HATE SPEECH DETECTION IN MASS MEDIA: IT-BASED AND PSYCHOLINGUISTIC INTERDISCIPLINARY APPROACH

Yuliya Krylova-Grek

G.S.Kostiuk Institute of Psychology of the NAPS of Ukraine, Kyiv, Ukraine.
National University of "Kyiv-Mohyla Academy", Kyiv, Ukraine.
yulgrek@gmail.com

Abstract

In the research, I examine the issue of the influence of the mass media on a person's consciousness from a psycholinguistic perspective. The core focus of my current research is hate speech in print media. The work was performed within an international project «Free voices: Promoting Independent Media in the Target Region» featuring the Crimean Human Rights Group. This paper presents an intermediate study of hate speech in the texts of online publications dated November 2020. To receive data I combined the author's method of psycholinguistic analysis with a help of a software bot. Our research consists of four stages. During the first stage, a team of journalists and volunteers with the help of a software bot carried out search query mining to find negative information. Later, applying the author's method of psycholinguistic analysis of the text, we scrutinized the material obtained to identify the type of hate speech, its techniques of influencing consciousness.

The examples of hate speech were split into three types based on the use of specific linguistic and graphic tools: #1 direct hate speech; #2 indirect (hidden) hate speech; #3 manipulative hate speech. The evidence from this study intimates that the media tend to use examples of hate speech of the second and third types and avoid direct discrimination. In addition, there were mixed types: the combination of the second and third types and the combination of the first with the third or second type. I assumed the third and mixed#2-3 types of hate speech to be the methods of psycholinguistic manipulation. I have obtained comprehensive results proving that the majority of selected texts (75.7%) widely exploit methods of psycholinguistic manipulation. The results of this study will be used for a project report as well as to develop training for pre-journalist students and human rights defenders.

Keywords

Media. Hate speech. Manipulation. Text. Types of hate speech. Psycholinguistic perspective.

INTRODUCTION

The exponential growth of information and communication technologies and mass access to sources of information greatly influenced media's impact on society. In this context, manipulation, fake news, and propaganda can represent a perfect setting for the spread of violence, xenophobia, and hatred. For example, Donald Trump's incorrect comments on social media during the 2020 election in the United States provoked his supporters to storm the Capitol

that cost several persons their lives (Donald Trump was a linchpin behind Capitol riots, 2021). Another striking example is the brutal decapitation of Samuel Paty, a French schoolteacher, after a hate campaign against him on social media (Samuel Paty: French schoolgirl admits lying about the murdered teacher, 2021).

The dark side of the media is mentioned in the 2017 annual report of the Council of Europe. It highlights the danger of the dissemination and amplification of poor-quality information that originates online. The authorities identified three different types of information disorder: mis-, dis-, and mal-information, which are differed based on the dimensions of harm and falseness: mis-information is when false information is shared, but no harm is meant; dis-information is when false information is knowingly shared to cause harm, and mal-information is when genuine information is shared to cause harm. The authors consider hate speech as mal-information that is used to cause harm and discriminate against a group of people on religion, race, and other grounds:«…people are often targeted because of their personal history or affiliations. While the information can sometimes be based on reality (for example targeting someone based on their religion) the information is being used strategically to cause harm (Wardle & Derakhshan, 2017: 20-21).

In recent years, there has been considerable interest in issues of the negative influence of mass media in humanities (psychology, linguistics, sociology, etc.), IT, and Computational linguistics.

A growing body of literature in psychology has examined the media's influence on an individual's consciousness (Leontiev, 2004; Strong, 1922; Giles, 2003, 2016, and others). For instance, Leontiev (2004) focuses on mass media advertising, whose words, images, meanings, and associations stimulate positive consumers' attitudes toward products or brands and induce them into highly advertisement engagement.

In linguistics, much work on the potential of speech propaganda and manipulation has been carried out by Bulyigina & Shmelev A.(1997), Elswah & Howard (2020), Aronson & McGlone (2009), Rizun, Nepyivoda & Kornieiev (2005), Arutyunova (1990), etc. Numerous studies have been published on the distortion of the meaning of concepts in media texts (McGlone, Beck & Pfiester (2006), Shmelev D. (2008), Keith & Burridge (1991), and others).

