DOI: https://doi.org/10.46991/TSTP/2023.3.2.040

Lost in “Transl-Hation”: Exploring the Impact of Machine Translation as an Intermediary Tool in Detecting Armenian Hate Speech

Lilit Bekaryan[*]
https://orcid.org/0000-0001-8189-340X
Yerevan State University

Abstract: As the pervasive spread of hate speech continues to pose significant challenges to online communities, detecting, and countering hateful content on social media has become a priority. Social media platforms typically use machine translation to identify the hateful content of the posts made in languages other than English. If this approach works effectively in identifying explicit hateful content in languages that are predominantly used on social media, its effect is almost insignificant when it comes to Armenian.

The present research investigates the effectiveness of machine translation as an intermediary tool in accurately identifying and addressing instances of Armenian hate speech posts retrieved from social networking websites. The study of hate speech posts and comments made by Armenian users in Armenian helps identify that it is often the absence of intricate cultural and linguistic nuances, as well as insufficient contextualized understanding, that impede with hate speech detection in Armenian.

Keywords: machine translation, hate speech, NLLB, hostile comments, reporting, social media

1. Introduction

The advent of the new millennium brought the emergence of social media platforms like MySpace, Twitter, and eventually Facebook or Meta. This transformative era witnessed a shift in dynamics, as individuals transitioned into social media users and started sharing updates about their personal lives and experiences and engage in discourse on topics including business, politics, social movements, education, entertainment, and even science.

When engaging in social media interaction, users tend to express a range of positive and negative sentiments depending on the context and the topic under discussion. Positive sentiments may include but are not limited to expressing joy or happiness over others’ accomplishments, supporting initiatives, extending congratulations, complimenting people on their achievements, etc. In this research, I call this

[*] lilitbekaryan@ysu.am

Received: 16.11.2023
Revised: 11.12.2023
Accepted: 22.12.2023

communication geared towards supporting other users *cooperative* communication. On social media, cooperative communication is usually manifested through expressive speech acts, like compliments, expressions of love and support, encouragement, praise, or positive reinforcement.

Conversely, when users use social platforms to critique the phenomena or behaviour that they find disagreeable, they engage in a type of communication that is marked by the absence of cooperation. In the context of this research, this communication is referred to as *non-cooperative*.

While supportive communication can help people build a sense of community and connection, non-cooperative communication may not only hinder effective communication but may also lead to behaviours that are intentionally negative, aggressive, or threatening. In these instances, non-cooperative communication may escalate into a discourse intended to hurt and intimidate the recipient, incite violence, and create an unsafe environment. Examples of non-cooperative communication include hate speech or discriminatory language towards a specific group or an individual, cyberbullying, cyberharassment, cyberhate posts and comments intended to inflict harm on others.

Social media platforms have come up with different policies and regulations to combat cyberhate posts and comments. For instance, Facebook prohibits hate speech and content that incites violence or discrimination against a particular group of people. Twitter and Instagram bar users from posting hate content that can promote harm against individuals or groups. When users violate these rules, their content is removed, and their user accounts may be terminated or suspended. To help users identify instances of hate content on social media, platforms are implementing features like content moderation, reporting posts and comments that comprise offensive language, in this way allowing social media users to block content that violates their community standards. When evaluating posts made in other languages than English, social media platforms use automatic translation to check the post for hate speech and offensive language.

The present study investigates how effective machine translation is as an intermediary tool in accurately identifying instances of hate speech posts and comments made in Armenian on social media. The research is entirely based on 50 samples of data retrieved through manual collection, given the absence of any hate speech detecting software in Armenian.

The research findings of the present study can further contribute to the linguistic studies unfolding in the field of hate speech linguistics and discourse, as well as may be invested in developing algorithms and tools to detect and flag hostile content posted in Armenian. All the posts and comments for the study are written in Armenian. The research data have been retrieved in the form of screenshots. User handles and the names of the targets have been removed from the examples presented in the paper out of ethical considerations. English translations are provided throughout.

2. Research Background

A recent study by Hootsuite indicates that 90% of social media users engage in social interaction with others on various platforms, including liking or commenting on posts, direct messaging, or following other users (Hootsuite 2022).

