CHATBOT ON HATE SPEECH

HIA POLAND GROUP 1

MAY LIM | RENA PITSAKI | EWA RODZIK

Our overarching goal is to reduce hate speech online as a step towards creating a safer and more peaceful environment for online users.

One comment at a time!

WARSAW | 3 JULY 2018

HUMANITY IN ACTION POLSKA

>> THE TEAM

Our team consists of three young, but passionate women determined to change Polish reality from three different sides of the world. We combine unique experiences, roots, backgrounds, and identities with a shared strong belief that nothing is impossible and that change will happen only through dedication, strong determination, bravery, but most importantly, through consistent action.

MAY LIM

May Lim graduated from the University of Washington in 2016 with bachelor's degrees in Political Science and Psychology. May has worked for local elected officials in both Seattle and LA and will be starting a master's program in public policy at UC Berkeley in August 2018.

RENA PITSAKI

Born and raised in Greece, Rena is an Art Historian and Art Curator. She is now completing her MSc studies on Cultural Technology and Communications, focusing on organising artistic projects, which highlight burning social issues.

EWA RODZIK

Born and raised in Warsaw, Ewa is a third year law student at University of Warsaw, with her academic interests including migration and refugee law, international humanitarian law and labor law. After graduation she plans to pursue her professional career as a human rights lawyer.

HUMANITY IN ACTION POLSKA

>> ABOUT THE PROBLEM

THE CHALLENGE

Our overarching goal is to reduce hate speech online as a step towards creating a safer and more peaceful environment for online users. Our general assumption is that by reducing the amount of hate speech online and encouraging an online culture based on respect, rather than hate, we can also reduce hate speech in society. We hope to achieve this by creating a chatbot targeting young male gamers. By using empathetic language, the chatbot will aim to build trust with the user in the preliminary stages of the conversation. The chatbot will then share educational and informational materials with the user regarding the harmful effects of hate speech on various populations. We will create a Facebook page which links to the chatbot in order for the chatbot to be more easily shared over social media.

TARGET GROUP

Our project targets male gamers in Poland between the ages of 12-17 years. Multiple studies (Oksanen et al, 2014; Hawdon, Oksanen & Räsänen, 2015) have shown that young individuals are highly prone to using hate speech in online games. A study done in 2017 by the Stefan Batory Foundation shows that the more individuals are exposed to hate speech, the more normalized hate speech becomes to them. As young game users become more desensitized to hate speech online, they will also become less affected by hate speech outside of the internet and will be more likely to use it themselves as well.

