HATE SPEECH, DISCRIMINATION, POLARIZING EVENTS. MANAGING PUBLIC REPORTING AND RESPONSIBLE COMMUNICATION IN CASE OF SECURITY THREATS[*]

Ileana-Cinziana SURDU[*]

Abstract

The internet may be seen, through the social networks, as a „public sphere", which invites to a democratic discourse. On the other hand, the internet can also be seen as a support of the echo chambers, which create an environment for reinforcing certain beliefs and discrimination, and for creating hate speech, which can lead to polarization.

The culture of communication is highly influenced by the impact of the social networks, determining an increasing pluralism and a certain level of unethical dissemination of information, in the absence of critical analysis. Sensitive aspects, like security threats, impose a special approach, so they would not have a very negative impact over the public. The generalization of these type of topics may lead to panic, fear, polarization, discrimination, and even violent attitude and behavior.

The professionals who can prevent or soften the negative reactions of the public are the first liners in the field of communication and journalism, like institutional spokespersons and journalists in the field of security and law enforcement. Thus, these communicators bear the responsibility of delivering accurate data and information, in an ethical manner. These are only two of the requirements in relation to their audience. Other skills, competences and knowledge are also a must, such as the ability to think critically, develop responsible reactions, or the literacy in negative phenomena and actions that may lead to violent behavior.

[*] Acknowledgement: This document was drafted with the financial support of the European Commission, as part of the CRESCEnt project ("Mind the gap in media CoveRagE and Strategic communication in CasE of security Threats – the development of critical thinking and responsible reaction" – CRESCEnt Project 2018-1-RO01-KA202-049449). The content of the present document is the exclusive responsibility of the author, and the National Agency and the European Commission are not responsible for the way that the information will be used.

[*]Researcher, "Mihai Viteazul" National Intelligence Academy, Bucharest, Romania, email: ileana.surdu@animv.ro; surdu.ileana@animv.eu

The present study is a theoretical approach which aims at contributing to the understanding of the factors that may determine the elaboration of media messages and articles in an accountable manner, when reporting on security threats or sensitive issues for the public. The analysis represents a contribution to the first steps towards the literacy of both the communicators and the audience in the field of hate speech, polarization, discrimination and other related phenomena and actions. Each type of phenomenon is analyzed through comparative definitions and characteristics of manifestation, followed by the analysis of the human rights perspective in dealing with it, the analysis of the legal framework at European level, possible counteraction approaches, main challenges and lessons learnt when addressing discrimination, hate speech and polarization.

Keywords: *hate speech, discrimination, polarization, and communication, public reporting.*

Arguments towards responsible public reporting

The power of the internet and the impact of social networks have a determinant influence over the culture of communication, thus, determining an increasing pluralism and the unethical dissemination of information, in the absence of the critical analysis of information. Actions like polarization or discrimination have led to the normalization of hate speech at European level in the recent years, fuelling radicalization, racism, xenophobia, toxic behaviours etc. Social networks can act as channels for communicating freely, within the public sphere, but also as echo chambers for reinforcing certain beliefs (Grömping, 2014).

Communicating news about security threats in a generalized manner may have a strong negative impact over the public and lead to different kinds of reactions, from panic and fear to polarization and violent behaviour. As such, institutional spokespersons and journalists in the field of security and law enforcement bear a high responsibility in relation to their audience, in order to deliver accurate information, in an accountable manner. As relevant communicators, the spokespersons and journalists are required to have crucial skills and competences, like critical thinking, responsible reaction, or the ability to identify fake news, polarizing discourses, or push and pull factors of radicalization that may lead to violent behaviour.

The present theoretical approach may contribute to the understanding of such factors and may determine an alignment of the social reality and the conveyed messages. Thus, the literacy of the communicators involved in such type of actions and elements represents the first step of the complex process of achieving instinctive responsible reactions in relation to the audience.

The analysis aims to provide a better understanding and use of communication techniques, dedicated to institutional spokespersons and journalists in the field of security and law enforcement and to relevant stakeholders, when dealing with hate speech, discrimination and polarizing events and reporting on security threats or sensitive topics.

The analysis of hate speech, discrimination and polarizing events contributes to the development of both individual and community capacities of spokespersons and journalists, in order to use media reporting conscientiously and ethically.

What is hate speech and which are its specific characteristics of manifestation?

Negative opinions and views expressed with respect to certain individuals or groups, in the absence of counteracting actions, tend to be generally accepted and integrated as „normal" attitude. (Pálmadóttir and Kalenikova, s.a.)

The term "hate speech" refers to negative acts and perspectives towards society, minorities, democracy etc., which may lead to violent actions. The expression of hate speech in different ways of manifestation and through different types of channels, has contributed to phenomena such as radicalization, racism, discrimination, polarization and hate crime. This has led to the promotion of hate narratives towards women and minority populations like LGBTQI, Roma, migrants, refugees, minority religious communities, but also towards political movements, governmental decisions, policies, or associated key personalities. The resulted action of hate speech may contribute to the weakening of democracy, of the equity among populations, of social cohesion, but may also lead to distrust in the law and to concrete violent acts. (Erasmus+ Virtual Exchange, s.a.)

According to the Recommendation no. 97(20) of the Council of Europe Committee of Ministers, hate speech is to be seen as: „all forms of expression which spread, incite, promote or justify racial hatred, xenophobia, anti-Semitism or other forms of hatred based on intolerance, including: intolerance expressed through aggressive nationalism and ethnocentrism, discrimination and hostility against minorities, migrants and people of immigrant origin". (Pálmadóttir and Kalenikova, s.a., p. 7)

The *European Court of Human Rights* refers to hate speech as: "all forms of expression, verbal or written, which spread, incite, promote or justify hatred based on intolerance (also on grounds of religion)". (Pálmadóttir and Kalenikova, s.a., p. 7)

ILGA Europe, an active organization in the field of equality and human rights, including countering hate speech, defines the term in relation to the concept of „hate crime":

"Hate speech is public expressions which spread, incite, promote or justify hatred, discrimination or hostility towards a specific group. They contribute to a general climate of intolerance which in turn makes attacks more probable against those given groups". (ILGA Europe)

Or: "Hate crime is any form of crime targeting people because of their actual or perceived belonging to a particular group. The crimes can manifest in a variety of forms: physical and psychological intimidation, blackmail, property damage, aggression and violence, rape, and murder". (ILGA Europe)

The above definitions underline the type of actions, which are grouped under the umbrella of hate speech: spreading, inciting, promoting, justifying. These actions have hatred, discrimination, hostility, and the characteristics of the targeted population as their triggers. They may result in the promotion of intolerance or in an attack by a third party. The targeted groups are usually the victims of hate crimes, which can manifest in the form of physical or psychological abuse, damage, aggression or even murder.