When observing users' behaviour online, clinicians and researchers noticed people doing and saying things in cyberspace they would not normally do in real life (Joinson 2001: 177; Suler 2003). John Suler refers to this phenomenon of individuals becoming less restrained and more open in their expression when engaging online as *disinhibition effect* and believes it to operate in two seemingly contradictory manners (Suler 2004: 321). He describes individuals demonstrating generosity, feeling enthusiastic about sharing very personal aspects of their life, and refers to this behaviour as *benign* disinhibition. Examples of benign disinhibition may include but are not restricted to divulging secret personal information, wishes and fears, or going out of one's way to help others.

However, the disinhibition effect also has its dark side, often manifested in the offensive language, harsh criticism, anger, or even hatred people demonstrate when engaging in online interaction. Suler tentatively labels this aspect of disinhibition as *toxic* disinhibition (ibid.).

Both types of disinhibition are believed to stem from a set of factors including dissociative anonymity, invisibility, asynchronicity, etc. Anonymity was first discussed by Plato in his myth of the Ring of Gyges, where the philosopher posits how a person is likely to act immorally when they know they will not be caught, and can get away with the crime (Plato 2017). Although the idea was put forward in a Socratic dialogue back in 375, it still holds true for modern times. Building on this, Suler argues that *anonymity* is the principal cause behind the disinhibition effect, since it extends the individuals an opportunity to reveal those sides of their identity online they would usually keep under guard in real life. This detachment from real life leads to a reduced sense of vulnerability and provides a sense of security, as individuals feel they do not have to own up to their behaviour in the context of online interaction and will not be punished for their hostile behaviour on social media. In fact, what happens quite often is that social media users feel emboldened to engage in a more open, at times antagonistic interaction and produce offensive language they would refrain from in real life interaction.

Research supports the role of anonymity in cyberhate crimes (Huang et al., 2020; Udris, 2014; Wachs et al., 2019; Wu et al., 2017).

Cyberhate crimes are defined as the crimes committed through the use of electronic communications technology to spread anti- Semitic, racist, bigoted, extremist or terrorist messages or information. These electronic communications technologies include the Internet, user- generated content, dating sites, blogs, on- line games, instant messages, and e- mail, as well as other computer and cell phone- based information technologies, such as text messages and mobile phones. Examples of cyberhate may include flaming, cyberbullying, stalking, sexting, etc. (Willard, 2007).

As indicated above, social media platforms have developed their own toolkit to fight hate speech and to discourage users from posting it. Meta, for instance, regularly updates its policies and algorithms to adapt to modern challenges, including hate speech.

3. Discussion

Most of us may have been 'guilty' of reporting hatred on social media at some stage in our life. However, there might have been situations, when reporting a comment that one thought to be explicitly hateful, we received a notification that the post did not go against community standards and is not subject to removal.

This happens because social media platforms do not always succeed in recognizing offensive content and labelling it as appropriate. This may be due to a range of reasons.

First, there might be some reluctance on part of the social platforms to take steps against hate speech due to certain political circumstances or their desire to maintain some level of free speech. For instance, back in March 2023, Meta platforms were reported to have made a temporary change to their hate policy permitting certain expressions of violence against the Russians and the Russian army in the context of the Russian invasion of Ukraine. Calls like 'death to Russian invaders' that would previously be considered unacceptable by Meta community standards were temporarily allowed unless they contained other targets, like political prisoners.

Language is another barrier. Until July 2022, most social media platforms, including Facebook, used statistical machine translation to identify the hateful content of the posts made in languages other than English. Statistical machine translation (SMT) uses statistical models that are trained on large bilingual corpora to generate translations. In phrase-based SMT, as introduced by Koehn, Och, and Marcu in 2003, translation units consist of consecutive sequences of a limited number of words, which may not necessarily be linguistic phrases. Hierarchical and syntax-based SMT, proposed by Chiang in 2007, involves modeling translation using context-free grammar. In SMT, models analyse a range of options and patterns to come up with the most possible translations. This would work with the so-called 'celebrity' languages, like Spanish, English or Russian, but was not very effective in case of low-resource languages.

