Identifying Online Hate Speech Using Artificial Intelligence

ÖSSZEFOGÁS
A GYŰLÖLET ELLEN

Research Report

Contents

Bevezető

The "CHAD: Countering Hate Speech and Hurtful Speech against Diversity: Roma, LGBTIQ, Jewish and Migrant Communities" is a two-year project carried out by RGDTS Nonprofit Kft., Budapest Pride, Haver Jewish Informal Educational Foundation and Political Capital with the aim of enabling the participating organizations and vulnerable, Roma, LGBTQ+, Jewish, migrant and Muslim communities to take effective action against intolerance, racism, xenophobia, homo- and transphobia, and discrimination . The project prepares them for the recognition, identification and monitoring of online hate speech and hurtful speech through exchanges of experiences, trainings and conferences. Based on the acquired knowledge, in the later stages of the project, the participants jointly develop counter- and alternative narratives and use them in the context of local actions, workshops, and social campaigns, sensitizing members of mainstream society to the consequences of hate speech. Another goal of the project is to create a community involving national and international organizations, experts, activists and decision-makers, which can fight online hate speech in the long term.

Methodology

In the first phase of the project, between August 2022 and February 2023, we monitored hate speech (incitement to hatred and violence) and hurtful expressions appearing online, targeting the Roma, LGBTQ+, Jewish, migrant and Muslim communities. We have created an almost completely automated (without human intervention, but with human validation) system which is suitable for finding texts that most likely contain hate speech or hurtful speech based on keywords (see Appendix III.: List of keywords) in Hungarian that most likely contain hate speech or offensive expressions (see Appendix I: Explanation of expressions). These texts were manually downloaded and sorted into Excel tables, along with their sources and dates of publication. In the next step, the volunteers previously trained for the task classified these texts according to a predetermined system of criteria, as hate speech or hurtful speech. The final result is a table that includes those texts (along with their sources) that most likely contain hate speech or hurtful speech.

In addition to the data set, an artificial intelligence algorithm was created that can learn from human decisions. Similarly to a human, if it encounters examples of hate speech and non-hate speech, it can tell the difference between them with high accuracy.

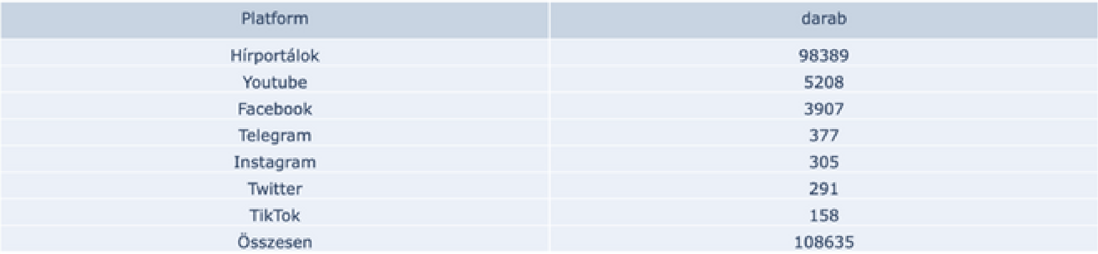
04

Data

We had two options for the monitoring: searching for hate speech on the Internet manually, or using some kind of computerized, i.e. automated, solution. In the absence of a sufficient number of specialists with adequate time, the search for these texts by people would only have been possible arbitrarily. In other words, we could have read online content randomly. This is a less efficient solution, but it is also very time-consuming. Computer-assisted "semi-automatic" hate speech monitoring is significantly more effective and is used in many places around the world. It is semi-automatic because at the moment the technology is not yet there to create an accurate and fully automatic solution in Hungarian without human expertise or intervention.

The created algorithms collected data from websites predetermined by the project partners; news portals, blogs, Facebook, TikTok, Twitter, Youtube, Instagram and Telegram pages (see Appendix II.: List of investigated websites), and then filtered and scraped (see Appendix I.: Explanation of terms) the texts that contain predetermined, jointly defined words or phrases that potentially refer to hate speech or insults directed at minority groups (LGBTQ+, refugee, Muslim, Roma, Jewish). (see Annex III.: List of keywords).

We downloaded the following amount of texts from the selected websites:



Platform	darab
Hírportálok	98389
Youtube	5208
Facebook	3907
Telegram	377
Instagram	305
Twitter	291
TikTok	158
Összesen	108635

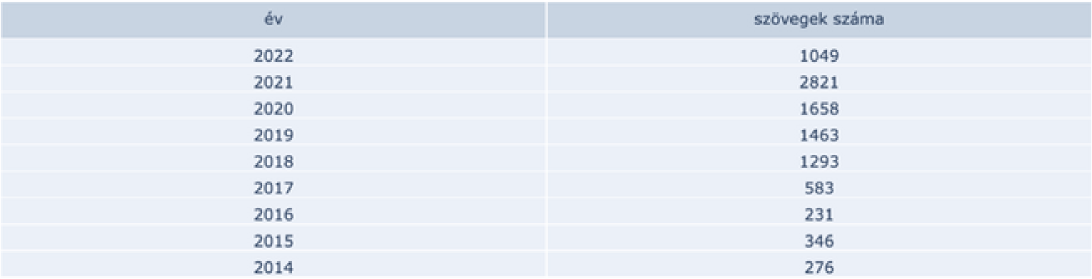
The table shows that more than 90% of all data was collected from news portals. The reason for this is, on the one hand, that among the selected pages there were some that had very few posts or comments overall, and on the other hand, for technical reasons, it was easiest to collect a large amount of text from these pages, while it was only sporadically possible from the others.

From certain portals it was quite difficult to download texts because they constantly change their program code, which must be decrypted regularly (even within days) in order to automatically download texts from them. We worked with an international team, because the scraping of each portal requires different expertise and experience.

A total of 108,635 texts were collected from the Internet. When we collect texts from the web, we download all texts without filtering. We performed a first filter on these texts using a keyword search. The number of texts filtered using the keywords is 11,354. We sent these texts to the monitoring team, who reviewed and flagged the texts they deemed hurtful or hate speech.

The examined period spans roughly 9 years from 2014 to May 2022. Texts published earlier or later are not included in the data.

