necessary to counter-argue, report or down-vote harmful content, but users can also **highlight good examples of positive comments**, (such as comments refuting racial prejudices)[36]. In the Facebook discussions, a "like" button may serve as such strategy because users, by "liking" positive materials (comments, posts, videos, pictures, etc.), support other users who are authors of certain positive material or constructive opinions, and express their agreement with the content. Highlighting positive online material might be an option when the user does not want to actively engage in the online discussion. This can be because for several different reasons, such as a fear of personal attacks from other discussants or a lack of time to write counter-arguments.

Internet Service Providers[37] as crucial actors in alleviating hate speech online

Internet Service Providers (ISPs) may play an important role in tackling hate speech online, as they may set up internal rules and policies for users that would ban certain content, such as racially or ethnically offensive content. The most common examples of such internal policies are **Terms of Service Agreements and Community Guidelines** that allow ISPs to remove hateful content by defining hate speech and other unacceptable behaviour online. Therefore, ISPs should be encouraged to use their own internal policies to determine and tackle hateful speech.

Furthermore, internal policies also play an educational role. Community standards stand not only to enable ISPs to ban and remove hateful content, but also to **educate users about unacceptable online speech** and inform them about their responsibility to protect other users from harm.[38]

Moreover, community standards are beneficial for ISPs too, since if online hate speech and cybercrime are not sufficiently regulated, they may discourage users from participating in the online public discussions or any other activities, as they do not want to be exposed to prejudices and hate but to discuss different topics at the certain quality level. There is also a struggle regarding the quality of comments, as constructive comments providing new insights and perspectives attract users, whereas hateful comments do the opposite[39].

Complementary to internal policies are the previously mentioned **reporting systems** (such as flagging, report buttons, complaint forms, and hotlines) that enable users to report violent content so ISPs can take an action in terms of removing any abusive content, warning violators that they have violated internal rules, or even blocking them in cases of severe or repetitive violations[40].

3.7 Filtering software

Other technological tools which may stand as a strategy to eliminate hateful content online (on web pages, not in social media) are filtering software that can block hateful material. One of the oldest well-known filtering software is Hatefilter released by the Anti-Defamation League launched in the US in 1998. This software not only sorts out and blocks hate sites, but also provides its users with information and education on hate and bigotry. The filtering tool is used mainly by parents and its main aim is to protect children from hateful material, examples of this filtering tool include "NetNanny", "SurfWact", "Parental Control" (in Slovak "Rodičovská Kontrola") or "Webwatcher". It is important to note that such software filters and blocks websites that have been reported by watchdog organisations which monitor hateful sites (e.g., Anti-Defamation League, Gay and Lesbian Alliance against Defamation).

Filtering tools can be categorised into **client-side** and **server-side.** Client-side filtering tools are initiated by users themselves (e.g., by parents, schools, workplaces), whereas the server-side ones are initiated and installed by the ISPs, which may, for example, block websites that have been labelled by watchdog organisations as containing hateful material.

Nonetheless, filtering software faces certain limitations, mainly in dealing with discussions on social media (such as Facebook and Twitter) where content is dynamic and constantly changing[41]. That is why Facebook and Twitter discussion platforms are highly criticised for failing to protect their users from hate content.

3.8 Summary

As hateful content is overwhelmingly invading the cyber world, a variety of actors including basic users, media representatives, and Internet Service Providers should intervene and participate in creating a safe online world. As presented above, a variety of tools might be used to achieve this goal. In the following chapters, we will present experimental testing of the effectiveness of two forms of counter-speech, namely fact-checking and personal experience, and the perspectives of the media representatives in the field of combating hate speech.

4 Counter-speech experiment on decreasing hate speech on Facebook

4.1 Overall purpose

In this experiment we, the authors of this publication, wanted to achieve two main aims. First, we wanted to identify which kinds of hate speech against the Roma appear the most in Facebook discussions. Second, we wanted to test how people react when somebody enters the Facebook discussions defending the Roma minority, and which strategies to counter the hate speech against the Roma appear to be the most effective.

4.2 Tested strategies

Although we were aware that there are more strategies for how hate speech can be countered (see chapter 3), to make this experiment more feasible and focused, we decided to test merely two main strategies. In the first strategy, we argued by using a *personal experience* with particular Roma people or a Roma person, which undermined the prejudicial claim of another discussant. In the second strategy, we argued by using *fact-checking* or research evidence, qualitative or quantitative findings from various well-respected researchers, research institutions or administrative data (such as from the Central Labour Office), which disprove the prejudicial claim.

Example of personal experience strategy:

"Everybody is kind of upset by the idea that he or she feeds somebody who is lazy to work. However, I want to point out that from my experience it is not always the case how it is presented here. I have friends and acquaintances of a Roma origin. One of these received a message from an employment agency that they do not accept the Roma. I just want to say by this that it is quite feasible that many Roma want to work or wanted to work but they have not been given even a chance to get employed."

