Computational Identification of Incivility

22.11.2024
2nd Budapest Methods Workshop

Political Communication Research Group
POLCOM
Department of Communication | University of Vienna

WWTF
Vienna Science and Technology Fund

Hossein Kermani

What is the problem?

- The prevalence of online incivility and hate
- Hate speech on Twitter increased after Musk bought Twitter
- The detriment effects of hate speech on democracy
- The need for identifying incivility on large scale
- Here, computational methods become necessary as many researchers employed them in online hate speech detection

Hate Speech Detection

- Different automated techniques have been used like Dictionary-based techniques, Supervised Machine Learning (SML) algorithms like BERT, and recently introduced Large Language models (LLMs) like ChatGPT.
- Each of which has their strengths and weaknesses.

- Dictionary-based models are straightforward and easy to understand as they work on a predefined list of hateful words. Such methods could be effective, especially in identifying clear-cut forms of incivility. However, the downside of these algorithms is their inability to understand more complex forms of incivility. Furthermore, there are words with different meanings, insulative and non-insulative, in each language, but these models cannot differentiate between the various lexical uses of language.
- SML tries to solve this problem by going beyond simply identification of hate speech based on specific words. These models are more complicated in design and use many layers to identify the relationships between words and sentences in a training dataset. Then, the trained model could be employed to code a new unseen dataset. Researchers found SML is more effective in detecting incivility. However, these models suffer from the lack of transparency and need large training datasets which have been deliberatively coded by human coders.
- ChatGPT as a zero shot model do not need the training dataset. It is a big benefit in comparison to other SML algorithms.

Other gaps

- Most of the studies focus empirically on mainstream languages like English and German. There is a severe lack of research in other languages like Farsi.
- Furthermore, studies show that automated algorithms work better on clear-cut concepts. In this case, text that includes offensive words could be detected by a machine model with higher reliability. Identification of subtle forms of incivility remains a challenge to date. This problem is intensified when it comes to Margin languages like Farsi.
- Finally, our knowledge of the performance of computational methods on different types of incivility is also uncertain.

What am I doing?

- I am comparing human-driven qualitative coding with three computational methods
- Unlike English, there are almost no dictionary of offensive words for Farsi. I will create such a dictionary
- In addition, I will train a BERT model on a huge Persian dataset and open source it.
- The comparison of the methods will enhance our understanding of the strengths and weaknesses of each method in hate speech detection

Theoretical debates

- Incivility is defined as impolite behavior that violates the polite norms of interpersonal communication (Coe et al., 2014; Gervais, 2014; Rains et al., 2017).
- While it has been studied as a stand-alone concept, hate speech could be understood as the most severe type of online incivility (Hameleers et al., 2022). European Commission (2019) defines hate speech as incitement to violence or hatred against a group, defined in relation to race, religion, or ethnicity.
- There are different classifications of incivility/hate speech.
- Coe et al. (2014) suggest 5 categories: Name-calling, Aspersion, Vulgarity, Lying, and the pejorative of speech (p. 661).
- The Anti-Defamation League proposes five levels: negative stereotypes, insults, discriminatory expressions, threats and genocide.
- Based on these works, I propose three categories:
    - Pejorative speech: A word or grammatical form expressing a negative or disrespectful connotation, a low opinion, or a lack of respect toward someone or something.
    - Insult: Attacking someone or a group with offensive words to convey hatred
    - Threatening messages: Clearly incite violence

Data and methods

- Empirical analyses is focused on the Women, Life, Freedom movement. Such a large scale crises provides a space where higher levels of harmful messages are shared on social media.
- I have collected all popular Farsi tweets (>1k likes in a day) from September 15 to November 15, 2022 (N= 36,255).
- Four coders coded the dataset qualitatively and discursively drawing on KhosraviNik's (2017) approach to Social Media Critical Discourse Studies (SM-CDS) in five rounds!

- Building reliable training datasets is a big challenge in SML. That is why I tried to educate four coders and draw on a solid discourse theory to code the dataset. In each round, intercoders' reliability score was measured. In the final round, it was 0.96 percent!
- Coders could code two discursive practices for each tweet including incivility.
- Having finished coding the dataset, I selected tweets including incivility as a train set for computational methods.

- Building reliable training datasets is a big challenge in SML. That is why I tried to educate four coders and draw on a solid discourse theory to code the dataset. In each round, intercoders' reliability score was measured. In the final round, it was 0.96 percent!
- Coders could code two discursive practices for each tweet including incivility.
- Having finished coding the dataset, I selected tweets including incivility as a train set for computational methods.

Qualitative results

	Frequency	Percentage
Uncivil tweets	15078	0.41

Incivility type	**Frequency**	**Percentage**
insult	9810	0.65
pejorative speech	3790	0.25
threatening messages	1478	0.1

Next steps

- I am collecting the offensive words in uncivil tweets to create the dictionary
- At the same time, I am training a BERT model on this sample. The model is running now. When it will be finished, I will use the classifier to code a dataset of 2,097,539 tweets. Then, human coders will validate and compare the results.
- The same procedure will be done with ChatGPT.

THANK YOU

Hossein Kermani is a MSCA post-doctoral researcher at the Political Communication Research Group of the University of Vienna. Hossein is studying social media, digital repression, computational propaganda and political activism in restrictive contexts, with particular attention to Iran. His research mainly revolves around a) the discursive power of social media in changing the microphysics of power and playing with the political and social structures, and b) the strategies that have been employed to manipulate and dismantle social media activism in non-democratic societies. In order to do so, Hossein is chiefly combining social and communication theories with computational techniques, in particular Social Network Analysis (SNA), Natural Language Processing (NLP), and critical discourse analysis.

Hossein has recently published in, among others, *New Media and Society, Big Data & Society, Information, Communication, and Society and Asian Journal of Communication*. His first book, social media research in Iran (in Farsi), was published in 2020. He is now working on his first English book, *Twitter activism in Iran*, which Palgrave Macmillan will publish in 2024.

Hossein Kermani