The temporal distribution of the texts is shown in the following table:



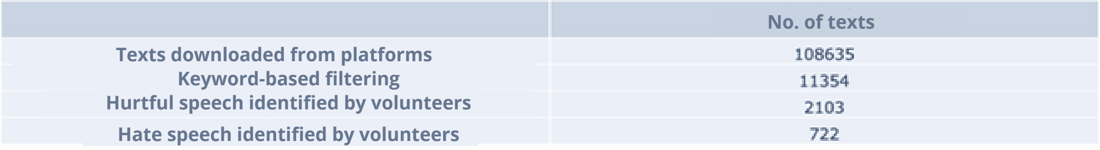
év	szövegek száma
2022	1049
2021	2821
2020	1658
2019	1463
2018	1293
2017	583
2016	231
2015	346
2014	276

It can be seen that there are more and more texts as we approach the present. This may also result from the fact that certain websites archive their content, and they can no longer be found and downloaded. The year 2022 is a partial year, so the number of texts there is less.

Results

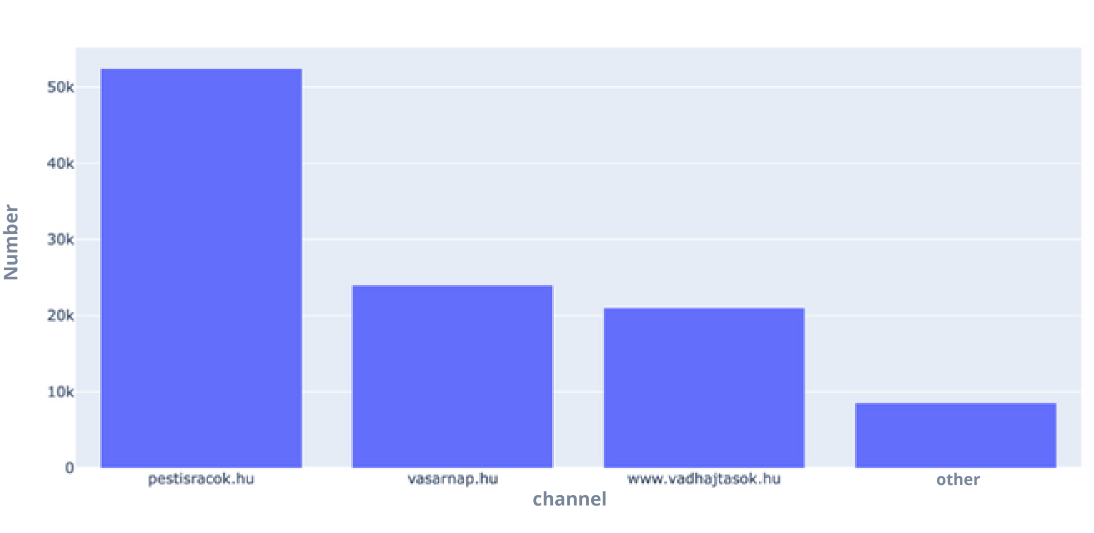
Each of the participating organizations and communities delegated volunteers (14 in total) who, after theoretical, legal and technical preparation, reviewed the texts and decided whether they contained hurtful speech or even hate speech. During the 6 months of monitoring we kept in constant contact with the volunteers not only to answer their technical questions, but also to resolve their tensions during the work.

The team deemed 2,103 texts offensive and classified 722 as hate speech. They found insults or incitement in 25% of the texts filtered by keyword search:



	No. of texts
Texts downloaded from platforms	108635
Keyword-based filtering	11354
Hurtful speech identified by volunteers	2103
Hate speech identified by volunteers	722

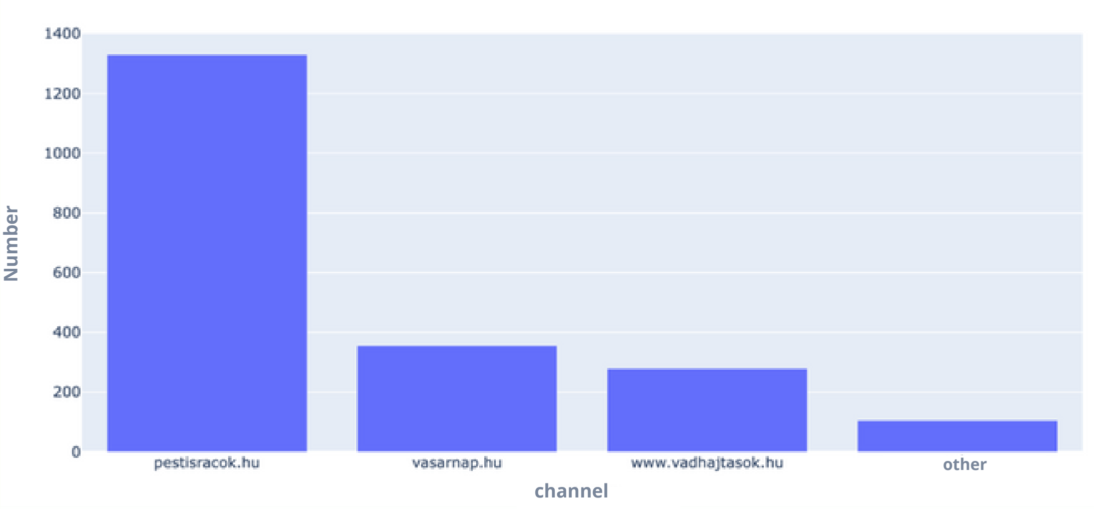
The following figure shows the sources of the texts sent to the monitoring team. More than 50% of the texts come from *pestisracok.hu*. It should be emphasized that all texts filtered by keyword search are visible here, and among the keywords there are expressions that are not considered offensive at all in a certain context, or that the author of the article or post is not responsible for it. For example, it happens that a journalist quotes a politician. The quote itself is classified as hate speech, but the journalist quotes the given statement to draw attention to its inciting nature. Another common case is the appearance of the word "Gypsy", which, depending on the context and the identity of the speaker, does not always project an insult or hate speech.