Hate speech cannot be identified only through the use of certain type of words, but also through the context in which it is used of using it, the expressed intention and the possibility to have negative outcomes (Pálmadóttir and Kalenikova, s.a.). Hate speech has as trigger certain characteristics of the targeted communities, like ethnicity, religiosity, gender, sexual orientations, and its main purpose is the humiliation, the disrespect, or the legitimization of discrimination and attack on them (ILGA Europe).

Hate speech differs from hate propaganda, which is defined by its main characteristic of being systematic and consistent to specific ideologies. On the other hand, hate speech is not systematic, or the people who express such content do not necessarily share the same ideology (Pálmadóttir and Kalenikova, s.a.).

During the past years the internet has become an important channel of disseminating hateful content on the grounds of these terms/ideas are repeated over and over again. This channel and the rapid development of IT facilitated the work of extremist groups: if the first hate site was launched in 1995, by 2012 there were already 15,000 such web sites, mostly with racist or xenophobic content (Pálmadóttir and Kalenikova, s.a.).

What is discrimination and how does it manifest?

Instability, especially in the financial, economic and labor field may lead to the discrimination of certain groups, racism and xenophobia (Pálmadóttir and Kalenikova, s.a.).

The act of discrimination refers to "treating a person unfairly because of who they are or because they possess certain characteristics". (EOC, 2019) According to the UK Equality Act 2010[1], discrimination may occur and is protected according to nine characteristics: age, gender, race, disability, religion, pregnancy and maternity, sexual orientation, gender reassignment, marriage and civil partnership (EOC, 2019).

Research regarding discrimination has shown that most Member States (MS) are implementing the EU principle of non-discrimination on grounds of nationality, but, at the same time, results have indicated that practitioners in the field, at the level of MS, don't know what procedure to apply if such a case occurs. The European Union has set as one of its

[1] The Equality Act 2010 provides the legal framework for protecting individuals in case of discrimination acts, comprising 116 pieces of legislation in the field (https://www.equalityhumanrights.com/en/equality-act-2010/what-equality-act).

objectives the safeguarding of non-discrimination and the implementation of the principle of equal treatments in relation to the victims' rights and on the grounds of „race, ethnicity, gender, disability, age, sexual orientation, gender, and religious orientation". (EPRS, 2017)

According to an evaluation of non-discrimination actions at the level of the EU's MS, the European Commission (EC, 2017) highlights the possible different types of discrimination:

- Assumed discrimination – it occurs when assumptions are made about a certain individual or a group, even if the facts are not true.
- Associated discrimination – it occurs when an individual associate with another person or group who present/s a certain characteristic.
- Multiple discrimination – it occurs when an individual or a group of individuals are discriminated against on multiple grounds, e.g. for being a Roma elderly woman.
- Intersectional discrimination – it occurs when an individual or a group of people are victims of discriminating acts on the grounds of several inseparable characteristics.
- Direct discrimination – it occurs when people feel "the need to demonstrate less favourable treatment", when there is „a requirement for comparison with another person in a similar situation but with different characteristics (e.g. ethnic origin, religion, sexual orientation), when there is "the opportunity to use a comparator from the past" etc. (EC, 2017, pg. 43); direct discrimination can be stated when a person is "treated worse than another person or other people because: you have a protected characteristic, someone thinks you have that protected characteristic (known as discrimination by perception), you are connected to someone with that protected characteristic (known as discrimination by association)". (Equality and Human Rights Commission, 2018)
- Indirect discrimination – it occurs when "there is a policy that applies in the same way for everybody but disadvantages a group of people who share a protected characteristic". (Equality and Human Rights Commission, 2018)
- Harassment – defined as „unwanted conduct relating to racial or ethnic origin, religion or belief, disability, age, or sexual

orientation with the purpose or effect of violating the dignity of a person and of creating an intimidating, hostile, degrading, humiliating or offensive environment". (Directive 2000/43/EC)

➢ Discrimination by perception – it occurs when people are treated unfairly because they are thought to belong to a certain group or to have certain characteristics (EOC, 2019)

➢ Victimization – it may happen when people who complain about being victims of discrimination, or who sustain victims of discrimination, are themselves treated badly because of this (EOC, 2019).

What is polarization and which are its specific characteristics of propagation through events?

The elements that define polarization are mainly focused on attitudes, rather than behaviors. DiMaggio, Evans and Bryson (1996) consider that polarization can be related to either the process or the state by which attitudes are being diverted to ideological extremes. The EPRS study (2019) highlights the distinction between polarization of the elite public and that of the general public.

There is little evidence that factors like exposure to news or to opposing views may lead to the spread of polarization among the media's public. On the other hand, there are studies which prove that the two elements may contribute to the already strong attitudes and views of people, regarding a certain aspect. The selection of news sources across Europeans differs by the countries' current situations (political, economic, social etc.), while research on the topic at the level of the United States of America shows higher degrees of partisan media coverage, news consumption and polarization (EPRS, 2019).

In understanding the phenomenon of polarization, the EPRS report (2019) takes into consideration both levels of production and consumption of news, and analyses four topics when targeting news production:

➢ News content

o Current European issues, like immigration, corruption, refugees etc., are reflected in the news according to the political leaning of the source.

  - While in Europe researchers do not show a high interest regarding news polarization, researchers in the US indicate a high degree of polarization in news media content.
- News media landscape
  - News outlets tend to become commercialized, especially the online ones, but the degree of polarization has not been correlated to this aspect.
  - While national newspapers cover diverse topics, local newspapers tend to present more homogenous content.
- Public news media
  - Public news media is adapting to the audience's behavior of online consumption, in order to reach the public and to reduce the consumption of polarized news.
  - Because public news media relies on social media, it can itself be a factor of polarization.
- Digital news media
  - Digital news media tend to present news so that it resonates with young people and the views of the targeted groups.