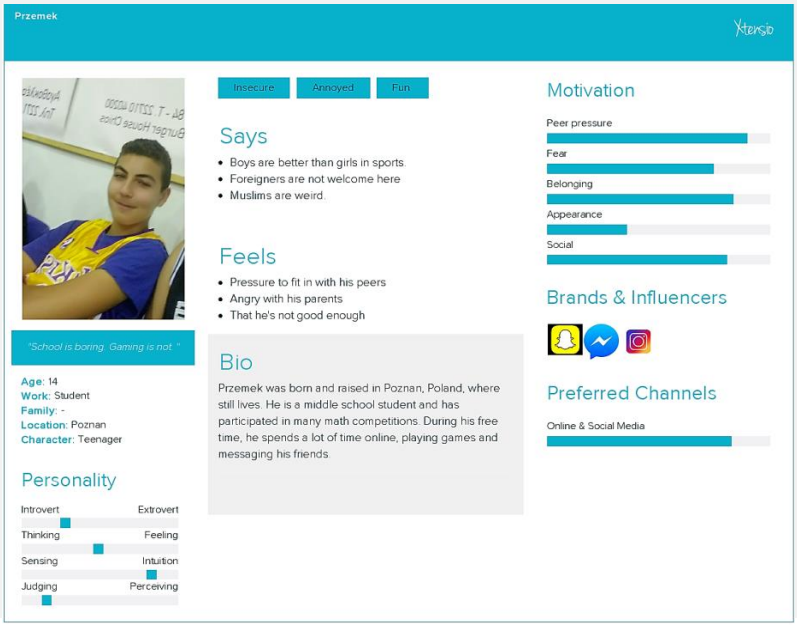
>>ABOUT THE SOLUTION

SOLUTION IDEA

To combat online hate speech made by young male gamers between the ages of 12-17 years, we tested a possible technological solution to educate and create awareness of the effects of hate speech among young online users. Therefore, we created a chatbot named **"Cursing is Half the Fun."** This chatbot will be linked to a Facebook page with the same name that falls into the category of "Gaming" on Facebook. The name is intended to attract young gamers, many of whom like to use curse words towards others on the internet. The use of cursing is sometimes linked to hate speech and other times not, but we believe it is important for young users to know the difference. Research has shown that individuals in our target group do not tend to understand the difference between hate and hate speech, so our chatbot aims to teach them the difference and to educate them on the effect of their hate speech on different groups. Our aim by doing so is to reduce the harmful effects of our target group's use of hateful language on online games.

IMPLEMENTATION PLAN

We started off by brainstorming for ideas using a persona named Przemek. We created this persona in order to understand the needs, pains, and gains of our target group. Przemek is a 14-year-old boy who enjoys playing online games with his friends. Attached below is a summary of Przemek's persona.



Przemek's pain is that he is not as good as his friends at gaming, which results in why he tries to match their behavior. His gain is playing games online and using the internet. A pain reliever for Przemek is to gain confidence and self-esteem, and a gain creator is for Przemek to keep using the Internet. We decided that a chatbot would be the best solution for Przemek. A chatbot incorporates the pain reliever by empowering Przemek with knowledge and education, helping him feel more secure and confident in his own thoughts instead of following along with his friends. It also incorporates the gain creator by giving him a reason to be online even more.

After coming up with our solution, we created language for the chatbot that would be both appealing to young users and also effective at educating young people on the negative effects of using hate speech online. We incorporated this language into relevant questions and answers for the chatbot. We supplemented this with other empathetic questions that would appeal to emotions, as well as some questions intended to make the user reflect and think about the impact of their words. We then inputted these answers and questions into a schema in order to create a prototype that could be tested among our target group. We created a Facebook page that would eventually link to the chatbot so that the link to the Facebook page could be posted to various Facebook threads and the comments of Youtube gaming videos that contained hate speech. The title of our Facebook page and chatbot is "Cursing is Half the Fun," intended to appeal to young users who typically use a great deal of cursing during online gaming.

>> ABOUT THE SOLUTION

QUESTIONNAIRE

IMPACT INDICATORS

After the process of creating our chatbot prototype, we tested our chatbot as a demo version through a schema so that our bot could be available for teenage boys living in Poland. This demo version allowed us to send our trial to them and get real results on how much time they spent on it, whether it was effective, and what their feedback was in general. Therefore, we created a questionnaire that followed up the conversation with the bot.

- Had you ever used a bot before in the past?
- Were there enough answers available as options for you?
- How much time did you spend with this bot?
- Did you find it old fashioned?
- Did you learn something new about hate speech?
- Would you share it with your friends?

>> ABOUT THE SOLUTION

PRELIMINARY RESULTS

After creating the chatbot, we tested it among five individuals in our target group using both a schema and a questionnaire to test the effectiveness of the chatbot. Among the five individuals in our trial, 4 out of 5 spent less than one minute on the chatbot. 3 out of 5 found that the language we used in our chatbot schema was not too old-fashioned. 3 out of 5 learned something new about hate speech from the bot, and 2 out of 5 said they would share with their friends.

Considering these results, we found that the chatbot was not effective in targeting individuals in our target group in terms of teaching them about hate speech and reducing their use of hate speech online. While this prototype was a good basis and trial for a hate speech-targeted chatbot, the chatbot must be improved in the future in order to be more attractive towards users in the target group.