There is a vast amount of literature on specific features of media representations of war and military conflicts (Pocheptsov (2001, 2019), Pack (2009), Kamalipour (2004), Galtung (1987), Dawes (2005)). Kamalipour (2004: 87-94) points out that since speech shapes our perception of reality, the age of information can be called the age of manipulation.

Chomsky & Herman (2002), Moscovici (1985), Browne & Keeley (2018), and others discuss the peculiarities of media propaganda as a mass media phenomenon. For example, Browne & Keeley (2018) draw attention to the rhetoric of propaganda speech, which ignores logical connections and does not provide empirical evidence for misleading information presented as factual. It ultimately leads to speech ambiguity and overload as well as makes speech difficult to analyze.

Chomsky & Herman (1998) developed a «propaganda model», which describes several filters applied to make news look like "fit to print". As commented by Chomsky (1998: 2), media information for broad audiences is in line with the interests of government and big organizations and is published to "marginalize the dissent."

In computational linguistics, the study and detection of hate speech explore by using natural language processing. For example, Schmidt and Wiegand (2017) studied the ways of the automatic detection of hate speech. All surveyed methods include common features that are usually used in the computer program to identify hate speech, such as a set of negative words

or expressions, using various complex features using ("dependency parse information", "features modeling specific linguistic constructs", "meta-information" and so on). At the same time the authors stressed that in most cases, computer program results can't be considered as full, because "they are only evaluated on individual data sets most of which are not publicly available" (Schmidt and Wiegand, 2017: 8-9). Taking into consideration the weak features of computer analysis of hate speech, we think that the best result researchers can receive if they combine computer and human inspection.

IT actively develops software to analyze text arrays, which parses text from unformatted content and unstructured data from social media, news reports, surveys, etc. to provide practical information like the mentioning frequency of a certain brand, person event, or counting particular words in the texts (e.g., programs for sentiment analysis and content analysis Semantrum, TABARI, Wordstat, JFreq etc.). However, some of these programs are paid ones, so they cannot be used in our study. Besides, each program is designed to perform certain tasks (monitoring posts with keywords, examining publication resonance, etc.), while this study is a report on hate speech on modern media that requires specific software, which will be discussed in the next section.

Considering hate speech as a product of media activity, we have paid special attention to the definition of the given concept. It is worth noting that there seems to be no general definition of hate speech for many reasons (Bartl et al, 2014, Benesh 2015, Saleem et al, 2017). Actually, the definition of hate speech depends on the legislation of each country, viewpoints on the issue, etc. (Howard, 2019: 96). Our study follows the definition of the Cambridge Dictionary, which interprets hate speech as "public speech that expresses hate or encourages violence towards a person or group based on something such as race, religion, sex, or sexual orientation" ("hatespeech". dictionary.cambridge.org.). In terms of media communication, hate speech is potentially dangerous as well since it frequently provokes acts of violence and hate crimes. For instance, notoriously famous the Rwandan genocide during the Rwandan Civil War in 1994 was preceded by anti-Tutsi propaganda in the local media (Yanagizawa-Drott, 2014; Melvern, 2004).

The increasing number of studies on this subject in Political and Social Sciences (Howard, 2019; Paasch-Colbergetal., 2021 and others), Legal Science (Waldron, 2012; Ghanea, 2012; Howard, 2019 and others), Media Communication (Bahador and Kerchner, 2019; Bahador, 2020 and others) are beneficial to both scholars and practitioners.

Freedom of expression is the core of human individuality, that's why hate speech issues have close relations with freedom of the press and legal aspects. In his recent research, Howard (2019: 93-94) shows much concern about distinguishing between freedom of expression and hate speech begin. The author notes that despite its relevance, this issue is not addressed at the legislative level in many countries, including the United States.

As indicated by Professor of law and philosophy Waldron (2012), each state should take appropriate measures to protect a person's dignity from hate speech by adopting special laws.

On 31 May 2016, the European Union cooperated with Facebook, Twitter, YouTube, and Microsoft to launch an online "code of conduct" aimed at fighting spreading xenophobia and racism across Europe. Věra Jourová, the EU commissioner for justice, consumers, and gender equality said that "The recent terror attacks have reminded us of the urgent need to address illegal online hate speech. This agreement is an important step forward to ensure that the internet remains a place of free and democratic expression, where European values and laws are respected." (Hern, 31 May 2016).