Polarization may also be the result of exposure to news, which can be "incidental" or "selective". While the incidental exposure happens as an incident, when trying to inform on other topics, the selective exposure implies the selection of topics, news, articles etc., in accordance with the people's previous interests. (EPRS, 2019) The media may increase the polarization level in case the audience manifests a dislike of the opposite views, and, at the same time, the media may contribute to the moderation of attitudes in the presence of convincing arguments. (EPRS, 2019)

With regard to the channels of propagation of polarization, studies have shown that social media platforms may facilitate the exposure to opposite views, especially concerning political topics, but with a lower impact on people who present a high degree of polarization (EPRS, 2019). Fletcher and Nielsen (2018) concluded as a result of their study on data from the 2017 Digital News Report that search engines used for news expose people to different type of views, but it didn't indicate a clear impact of polarization. Flaxman et al.

(2016), though, found that people who use search engines for news are more ideologically dispersed and polarized than the ones who use social platforms, or both social platforms and search engines.

What is the human rights framework in dealing with hate speech, discrimination and polarizing events?

The main element of human rights is “equality for all persons” (Pálmadóttir and Kalenikova, s.a.). The act of hate speech has an important impact over the act of discrimination. It can lead to prejudice and social division. Mass media plays an important role in spreading and stopping the dissemination of hate speech among certain groups, through the messages they communicate. Nazi Germany and former Yugolsavia are two examples of the involvement of media in spreading hate speech, which has contributed to conflicts and mass murders against national minority groups (Pálmadóttir and Kalenikova, s.a.).

The European Union Charter of Fundamental Rights highlights the freedom of expression as a central value. The democratic character of the MS implies the availability of verified information, which citizens can use in understanding the political facts through their own critical lenses. Actions like disinformation[2], hate speech, discrimination, polarizing events etc. interfere with the aimed desire for democratic processes of thought and analysis. (European Commission, December 2018)

The freedom of expression is protected through a series of international instruments (e.g. UDHR, ECHR), which permit the dissemination of any opinion in any type of environment without any restrictions. Apart from this right, there are others which are also being addressed, like debating, sharing information, or analysis of political facts (Pálmadóttir and Kalenikova, s.a.).

Studies have shown that hate speech and hate crime are often not reported by the victims, because of the discomfort they have to face, especially if they have been attacked on the grounds of their sexuality (FRA, 2009).

[2] Disinformation is here defined as “verifiably false or misleading information that is created, presented and disseminated for economic gain or to intentionally deceive the public, and may cause public harm” (European Commission, April 2018).

Hate speech is being addressed across European countries through restrictive measures regarding the type of messages legally allowed. How do these limits address the freedom of speech, though, in a time of free access to channels of both expression and information? The perspective of regulating hate speech is at the intersection of the freedom of speech and of authoritarianism (Erasmus+ Virtual Exchange, s.a.).

The requirement to respect human rights imposes the established standards, at an international level, of the quality of life, highlighting the necessity of equality and dignity. In this context, hate speech manifestation is considered in relation with the violation of human rights (Erasmus+ Virtual Exchange, s.a.).

"The Universal declaration of Human Rights" (UDHR), adopted by the United Nations after World War II with the aim of preventing the spread of intolerance and hatred, has contributed to the process of combating discrimination based on race, xenophobia and other forms of intolerance. The UDHR protects the people's freedom of opinion and expression. The CERD Committee has paid special attention to discrimination based on race, hate speech and derogatory speech, stipulating the right of the victims to be compensated (Pálmadóttir and Kalenikova, s.a.).

What does the legal framework on hate speech, discrimination and polarization state?

In the context of the 2019 European, national and local elections, The European Union has developed an "Action Plan against Disinformation". The document establishes the allocation of the necessary resources in the field, the creation of a "Rapid Alert System" and the monitoring of the "Code of Practice" of online industry (EC, Press Release, 2018): "Healthy democracy relies on open, free and fair public debate. It's our duty to protect this space and not allow anybody to spread disinformation that fuels hatred, division, and mistrust in democracy". (HR Federica Mogherini, EC, Press Release, 2018)

"We need to be united and join our forces to protect our democracies against disinformation. (…). To address these threats, we propose to improve coordination with Member States through a Rapid

Alert System, reinforce our teams exposing disinformation, increase support for media and researchers, and ask online platforms to deliver on their commitments. Fighting disinformation requires a collective effort". (Andrus Ansip, Vice-president responsible with Digital Single Market, EC, Press Release, 2018)

The Action Plan against Disinformation focuses on four aspects which may contribute to counter disinformation, by capacitating the MS and the inter-state cooperation (European Commission, December 2018):

- Improving detection capabilities – this will be tackled by reinforcing the Strategic Communication Task Forces, the EU Hybrid Fusion Cell in the European External Action Service (EEAS) and the MS with specialized human resources and tools; also, EEAS allocates a significant budget for raising awareness regarding disinformation (1,9 mil Euro in 2018 and an estimative budget of 5 mil Euro in 2019).
- Coordinating the response between EU institutions and MS – the action includes the implementation of a Rapid Alert System, in order to better share the information between them in real time.
- Monitoring the implementation of the "Code of Practice by the online platforms" – the commitments made by the online industry include the insurance of the transparency of political advertising, closing fake accounts, working on identifying bots and disinformation content, or promoting fact-checked content.
- Empowering citizens through awareness and media literacy campaigns – these actions will include the empowering of local fact-checkers and researchers to identify disinformation content on social platforms.