Number
50k
40k
30k
20k
10k
0
pestisracok.hu
vasarnap.hu
www.vadhajtasok.hu
other
channel

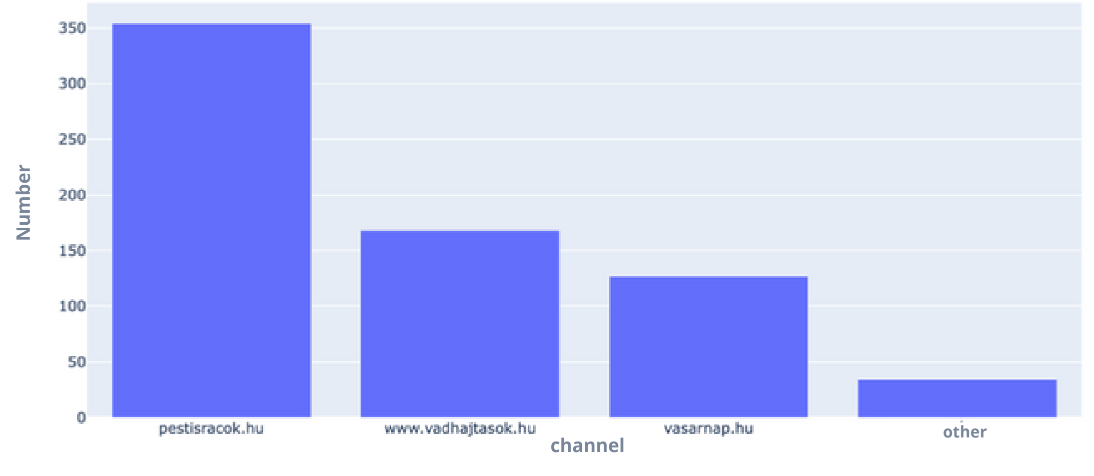
Number of texts containing hate speech and insults by channel

The following figures show the number of texts that our team found to be hurtful or contain hate speech, broken down by source. It can be seen that pestisracok.hu leads in both categories. Vadhajtasok.hu and vasarnap.hu contain quite a lot of insults and hate speech for certain groups, which is why they are shown in separate columns here as well.

Number of texts containing hurtful speech on each channel:



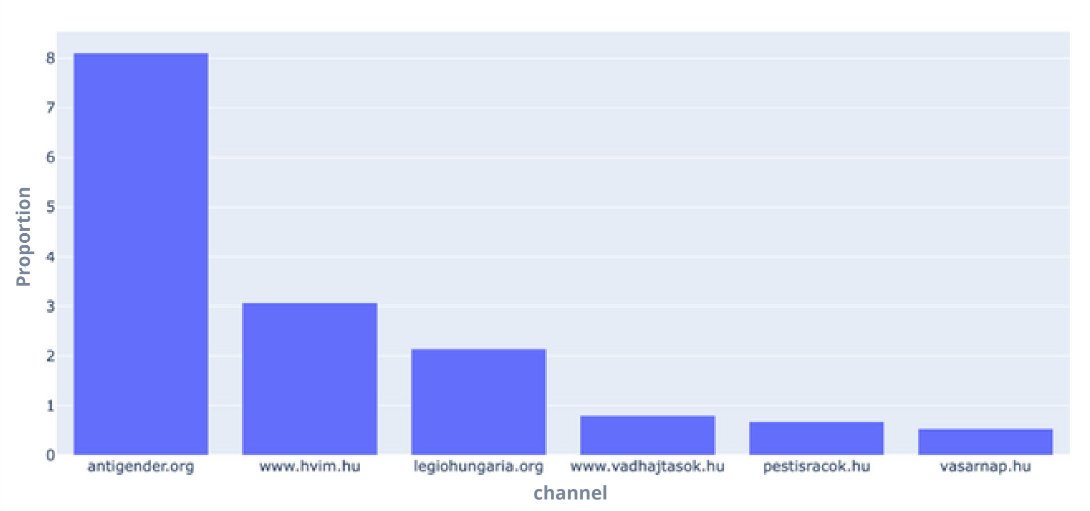
Number of texts containing hate speech on each channel:



Proportion of texts containing hate speech and insults on different platforms

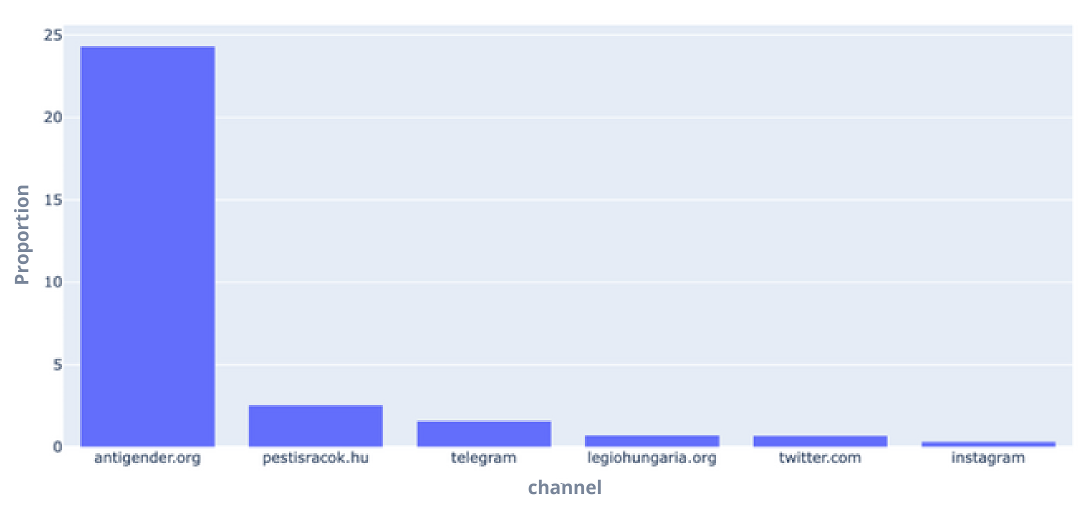
The figure below shows us in what proportion each website contains hate speech. For example, almost 8% of texts from antigender.org were classified as hate speech - this means that 92% of the texts found here are not hate speech.

The proportion of texts containing hate speech compared to all texts on the given portal:



This figure gives us a more nuanced picture, it shows us that although texts from pestisracok.hu contain the biggest number of probable hate speech, they "only" make up less than 1% of all texts on pestisracok.hu. It can be seen that even pestisracok.hu and vasarnap.hu are below 1%. However, the rate of hate speech on antigender.org, hvim.hu and legiohungaria.org is several times higher than 1%. It can therefore be said about these pages that they contain a lot of hate speech.