Based on the relevance of this issue, researchers explore the features of hate speech in comments and posts on social networks (Chen, 2011; Paasch-Colbergetal, 2021; Schmidt and Wiegand, 2017, etc.).

At the same time, media texts can contain indirect hate speech without making direct calls to violence and insults. Badak Bahador (2020) suggested a hate-speech intensity scale grounded on the in-group and out-group member criteria. He identified three strategies of hate speech, which characterizes the attribution of a person to either in-group or out-group member: 1) Dehumanization and demonization; 2) Violence and incitement; 3) Early warning. The first and second strategies imply the use of negative and offensive words and direct incitement to violence, while the third one is seen as the initial stage of dehumanization, which walks hand in hand with the demonstration of superiority. We are convinced that the study of the third category should gain more attention since at this stage it is more likely to prevent the spread of intolerance and calls for violence.

In January 2019, Bahador and Kerchner analyzed the USA media landscape and introduced a 6-point color-coded hate speech intensity scale. The researchers say about "variations of intensity are distinguishable of hate speech". The scale varies from basic level "disagreement" to the highest level "that provokes "violence" and "death" (Bahador and Kerchner, 2019: 5-8). This research demonstrated words and topics that are used in USA media. At the same time, the problem of hate speech in media should be thoroughly studied in other countries as well as in general. It is important to find common tools and methods of hate speech spreading.

Thus, hate speech was a subject of study, but we consider this phenomenon as interdisciplinary research at the intersection of psychology, linguistics, and IT technology. Our work aimed to widen current knowledge of hate speech in written media texts.

In this paper, we consider hate speech as a way of psycholinguistic manipulation of consciousness, to detect which we designed a special artificial intelligence-based bot.

The work was performed within an international project "Media freedom" featuring The Crimean Human Rights Group, where the author is involved as an expert in psycholinguistics.

In order to analyze the impact of hate speech on an individual's consciousness, we combined a computer bot program and the author's method of psycholinguistic analysis.

Aim. The present paper aims to examine hate speech in digital media, identify its type and methods of influencing an individual's consciousness by the author's method of psycholinguistic analysis and AI-based bot.

Hypothesis. It is hypothesized that the psycholinguistic approach combined with the software bot optimizes the search for hate speech and helps to avoid wasting time on routine work when selecting texts. In our opinion, this innovative approach gives experts more time to analyze information, as well as increases the efficiency of work. Moreover, our approach will be useful in advanced media literacy courses for journalists and pre-journalist students.

METHODS

Since looking through the texts on the Internet is time-consuming and labor-intensive, we decided to develop artificial intelligence software for selecting texts that are likely to contain hate speech. The bot optimizes the expert's work by surveying a large array of information to find a pool of texts that can incite xenophobia and hatred.

Thus, it was decided that the best procedure for this investigation is to join the author's psycholinguistic method of textual analysis and ICT, namely artificial intelligence-based bot

for selecting texts following the given parameters. The bot performs lexical analysis of content. Keyword selection strategy was based on the semantic meaning of the words in phrases and the entire publication with a negative tone.

Psycholinguistic analysis of the text is the author's methodology developed by Yu.M. Krylova-Grek. Now it is being registered in the patent office. Since our study is carried out within a long-term international project, the paper presents its intermediate results, which outlines the main approaches and opportunities for interdisciplinary interaction of humanities and computer sciences.

I. Setting up an AI-based software bot.

The major search parameters are as follows:

1. Online resource name. We specified the names of news agencies that specialize in current news. The initial cohort includes the nine most popular online media, whose traffic ranges from 1 million to 15 million visitors monthly.

2. Keywords. We entered words that usually accompany texts with hate speech. To single out keywords, we developed Hate dictionary based on the careful analysis of publications in selected media. Monitoring and word selection took place in 2017-2018. The dictionary includes 400 words. It should be noted the dictionary is constantly supplemented and changed due to the emergence of new narratives, concepts, and words.

3. Search period. We set the time interval: date and year. We strongly believe that a month is the most appropriate search period that can provide us with the required amount of data. Later, we plan to combine, compare and correlate the results obtained with the media-covered events.

II. *Psycholinguistic analysis of texts*. Text selection was followed by a psycholinguistic analysis of the data to determine and justify the presence of hate speech in these media. The analysis is a part of an innovative author's methodology, which assists in identifying both direct and manipulative hate speech that does not contain direct insults and calls for gender, racial or religious intolerance, but forms a negative attitude towards certain groups and individuals.