The Rapid Alert System (RAS) represents one of the four pillars of the Action Plan against disinformation proposed in December 2018 by the European Council. This digital platform has as its main purpose the coordination of actions and responses regarding disinformation, at the level of EU institutions and the MS. The RAS has among its main sources of information open-sources, academia, fact-checkers, and online platforms. It brings together 28 national contact points, which contribute with information, best practices, analyses, trends and insights to countering disinformation. The outcomes foreseen by the RAS include raising awareness on disinformation among the general public, identifying cases of disinformation in the online, empowering the civil society and the professionals involved, ensuring a coordinated response and responsibility. (https://eeas.europa.eu/headquarters/headquarters-Homepage/59644/factsheet-rapid-alert-system_en)

The Code of Practice against Disinformation has been signed in October 2018 by the online industry (platforms – Facebook, Google, Twitter, Mozilla and, from May 2019, Microsoft –, social networks, advertisers etc.), agreeing on counteracting disinformation and fake news in the online environment (EC, June 2019). The first monitoring report on implementing the code of practice has registered a significant progress in eliminating fake accounts and making less visible disinformation sites. The European Commission has stressed the necessity to ensure the transparency of ads, to allow access for documentation and research and to sustain the collaboration of MS and the Rapid Alert System. The implementation of the Code is to be monitored during the first year, followed by possible standardization actions proposed by the EC. (EC, January 2019)

In order to prevent and counter illegal hate speech in the online environment, in May 2016 the European Commission has signed with Microsoft, Facebook, Twitter and YouTube the "Code of conduct on countering illegal hate speech online". In 2018, Instagram, Google+, Snapchat and Dailymotion, have joined in the agreement and in 2019 Jeuxvideo.com also became a member of the "Code of conduct". The actions developed by these IT companies in order to respect the "Code of conduct" are being monitored by established EU organizations in different MS, based on a standard procedure. The evaluation has shown that the companies have managed to act rapidly to eliminate racist and xenophobic hate speech and the last reports show that approximately 89% of the flagged content is being evaluated within 24 hours and approximately 72% of the illegal hate speech is being deleted (https://ec.europa.eu/info/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/countering-illegal-hate-speech-online_en). The IT companies that have signed the Code of conduct have taken on board different entities with the role of flaggers of (illegal) hate speech: within the first year of implementation 106 NGOs have joined the mission of Facebook and Twitter, at the level of 21 countries. Likewise, national contact points have been established, in order to facilitate the collaboration of the IT companies that signed the Code of conduct and the national competent authorities. The nine companies that signed the Code of conduct cover approximately 96% of the EU market share of online platforms susceptible to hate speech content. (EC, February 2019)

In 1965, the United Nations adopted the "UN Convention on the Elimination of All Forms of Racial Discrimination" (CERD) as a response to anti-Semitic attacks in Germany and colonialism. CERD promotes the eradication of incitement and discrimination on racial arguments and forces the parties of the convention to condemn hate speech, hate crimes and racial discrimination (Pálmadóttir and Kalenikova, s.a.).

The European Convention on Human Rights (ECHR) also addresses the issue of non-discrimination of people (on grounds of sex, race, color, spoken language, religion, political orientation) in relation to the fundamental human rights and freedoms stated within the document. In 2000, Protocol no. 12 to the ECHR added the prohibition of discrimination for benefiting of any legal right within the national law.

The European Social Charter (1996) set the prohibition of discrimination in relation to employment and gender, aiming to install equal treatment of individuals. Other European conventions also address the issue of discrimination, such as: the CoE "Framework Convention for the Protection of National Minorities", the CoE "Convention against Trafficking" and the CoE "Convention on Access to Official Documents". Also, the CoE Convention on Cybercrime, through the Protocol on Xenophobia and Racism prohibits the dissemination of racist or xenophobic content in the online environment (Pálmadóttir and Kalenikova, s.a.).

All EU MS incriminate physical assault, and when it has a discriminative or hate-related reason, the crime may be considered even more dangerous. To this matter, the EU has adopted since

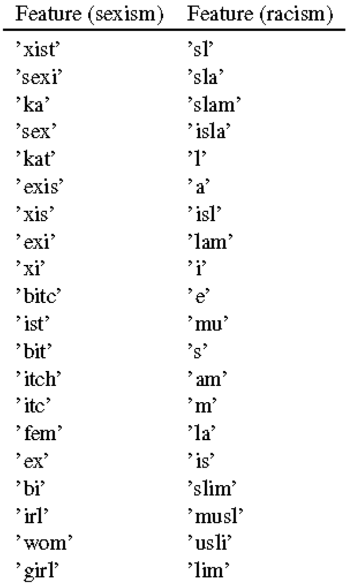
November 2008 a decision against inciting to hate crime on the basis of racism or xenophobia (JO L 328.20068). (FRA, 2009)

The Racial Equality Directive of the European Union prohibits discrimination on the grounds of ethnicity and race in different fields, like education, employment, healthcare, supply, social protection. Moreover, the Employment Equality Directive prohibits the discrimination on the grounds of „religion, disability, age, and sexual orientation within the labor market”. (EC, 2017)

How to counter hate speech, discrimination and polarization?

A critical analysis of hate speech leads to the necessity of promoting alternative narratives. The Internet has proven to be an efficient channel of communication for hate speech discourse (Eadicicco, 2014; Kettrey and Laster, 2014), burdening the social media platforms with the task of detecting and erasing such content (Moulson, 2016), while respecting the freedom of speech (Waseem and Hovy, 2016). The disastrous outcomes of hate speech, like hate crime, highlights the importance of detecting and managing hate speech discourse, narratives and actions (Hate Speech Watch, 2014).

In order to identify hate speech, it is absolutely necessary to critically analyze the messages, because they don't necessarily include pre-established hate speech terms (McIntosh, 2003; DeAngelis, 2009). Waseem and Hovy (2016) propose as an efficient method in



Feature (sexism)	Feature (racism)
'xist'	'sl'
'sexi'	'sla'
'ka'	'slam'
'sex'	'isla'
'kat'	'l'
'exis'	'a'
'xis'	'isl'
'exi'	'lam'
'xi'	'i'
'bitc'	'e'
'ist'	'mu'
'bit'	's'
'itch'	'am'
'itc'	'm'
'fem'	'la'
'ex'	'is'
'bi'	'slim'
'irl'	'musl'
'wom'	'usli'
'girl'	'lim'

Table 1. The most indicative character n-gram features for hate-speech detection
Source: Waseem and Hovy (2016, p. 92)

detecting hate speech *the n-gram model*. The model results with the probabilistic prediction of the items in a sequence of words (Jurafsky and Martin, 2018). Waseem and Hovy (2016) have considered more efficient for their analysis the use of character n-grams instead of word n-grams, in correlation to gender associated features and location.