The proportion of texts containing hurtful speech compared to all texts on the given portal:

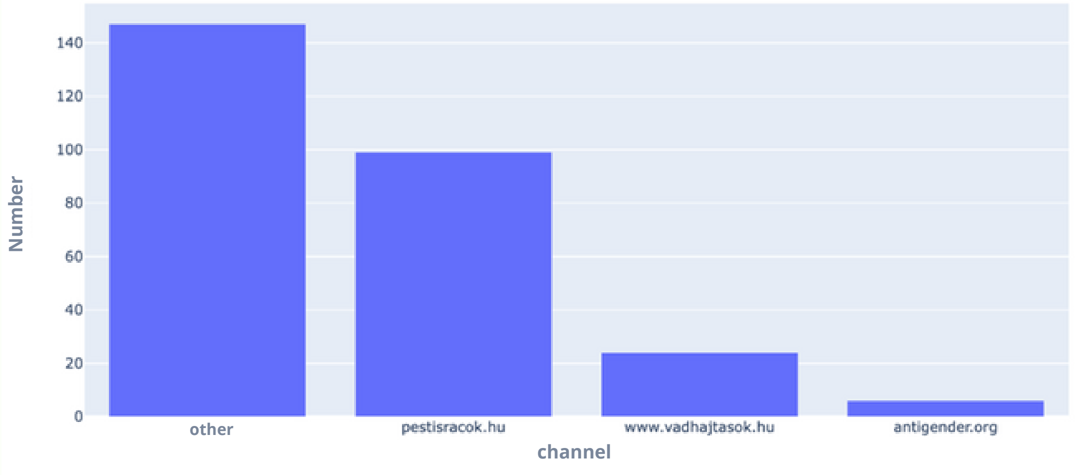


When looking at hurtful speech, antigender.org leads the way with a rate of nearly 25%. This means that the team found the quarter of texts from antigendender.org to be offensive. pestisracok.hu came in second place, where roughly every fortieth text is considered hurtful.

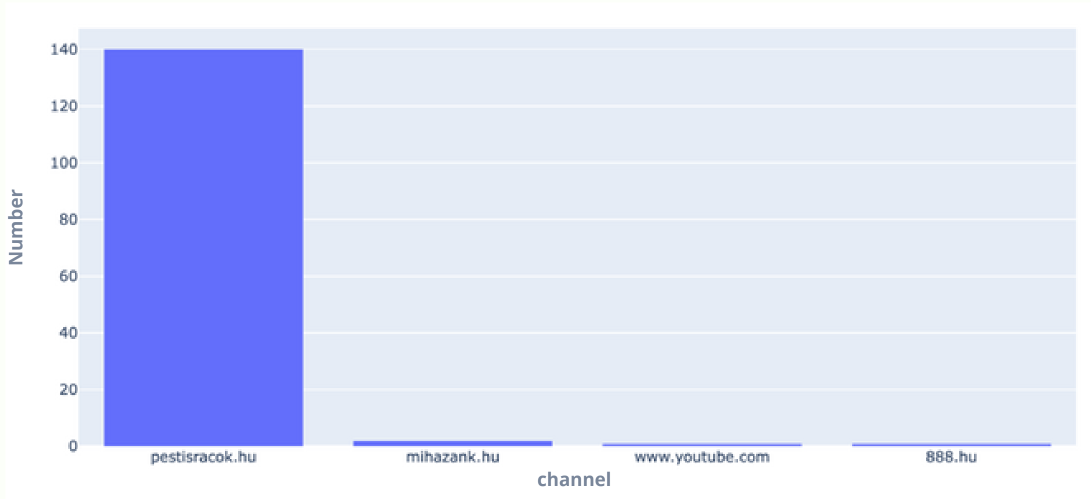
The number of texts containing hate speech against certain groups on different channels

Each of the keywords is related to a group or widespread stereotypes about the group, and in some cases to conspiracy theories. If we look at the targets of the hate speech appearing on various pages, we get an interesting result. We see that pestisracok.hu contains a lot of hate speech against all groups. However, vadhajalsok.hu has published a lot of hate speech against LGBTQ+, Roma and Jewish groups, but none targating refugees and Muslims. Among these, the Jewish group stands out, because the team found the most hate speech against them on vadhajtasok.hu.