To conduct the foregoing textual analysis, hate speech was divided into three types, which are characterized by the use of specific linguistic and graphic tools:

Type #1direct hate speech;

Type # 2 indirect (hidden) hate speech;

Type # 3 manipulative hate speech

III. *Case study.* The present paper describes an interim study of hate speech found in the texts of nine online media in November 2020. We were looking for hate speech directed against certain national and religious groups living in Ukraine.

There were four stages. На першому етапі a team of journalists and volunteers with the help of a software bot carried out search query mining to find negative information according to key words.

Next (second stage), we looked through an array of texts to find articles that did not contain hate speech, but factual retrospectives of military events in the last century. We considered them as error deviation and did not consider in further work.

During the third stage, with applying the author's method of psycholinguistic analysis of the text, we scrutinized the material obtained to identify the type of hatespeech, its methods and techniques of influencing the consciousness.

At the fourth stage, we drew conclusions, which will be used in our further studies.

The texts that consist hate speech we split into three types based on the use of specific linguistic and graphic tools:

#1 direct hate speech: incitement to hatred through the use of obscenities, direct insults, calls to action on discrimination and violence, etc.;

#2 indirect (hidden) hate speech: opponents' dehumanization and marginalization, demonstration of disrespect, contempt for another ethnic group, culture, religion, distortion of historical facts.

#3 manipulative hate speech: employing means of influencing the individual's emotional state, whipping up negativity, opinions of biased "experts", amplification of information by non-linguistic means (for example, a photo that no relation to the event), misleading narratives.

RESULTS

The initial sample was composed of 306 texts. Approximately 70 % of the sample (215 texts) was selected for further analysis, while 91 texts were classified as an error because they did not contain any signs of hate speech. Hence, the efficiency of the bot was equal to 70.3%. It is a rather high value of software efficiency and effectiveness in finding hate speech in online media.

So, the third stage implies the textual analysis of 215 texts with hate speech. The psycholinguistic analysis made it possible to process all the texts and categorize them into three groups depending on the type of hate speech.

The array of texts for the specified period was lacking texts of # 1 type. At the same time, the results demonstrated that in most cases, media contained hate speech that belonged to the second and third types and avoided direct insults and discriminatory statements. Thus, contemporary media tend to keep distance from direct hate speech, but widely employ hate speech of the second (18.3%) and the third type (57.3%).

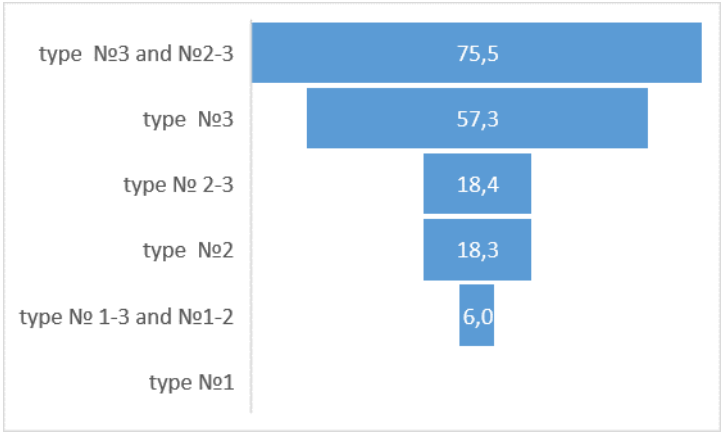
To spread the second type of hate speech, modern media predominantly exploit contempt, ridicule, and arrogance as well as opposing, deliberate exaggeration, or bracketing, which gives the word a figurative meaning "so-called", affirmative forms to elucidate events and refer to historical facts, etc.

As for the third type of hate speech, contemporary media mostly use tools and techniques aimed at affecting the individual's emotions, creating strong associations, and forming negative attitudes that the reader will consider a self-formed thought. In particular, such texts compile negatives information, references to non-professionals under the guise of experts who comment on the event without appropriate background, education, and experience in this field, distribution of fact-like fakes, and use photos that actually have nothing to do with the event. To validate the photo, we employed special photo verification programs (Krylova-Grek, 2019a; Krylova-Grek, 2019b).