Aiming to understand the role of social media in the polarization process, Beam, Hutchens and Hmielowski (2018) conducted a three-wave online survey during the US Presidential Elections of 2016. The results showed that news disseminated on Facebook have registered a decreasing polarization effect, especially as a result of posting cross-cutting news or pro-attitudinal information. The authors considered that Facebook might be used as an instrument of depolarization.

Results have shown that education is the main element in combating and preventing acts of discrimination, hate speech and social polarization. As a result, UNESCO (2019) proposes five ways to counter hate speech in the media, by imposing ethics and self-regulation:

- ➢ "Education on media ethics" – during the last years, the emergence of social media has determined the creation of online platforms and has facilitated the dissemination of hate speech; UNESCO considers that education on media ethics and the important role of spokespersons and journalists in promoting peace is a first step in countering hate speech; the process has to start with the introduction into political, social and cultural rights of individuals, and has to continue with awareness in relation to the responsibilities that derive from the freedom of the press;
- ➢ "Encourage conflict sensitive reporting and multicultural awareness campaigns" – the approach is destined to eliminate the fallacy of "us" versus „them"; in this respect, journalists are to develop skills for reporting on sensitive issues, and to learn about different cultures and traditions.
- ➢ "Regulate social media" – media laws and ethics can contribute to the regulation of social media without trespassing the freedom of the press.

- "Encourage victims and witnesses to report hate speech related crimes" – it is important that victims know where to report the experience, so it can be countered.
- "End impunity against hate crimes" – UNESCO proposes to tackle the impunity against hate crimes by establishing units of monitoring and evaluation of hate speech; the units would have the responsibility of disseminating the evaluations to stakeholders and the civil society. (Jamekolok, P. A., 2019)

Media, especially the visual, is seen as an important instrument in shaping public opinion. As such, media can be a tool in promoting human rights, combating hate speech and violence, and creating group and social cohesion. At the same time, media can propagate intolerance and hatred. To have a positive contribution to the battle against hate speech, social polarization, and discrimination, media should report "factually and accurately", "draw upon professional codes of conduct within their different media sectors", "provide in-house training or opportunities for outside training for their media professionals at all levels, on professional standards on tolerance and intolerance as well as a multi-ethnic journalistic team". Media can also be used as a channel for public debate, facilitating the dialogue between communities. This is a must in a democratic society (Pálmadóttir and Kalenikova, s.a.).

The media literacy of individuals is also important in the fight against intolerance. The internet has become a more accessible channel of information, so individuals need to be taught about how the media works and how to critically analyze the information. "Media literacy is the ability to access, analyze, evaluate, and create media; from television, radio, Internet, newspapers, social media, and all other forms of media and to use them in a responsible and critical manner". (Pálmadóttir and Kalenikova, s.a., p. 23)

Article 19 (2018) proposes the counteracting of hate speech by implementing a series of measures at legislative level, in relation to the human rights and the right to free expression, proposing, at the same time, a series of literacy actions:

- Providing trainings on human rights applicable to hate speech, dedicated to law enforcement, judiciary and public entities; the main tool for training and further use should be a guideline based on the human rights regulations.

- Elaborating regulatory framework for the media, in order to ensure its diversity.
- Elaborating and implementing clear policy guidelines in relation to hate speech.
- Media outlets should guarantee the media reporters with resources for validation of information, in order to present accurate data; this process may include trainings for media in relation to hate speech and the provision of the proper technical equipment.
- Journalists' organizations should prepare proper responses for journalists to use in case of hate speech and freedom of expression; this may include a code of conduct or training on ethics and human rights.

Main challenges when addressing hate speech, discrimination and polarization

When addressing hate speech, discrimination and polarizing events, the news consumption habits and attitudes of the public tend to become a challenge, especially regarding three aspects:

- the prevalence of online news media: Europeans have developed a behavior of consuming news online, because of the possibility to access various sources in a short time, based on their interests (EPRS, 2019);
- the use of social media platforms as news sources: the information on social media may lead to a higher degree of exposure to opposite political views, and few studies have indicated a higher degree of polarization in case of news consumption on social media, while others, conversely, showed de-polarization (EPRS, 2019);
- the consumption of populist news: the exposure to populist trusts has proven to have an impact only on people with populist views, without having a significant influence on those with no views to this regard (EPRS, 2019).

The EPRS study (2019) states that peoples' attitudes in the UK and Southern European countries have been more influenced by politics than in the Western and Northern Europe. The data doesn't show, though, that a selective exposure has a polarizing effect, but it can have a strengthening effect over the public with an already formed opinion.

A significant challenge in preventing and countering hate speech on the internet is the possibility of not being able to localize the source of the message/act. Also, the national legislations differ, so not all messengers of hate speech can be punished according to a standardized regulation, nor can the same ethical guidelines be implemented in all the countries. Thus, cooperation and coordination of responses and stakeholders, including private suppliers of internet services, are seen as main measures in preventing and combating hate speech and propaganda (Pálmadóttir and Kalenikova, s.a.).

Case studies and lessons learnt

European studies on polarization have connected the process elements to topic, source, frame and tone of the news, and, particularly, to the coverage of political issues. For example, in what concerns immigration, the UK newspapers have used as sources the Government or other official entities, and experts in the field (like research institutes and think tanks). Balch and Balabanova (2011) show that while the right-wing trusts used these types of sources to correlate immigration to a dangerous situation, the left-wing ones used it to denounce associated presuppositions.

Another topic of interest at European level, corruption, has led to the association of this issue with polarization. A comparative analysis regarding the level of press-freedom in UK, France and Italy, in relation to the commercialization character of the media, the target segmentation and the influence of politics, has indicated that the topic of corruption has been covered to a higher extent in Italy, than in the UK or France. It also covered, to a higher extent, topics regarding local politicians and used dramatic tones. Each newspaper that was analyzed targeted a specific market segment, by addressing corruption in such a way so it would attract its own audience. (Mancini, Mazzoni, Cornia, and Marchetti, 2017)

The analysis of the two main Spanish newspapers (El País and El Mundo) showed that they both rely and promote official sources and dominant political parties, especially during economic crisis and elections. The two focused more on the opposition than on the allies, considering that this approach would make the news more appealing to the public. (Bonafont and Baumgartner, 2013)

The EPRS study (2019) emphasizes the fact that media platforms can influence each other's lines of topics, as an effect of the so called "intermedia agenda setting". The study conducted by Cushion, Kilby, Thomas, Morani and Sambrook (2018), during the 2015 UK election campaign, showed that, despite declaring that broadcast news haven't been influenced by right and left wing newspapers' coverage, the TV news reflected in the newspapers' agenda.