We also reflected that an alternative medium could be more effective than chatbots in order to reduce hate speech online among Polish youth. To make an effective chatbot that understands the complex content that users in the target group will input regarding hate speech and for the chatbot to be able to answer appropriately to each answer from the user, much more financial resources, human resources, and time must be needed in order to improve the NLP (natural language processing) technology of the chatbot.

FINAL REFLECTIONS

This project was a learning experience for our group. We learned more about hate speech and more about the use of hate speech among young gamers, as well as the effect of this hate speech on minority target groups.

However, we found difficulty in finding a tech-savvy and interesting method to appeal to young gamers in our target group, especially given that our funding from our organization was already limited to creating a chatbot.

From our findings, we suggest that HIA Poland find alternative methods of using technology as a method of reaching young gamers and raising awareness about the harmful effects of hate speech in order to reduce hate speech both online and in the real world.

>> PICTURES

>> Visualisation for Pitch & Pizza Night

Której z poniższych fraz używasz najczęściej?

Ty ... wróć do kraju, z którego przybyłeś, nie chcemy Cię u nas w Polsce.

Hej dziewczynko/ kochanie, idź do kuchni, gdzie Twoje miejsce i zrób mi kanapkę!

Ty Żydzie!

Jesteś taki głupi, kto Ci w ogóle pozwolił grać w tę grę!

Ale z Ciebie pedał.

Więcej niż jednej.

Czy kiedykolwiek czułeś się wykluczony lub odsunięty ?

Tak, wiele razy.

Czasami.

Nie, nigdy.

Myślisz, że Twoja rodzina i przyjaciele kiedykolwiek czuli się wykluczeni lub odrzuceni?

Tak.

Nie.

Powiedziałbyś to samo do chłopca – Twojego kolegi, brata, wujka, czy Taty?

O nie, miejscem chłopaka nie jest kuchnia. Muszą robić lepsze i ważniejsze rzeczy.

Obejrzyj wideo, w którym jeden z najlepszych kucharzy na świecie - Gordon Ramsay – pokazuje, jak zrobić perfekcyjnego burgera!

https://www.youtube.com/watch?v=iM_KMYuII_s

Tak, wszyscy jesteśmy równi i mamy równe prawa.

Uchodźcy i migranci czują się wykluczeni głównie przez brak akceptacji ze strony polskiego społeczeństwa, a więc także Ciebie, Twoich znajomych i rodziny.

Zgodnie z badaniami przeprowadzonymi przez Centrum Badań nad Uprzedzeniami 2,86 (zgodnie z przyjętą skalą) osób badanych odpowiedziało, że los migrantów i uchodźców, przybywających do Polski powinien polegać na „wykluczeniu", a aż 4, 09 osób wybrało „przemoc fizyczną" rozumianą na przykład jako: „Uchodźcy powinni być najpierw umieszczani w obozach przejściowych, w których poddani byliby selekcji pod względem potencjalnych zagrożeń jak i potencjalnej przydatności w Polsce."

Oznacza to, że Polacy świadomie zgadzają się na wykluczenie społeczne uchodźców i migrantów.

Zastanów się, w jakich sytuacjach Ty lub Twoi bliscy czuli się wykluczeni. Dlaczego? Ponieważ ubierasz się inaczej? Ponieważ Twoja rodzina ma mniej pieniędzy? Ponieważ masz najlepsze oceny? Ponieważ Twoi rodzice uważają inaczej? Ponieważ nie masz najnowszego modelu telefonu? Ty masz wybór. Możesz zmienić swoją rzeczywistość. Migranci i uchodźcy nie. Nie mają dokąd wrócić. Stracili bliskich. Chcą normalnie żyć i być szczęśliwi. Tak jak Ty. Możesz to zmienić. Pomóż.

Zastanów się, co rozumiesz poprzez „wykluczenie" oraz w jakich sytuacjach Ty lub Twoi bliscy czuli się wykluczeni. Dlaczego? Ponieważ ubierasz się inaczej? Ponieważ Twoja rodzina ma mniej pieniędzy? Ponieważ masz najlepsze oceny? Ponieważ Twoi rodzice uważają inaczej? Ponieważ nie masz najnowszego modelu telefonu?