Besides, modern media demonstrate a tendency to use hate speech of a mixed type. For instance, slightly under a fifth of the sample (18.4%) was represented by type #2-3 (a combination of the second and third types). This type provokes negative attitudes towards a certain group of people with the help of indirect. Less common (6 %) are #1-3 and #1-2 mixed types (combination of the first with the third and second type, respectively). These types drive a negative word-characteristic to an absolute axiom, which influences the content of the whole text (Figure 1).

Since the third and mixed types of hate speech have an indirect and hidden impact on the individual's consciousness, we consider #3 and #2-3 types as the methods of psycholinguistic manipulation.

Figure 1: Percentage (%) distribution of hate speech in the media in November 2020



(Source: Own)

DISCUSSION AND CONCLUSION

The interdisciplinary approach proved the effectiveness of combining artificial intelligence-based software with psycholinguistic research. Since bot efficiency exceeded 70 %, it was a powerful tool for optimizing journalists' and experts' analysis of hate speech.

The evidence from this study has revealed the techniques that are currently used by online news outlets to discriminate against certain social groups based on nationality and religion. The analysis of the largest Ukrainian online media has shown that hate speech is mainly presented by indirect influence and manipulation. The selected array of texts for November 2020 contained 75.7% texts (57.3% + 18.4%) that used diverse methods of psycholinguistic manipulation. Taken together, these findings implicate an important role of media in shaping negative public opinion.

On the other hand, our investigations into this area are still ongoing, thus the results presented in the paper are interim.

However, considerable progress has been made in manifesting media trends and novel ways of influencing public opinion in the context of marginalization and hatred of certain groups in society.

The strong point of our study lies in specifying common methods and techniques of affecting audiences' consciousness.

The results of this study will be used for a project report as well as to develop training for pre-journalist students and human rights defenders.

ACKNOWLEDGEMENT

I would like to express my special thanks of gratitude to the civil organization «The Crimean Human Rights Group» for selecting and providing research materials.

REFERENCES

Aronson, J.and McGlone, M. S. (2009). Social identity and stereotype threat. In T. Nelson, *Handbook of Stereotyping, Prejudice, and Discrimination Research.* Psychology Press, NewYork. 153-178.

Arutyunova, N. D. (1990). Metafora I diskurs [Metaphor and discourse]. In Arutyunova, N.D., Teoriya metafory [Theory of metaphor]. Progress, Moscow, 5-33.

Bahador, B. and Kerchner, D. (2019). *Monitoring hate speech in the US media* (Working Paper). The George Washington University, Washington, DC.

Bahador, B. (2020). *Classifying and Identifying the Intensity of Hate Speech.* Available at: https://items.ssrc.org/disinformation-democracy-and-conflict-prevention/classifying-and-identifying-the-intensity-of-hate-speech/ (Accessed 6 April 2021).

Bartlett, J., Reffin, J., Rumball, N., and Williamson, S. (2014). Anti-social media. *Demos.* Available at https://www.demos.co.uk/files/DEMOS_Anti-social_Media.pdf?1391774638 (Accessed 6 April 2021).

Browne, M. N., and Keeley, S. M. (2018). *Asking The Right Questions: A Guide to Critical Thinking*.12th ed. Pearson.

Bulyigina T. and Shmelev A. (1997). *Linguistic conceptualization of the world (based on the material of Russian grammar) [Yazykovaya konczeptualizacziya mira (na materi ale russkoj grammatiki]*.. Shkola Yazyki russkoj kultury, Moscow.

Chen, Ying, (2011). *Detecting Offensive Language in Social Medias for Protection of Adolescent Online Safety. Ph.D. Dissertation.* The Pennsylvania State University.

Dawes, J. (2005.) *The language of war: literature and culture in the US from the Civil War through World War II*. Harvard University Press.

Elswah, M. and Howard, P. (2020). "Anything that Causes Chaos": The Organizational Behavior of Russia Today (RT). *Journal of Communication*. 70(5), 623–645.

Galtung, J. (1987). Language and war: is there a connection? *Current Research on Peaceand Violence*, 10(1*)*, 2-6.

Ghanea, N. (2012). The Concept of Racist Hate Speech and its Evolution overtime. *UN CERD. Thematic Discussion on Racist hate speech*. Available at: https://www.ohchr.org/documents/hrbodies/cerd/discussions/racisthatespeech/nazilaghanea.pdf (Accessed April 6 2021).