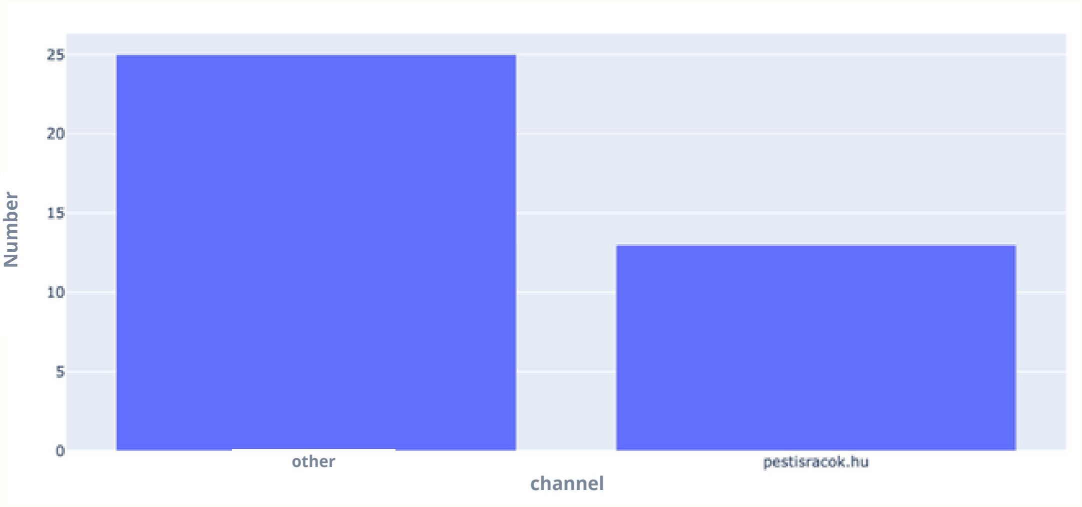
LGBTQ+



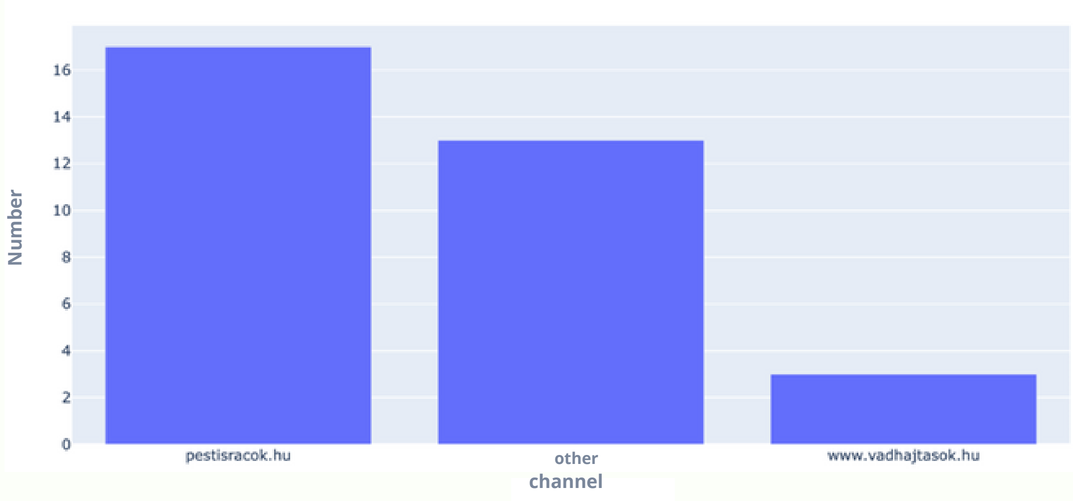
Refugee



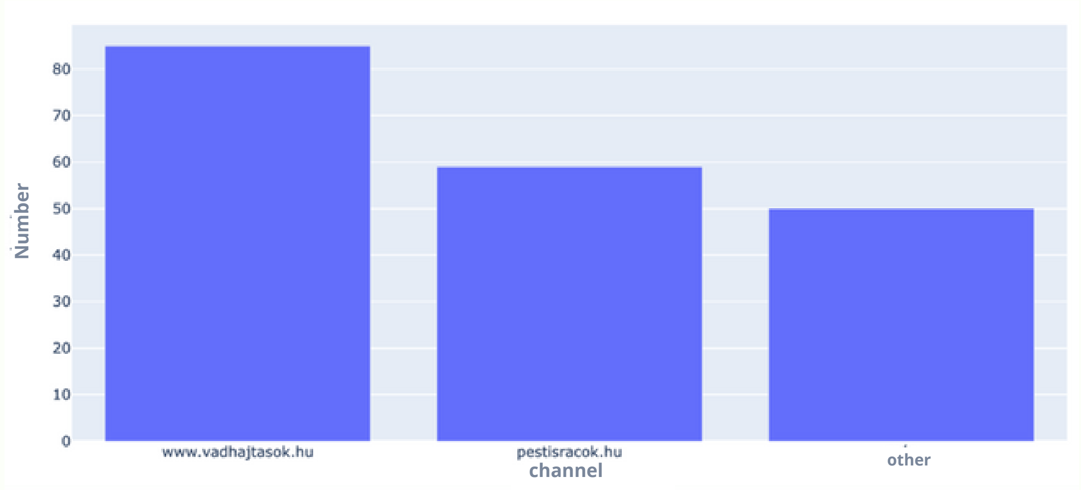
Muslim



Roma

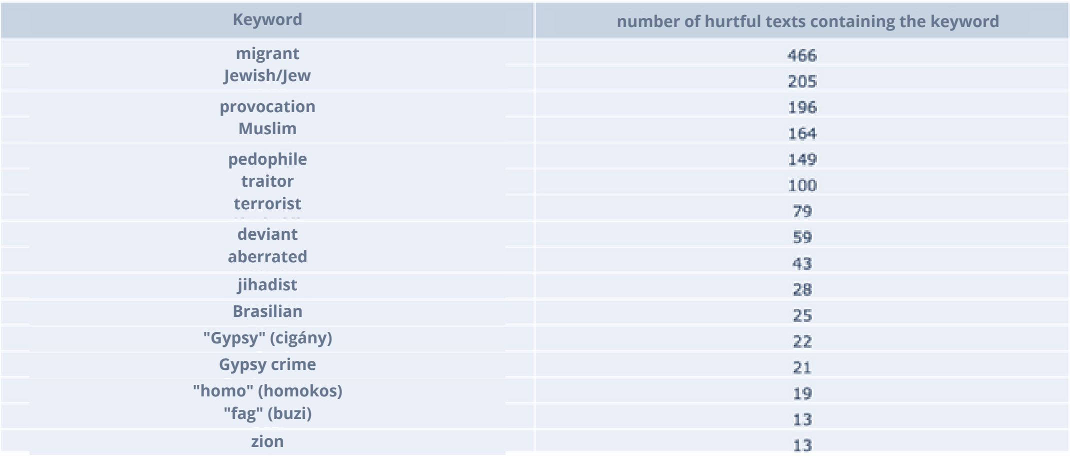


Jewish



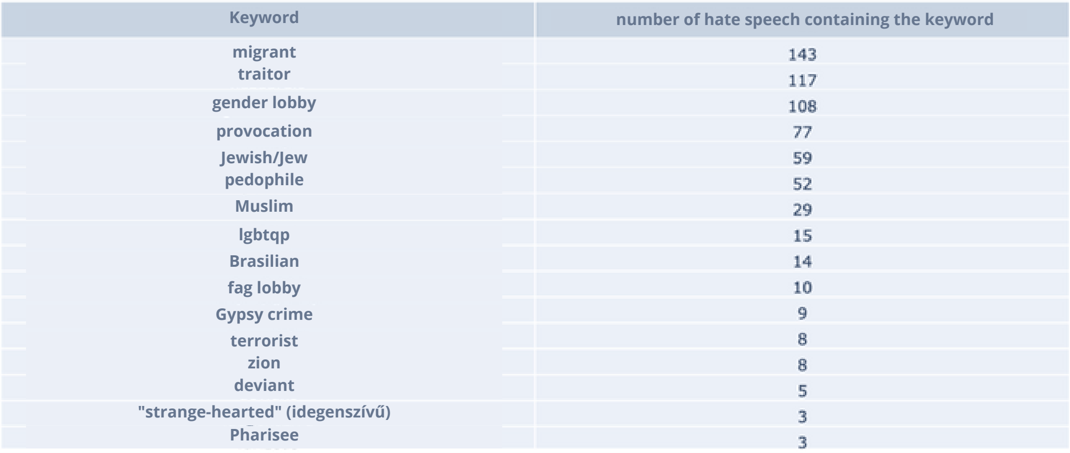
The number of texts containing insults and hate speech for each keyword

IThis table shows how many hurtful speech are connected to each keyword. This table is not yet broken down into groups, so keywords against LGBTQ+, refugee, Muslim, Roma and Jewish groups appear together.