Wciąż uważasz, że nigdy Ty lub Twoja rodzina i przyjaciele nie czuli się odrzuceni?

Nie, wydaje mi się, że większość osób doświadczyła kiedyś wykluczenia lub odrzucenia.

Tak, ani ja, ani moja rodzina nigdy nie doświadczyliśmy wykluczenia lub odrzucenia.

Zgodnie z badaniami przeprowadzonymi wśród Polaków, 48% mężczyzn uważa, że podział obowiązków (np. gotowanie, sprzątanie) pomiędzy nimi a kobietami w domu jest równy. Takie samo zdanie wyraża już tylko 27% kobiet. To samo źródło podaje, że *„W rzeczywistości, na prace domowe mężczyźni poświęcają dziennie średnio 2 godziny i 48 minut. Kobiety 4 godziny i 33 minuty."*

Niezależnie od praktyki, Konstytucja Rzeczypospolitej Polskiej, a więc najważniejszy akt prawny w Polsce, gwarantuje kobietom i mężczyznom równe prawa oraz nakłada na nich takie same obowiązki.

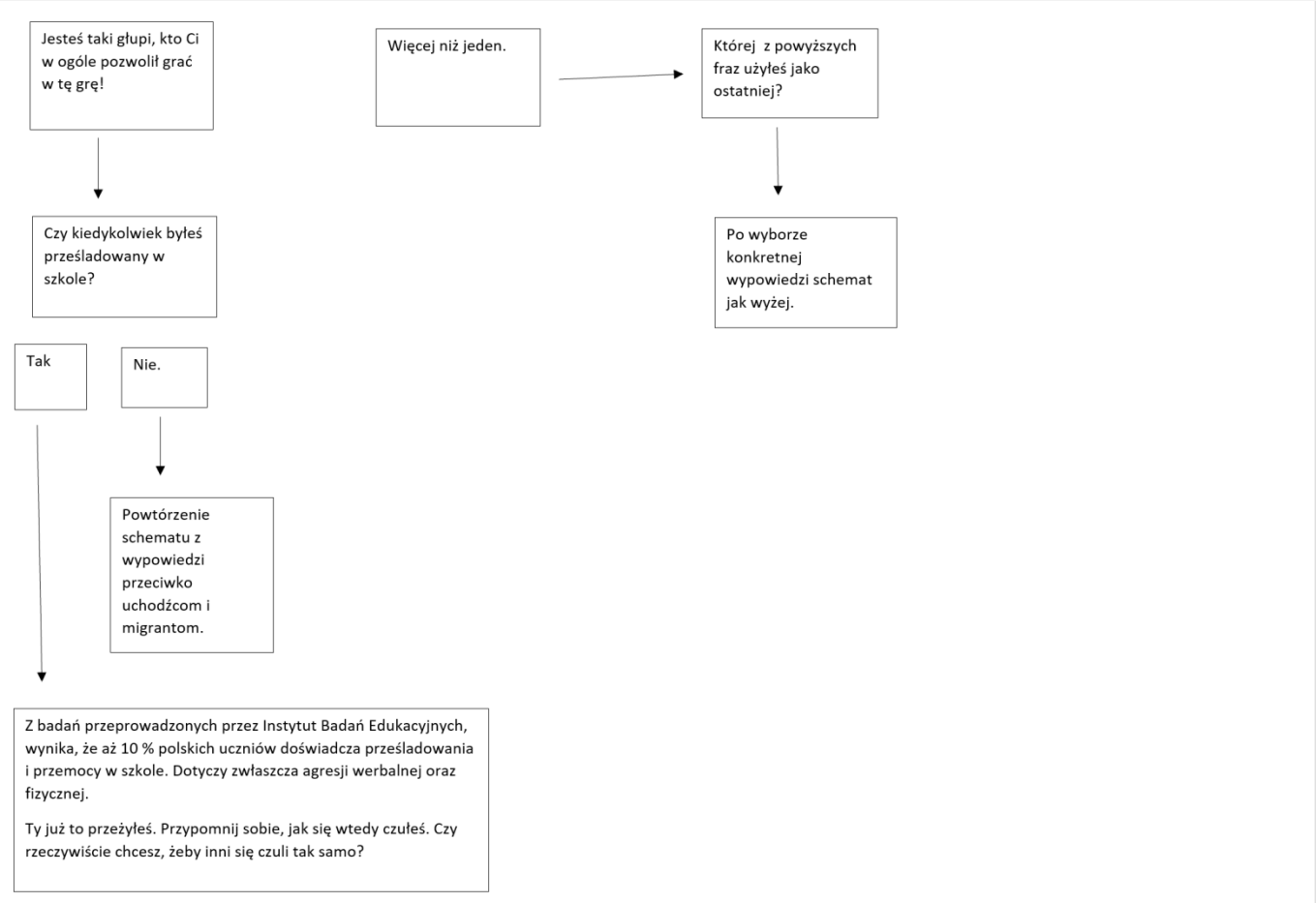
Zatem, zanim skomentujesz, zastanów się: czy rzeczywiście to jest Twoja opinia, czy może po prostu powtarzasz to, co mówią koledzy? Chcesz być taki jak oni, dopasować się. To zrozumiałe, ale pomyśl, „ po co na siłę starać się dopasować, skoro urodziłeś się, by się wyróżnić?".