The 2017 Reuters Institute Digital News Report (Newman, Fletcher, Kalogeropoulos, Levy, and Nielsen, 2017 apud EPRS, 2019) highlighted an approach of measuring news audience polarization, based on the level of left – or right-wing beliefs of a news outlet's public. The study implied a quantitative measure on a seven-step scale from "very left wing" to "very right wing", which has been correlated to the type of news outlets read during the last week. The study has been implemented in 22 countries. The data resulted with "the average political leaning of the population" and "the average political leaning of the audience for each news outlet", which indicated the partisanship level of the audiences and the level of polarization of the online audiences per country (reported to the standard deviation of the resulted scores for each news outlet, at the level of each country) (see figure 1). According to the results presented in figure 1, news audience polarization is higher in the USA (5.93) than in any other country included in the sample, and it may have a smaller impact on the European audience.

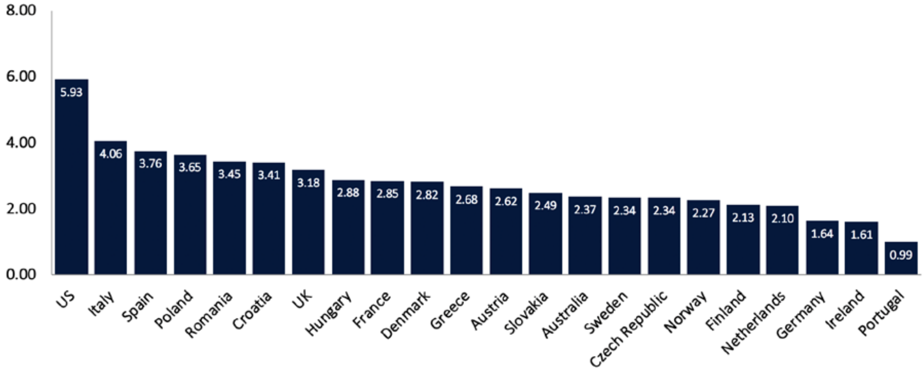


Figure 1. Level of online news audience polarization
(Source: Newman et al., 2017, apud EPRS, 2019, p. 29)

Trilling et al. (2017) have also studied the effect of news on polarization, using as topic the immigration situation in the Netherlands. The experiment conducted analyzed the impact of positive and negative news regarding immigration on the subjects' attitudes. The experiment included the measure of attitude before and after the exposure to the articles, and the assignation versus the free choice of articles. Those who could choose the articles selected the ones in line with their previous attitudes; those who were assigned articles with positive content towards immigration tended to express a more positive attitude, while those who were displayed negative articles did not register any change of attitude towards immigration.

Waseem and Hovy (2016) analyzed over a two months period of time, 16,914 tweets, out of which 3,383 contained sexist content, 1972 racist content, and 11,559 contained other different types of content. The process followed an initial manual analysis of terms associated with religious, gender, ethnic and sexual minorities, followed by an automatic process of collecting English tweets by using API[3]. The data

[3] API is the acronym for "application programming interfaces". APIs allow users to post tweets, to search for certain content using keywords, or monitor certain Twitter accounts (Source: https://help.twitter.com/en/rules-and-policies/twitter-api accessed on 18.07.2019).

set has been annotated manually and validated by a gender-studies expert, in order to eliminate any type of biases. The authors have proposed a model of identifying hate speech within tweets, considering that a message is offensive if it:

"1. uses a sexist or racial slur.

2. attacks a minority.

3. seeks to silence a minority.

4. criticizes a minority (without a well-founded argument).

5. promotes, but does not directly use, hate speech or violent crime.

6. criticizes a minority and uses a straw man argument.

7. blatantly misrepresents truth or seeks to distort views on a minority with unfounded claims.

8. shows support of problematic hash tags. E.g. #BanIslam, #whoriental, #whitegenocide.

9. negatively stereotypes a minority.

10. defends xenophobia or sexism.

11. contains a screen name that is offensive, as per the previous criteria, the tweet is ambiguous (at best), and the tweet is on a topic that satisfies any of the above criteria". (Waseem and Hovy, 2016, p. 89)

Aiming to study hate speech tweets in relation to demographic distribution, Waseem and Hovy (2016) have used proxy data (gender-associated names of profiles, or gender specific pronouns and nouns) in their analysis. The results have indicated a high prevalence of male users as being active in hate speech; the gender characteristic has resulted to be statistically significant only in relation to location.

The Office of the OSCE Representative on Freedom of the Media organized on December 18th, 2014 in Vienna, Austria, a conference with the theme "Freedom of Expression for Tolerance and Non-Discrimination". The purpose of the event was to raise awareness regarding the relationship between freedom of expression, tolerance and non-discrimination, and it was addressed to international experts in the field, academia, and OSCE institutions. (https://www.osce.org/fom/127110)

The United Nations have acted against racism and discrimination over three decades between 1973 and 2003, which resulted in three global conferences. The third one, held in Durban in 2001, focused on developing a monitoring system of the actions of the MS and has resulted in an "Intergovernmental Working Group on the Effective Implementation of the Durban Declaration and Programme of Action" (DDPA). The DDPA contains measures for combating issues raised during the Durban conference, like racism, discrimination, xenophobia and intolerance (Pálmadóttir and Kalenikova, s.a.).

As a result of revising the DDPA and the organization of workshops with 45 experts from different areas in the field of incitement to hatred from a legislative, judicial and policies perspectives, the Rabat Plan of Action was elaborated in February 2013. The plan highlights the responsibility of communities and leaders, media actors and civilians to manifest and promote tolerance and communication, hence managing to determine the collaboration between different type of entities – academia, journalists, NGOs – for the purpose of ensuring the freedom of speech while removing hateful content (Pálmadóttir and Kalenikova, s.a.).