Example of fact-checking strategy:

"I don't understand how you can believe that the Roma receive higher social benefits than other people who are not Roma. No so-called "Roma benefits" exist. Not even meals in schools are provided exclusively for Roma children. All children receive the meals if they come from socially disadvantaged families. You can easily google all this information through upsvar.sk [the official website of the Central Labour Office]. By the way, the level of social benefits for an individual is 61.60 Euro per month, for an individual with 1-4 children is 117.20 Euro, for a couple with 1-4 children is 160.40 Euro and a couple with more than 4 children is 216.10 Euro and it does not matter if the family has 5 or 10 children. Besides that, an individual or a family may receive a housing benefit, but many Roma are not recipients of these, since for that they would need to officially own a flat or house or have a lease, which they do not in the segregated Roma settlements. Link http://www.upsvar.sk/socialne-veci-a-rodina/hmotna-nudza/pomoc-v-hmotnej-nudzi.html?page_id=363675"

In this experiment, using the above-mentioned strategies, we were testing two hypotheses:

(1) Entering into the Facebook discussions with pro-Roma arguments will increase the overall proportion of the pro-Roma comments (while not including our comments when calculating the overall proportion). In other words, defending the Roma minority will motivate other people with pro-Roma attitudes to enter the discussions as well.

(2) The strategy of personal experience will be more efficient in eliminating hate speech against the Roma, as opposed to the fact-checking strategy.

To test these two hypotheses, we decided to analyse a total of 60 discussions, each under one Facebook post related to the Roma topic (for example, the news media article about the high unemployment of the Roma posted on Facebook). All of these Facebook posts were in the period between April 2016 and January 2017. In 30 of these discussions, we entered with the first, second, or both strategies. We call this set of 30 discussions an "intervention group". We did not enter the remaining 30 Facebook discussions related to the Roma topic, and considered these as a "control group". This way we could compare the discussions in the intervention group and the control group to test the first hypothesis. In order to test the second hypothesis, we compared the individual comments of ours only within the intervention group. In 10 of these discussions, we exclusively used the personal experience strategy. In 10 discussions, we exclusively used the fact-checking strategy. And in the remaining 10 discussions, we used both of these strategies.

In an attempt to test the second hypothesis, whether one strategy is more effective than the other, we needed to first define what to consider as "effective" or "successful" intervention. Obviously, as a failed attempt to eliminate hate speech, we considered a situation when the discussant (the author of the original comment or some other discussant within the same thread of the discussion) reacted to our comment with another hateful or prejudicial comment. As a successful or effective intervention, we considered the following situations:

- the discussant fully acknowledged our argument defending the Roma;
- the discussant partly acknowledged our argument defending the Roma;
- no discussant within the particular discussion thread reacted to our comment, thus, the hate speech stopped.

This last situation is the most problematic to consider as a success since we do not know the reasons the discussants ceased to continue in the discussion. It could be because they were persuaded by our arguments, but it could also be because they did not notice our comment or just did not feel like continuing although not changing their hateful attitude towards the Roma at all. Nonetheless, the overall aim of eliminating hate speech in the Facebook discussion space was fulfilled, even in this situation, since the hate did not continue. Hence, we did consider this situation as a success as well.

4.3 Platform for experiment

A team of four researchers participated in this experiment. As a research team, we constantly coordinated our procedures. Initially, we tried to intervene through our authentic Facebook profiles, however, in order to protect ourselves from personal attacks of other discussants, we decided to create one fake profile which we all used to intervene instead.

We intervened in discussions under Facebook posts which were somehow related to the topic of the Roma minority in Slovakia. These posts were published either by a particular Member of the Parliament of the Slovak Republic or by various Slovak news media Facebook pages, such as SME, aktuality.sk, RTVS, Topky, since these were the most popular news Internet profiles in Slovakia.[42] Neither politicians nor news media post very often about the Roma. That is why we practically entered all Facebook discussions which we noticed during the research period and were available to intervene. We were not selective in any way about these posts. The control group discussions consisted of all posts which we missed in that research period (because the authors were involved in other research projects or took vacations and were not always available to intervene). Table 1 illustrated the sources of all Facebook posts, which we included in the intervention or control group discussions. The various online news media sources slightly outnumbered the Facebook discussions initiated by the politicians.

The Facebook space is a very lively and constantly changing environment which brought about a number of challenges to this research, and challenged the possibility of drawing any generalised research conclusions. Here are the most salient challenges which contest the validity of this research:

- the sources of Facebook posts (Table 1) were very different in the composition of their followers

(socio-economic background, educational attainment, electoral preferences, etc.);
- the content of the posts significantly varied, some posts already contained a hate speech, some used terms with particular emotional connotations e.g., "the Roma *gang* attacked the tourist", and others used relatively neutral terms and content e.g., about the policies of housing;
- the length of discussions and number of comments within the studied discussions varied significantly (the shortest discussion had 35 comments while the longest had 523 comments), as well as the visibility of these discussions through being shared and liked on Facebook.

4.4 Results

Intervention increases the proportion of pro-Roma comments (Hypothesis 1)

In 60 examined Facebook discussions, there were 7,586 comments (4,027 in the intervention group and 3,559 in the control group) which we analysed. Each comment we coded either as (1) pro-Roma comment; (2) anti-Roma comment; (3) mixed comment (containing both pro-Roma and anti-Roma opinions); or (4) irrelevant comment (not related to the topic of Roma minority).