Zgodnie ze Narodową Strategią Integracji, do grup szczególnie narażonych na wykluczenie społeczne należą między innymi:

- dzieci i młodzież ze środowisk zaniedbanych
- imigranci
- osoby żyjące w bardzo trudnych warunkach mieszkaniowych
- ofiary patologii życia rodzinnego np. przemocy w rodzinie, alkoholizmu.

Jak widzisz bardzo wiele osób jest narażonych lub doświadcza wykluczenia społecznego. Rozejrzyj się. Może w Twojej klasie, na osiedlu, na zajęciach tanecznych lub treningu piłkarskim, jest ktoś taki? Uśmiechnij się do niego. Porozmawiaj. Zapytaj, czy czegoś nie potrzebuje. Nie trzeba wiele, by zmienić czyjeś życie. Masz moc, by zmienić świat, zmieniając życie poszczególnych osób. Każde działanie, każdy ruch się liczy.

>> Chatbot Schema (Part 1)



>> Chatbot Schema (Part 2)

>> BIBLIOGRAPHY

1. Hawdon, James & Oksanen, Atte & Räsänen, Pekka. (2015). Online Extremism and Online Hate: Exposure Among Adolescents and Young Adults in Four Nations. Nordicom Information. 37. 29-37.

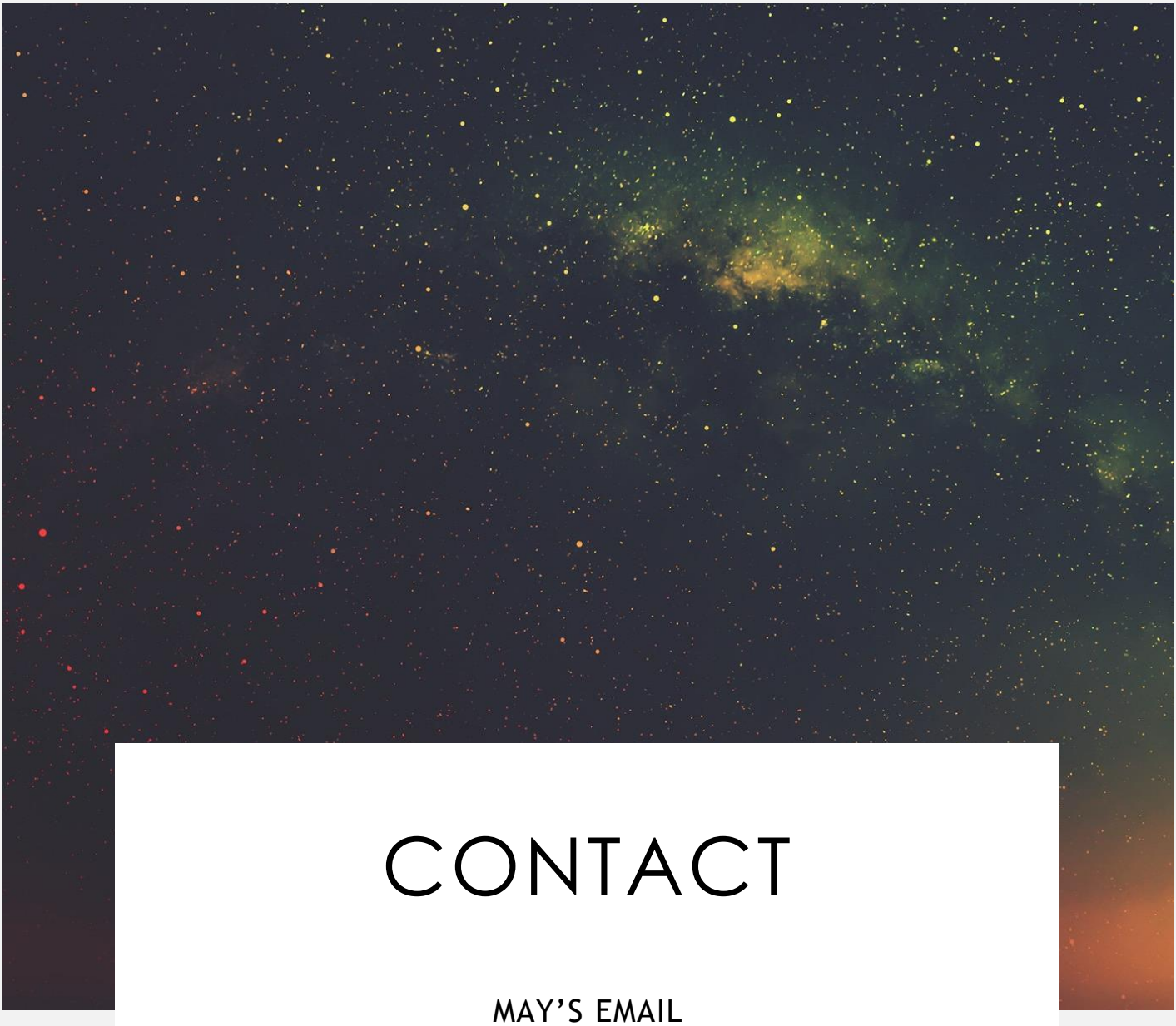