In response to an investigation by the human rights group Global Witness, which revealed Meta's failure to adequately moderate harmful content in various languages, including Amharic (the predominant language in Ethiopia) and Burmese (spoken in Myanmar), Meta implemented a significant change to its translation system in July 2022. The new system, known as NLLB-200, is powered by artificial intelligence. "No language left behind" indicates Meta's attempt to invest heavily in identifying hateful content not only in most widely spoken languages but also in low-resource languages (see Meta webpage). NLLB supports 200 languages and goes through the stages of automatic dataset construction, training, and evaluation. The final stage implies detecting and filtering out profanity and offensive content. Meta builds 'toxicity lists' typically comprising offensive and harmful words and/or phrases for supported languages. These lists help identify and flag potentially offensive language within a corpus of text data.

Based on my analysis of factual data, the NLLB-200 tools exhibit heightened proficiency in recognizing explicit hateful content within low-resource languages. However, their effectiveness appears limited when it comes to detecting the implicit hatefulness embedded in the content. Let me illustrate this through an example based on a comment that was made on an Armenian reporter’s page. The comment was initially posted in Armenian, and the English direct translation of the comment is provided.

Example 1

> «Մարդուդ» հիմիկվա տեղը ավելի անվտանգ ա։
>
> **English Translation**
>
> Your “man” is in a safer place now.

If one considers this comment without knowing the circumstances under which it was generated, one can notice no hostility in its message. However, when probing further into the historical background of the context, one learns that the target of the comment, is an Armenian reporter and a human rights advocate who has always been defending the rights of life-termers. Back in 2013, she married a life-termer, who is still serving in prison. When the couple failed to publicise their marriage, there was a barrage of criticism against the journalist for falling in love with a man, who was labelled as a life-termer and a criminal. It has been more than fifteen years that the journalist has been trying to prove that her husband deserves pardon after having served about thirty years in prison.

The comment is hateful because it comprises an implicit threat; the comment maker implies that if the woman’s spouse comes out of prison, he is most likely to face harm.

Does the English translation of the sentence “Your man is in a safer place now,” convey the threat? It does not. In Armenian, especially when we consider the context, under which the post was generated, the post sounds like a threat. Languagewise, it features an extremely informal tone through the use of an informal possessive pronoun “մարդուդ”, although the user addresses a woman they presumably have not met and do not know. It is also interesting to note the use of the word “man”/ մարդ/ rather than ‘husband’ /ամուսին/, and the fact that they placed the word in scare quotes. The target reported the comment but since its hatefulness was lost in translation, Meta decided that it did not go against their community standards and refused to delete it.

Apparently, some users producing hostile posts and comments are well aware of these technical impediments and deliberately employ strategies to have their posts and comments “survive” hate speech filters in the media. Let us consider a pertinent example, where the user resorts to transliteration to post an offensive comment on YouTube.

The English translation of the comment is provided below the example with its original spelling and punctuation retained.

Example 2

айс вирусин петке варел мохит лцнел ахстев гет тох гна асни газах, Ир еркир
English Translation
This virus should be set on fire the ashes should be dropped into aghstev let her go to gazakh, Her country.

As we can see, the comment writer refers to the reporter as “a virus”. The use of the word “virus” in the context of 2023, the year succeeding almost three years of pandemic filled with pain and despair, makes the comment sound even more offensive. The comment clearly comprises a violent call to action, which is to kill the reporter by having her set on fire. Moreover, the writer suggests dropping her remains to Aghstev River. This reference is not accidental, either. The Agshtev is a transboundary river in Armenia and Azerbaijan. One of the cities it flows through is Qazakh, a city and the capital of the Gazakh District in Azerbaijan. In her reference to Gazakh, the writer implies that the reporter must be of Azeri descent. This example is a good illustration of an explicit call to violence combined with implicitly expressed ethnic slur.

As it has been mentioned, the comment survived a removal by YouTube Community Standards because of the use of transliteration in crafting their message, a strategy that makes it hard for social media algorithms to recognize and identify the text as an expression of hate speech for the simple reason that they could not translate it.

The study of the retrieved examples shows that in some cases, it is the use of idiomatic expressions or colloquial language in the posts that makes it challenging for artificial intelligence tools to identify the meaning of the message, let alone its hatefulness.

Let us consider the following example, the English translation of which was made by the AI.

Example 3

Աչքիս մի չորս անգամ պատվաստված ա մեր քուրը, բան չի ջոգում:
AI translation
My eye has been inoculated four times with our scabies, but nothing is working.
Human translation
Our sis must have been vaccinated four times. She doesn’t get anything.