In August 2018, in Bucharest, and in January 2019, in Berlin, as part of the project "*Like Share Diversity! Log Out Hate Speech!"*, a campaign dedicated to youngsters was implemented, aiming to promote diversity. The campaign included non-formal education activities which addressed the way the hate speech works as a social phenomenon. The youngsters had the possibility to learn about efficient ways of reacting to hate speech, especially in the online, and to accept diversity. The project was implemented by two partner NGOs, one from Romania and one from Germany, targeting to create an intercultural civic frame of education for the young generation. The project started as a response to the discrimination wave against vulnerable groups, through propaganda, hate speech and disinformation. (STIRI.ONG, 2019)

Conclusions

The present theoretical approach aimed at contributing to the understanding of the factors that may determine the alignment of the social reality to the message transmitted in case of security threats or of sensitive issues. Targeting to contribute to the development of individual and community capacities of institutional spokespersons and journalists in the field of security and law enforcement, but also of other stakeholders, in order to use media reporting in an ethical and responsible manner, the analysis discussed the three phenomena addressed from the perspective of manifesting characteristics and definitions. Furthermore, the study discussed the human rights perspective, the legal framework, possible counteraction and preventive actions, main challenges and lessons learnt when addressing hate speech, discrimination and polarization.

Education resulted as the main element of prevention and counteraction of such negative phenomena, of both the communicator and the audience. A series of main challenges when addressing hate speech, discrimination, polarization and other similar actions, have also been highlighted, all in relation to the characteristics of the internet and online channels – the speed of circulating a message, the variety of sources of information, the creation of echo chambers, the anonymity of the source etc.

References:

1. Balch, A., & Balabanova, E. (2011). "A System in Chaos? Knowledge and Sense-Making on Immigration Policy in Public Debates". Media, Culture & Society, 33(6), 885–904.
2. Beam, M. A., Hutchens, M. J., & Hmielowski, J. D. (2018). "Facebook News and (De)polarization: Reinforcing Spirals in the 2016 US Election". *Information, Communication & Society*. 0(0), p. 1-19.
3. Bonafont, L. C., & Baumgartner, F. R. (2013). "Newspaper Attention and Policy Activities in Spain". *Journal of Public Policy*. 33(1), 65-88
4. Cushion, S., Kilby, A., Thomas, R., Morani, M., & Sambrook, R. (2018). "Newspapers, Impartiality and Television News: Intermedia Agenda-Setting during the 2015 UK General Election Campaign". *Journalism Studies*. 19(2), p. 162-181.
5. DeAngelis, T. (February 2009). "Unmasking *racial micro aggressions*". *Monitor on Psychology*, 40(2):42.
6. DiMaggio, P., Evans, J., & Bryson, B. (1996). "Have American's Social Attitudes Become More Polarized?" *American Journal of Sociology*. 102(3), p. 690-755.

7. European Commission. (April, 2018). “Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions. Tackling online disinformation: a European Approach”. [Online]. Available at https://ec.europa.eu/transparency/regdoc/rep/1/2018/EN/COM-2018-236-F1-EN-MAIN-PART-1.PDF. Accessed on 23.07.2019.
8. European Commission. (December 2018). “Joint Communication to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions. Action Plan against Disinformation”. Brussels. [Online]. Available at https://eeas.europa.eu/headquarters/headquarters-homepage/54866/action-plan-against-disinformation_en. Accessed on 23.07.2019.
9. European Parliamentary Research Service (EPRS). Ex-Post Evaluation Unit. (December 2017). “The Victims’ Rights Directive 2012/29/EU. European Implementation Assessment”. [Online]. Available at http://www.europarl.europa.eu/RegData/etudes/STUD/2017/611022/EPRS_STU(2017)611022_EN.pdf. Accessed on 23.07.2019.
10. European Parliamentary Research Service (EPRS). “Scientific Foresight Unit (STOA). (March 2019)”. *Polarisation and the news media in Europe*. [Online]. Available at https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2019-03/Polarisation_and_the_news_media_in_Europe.pdf. Accessed on 23.07.2019.
11. Grömping, M. (2014). “Echo Chambers. Partisan Facebook Groups during the 2014 Thai Election”. *Asia Pacific Media Educator*. 24(I), 39-59. DOI: 10.1177/1326365X14539185
12. Heather Hensman Kettrey and Whitney Nicole Laster. (2014). “Staking territory in the world white web: An exploration of the roles of overt and color-blind racism in maintaining racial boundaries on a popular web site”. *Social Currents*. 1(3):257–274. DOI: 10.1177/2329496514540134
13. Flaxman, S., Goel, S., & Rao, J. M. (2016). “Filter bubbles, echo chambers, and online news consumption”. *Public Opinion Quarterly*. 80, p. 298-320.
14. Fletcher, R., & Nielsen, R. K. (2018). “Automated Serendipity: The Effect of Using Search Engines on the Diversity and Balance of News Diets”. *Digital Journalism*. 8(6), p. 976-989.

15. Jurafsky, D. & Martin, J. (September 2018). Chapter 3. "N-gram Language Models". In *Speech and Language Processing*. [Online]. Available at https://web.stanford.edu/~jurafsky/slp3/3.pdf. Accessed on 19.07.2019.
16. Mancini, P., Mazzoni, M., Cornia, A., & Marchetti, R. (2017). "Representations of Corruption in the British, French, and Italian Press". *The International Journal of Press/Politics*. 22(1), p. 67-91.
17. McIntosh, P. (2003). Chapter "White privilege: Unpacking the invisible knapsack". In *Understanding prejudice and discrimination*. Plous, S. (Ed.). (2003). New-York: McGraw- Hill. p. 191–196.
18. "Official Journal of the European Union". (December 2008). "Council Framework Decision 2008.913.JHA of 28 November 2008 on combating certain forms and expressions of racism and xenophobia by means of criminal law". *L328/55.* [Online]. Available at https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L: 2008: 328:0055:0058:en:PDF. Accessed on 19.07.2019.
19. Pálmadóttir, J. A., Kalenikova, I. (s.a.). "Hate speech; an overview and recommendations for combating it". Icelandic Human Rights Centre. [Online]. Available at http://www.humanrights.is/static/ files/Skyrslur/Hatursraeda/hatursraeda-utdrattur.pdf. Accessed on 23.07.2019.
20. Trilling, D., van Klingeren, M., & Tsfati, Y. (2017). "Selective Exposure, Political Polarization, and Possible Mediators: Evidence from the Netherlands"*. International Journal of Public Opinion Research*. 29(2), p. 189-213.
21. Waseem, Z. & Hovy, D. (2016). "Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter"*. Proceedings of NAACL-HLT 2016.* p. 88-93, San Diego, California, June 12-17, 2016. 2016 Association for Computational Linguistics. Available at https://www.aclweb.org/anthology/N16-2013. Accessed on 10.07.2019.