As evidenced in Table 2, the irrelevant comments were represented the most in both the intervention and control group. In the intervention group, we contributed with 362 pro-Roma comments out of the total of 4,027 comments in this group. Although the overall proportion of pro-Roma comments was rather low in both the intervention and control group, while excluding our comments from the calculation, the representation of pro-Roma comments was higher in the intervention group (7.3%) in comparison to the control group (4.5%). This difference is statistically significant at the probability level of 0.0117. From this result, we may infer that our first hypothesis was confirmed. Hence, we may state that contributing to the Facebook discussions with pro-Roma arguments motivates other people to enter the discussions and defend the Roma as well. In other

Table 1: Sources of studied Facebook discussions

Source	Control group (no. of posts)	Intervention group (no. of posts)
Aktuality (online news media)	9	15
Topky (online news media)	8	2
Boris Kollár (MP for We Are Family – Boris Kollár)	5	5
Milan Krajniak (MP for We Are Family – Boris Kollár)	3	2
Lucia Nicholsonová (MP for Freedom and Solidarity)	2	0
Alojz Hlina (former MP, Chairman of Christian Democratic Movement)	2	0
Milan Uhrík (MP for Kotleba - People's Party Our Slovakia)	1	0
SME (print and online news media)	0	3
RTVS (TV, radio and online media)	0	2
Lucia Žitňanská (Minister of Justice of the Slovak Republic, Most-Híd)	0	1
Total	**30**	**30**

words, it is worthwhile defending the discriminated social group in the online platforms as it not only increases the overall visibility of the inclusive attitudes but also motivates others to contribute to this visibility.

No difference in effectiveness between two tested strategies (Hypothesis 2)

In 30 discussions within the intervention group, we used either the strategy of personal experience or fact-checking in the individual comments. The total number of fact-checking comments was significantly higher than the personal experience comments (195 for fact-checking and 79 for personal experience). The researchers were more skilled in using facts and evidence than in creating fake personal experiences if not having a real one related to the discussed topics. This significant imbalance in the number of comments representing one or the other strategy introduces limitations to generalising the research findings.

By comparing the comments that used the personal experience strategy and those using the fact-checking strategy, we discovered that there was no statistically significant difference in terms of their effectiveness. The personal experience comments stopped the hate in 33.33% of cases, while the fact-checking comments stopped them in 36.27% of cases. The personal experience comments made the discussants partially acknowledge their arguments in 2.56% of cases, while in the case of the fact-checking comments, it was 4.15% of cases. None of the discussants ever fully acknowledged our pro-Roma arguments. From all these results, while being aware of the methodological limitations of this research, we may infer that both the personal experience and the fact-checking strategies have similar potential to be effective in eliminating the online hate speech, thus, neither of them proved to be more effective than the other.

Discussions overwhelmed by anti-Roma prejudices

In 60 Facebook discussions, which we included in this research, we monitored the frequency of various types of anti-Roma comments. These comments could be divided into three groups:

1) What are the Roma like and how do they behave?
2) Which policies and practices towards the Roma are in place at present?
3) Which policies and practices towards the Roma should be in place in the future?

Figure 1 shows the various types of anti-Roma comments and which addresses each one of these questions. As obvious from this list of types, almost no discussants used any research evidence or facts to support their arguments. This was the case for both the anti-Roma as well as the pro-Roma comments. The discussants either used their personal experience to support their arguments or they merely used general claims presenting

Table 2: Proportion of pro-Roma and anti-Roma comments

Intervention group					
our comments	pro-Roma (without ours)	anti-Roma	mixed	irrelevant	**total**
362	293	1203	78	2091	**4027**
9 %	7,3 %	29,9 %	1,95 %	51,9 %	**100%**

Control group				
pro-Roma	anti-Roma	mixed	irrelevant	**total**
160	1581	99	1719	**3559**
4,5%	44,4 %	2,8 %	48,3 %	**100%**

these as well-known and taken-for-granted truths. The trouble was that most of these "truths" were merely "myths".

What are the Roma like and how they behave?

Figure 1 demonstrates that there are several characteristics and behaviours attached to the Roma:

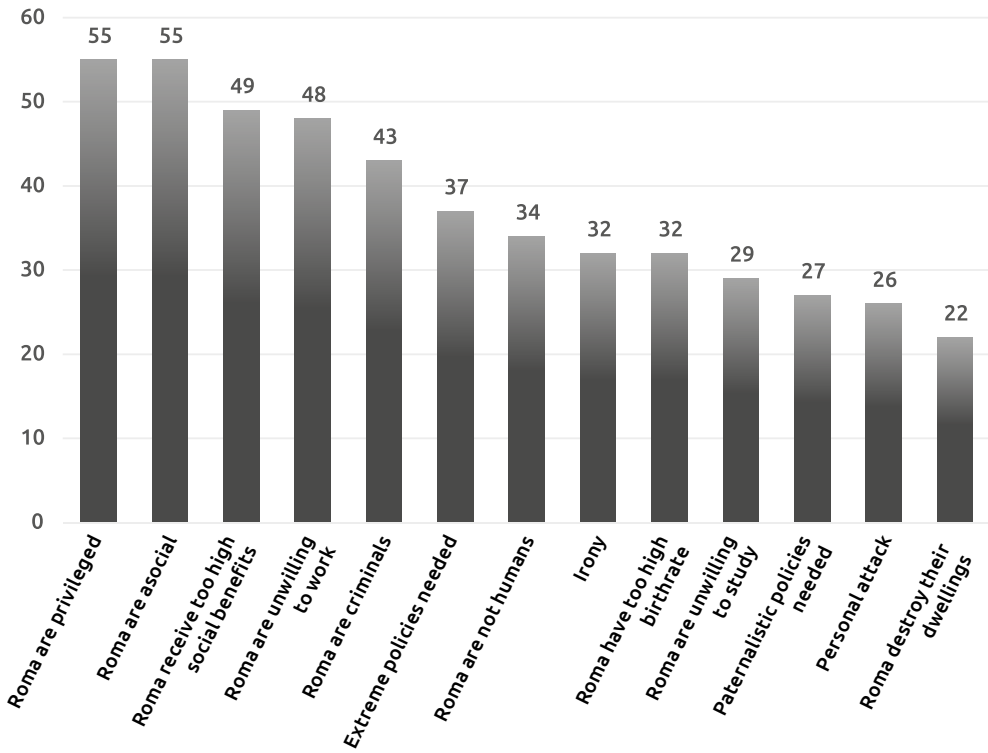
- asocial (mentioned 55 out of 60 discussions);
- unwilling to work (mentioned 48 out of 60 discussions);
- criminals (mentioned 43 out of 60 discussions);
- likened to animals (e.g., rats, pigs), insects (e.g., cockroaches, parasites) or things (mentioned 34 out of 60 discussions);
- having too many children (mentioned 32 out of 60 discussions);
- unwilling to study (mentioned 29 out of 60 discussions);
- destroying their dwellings (mentioned 22 out of 60 discussions).