Keyword	number of hurtful texts containing the keyword
migrant	466
Jewish/Jew	205
provocation	196
Muslim	164
pedophile	149
traitor	100
terrorist	79
deviant	59
aberrated	43
jihadist	28
Brasilian	25
"Gypsy" (cigány)	22
Gypsy crime	21
"homo" (homokos)	19
"fag" (buzi)	13
zion	13

The following table shows the number of texts containing hate speech, belonging to the keywords:



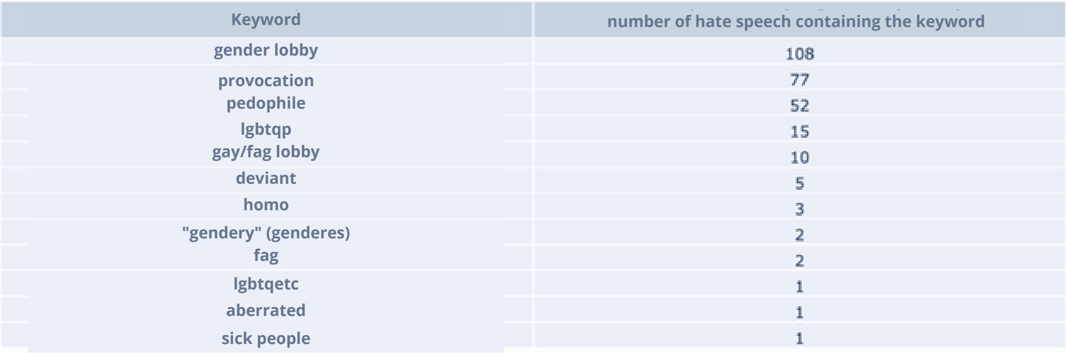
Keyword	number of hate speech containing the keyword
migrant	143
traitor	117
gender lobby	108
provocation	77
Jewish/Jew	59
pedophile	52
Muslim	29
lgbtqp	15
Brasilian	14
fag lobby	10
Gypsy crime	9
terrorist	8
zion	8
deviant	5
"strange-hearted" (idegenszívű)	3
Pharisee	3

The above tables are not suitable for comparing the number of hate speech directed against different groups, and for drawing conclusions about which groups are most exposed to hatred in Hungary. Determining offensive speech and hate speech is a rather subjective task due to the broad interpretation of the legislation, and although the volunteers were also trained on the legal interpretation of incitement against the community, it is conceivable that certain groups were monitored by people who are more sensitive to insults than other groups.

It is useful to divide the above table into groups. In this way the keywords used against each group (and their frequency) in the texts classified as hate speech becomes visible.

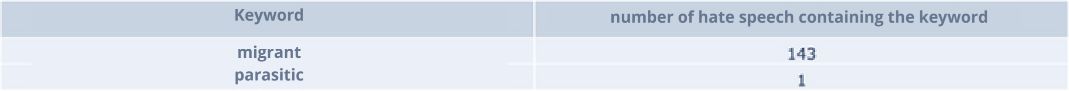
Number of texts containing insults and hate speech related to keywords, broken down by group

Keywords related to LGBTQ+ people found in hate speech:



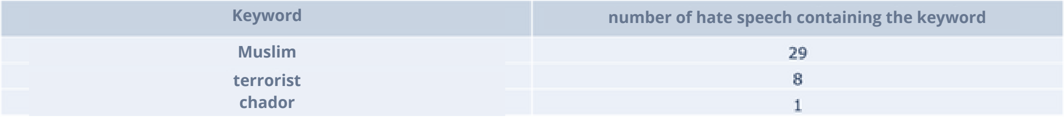
Keyword	number of hate speech containing the keyword
gender lobby	108
provocation	77
pedophile	52
lgbtqp	15
gay/fag lobby	10
deviant	5
homo	3
"gendery" (genderes)	2
fag	2
lgbtqetc	1
aberrated	1
sick people	1

Keywords related to regugees found in hate speech:



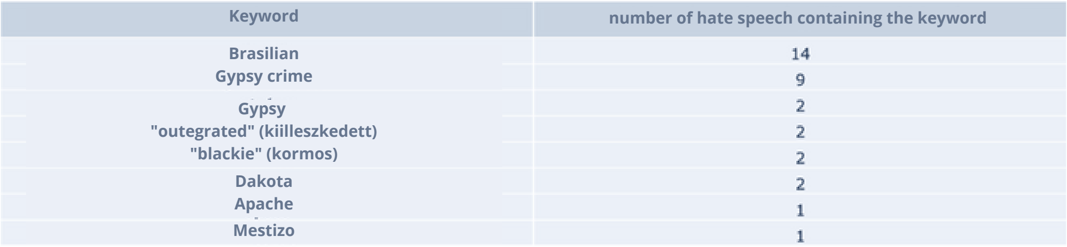
Keyword	number of hate speech containing the keyword
migrant	143
parasitic	1

Keywords related to Muslim people found in hate speech:



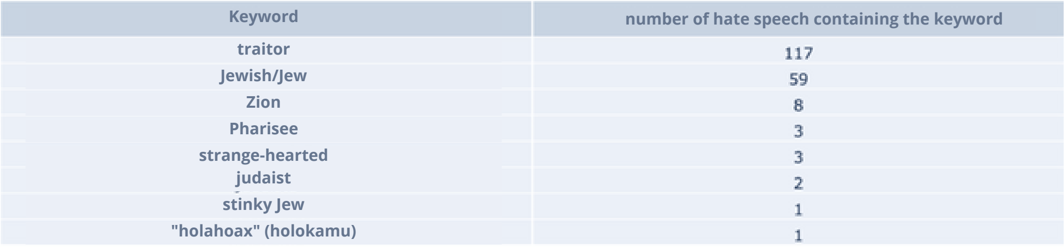
Keyword	number of hate speech containing the keyword
Muslim	29
terrorist	8
chador	1

Keywords related to Roma found in hate speech:



Keyword	number of hate speech containing the keyword
Brasilian	14
Gypsy crime	9
Gypsy	2
"outegrated" (kiilleszkedett)	2
"blackie" (kormos)	2
Dakota	2
Apache	1
Mestizo	1

Keywords related to Jewish people found in hate speech:



Keyword	number of hate speech containing the keyword
traitor	117
Jewish/Jew	59
Zion	8
Pharisee	3
strange-hearted	3
judaist	2
stinky Jew	1
"holahoax" (holokamu)	1

Creation, training and results of the artificial intelligence model

We created a neural network-based artificial intelligence algorithm that can classify whether individual texts contain hate speech or not. The results are quite good. To teach the model we used texts and their corresponding labels (hate speech, hurtful speech, none).

