Our present work relates to several prior computational studies focusing on two main topics. The first set tries to predict the antisocial behavior in various social media outlets or forums, and find the relation between group norms and communication behavior. A variety of methods have been proposed for cyberbullying and hostility detection. These methods mostly approach the problems by treating it as a classification problem, where the comments been classified as antisocial or not.

Some of these methods have used their data to come up with forecasting methods which can predict if the conversation will go wrong [132][124], predicting the intensity of it in 5-15 next comments[124], or even further predicting whether a user will get banned in the future or not[130][152]. However, others only detect these types of behavior in the existing comments rather than attempting to predict them. [134][16].

Researchers usually use different types of social media such as Twitter, Facebook, Instagram or forums such as Reddit or Wikipedia as their data source. Next, the categorization happens in two ways in general. Either custom groups are defined and data is being categorized manually applying one or a couple of groups[134]. Another approach is using different unsupervised methods such as K-Special Centroid to identify the various types of conversations[124][132]. Also, it's possible to only use the plain text for the data categorization. [16]

With data categorization, important features will be extracted where they are essential in the classification stage. While a group of studies used linguistical features such as Unigram[134], Word2vec, and etc [124], others have used more specific features such as the number of comments in each thread[134], deciding based on the time order of comments, -- if it's the final comment or not [124], the number of people participated in the conversation, their gender, age, and number of followers or following [16], or even the effects of the authors' mood [130-131].
On the contrary, some have employes the features related conversation content. For instance, Vulgarities [16] or the level of politeness[132].
However, these methods are different to one or another for feature extraction, in the end, they applied a combination of features for more efficient extraction.

Comments are not the only content available to be considered as the data source. Another article focused on people's actions and discussed the fact that although some people may not always write a comment. It's possible they interact via liking the post itself or the existing comments (if liking the comment is available in that specific social media) which can result in an elevated discussion with more and more comments in the future. This situation might even start or stop a hostile conversation.[152]

The second set of computational studies emphasized finding the different semantic variation of words, linguistic change and evaluation of users and communities, and in general, find the words meaning base on their contextual users.[138][139][144]

Some researchers looked for nouns and generated graphs to keep track of words that appear together in the context. The graph is a co-occurrence graph, thus the nodes represent the terms and the edge between them indicates their co-occurrence in a context. As the number of edges between two nodes increases, their weight and importance will increase as well. Therefore, they pruned the graph and only keep the ones which were higher than the threshold. They have used the number of shared neighbors as their similarity function and have tried to detect different communities using clustering.[138] Also, the linguistic features can be applied to differentiate semantic variation of a word [139].

Others have tried to extract norms and hidden rules inside a big community by defining micro, meso, macro as the different levels of the norm. After collecting the data and preprocessing phase, the k-mean clustering has been used on the prediction matrix. As the result, they ended up with the cluster of subreddits which share norms among themselves at three different levels (macro, meso, and micro). They have used topic modeling and open coding to extract these norms.[150]

However, some articles have worked on the users' level adaptability during their lifecycle, which started with users writing their first comment and ended by leaving the website. They have realized that at the beginning the users are more flexible and they mostly try to learn the norm of the community. They also realized that when the users stop learning they mostly leave the website as well.[144]

[16] Mohammed Ali Al-garadi, Kasturi Dewi Varathan, Sri Devi Ravana: "Cybercrime detection in online communications: The experimental case of cyberbullying detection in the Twitter network"

[124] Ping Liu, Joshua Guberman, Libby Hemphill, Aron Culotta: " Forecasting the presence and intensity of hostility on Instagram using linguistic and social features"

[131] Justin Cheng, Michael Bernstein, Cristian Danescu-Niculescu-Mizil, Jure Leskovec: " Anyone Can Become a Troll: Causes of Trolling Behavior in Online Discussions"

[132] Justine Zhang, Jonathan P. Chang, Cristian Danescu-Niculescu-Mizil, Lucas Dixon, Nithum Thain, Yiqing Hua, Dario Taraborelli: "Conversations Gone Awry: Detecting Early Signs of Conversational Failure"

[133] Elaheh Raisi, Bert Huang: "Cyberbullying Detection with Weakly Supervised Machine Learning"

[134] Amy X. Zhang, Bryan Culbertson, Praveen Paritosh: "Characterizing Online Discussion Using Coarse Discourse Sequences"

[138] David Jurgens: "Word Sense Induction by Community Detection"

[139] Marco Del Tredici, Raquel Fernandez: "Semantic Variation in Online Communities of Practice"

[140] Kartik Sawhney, Marcella Cindy Prasetio, Suvadip Paul: "Community Detection Using Graph Structure and Semantic Understanding of Text"

[144] Cristian Danescu-Niculescu-Mizil, Robert West, Dan Jurafsky, Jure Leskovec, Christopher Potts: "No Country for Old Members: User Lifecycle and Linguistic Change in Online Communities"

[150] Eshwar Chandrasekharan, Mattia Samory, Shagun Jhaver, Hunter Charvat, Amy Bruckman, Cliff Lampe, Jacob Eisenstein, Eric Gilbert: "The Internet’s Hidden Rules: An Empirical Study of Reddit Norm Violations at Micro, Meso, and Macro Scales"

[152] JUSTINE ZHANG, CRISTIAN DANESCU-NICULESCU-MIZIL, CHRISTINA SAUPER, SEAN J. TAYLOR: "Characterizing Online Public Discussions through Patterns of Participant Interactions"