Giles, D .C. (2003). *Media psychology*. Lawrence Erlbaum Associates Publishers, Mahwah, N.J.

Giles, D. C. (2016). Observing real-world groups in the virtual field: The analysis of online discussion. *British Journal of Social Psychology*, 55(3), 484-498.

Herman, E and Chomsky, N. (2002). *Manufacturing Consent: The Political Economy of the Mass Media.* Pantheon Books.

Hern, A. (2016). Facebook, YouTube, Twitter and Microsoft sign EU hate speech code. *The Guardian*, 31.5. Available at: https://www.theguardian.com/technology/2016/may/31/facebook-youtube-twitter-microsoft-eu-hate-speech-code (Accessed April 6 2021).

Howard, J. W. (2019). Free Speech and Hate Speech. *Annual Review of Political Science*. 22, 93–109.

Kamalipour, Y. R. (2004). Language, media, and war: Manipulating public perception. In Berenger, R. D, *Global Media Go toWar*. Marquette Books, Spokane, WA, 87-94.

Keith, A. and Burridge, K. (1991). *Euphemism & Dysphemism: Language Used as Shieldand Weapon*. Oxford University Press.

Krylova-Grek Y. (2019). 'Advanced Information Technology Toolsf or Media and Information Literacy Training', in *ICTERI Workshops*, pp. 229-240.

Krylova-Grek, Y. (2019.) 'Photo verification with mobile internet devices as the way to improve media literacy', in *Proceedings of the 14th international conference DisCo 2019: Overcoming the challenges and Barriers in Open Education*, pp 271-279.

Leontiev, D. (2004). *The hidden emotional content of the texts of mass media and methods of its objective diagnostic* [Skryitoe emotsionalnoe soderzhanie tekstov SMI i metodyi ego ob'ektivnoy diagnostiki], Smyisl, Moscow.

McGlone, M. S., Beck, G. A. and Pfiester, R. A. (2006). Contamination and camouflage in euphemisms. *Communication Monographs,* 73, 261-282.

Melvern, L. (2004). *Conspiracy to Murder: The Rwandan Genocide*. Verso.

Moscovici, S. (1985). *The Age of the Crowd: A Historical Treatise on Mass Psychology.* Cambridge University Press.

Paasch-Colberg, S., Strippel, C., Trebbe, J., Emmer, M. (2021). From Insult to Hate Speech: Mapping Offensive Language in German User Comments on Immigration. *Media and Communication,* 9(1), 171-180.

Pack, J. L. (2009). Book Review: The Language of War by Steve Thorne, 2006. Routledge, London and New York, *Language and Literature*, 18(4), 408–410.

Pocheptsov, G. G. (2001). *The theory of communication* [Teoriya kommunikatsii]. Reselbuk, Moscow.

Pocheptsov, G. G. (2019). (*Des)information.* Palivoda, Kiev.

Rizun, V. V., Nepyivoda, N. F., Kornieiev, V. M. (2005). *Linguistics of influence. monograph* (Linhvistyka vplyvu. Monohrafiia). Vidavnicho-polygraphic center "Kiev University".

Saleem, H. M., Dillon, K. P., Benesch, S., and Ruths, D. (2017). *A Web of Hate: Tackling Hateful Speech in Online Social Spaces.* Available at: https://arxiv.org/abs/1709.10159 (Accessed November 1, 2020).

Schmidt, A. and Wiegand, M. (2017). 'A survey on hate speech detection using natural language processing', in *Proceedings of the Workshop on Natural Language Processing for SocialMedia (SocialNLP'17)*, pp. 1-10.

Shmelev, D. (2008). Euphemism [evfemizm]. In Karaulov, Y. N. *Russkij yazyk. Enciklopediya* [Russian language. Encyclopedia]. 2nd ed. Bolshaya ros. Enciklјpediya. Drofa, Moscow.

Strong, E. (1922). Control of Propaganda as a Psychological Problem. *The Scientific Monthly*, 14(3), 234-252.

Waldron J. (2012). *The Harm in Hate Speech*, Harvard University Press, Cambridge, MA and London, England.

Yanagizawa-Drott, D. (2014). Propaganda and Conflict: Evidence from the Rwandan Genocide , *The Quarterly Journal of Economics,* 129(4), 1947–1994.