When the AI-generated translation appears nonsensical, the human translation reveals an offensive comment directed towards a woman. Moreover, the comment makes assumptions about her mental capacity solely based on her vaccination status. The comment not only contains offensive content but also has the potential to contribute to conspiracy theories related to vaccination, posing a threat to public health and safety.

4. Conclusion and Recommendations

Apparently, it is often the absence of intricate cultural and linguistic nuances, as well as insufficient contextualized understanding that impede with hate speech detection in Armenian. Improving the quality of translation of Armenian hateful content on social media is a complex task that involves both technological and policy considerations.

To begin with, social media sites should recognize the evolving nature of nature and invest in refining their translation models to ensure more accurate rendering of content.

As the listed examples show, hatefulness of the message may often rely on the subtleties that are lost in translation. Hence, it is also important to focus on improving the contextual understanding of the translation.

Encouraging user feedback and reporting mechanisms is another vital aspect of addressing translation challenges. Social media community should feel empowered to have their own investment in identifying and removing inappropriate content.

Finally, it would help to have a team of human translators working with AI tools to make a more reliable evaluation of content for potential harm and offence.

References

Chiang, David. 2007. “Hierarchical Phrase-Based Translation.” *Computational Linguistics* 33 (2), 201-228.

Hootsuite. 2022. *2022 Digital Trends Report*. Accessed January 10, 2023. https://www.hootsuite.com/resources/digital-trends

Huang, Chiao Ling, Zhang, Sining, and Yang, Shu Ching. 2020. “How Students React to Different Cyberbullying Events: Past Experience, Judgment, Perceived Seriousness, Helping Behavior and the Effect of Online Disinhibition.” *Computers in Human Behavior*, 110, 106338.

Hutson, Elizabeth. 2016. “Cyberbullying in Adolescence”. *Advances in Nursing Science*, 39(1): 60-70.

Joinson, Adam N. 2001. “Self-Disclosure in Computer-Mediated Communication: The Role of Self-Awareness and Visual Anonymity.” *European Journal of Social* Psychology 31: 177–192.

Koehn, Philipp, Och, Franz J., and Marcu, Daniel. 2003. “Statistical Phrase-Based Translation.” *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*. Stroudsburg: Association for Computational Linguistics, 48-54. Accessed 15 November, 2023. http://aclweb.org/anthology-new/N/N03/N03-1017.pdf.

Meta. *No Language Left Behind*. Accessed January 12, 2023. https://ai.meta.com/research/no-language-left-behind/

Plato. 2017. “The Republic” (B. Jowett, Trans.) *Green World Classics*. (Original work published in 375.)

Suler, John. 2003. “The Psychology of Cyberspace”. Accessed 1 September, 2023. www.rider.edu/suler/psycyber/ psy-cyber.html

Suler, John. 2004. “The Online Disinhibition Effect.” *CyberPsychology & Behavior*. 7 (3): 321–326.

The Guardian, March 11, 2022. Accessed 14 September, 2023. https://www.theguardian.com/technology/2022/mar/11/facebook-and-instagram-let-users-call-for-death-to-russian-soldiers-over-ukraine

Udris, Reinis. 2014. “Cyberbullying among High School Students in Japan: Development and Validation of the Online Disinhibition Scale.” *Computers in Human Behavior*, 41: 253-261.

Wachs, Sebastian, Wright, Michelle F., and Vazsonyi, Alexander T. 2019. “Understanding the Overlap between Cyberbullying and Cyberhate Perpetration: Moderating Effects of Toxic Online Disinhibition.” *Criminal Behaviour and Mental Health*, 29(3): 179-188.

Willard, Nancy E. 2007. *Cyberbullying and Cyberthreats: Responding to the Challenge of Online Social Aggression, Threats, and Distress*. Champaign: Research Press.

Wu, Sheng, Lin, Tung-Ching, and Shih, Jou-Fan. 2017. “Examining the Antecedents of Online Disinhibition.” *Information Technology & People*, 30(1): 189-209.

Conflict of Interests

The author declares no ethical issues or conflicts of interest in this research.

Ethical Standards

The author affirms this research did not involve human subjects.