The discussants did not acknowledge that there is a vast majority of integrated Roma in society, but constantly framed these as "exceptions". They also often presented the Roma as unchangeably "unadjustable" to the majority, and as asocial criminals. That it is in their DNA, they were born with it. Sometimes even the Roma-identified discussants presented some hateful attitudes towards other Roma.

Which policies and practices towards the Roma are in place at present?

Figure 1 shows that the discussants are overwhelmingly convinced that the Roma are a privileged group, meaning that they have the better access to public services (such as tuition-free kindergartens, free lunches in the school, new apartments for free, free medicine), and receive higher amount of social benefits or extra social allowances. This happened in 55 Facebook discussions out of 60. In 49 cases, the discussants used false information about the level of social benefits.

Figure 1: Types of anti-Roma comments



They significantly overstated this level, in some cases it was five-times higher than the actual level.

Discussants usually denied the existence of discrimination against the Roma and when they did admit it, they defended it as justifiable since the people who discriminated the Roma probably acted on the basis of their previous negative experiences with them. They attached "collective guilt" to all Roma.

The discussants often ignored the most basic sociological knowledge that social phenomena are very complex and contextual matters, which are influenced by a myriad of psychological, social, cultural, and political factors. No one single policy measure, even if at the state level, can solve and radically change such a complex phenomenon as the segregation and discrimination of the Roma. This phenomenon can only be changed through a long-term, gradual process requiring dozens of policies and actions at all levels of society (individual, communal, regional, state, international) and spheres of life (employment, education, housing, health, culture, etc.). For failing to see the overall situation of the Roma as this complex, and for not understanding that one policy or particular programs or activities cannot bring about a quick and radical social change, the discussants also tended to blame the politicians and NGOs for wasting the resources spent on Roma inclusion.

Which policies and practices towards the Roma should be in place in the future?

Seeing the social phenomena over-simplistically led some discussants to propose extremist policies to exterminate or expel the Roma. This occurred in 37 Facebook discussions out of 60. The over-simplistic perspective also led many discussants to position themselves in a superior and paternalistic position as knowing what is the best for the Roma and to the conviction that punishing them more strictly will solve the problem. This occurred in 27 discussions out of 60.

While some discussants presented a racist opinion, but distanced themselves from being called a racist, a group of discussants started to emerge who proudly presented themselves as racists. When the discussants on both sides ran out of arguments, they often slipped into vulgar verbal attacks (in 26 discussions out of 60) or irony (in 32 discussions out of 60). The sarcastic or ironic comments seemed to stir the emotions of discussants on both sides and did not seem to be effective at all.

Reactions to our interventions supporting the Roma

Many discussants reacted to mentioned research evidence with aversion and questioned the research as something that did not reflect the lived experience of people living in proximity to Roma settlements. They also often questioned the researchers' credibility as not having the direct experience with the Roma.

Using positive examples of personal experiences with the Roma was not persuasive for many discussants either. They did acknowledge that there are many Roma, who do not represent the negative stereotypical image of a Roma, but framed these merely as "exceptions".

In several cases, when the discussants partially acknowledged our arguments undermining one prejudice in a particular area, for instance, the level of social benefits or unwillingness to work and study, the very same discussants mentioned several other prejudices as a follow-up. While using a metaphor of an onion, although in some cases we managed to partially to refute one particular prejudice and remove one layer of an onion, the discussant did not change her/his overall attitude towards the Roma and presented several other prejudices, so the onion remained to exist. In order to change the overall attitude of a person, we would need to continue discussing and endeavour to remove all the remaining layers of the *onion*.

When the discussants proposed some extremist and various narrow-minded policies, such as cancelling all social benefits, and when we thoroughly explained what impacts such a policy would have on the Roma population as well as the entire population in the country, the discussants often did not continue in the discussion. There is a chance to interpret this phenomenon as suggesting they were at least partially persuaded by our arguments and explanations.

4.5 Summary

It is worthwhile entering the Facebook discussions to present the pro-Roma attitudes, regardless of using the fact-checking or personal experience strategy because it motivates other pro-Roma discussants to express their opinion. Discussants with anti-Roma attitudes are mostly convinced that the Roma are privileged and that they are inherently asocial. They see overall situation and impact of policies over-simplistically as a straightforward issue.