Testing the system on new texts we get the following results:

- True Negative (TN): 2,472 pieces of text are not hate speech, and the artificial intelligence does not classyfy them as such either
- True Positive (TP): 44 texts contain hate speech according to the artificial intelligence, and these were classified as hate speech by the monitoring team as well
- False Positive (FP): 250 texts are deemed to be hate speech by the artificial intelligence , although according to the team, they are not
- False Negative (FN): the algorithm classifies 6 pieces of text as not hate speech however, according to the monitors, they are

This means that by using artificial intelligence the human capacity needed to monitor hate speech can be significantly reduced in the future, because the monitoring team only has to check and validate texts that contain hate speech according to the artificial intelligence.

The Limitations of the Research

The list of resources, i.e. the surfaces to be motorized, was compiled by the participating organizations in the initial phase of the project. This mostly included portals, pages, and channels on which offensive content and hate speech can be expected to appear, but also a small number of general news portals. Since there is currently no technical possibility to examine all Hungarian-language articles and posts appearing on the World Wide Web, this narrowing greatly limited the research.

The keywords used were on the one hand borrowed from a previous online hate speech monitoring project, and on the other hand some were collected from the experiences of the participating organizations and volunteers. Looking at the results and the list, it is clear that there are many words and phrases that, although they imply insult or incitement, did not appear in the texts, or if they did appear, their conexts were not related to the corresponding groups. The list of keywords should therefore be reviewed from time to time during the scraping and monitoring, and a new list should be prepared during a possible continuation or further development, including the content of the hare speech and hurtful speech found during this research. However, with the training and application of artificial intelligence, the keyword list is pushed into the background, since there is no need to pre-filter the texts, the algorithm can quickly review all of them.

Although we have the legal concept of incitement against members of certain communities in force in Hungary, whether a text is really suitable for inciting hatred, strengthening it, or even calling for violence depends heavily on the context of the text, current events in the world and the values held by the audience - the sensitivity, identity, but also the current state of mind of the reader/listener. It is therefore very subjective what someone who is not an expert on the subject considers to be incitement, insult, or none of these. The results are also influenced by the fact that the collected texts were classified by 10-13 young people, most of them belonging to at least one of the affected groups.

Annexes

Annex I.: Explanation of terms

Hate speech

In this report hate speech refers to Section 332 of the Hungarian Criminal Code:

"Incitement Against a Community

Any person who before the public at large incites hatred against:

a) the Hungarian nation;

b) any national, ethnic, racial or religious group; or

c) certain societal groups, in particular on the grounds of disability, gender identity or sexual orientation;

is guilty of a felony punishable by imprisonment not exceeding three years."

Hurtful speech

Statements which cannot be considered hate speech legally, but are offensive and hurtful to certain groups. Incitement is not conspicuous in these texts, but they contain generalizations, degrading expressions, or false statements that are offensive to the members of the given group.

Scraping

Downloading data (in this case texts, their creation date, etc.) from various websites with a programmed solution.

Group

LGBTQ+ people, refugees, Muslim people, Roma, Jewish people

Appendix II.: List of investigated websites

Wbsites
https://888.hu
https://antigender.org
https://hvim.hu
https://legiohungaria.org
https://mihazank.hu
https://pestisracok.hu
https://vadhajtasok.hu
https://zoldinges.net

Facebook
Borgula András facebook:
https://www.facebook.com/BorgulaAndras

Dúró Dóra facebook:
https://www.facebook.com/durodora

Fidesz facebook:
https://www.facebook.com/FideszHU

Gyurcsány Ferenc facebook:
https://www.facebook.com/gyurcsanyf

Haver Foundation:
https://www.facebook.com/HaverAlapitvany

Mérce facebook:
https://www.facebook.com/magyarinfo/posts/10159992878893467

Tibi atya:
https://www.facebook.com/tibiatya

Instagram

Mi Hazánk Mozgalom instagram
https://www.instagram.com/mi_hazank_mozgalom/

Mi Hazánk Ifjai instagram
https://www.instagram.com/mihazankifjai/?hl=de

Telegram

Hatvannégy Vármegye Ifjúsági Mozgalom telegram
https://t.me/varmegye

Légió Hungária telegram
https://t.me/legiohungaria

Mi Hazánk Ifjai telegram
https://t.me/mihazankifjai

Blood And Honour Hungary telegram
https://t.me/BloodandHonourHungary

Mi Hazánk Mozgalom telegram
https://t.me/mihazank

TikTok

Hatvannégy Vármegye tiktok:
https://www.tiktok.com/@hatvannegyvarmegye

Mi Hazánk Mozgalom tiktok:
https://www.tiktok.com/@mihazank?lang=hu-HU

Twitter

Torockzai László twitter:
https://twitter.com/ToroczkaiLaszlo

Mi Hazánk Mozgalom twitter:
https://twitter.com/MiHazank

Youtube

Budaházy György youtube:
https://www.youtube.com/channel/UC6DulUMd-hi925V4uTn7smw

Budapest Pride youtube:
https://www.youtube.com/channel/UCvZtOhw04DojlZXcK_EmuBw

Kontra youtube:
https://www.youtube.com/c/Kontrahu/videos

Hatvannégy Vármegye Ifjúsági Mozgalom youtube:
https://www.youtube.com/channel/UCvZtOhw04DojlZXcK_EmuBw

Haver Alapítvány youtube:
https://www.youtube.com/results?search_query=haver+alap%C3%ADtv%C3%A1ny

Magyar Harcos youtube:
https://www.youtube.com/user/betyarosvideok?app=desktop

Mi Hazánk Mozgalom youtube:
https://www.youtube.com/c/MiHaz%C3%A1nkMozgalom

Mi Hazánk Ifjai youtube:
https://www.youtube.com/channel/UCILN_gM2b3Truz8_nvpJt3Q

Annex III.: List of keywords

Muszlim: muszlim, terrorista, burkás, csador, dzsihadista, hárem, nőgyűlölő, mocskos terrorista, muszlim terrorista, iszlám terrorizmus, fekete járvány, bevándorló, dzsihád, iszlám erőszak, szent gyűlölet, tevebaszó, iszlamista terrorista, felrobbantotta magát, dzsihád, iszlámfasiszta, saria, káfir, hidzsáb, sharia, abroszfejű.

