Toxic comment classification

What?

- A sentiment analysis through classifying online conversation by toxicity.
- Toxic: disrespectful (insult, threat, *etc.*)

Why?

- Improve comments quality.
- Help website evaluate and decide its content

Compared to directly searching swear words:

Accurate	Sometimes sentence without swear word still has toxic meaning.
Informative	Understand the overall meaning of the comment

Example:

Why the edits made under my username Hardcore Metallica Fan were reverted?	✓ **Safe**
Nonsense? kiss off, geek. what I said is true. I'll have your account terminated.	✗ **Toxic**

Who?

- All websites/ forums/ apps that allow users to have conversation
- Twitter, Reddit …

Toxic comment classification

How?

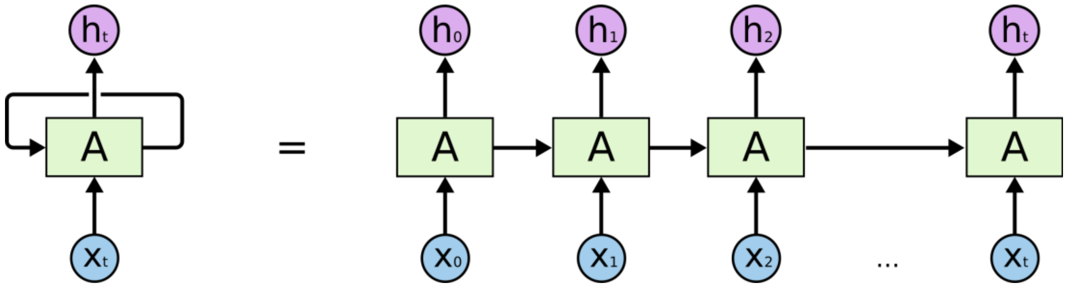
- Trained more than 1 million labeled comments by a long short term memory (LSTM) neural network
- Sequence input; Considering information from previous words with different weights

Other Training sets: Many public human labeled datasets Wikipedia Detox, Dkhate …

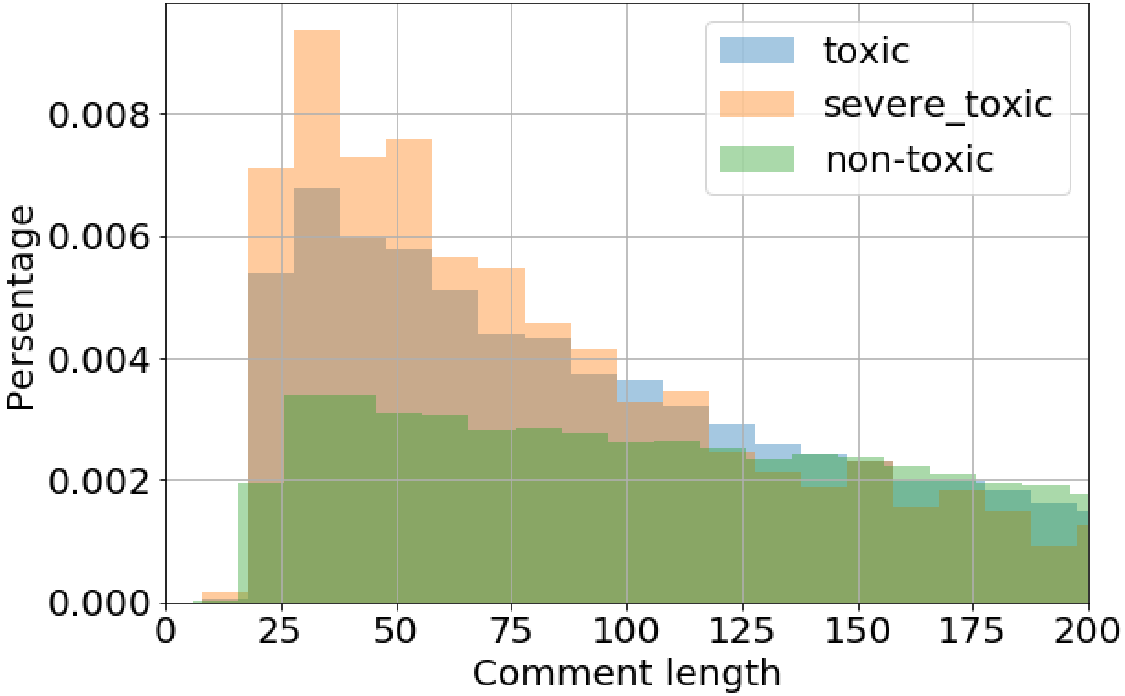
Next?

- Fine tune the parameters of the model
- More analysis on the result

LSTM:



Toxicity vs comment length


toxic
severe_toxic
non-toxic
Persentage
Comment length