5 Regulation of online discussions and elimination of hate speech by the Slovak media

5.1 Interviewed participants of the study

Since cyber hate speech is a relatively new phenomenon with severe consequences on the targeted groups, it is important to explore what role the media should play regarding reducing hateful comments in the online discussions they administer. On this note, during February 2017, five semi-structured interviews with news media representatives responsible for regulation of online discussions of SME, Aktuality.sk, Trend, Denník N and Pravda were conducted, and one questionnaire with open questions was filled out by Topky with the aim to examine what media do in terms of elimination of hate speech on their websites and what strategies they use to regulate online discussions (including Facebook discussions). The sample of the news media consists of the most popular news media in Slovakia[43].

5.2 Results

The main findings show that the news media monitor hate speech in online discussion and, in this respect, use different tools and strategies. Posts are assessed regularly by employees of particular media who are responsible for regulation of online discussions, most commonly by editors, chief-editors, editors of social networks, or other administrators. While regulation is exclusively their task, it is not the only task they deal with. Therefore, considering the workload of other tasks, the time spent on regulation of discussions and elimination of hate speech online is limited. Even though regulating discussions requires more working hours, it does not allow regulators to engage in this activity on a full-time basis. The time dedicated to it is on average approximately one hour a day.

When assessing the content of discussion comments, editors of social networks most often rely on written guidelines that are usually accessible on their media websites which serve primarily as rules for discussants for appropriate participation in online discussion. If guidelines are missing, making decisions over the content is dependent on the subjective discretion of the regulator. In addition to this, five out of six interviewed media own such guidelines for discussion. Yet, none of the respondents has developed a separate formal document serving the regulators of discussions as instructions. Regulators monitor a whole discussion including assessments of posts reported by other discussants, with the exception of SME, in which a regulator focuses merely on posts reported by other discussants.

In terms of regulating online discussions, several strategies are used to different extents by the media to eliminate hateful content. The media also use some strategies that can be utilised by discussants so they can actively contribute to the elimination of hate speech on social media. Here are all mentioned strategies used by researched Slovak news media:

1) **Removal of posts**
   is a strategy used by all interviewed media in countering online hate speech.
2) **Closing discussions**
   is used if news articles carry a sensitive topic such as those about the Roma, immigrants, etc., and/or ad hoc in the case of a discussion that is infected by hate speech. Closing discussions can be used as a default setting for all posted news articles, too. For instance, Denník N opens discussions only under a few of its posts per day. That may prevent hateful content.

3) **Blocking access of particular contributors to discussions**
   is used by media such as Aktuality.sk, in the case where a discussant breaks the rules of discussions. Banning can be short or long-term.
4) **Tracking the history of discussants**
   serves as an instrument for checking on the history of commenting of particular discussants. History of comments is mostly checked by a regulator if the rules of discussion were broken. Regulators can check if and how frequently a discussant previously violated the rules. Findings can be considered when deciding over a ban.
5) **Communication with discussants outside of discussion**
   is not used on a regular basis in most of the media. For instance, the regulators can provide discussants with an explanation related to a ban or content removal, and can notify them about inappropriate content they posted.
6) **Reporting posts**
   is a strategy used by the discussants. Readers can report posts with offensive content through a reporting system located on the webpage of discussions, for example Topky.sk. Reported posts are then assessed by a regulator and are removed if considered breaking the rules of discussion.
7) **Rating comments**
   is, for instance, available on the website of SME and is used by discussants. Readers, through ratings, express their opinions about comments posted by other discussants. Rating systems can have a form of plus and minus signs (where plus means good and minus the opposite) available above or below a comment posted by a discussant.
8) **Hiding posts on the wall**
   is used by regulators of some media such as Trend for posts that are rated negatively. As a result, hidden posts must be clicked through to be available for reading.
9) **Blacklisting**
   is a strategy used by most of the interviewed media, e.g. by Pravda. A blacklist contains forbidden words recorded into a special system and can be updated at any time. The system keeps all posts containing words on the blacklist inaccessible for other discussant.
10) **Moderation of discussions**
    is a rarely implemented strategy in practice, meaning leading discussions by a responsible media employee. Although all interviewees are aware of the importance of the strategy, moderation is used occasionally only by Denník N. Its journalists try to be present in discussions they are authors of. This happens sporadically, though.
11) **Fact-checking**
    is a particular form of moderation of discussions which is also rarely used. Fact-checking is usually used for defending the content (for example, correctness of the data in the article) of the journalist's posts/news articles or when a mistake is found. Fact-checking is used only by SME and Trend.

5.3 Barriers the media experience

Even though the media feel responsible for the content in online discussions (both on their websites and the Facebook platform they utilise) and are aware of a need for regulating discussions more effectively, they do not fully succeed in it. The media faces a set of barriers in doing more systematic regulation of online discussions. One of the most common obstacles seems to be a lack of human resources for dealing with online discussion regulation. Regulation of discussions is time-consuming and regulators are limited in this respect. Additionally, the hiring of more staff or/and creating positions made exclusively for regulating the discussions requires a high financial input by the media, which is seen problematic especially in the case of small media.