Muslim(rough translations): Muslim, terrorist, burqa, chador, jihadist, harem, misogynist, filthy terrorist, Muslim terrorist, Islamic terrorism, black plague, immigrant, jihad, Islamic violence, holy hate, camel fucker, Islamist terrorist, Blown himself up, Islamofascist, sharia, kafir, hijab, sharia, cloth head.

Zsidó: zsidó kutya, biboldó, görbe orrú, ferde orrú, büdös zsidó, bibsi, judaista, zsidóskodik, fukar zsidó, nyalka tetű, koszos zsidó, hazaáruló, soros fajta, zsidó komcsi, zsidóvagyon, menjenek haza izraelbe, zsidóbűnözés, zsidó lobbi, goj kizsákmányolása, bezsidó, bolsevizmus árja, cion bölcsei, cionista evolúció, csúfot űznek a hitből, degenerált világfajta, degenerált zsidó, elkorcsosítás, elzsidósítás, farizeus, galicianerek, izrael-teológia, idegenszívű, jézus árja, jézus-gyilkosok, jézus szkíta, Júdás zsidó, judaizmus gyakorlatiassága, judaizmus szabadkőművesség, judeobolsevizmus, judeo-kereszténység, judeokrisztianizmus, kiválasztott faj, kozmopolita internacionalizmus, krisztus-gyilkosok, liberális zsidó álnokság, talmud értelmetlenség, talmud másodlagos, tízedet szednek testvéreiktől, újra megfeszítik isten fiát, zsidó materializmus, zsidó morál, zsidó nacionalista, holokamu, zsidrák, cion.

Jewish (rough translations): Jewish dog, biboldo, crooked nose, smelly Jew, bibsi, Judaist, Jew-doing, stingy Jew, sleazy louse, dirty Jew, traitor, Soros breed, Jewish commie, Jewish wealth, go home to Israel, Jewish crime, Jewish lobby, exploitation of goy , Aryan of Bolshevism, sages of Zion, Zionist evolution, making a mockery of the faith, degenerate cosmopolitan type, degenerate Jew, degenerate, Judaization, Pharisee, Galicians, Israel-theology, foreign-hearted, Jesus-killers, Jesus the Scythian, JudasJew, practicality of judaism, judaist freemasonry, judeobolshevism, judeo-christianity, judeo-christianism, chosen race, cosmopolitan internationalism, Christ-killers, liberal jewish deceit, talmudic nonsense, talmudic secondary, tithing from their brothers, crucifying the son of god again, jewish materialism, jewish morality, jewish nationalist, Holocaust, Jews, zion.

Roma: génhulladék, orkok, kormos, tetves cigány, napbarnított, tolvaj cigány, apacs, vinettu, mesztic, rezesek, kolompár, rézműves, dakota, brazil, cigánybűnözés, cigány bűnözés, cigány anyád, büdös cigány, rohadt cigány, tolvaj cigány, cigu, köcsög cigány, cigó, rohadt cigány, koszos cigány, cigánymentes, gyilkos cigány, lakatos nintendó, kannás, kanalas, ármándó, lakatos ármándó, degenerált cigányok, anyabaszó cigányok, mocskos élősködők, segélyesek, dizsipszi, gypsy, dzsipó, családipótlékosok, cigány tetvek, hosszú szoknyások, rózsás szoknyások, ómigráns, kiilleszkedett, svéd, rejtett erőforrás, rejtett erőforrásék.

Roma (rough translations): genetic waste, orcs, sooty, lousy gypsy, tanned, thieving gypsy, apache, Winnetou, mestizo, copper-dealers, coppersmith, dakota, brazilian, gypsycrime, gypsy crime, your gypsy mother, stinky gypsy, rotten gypsy, thieving gypsy, bastard gypsy, gypsy, bloody gypsy, dirty gypsy, gypsy-free, murderous gypsy, spoon maker, Armando, degenerate gypsies, motherfucking gypsies, dirty parasites, free-riders, gypsies, gypos, family benefitters, gypsy lice, long skirts, rosey skirts, old-migrant, outegrated, Swedish, hidden resources.

LGBTQ+: buzeráns, mocskos buzik, fartúró, fasszopó, transzi, köcsög, kötsög, köcsögfelvonulás, ratyifelvonulás, langyifelvonulás, langyi, bőrtangás, ferde hajlam, genetikai hulladék, genetikai selejt, beteg emberek, beteg állatok, evolúciós zsákutca, abnormális, magamutogatók, vonaglás, provokáció, lmbtqp, lmbtqstb, buzilobbi, genderlobbi, genderes, apache helikopter, pedofil, pedofil buzi, deviáns, homokos, seggbaszó, aberrált, deviánslobbi, fajti, buziterror, ratyi, divatbuzi, divatból.

LGBTQ+ (rough translations): faggot, dirty faggot, cocksucker, trans, faggot, , faggot parade, faggot race, gay parade leather thongs, skewed tendency, genetic waste, genetic reject, sick people, sick animals, evolutionary dead end, abnormal, exhibitionists, provocation, lmbtqp, lmbtqetc, gaylobby, genderlobby, gendered, apache helicopter, pedophile, pedophile faggot, deviant, homo, aberrant, deviant lobby, homo terror, fashion faggot, fashionable.

Menekült: migráns, migráncs, áradat, állati horda, betolakodó, menekült, élősködő, mohamedán fenevad, határsértő, hívatlan vendég, népvándorlás, néger, idegen, migráns horda, jövevény, kontinensfoglaló, megszálló, újparazita, sötét horda, honfoglalók, új európaiak.

Refugee (rough translations): migrant, migranch, flood, animal horde, invader, refugee, parasite, Muslim beast, border invader, uninvited guest, mass migration, negro, alien, migrant horde, newcomer, continent invader, invader, new parasite, dark horde, squatters, new Europeans.

POLITICAL CAPITAL
POLICY RESEARCH & CONSULTING INSTITUTE

ÖSSZEFOGÁS
A GYŰLÖLET ELLEN

Phiren Amenca
ROMA-GADJE DIALOGUE THROUGH SERVICE

BUDAPEST
PRIDE

Funded by the
European Union

haver

Funded by the European Union. Views and opinions expressed are however those of the author(s) onlyand do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the granting authority can be held responsible for them.”