Online resources:

22. Article 19. (2018). "Responding to 'hate speech': Comparative overview of six EU countries". [Online]. Available at https://www.article19.org/wp-content/uploads/2018/03/ECA-hate-speech-compilation-report_March-2018.pdf. Accessed on 18.07.2019.

23. Directive 2000/43/EC. [Online]. Available at https://eur-lex.europa.eu/legal-content/en/TXT/?uri=CELEX%3A32000L0043. Accessed on 27.05.2019.
24. Eadicicco, L. (October 2014). "This female game developer was harassed so severely on twitter she had to leave her home". [Online]. Available at https://www.businessinsider.com/brianna-wu-harassed-twitter-2014-10. Accessed on 18.07.2019.
25. EOC. (2019). "What is Discrimination?" [Online]. Available at https://www.eoc.org.uk/what-is-discrimination/. Accessed on 22.07.2019.
26. Equality and Human Rights Commission. (2018). "What is direct and indirect discrimination?" [Online]. Available at https://www.equalityhumanrights.com/en/advice-and-guidance/what-direct-and-indirect-discrimination. Accessed on 22.07.2019.
27. Equality and Human Rights Commission. (2019). "What is the Equality Act? An introduction to the Equality Act 2010". [Online]. Available at https://www.equalityhumanrights.com/en/equality-act-2010/what-equality-act. Accessed on 22.07.2019.
28. Erasmus+ Virtual Exchange. (s.a.). "Countering Hate Speech in Europe". [Online]. Available at https://europa.eu/youth/erasmusvirtual/. Accessed on 27.05.2019.
29. European Commission. "The EU Code of conduct on countering illegal hate speech online. The robust response provided by the European Union". [Online]. Available at https://ec.europa.eu/info/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/countering-illegal-hate-speech-online_en. Accessed on 21.08.2019.
30. European Commission (EC). "Press Release. A Europe that Protects: The EU steps up action against disinformation". (December 2018). [Online]. Available at https://ec.europa.eu/commission/presscorner/detail/ga/ip_18_6647. Accessed on 23.07.2019.
31. European Commission (EC). (2017). "A comparative analysis of non-discrimination law in Europe 2017". [Online]. Available at https://publications.europa.eu/en/publication-detail/-/publication/36c9bb78-db01-11e7-a506-01aa75ed71a1. Accessed on 23.07.2019.
32. European Commission (EC). (January 2019). *Code of Practice against Discrimination*. [Online]. Available at https://ec.europa.eu/

commission/news/code-practice-against-disinformation-2019-jan-29_en. Accessed on 23.07.2019.

33. European Commission (EC). (February 2019). "How to Code of Conduct helped countering illegal hate speech online". [Online]. Available at https://ec.europa.eu/info/sites/info/files/hatespeech_infographic3_web.pdf. Accessed on 23.07.2019.
34. European Commission (EC). (June 2019). "Code of Practice on Disinformation". [Online]. Available at https://ec.europa.eu/digital-single-market/en/news/code-practice-disinformation. Accessed on 23.07.2019.
35. European Union Agency for Fundamental Rights (FRA). (2009). „Discursul de instigare la ură și infracțiunile motivate de ură împotriva persoanelor LGBT". [Online]. Available at https://fra.europa.eu/sites/default/files/fra_uploads/1226-Factsheet-homophobia-hate-speech-crime_RO.pdf, Accessed on 22.07.2019.
36. "Factsheet: Rapid Alert System". (15.03.2019). [Online]. Available at https://eeas.europa.eu/headquarters/headquarters-Homepage/59644/factsheet-rapid-alert-system_en. Accessed on 22.07.2019.
37. Hate Speech Watch. (June 2014). "Hate crimes: Consequences of hate speech". Available at http://www.nohatespeechmovement.org/hate-speech-watch/focus/consequences-of-hate-speech. Accessed on 18.07.2019.
38. ILGA Europe. "Hate crime &hate speech". [Online]. Available at https://www.ilga-europe.org/what-we-do/our-advocacy-work/hate-crime-hate-speech. Accessed on 22.07.2019.
39. Jamekolok, P. A. (2019). "5 ways to counter hate speech in the Media through Ethics and Self-regulation". UNESCO. [Online]. Available at https://en.unesco.org/5-ways-to-counter-hate-speech. Accessed on 23.07.2019.
40. Moulson, G. (February 2016). "Zuckerberg in Germany: No place for hate speech on Facebook". Available at https://www.businessinsider.com/ap-zuckerberg-no-place-for-hate-speech-on-facebook-2016-2. Accessed on 10/07/2019.
41. No Hate Speech Movement. "Communicating on Migrant Integration". [Online]. Available at http://www1.oecd.org/migration/netcom/campaigns-tools-platforms/no-hate-speech-movement.htm. Accessed on 21.08.2019.
42. STIRI.ONG. (February 2019). „Campanie a tinerilor împotriva discursului instigator la ură". [Online]. Available at https://www.

stiri.ong/ong/civic-si-campanii/campanie-a-tinerilor-impotriva-discursului-instigator-la-ura. Accessed on 21.08.2019.

43. The Organization for Security & Co-operation in Europe (OSCE). "Hate speech." [Online]. Available at https://www.osce.org/representative-on-freedom-of-media/106289. Accessed on 23.07.2019.
44. The Organization for Security & Co-operation in Europe (OSCE). "Discussion on Freedom of Expression for Tolerance and Non-Discrimination". [Online]. Available at https://www.osce.org/fom/127110. Accessed on 23.07.2019.