5.4 Recommendations to improve the regulation

In terms of improving the regulation of online discussions and the effective elimination of hate speech online, interviewed media recommended a number of measures. First, they proposed increasing the number of employees. Second, they suggested improving law enforcement, particularly to introduce more thorough investigation resulting in sanctions, if relevant. Third, the introduction of mandatory fees for posting a certain number of comments/posts in online discussions was

mentioned as another possible solution to limit the access and the number of comments in discussion for easier regulation. Fourth, unification of rules of discussions amongst media would help the media to unify in their regulation process and their approach to hate speech on the Internet. Last but not least, the media representatives pointed out a need for an improvement in regulating online discussions by Facebook itself. Facebook contains a lot of hateful posts, many of which remain online even after they were reported. Usually, posts and comments are removed from Facebook based on an assessment of the extent of likelihood that the threat contained in the post/comment would lead to physical attack or abuse. In addition, an assessment done by a non-native speaker of the particular language the post is written in may not be fully understood by a regulator. Such policies of the Facebook complicate its effectiveness in combating hate speech on the Internet.

5.5 Summary

The Slovak news media use most available strategies to regulate online discussions, including the Facebook discussions. Nonetheless, it seems that the elimination of hate speech in the online discussions is not their financial and personnel priority. The regulators of the discussions spend only very limited time on this task and are not adequately trained to do it effectively.

References

- Banks, J. (2010). *„Regulating hate speech online“*. International Review of Law, Computers & Technology, 24(3), 233-239.
- Boeckmann, R. J., & Turpin-Petrosino, C. (2002). *Understanding the harm of hate crime*. Journal of Social Issues, 58(2), 207-225.
- Citron, D. K. & Norton H. (2011). *„Intermediaries and hate speech: Fostering digital citizenship for our information age“*. Boston University Law Review, 91(4), 1435 – 1484.
- Cohen-Almagor, R. (2011). *„Fighting hate and bigotry on the Internet“*. In: *Policy & Internet*, Volume 3, Issue 3
- Delgado, R., & Stefancic, J. (2004). *Understanding words that wound*. Boulder: Westview Press.
- Diakopoulos, N. & Naaman, M. (2011). *„Towards quality discourse in online news comments“*, In: *Proceedings of the ACM 2011 conference on Computer supported cooperative work*, 2011
- Durrheim, K., Greener, R. and Whitehead, K.A. (2015). *„Race trouble: Attending to race and racism in online interaction“*. British Journal of Social Psychology, 54, 84-99.
- Foxman, A.H & Wolf, Ch. (2013) *Viral hate. Containing its spread on the Internet*. New York: Palgrave Macmillan.
- Hrdina, M. (2016). *Identity, activism and hatred: Hate speech against migrants on facebook in the Czech republic in 2015*. Naše společnost, 14 (1), 38-47
- Leets, L. (2002). *Experiencing hate speech: Perceptions and responses to anti-semitism and antigay speec*h. Journal of Social Issues, 58(2), 341-361.
- Leets, L., & Giles, H. (1997). *Words as weapons - When do they wound? Investigations of harmful speech*. Human Communication Research, 24(2), 260-301.
- Liao, Q. Vera, Wai-Tat Fu (2013). *Beyond the filter bubble: interactive effects of perceived threat and topic involvement on selective exposure to information. Proceedings of the SIGCHI Conferehence on Human Factors in Computing Systems.* Available at: https://dl.acm.org/purchase.cfm?id=2481326&CFID=953875486&CFTOKEN=97626206
- Maitra, I. (2012). *Subordinating speech*. In I. Maitra & M. K. McGowan (Eds.), *Speech and harm: Controversies over free speech* (s. 94-120). Oxford: Oxford University Press.
- McGonagle, T. A (2012). *„Survey and critical analysis of Council of Europe strategies for countering „hate speech““*. In: Herz, M. & Molnar, P. (2012). The content and context of hate speech. Cambridge University Press.
- Mosher, D., & Proenza, L. (1968). *Intensity of attack, displacement and verbal aggression*. Psychonomic Science, 12, 359-360.

- Molnar, P. (2012) *Responding to „hate speech with art education and the imminent danger test"* In: Herz, M. & Molnar P. (2012). *The content and context of hate speech: Rethinking regulation and responses*. New York: Cambridge University Press.

- Nenadović, M. (2013). *Applied debate: A weapon of fighting discrimination and building understandin*g. 4th International conference on argumentation, rhetoric, debate, and the pedagogy of empowerment. Available at: http://d2ivco2mxiw5i2.cloudfront.net/app/media/5207

- Nielsen, L. B. (2002). *Subtle, pervasive, harmful: Racist and sexist remarks in public as hate speech*. Journal of Social Issues, 58(2), 265-280.

- Simpson, R. M. (2013). *Dignity, harm, and hate speech*. Law and Philosophy, 32, 701-728.

- Sorial, S. (2015). *Hate speech and distorted communication: Rethinking the limits of incitement*. Law and Philosophy, 34, 299-324.

- Stevens, T. & Neumann, P. R. (2009). *Countering online radicalisation: A strategy for action*. London: The International Centre for the Study of Radicalisation and Political Violence and the Community Security Trust.

- Timmermann, W. K. (2005). *The relationship between hate propaganda and incitement to genocide: A new trend in international law towards criminalization of hate propaganda?* Leiden Journal of International Law, 18, 257-282.

