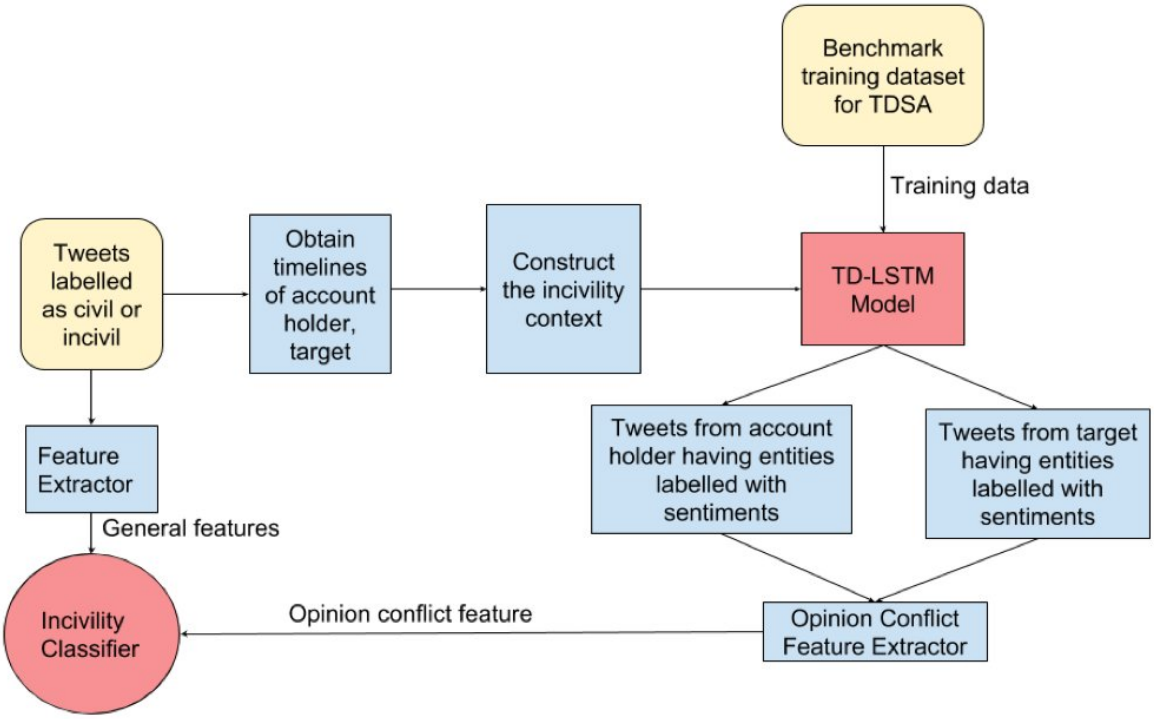
A new convolutional neural network model to detect abuse and incivility on Twitter

October 3 2018, by Ingrid Fadelli



Schematic of the steps for incivility detection. The yellow colored blocks represent inputs, the red colored blocks represent the classifiers and the blue colored blocks represent the intermediate steps. Credit: Maity et al.

Researchers at Northwestern University, McGill University, and the Indian Institute of Technology Kharagpur have recently developed a character level convolutional neural network (CNN) model that could help to detect abusive posts on Twitter. This model was found to outperform several baseline methods, achieving an accuracy of 93.3 percent.

In recent years, abusive behavior on online platforms has been rising exponentially, particularly on Twitter. Social media companies are hence seeking effective new methods to identify this behavior in order to intervene and prevent it from causing serious harm.

"Twitter, which was initially conceived as an 'e-town square,' is turning into a mosh pit," Animesh Mukherjee, one of the researchers who carried out the study, told *Tech Xplore*. "A rising number of cyber-aggression, cyberbullying and incivility cases are being reported every day, many of which severely affect users. In fact, this is one of the main reasons why Twitter is losing its active follower base."

Online content can spread rapidly and reach very wide audiences, so cases of online abuse often drag on for long periods of time with highly detrimental effects. The victim or victims, as well as other sensitive bystanders, might end up reading the offender's words countless times before these finally disappear from Twitter. This is why it is important for social media platforms to detect this content effectively and rapidly, performing timely interventions to remove it.