2. Lee, J. (2018, June 05). Chatbots were the next big thing: What happened? Retrieved from https://blog.growthbot.org/chatbots-were-the-next-big-thing-what-happened

3. Oksanen, Atte & Hawdon, James & Holkeri, Emma & Näsi, Matti & Räsänen, Pekka. (2014). Exposure to Online Hate among Young Social Media Users. Sociological Studies of Children and Youth. 18. 253 - 273. 10.1108/S1537-466120140000018021.

4. Sacirby, O. (2014, May 06). REPORT: Internet hate speech can lead to acts of violence. Retrieved from www.washingtonpost.com/national/religion/report-internet-hate-speech-can-lead-to-acts-of-violence/2014/05/06/9a9d9e60-d52c-11e3-8f7d-7786660fff7c_story.html

5. Winiewski, Mikołaj, et al. (2017). Contempt Speech and Hate Speech in Poland 2016. Stefan Batory Foundation.

6. Włodarczyk, J. (n.d.). Mowa nienawiści w internecie w doświadczeniu polskiej młodzieży. Fundacja Dzieci Niczyje.



CONTACT

MAY'S EMAIL

limxmay@gmail.com

RENA'S EMAIL

rena.pitsaki@gmail.com

EWA'S EMAIL

erodzik@gmail.com

FACEBOOK PAGE

www.facebook.com/HIAPolska

HUMANITY IN ACTION POLSKA