- Titley, G., Keen, E. & Foldi, L. (2012). *Starting points for combating hate speech online: Three studies about online hate speech and ways to address it*. Council of Europe, British Institute of Human Rights.

- van Laer, T. 2013. *„The means to justify the end: Combating cyber harassment in social media"*, Journal of Business Ethics, Volume 123, Issue 1, 85 – 98.

- Velšic, M. (2016). *Mladí ľudia v kyberpriestore: Šance a riziká pre demokraciu [Young people in cyberspace – chances and risks for democracy]*. Bratislava: Inštitút pre verejné otázky

- West, C. (2012). *Words that silence? Freedom of expression and racist hate speech*. In I. Maitra & M. K. McGowan (Eds.), *Speech and harm: Controversies over free speech* (222-248). Oxford: Oxford University Press

Notes

1 See Simpson, R. M., *Dignity, harm, and hate speech*, Law and Philosophy 32, p. 701 (2013)

2 Ibid. p. 702

3 See Banks, J., *Regulating hate speech online*, International Review of Law, Computers & Technology, 24(3), p. 2 (2010)

4 See Simpson, R. M., *Dignity, harm, and hate speech*, Law and Philosophy 32, p. 701 (2013)

5 See Foxman, A.H & Wolf, Ch., *Viral hate. Containing its spread on the Internet*. New York: Palgrave Macmillan, p. 16 (2013)

6 See Sorial, S., *Hate speech and distorted communication: Rethinking the limits of incitement*, Law and Philosophy 34 (2015)

7 See Leets, L., & Giles, H., *Words as weapons - When do they wound? Investigations of harmful spee*ch, Human Communication Research, 24(2) (1997)

8 See Leets, L., *Experiencing hate speech: Perceptions and responses to anti-semitism and antigay speech*, Journal of Social Issues 58(2) (2002)

9 See for example Hrdina, M., *Identity, activism and hatred: Hate speech against migrants on Facebook in the Czech Republic in 2015*, Naše společnost, 14 (1), p. 39 (2016) or Foxman, A.H & Wolf, Ch., *Viral hate. Containing its spread on the Internet*. New York: Palgrave Macmillan (2013)

10 See Hrdina, M., *Identity, activism and hatred: Hate speech against migrants on Facebook in the Czech Republic in 2015*, Naše společnost, 14 (1) (2016) or Liao, Q. Vera, Wai-Tat Fu, *Beyond the filter bubble: interactive effects of perceived threat and topic involvement on selective exposure to information* (2013), available at: https://dl.acm.org/purchase.cfm?id=2481326&CFID=953875486&CFTOKEN=97626206

11 See Council of Europe, *No hate speech movement survey on cyber hate speech* (2016), available at: http://www.nohatespeechmovement.org/survey

12 See Velšic, M., Mladí ľudia v kyberpriestore: *Šance a riziká pre demokraciu [Young people in cyber-space – chances and risks for democracy]*. Bratislava: Inštitút pre verejné otázky (2016)

13 See Nielsen, L. B., *Subtle, pervasive, harmful: Racist and sexist remarks in public as hate speech*, Journal of Social Issues 58(2) (2002)

14 See Delgado, R., & Stefancic, J., *Understanding words that wound*, Boulder: Westview Press (2004)

15 See for example Delgado, R., & Stefancic, J., *Understanding words that wound*, Boulder: Westview Press (2004); Boeckmann, R. J., & Turpin-Petrosino, C., *Understanding the harm of hate crime*, Journal of Social Issues 58(2) (2002) or Maitra, I., *Subordinating speech*, In I. Maitra & M. K. McGowan (Eds.), *Speech and harm: Controversies over free speech*, Oxford: Oxford University Press (2012)

16 See Boeckmann, R. J., & Turpin-Petrosino, C., *Understanding the harm of hate crime*, Journal of Social Issues 58(2), p. 222 (2002)

17 Ibid.

18 See Simpson, R. M., *Dignity, harm, and hate speech*, Law and Philosophy 32, p. 718 (2013) or Sorial, S. *Hate speech and distorted communication: Rethinking the limits of incitement*, Law and Philosophy 34, p. 306 (2015)

19 See West, C. (2012). *Words that silence? Freedom of expression and racist hate speech*. In I. Maitra & M. K. McGowan (Eds.), *Speech and harm: Controversies over free speech* (pp. 222-248). Oxford: Oxford University Press (2012)

20 See Citron, D. K. & Norton H., *Intermediaries and hate speech: Fostering digital citizenship for our information age*, Boston University Law Review 91(4) (2011)

21 See Foxman, A.H & Wolf, Ch., *Viral hate. Containing its spread on the Internet*, New York: Palgrave Macmillan (2013)

22 See Durrheim, K., Greener, R. and Whitehead, K.A., *Race trouble: Attending to race and racism in online interaction*,British Journal of Social Psychology 54 (2015).; Molnar, P., *Responding to hate speech with art education and the imminent danger test* In: Herz, M. & Molnar P., *The content and context of hate speech: Rethinking regulation and responses*, New York: Cambridge University Press (2012) or Foxman, A.H & Wolf, Ch., *Viral hate. Containing its spread on the Internet*, New York: Palgrave Macmillan (2013)

23 See Foxman, A.H & Wolf, Ch., *Viral hate. Containing its spread on the Interne*t, New York: Palgrave Macmillan (2013) or Titley, G., Keen, E. & Foldi, L., *Starting points for combating hate speech online: Three studies about online hate speech and ways to address it*, Council of Europe, British Institute of Human Rights (2012)