"We set out with the objective to develop a mechanism that can automatically detect uncivil tweets early, before they can make severe damage," Mukherjee said. "We observed that most often, a victim/target is attacked after expressing strong sentiments toward certain named entities. This led us to the central idea of leveraging opinion conflicts to detect uncivil tweets."

Mukherjee and his colleagues realized that abusive posts are often correlated with opinion differences between the offender and the target, particularly opinions about a renowned public figure or entity. They hence incorporated entity-specific sentiment information into their CNN model, hoping this would improve its performance in detecting abusive content.

account holder's tweet:
@user1 Enjoy prison a$$hole!
account holder's context tweets:
@user5 @user6 You sir, are just another clueless Trump lemming.
@user7 @user8 Seriously, get your head out of Trump's ass already. Go watch your Fox News & Friends and eat your jello.
@user8 The video of what your boyfriend said: Trump labels US justice system 'laughingstock' @CNNPolitics https://t.co/QNa2jqAYsE
@user9 if the Devil was running as a Republican, would you still vote for him? Your morals and priorities are so screwed up.
@user5 Seriously, let it f**king go. You are worse that a scorned girlfriend bringing up decades of shit that does not matter. You are the BIGGEST LOSER of all time.
@user11 Trump idiot lemmings are condemning the outrage over slavery and agreeing w/the idiot Kelly about praising Lee? Clueless losers
target's context tweets:
@user10 You are truly stupid. Trump is the first President to come into Office supporting marriage equality
Strange that the #fakenews media never gets stories wrong in favor of Trump. It's almost like they do it on purpose
According to HuffPo, President Trump is effective, but they don't like it. Donald Trump's relentless focus on tax cuts, deregulation and draining the swamp is great for job growth... with minorities

and so on ...

In the example of incivility context cited below, we observe that the target tweets positively about Donald Trump and US Economy. However, the offender (account holder) tweets negatively about Trump and positively about President Obama. We can observe that there is a conflict of opinion between the target and the account holder as the sentiments expressed toward the common named entity Donald Trump is opposite. Going through the entire exchange of messages, we find that this opinion conflict eventually leads to an uncivil post. Credit: Maity et al.

"The character level CNN tries to automatically extract patterns from

uncivil tweets that distinguish them from other tweets," Pawan Goyal, another researcher who carried out the study, told Tech Xplore. "We also chose to use character-level embedding, rather than word-level embedding. As tweets are usually small, only contain a few words, and have a lot of spelling variations, character-level models are found to be more robust than word-level models."

This character level CNN model outperformed the best baseline method by 4.9 percent, achieving an accuracy of 93.3 percent in detecting uncivil tweets. The researchers also carried out a post-hoc analysis, taking a closer look at behavioral aspects of offenders and victims on Twitter, hoping to better understand incivility incidents.

This analysis revealed that a sizable portion of users were repeated offenders who had harassed targets over 10 times. Similarly, some targets had been been harassed by different offenders on a several different occasions. "The most interesting finding of this study is that opinion conflicts strongly correlate with uncivil behavior on Twitter," Mukherjee said. "This single feature tied with the char-CNN-based deep neural model can be very effective in identifying uncivil tweets early."

In the future, the CNN model devised by Mukherjee and his colleagues could help to counteract and reduce abusive content on Twitter. The researchers are now trying to develop similar models to detect hate speech on Twitter, as well as on other social media platforms.

"Meanwhile, we are also studying how hate speech spreads on social media, as well as investigating how different methods of countering hate speech could help to tackle this vicious online phenomena," Mukherjee said.

More information: Opinion conflicts: An effective route to detect incivility in Twitter. arXiv:1809.00289v1 [cs.SI].

arxiv.org/abs/1809.00289

© 2018 Tech Xplore

Citation: A new convolutional neural network model to detect abuse and incivility on Twitter (2018, October 3) retrieved 3 May 2024 from https://techxplore.com/news/2018-10-convolutional-neural-network-abuse-incivility.html

This document is subject to copyright. Apart from any fair dealing for the purpose of private study or research, no part may be reproduced without the written permission. The content is provided for information purposes only.