24 See Titley, G., Keen, E. & Foldi, L., *Starting points for combating hate speech online: Three studies about online hate speech and ways to address it*, Council of Europe, British Institute of Human Rights (2012)

25 See Foxman, A.H & Wolf, Ch., *Viral hate. Containing its spread on the Internet*, New York: Palgrave Macmillan (2013) or Titley, G., Keen, E. & Foldi, L., *Starting points for combating hate speech online: Three studies about online hate speech and ways to address it*, Council of Europe, British Institute of Human Rights (2012)

26 See Titley, G., Keen, E. & Foldi, L., *Starting points for combating hate speech online: Three studies about online hate speech and ways to address it*, Council of Europe, British Institute of Human Rights, p. 63(2012),

27 Ibid, p. 64

28 Ibid, p. 69

29 Ibid. p. 74

30 See Foxman, A.H & Wolf, Ch., *Viral hate. Containing its spread on the Internet*. New York: Palgrave Macmillan (2013)

31 See Nenadović, M., *Applied debate: A weapon of fighting discrimination and building understanding*. 4th International conference on argumentation, rhetoric, debate, and the pedagogy of empowerment (2013) available at: http://d2ivco2mxiw5i2.cloudfront.net/app/media/5207

32 See Foxman, A.H & Wolf, Ch., *Viral hate. Containing its spread on the Internet*. New York: Palgrave Macmillan (2013)

33 See for example Leets, L., *Experiencing hate speech: Perceptions and responses to anti-semitism and antigay speech*, Journal of Social Issues, 58(2) (2002) or van Laer, T. *The means to justify the end: Combating cyber harassment in social media*, Journal of Business Ethics, 123(1) (2013)

34 See van Laer, T. *The means to justify the end: Combating cyber harassment in social media*, Journal of Business Ethics, 123(1) (2013)

35 #somtu or „som tu" means „I am here" expressing presence and support when combating hate comments in the facebook discussions. The Slovak initiative #somtu is similar to the Swedish one #jagärhär (meaning „I am here" as well) having the same purpose.

36 See Foxman, A.H & Wolf, Ch., *Viral hate. Containing its spread on the Internet*, New York: Palgrave Macmillan, p. 140 (2013)

37 Internet Service Providers (ISPs) or Online Service Providers (OSPs) present actors, predominantly organisations, that provide services related to accessing and using various Internet services. The most common Internet services are electronic mail services, news articles and videos (such as newspapers Guardian, Independent), downloadable materials and programmes (games, movies), e-shopping, entertainment content such as videos and pictures (Youtube), but also social networking site including chatrooms (e.g., Facebook, Twitter) etc. They can be commercial, privately- or state-owned, non-profit etc.

38 See Citron, D. K. & Norton H., *Intermediaries and hate speech: Fostering digital citizenship for our information age*, Boston University Law Review, 91(4), p. 1441 (2011)

39 See Diakopoulos, N. & Naaman, M., *Towards quality discourse in online news comments*, In: Proceedings of the ACM 2011 conference on Computer supported cooperative work, (2011) or Citron, D. K. & Norton H., *Intermediaries and hate speech: Fostering digital citizenship for our information age*, Boston University Law Review, 91(4), p. 1441, (2011)

40 See Banks, J., *Regulating hate speech online*, International Review of Law, Computers & Technology, 24(3), p. 233-239 (2010); McGonagle, T. A *Survey and critical analysis of Council of Europe strategies for countering „hate speech"*, In: Herz, M. & Molnar, P. , *The content and context of hate speech*. Cambridge University Press (2012) Titley, G., Keen, E. & Foldi, L., *Starting points for combating hate speech online: Three studies about online hate speech and ways to address it*, Council of Europe, British Institute of Human Rights (2012)

41 See Cohen-Almagor, R., *Fighting hate and bigotry on the Internet*, Policy & Internet, 3(3) (2011) or Stevens, T. & Neumann, P.R., *Countering online radicalisation: A strategy for action*, London: The International Centre for the Study of Radicalisation and Political Violence and the Community Security Trust (2009).

42 The research sample was established based on findings of 3rd and 4th quartal 2015 survey of the most popular media conducted by MEDIAN SK available at:https://medialne.etrend.sk/internet-grafy-a-tabulky.htm

43 Ibid.

About authors

Lucia Kováčová

is a researcher at the Slovak Governance Institute in Bratislava. She earned her MA degree in Public Policy at the Central European University in Budapest with the specialisation in equality and social justice. Her research interests are labour integration, social economy and inclusive education of disadvantaged children and youth.

Jozef Miškolci

earned his doctoral degree in education at the University of Sydney in Australia in 2014. His main areas of research are inclusive education, comparative education, educational policy, and human rights in education. At present, he works as a researcher at the Faculty of Education of the Comenius University in Bratislava and partly in the Slovak Governance Institute.

Edita Rigová

is a junior researcher at the Slovak Governance Institute. She earned her MPA degree in Public Administration at the IDHEAP – Swiss Graduate School of Public Administration at the University of Lausanne in Switzerland. Edita completed an internship at the European Union Agency for Fundamental Rights in Vienna where she worked in the sector of Roma and Migrant Integration within the Equality and Citizens' Rights department. Her research interest is Roma inclusion and Roma-related inclusive policies.