Automatic moderator

Identify and classify toxic online comments

Maxime Guillaume
Sophea Ly
Guilherme Razet
Yi-ting Tsai

Table of contents

Abstract

Context	Toxic comments: violence, hostility and discrimination;
Need	How to moderate toxic comments ?
Problem	How to identify them ?
Solution	Create a tool to detect toxic comments.

Formalization

$$\forall comment \in Comments,$$

$$\forall category \in \{toxic, severe\_toxic, obscene, threat, insult, identity\_hate\}$$

$$\mathscr{R}(comment, category) \in [0, 1]$$

where $\mathscr{R}$ is a regression function.

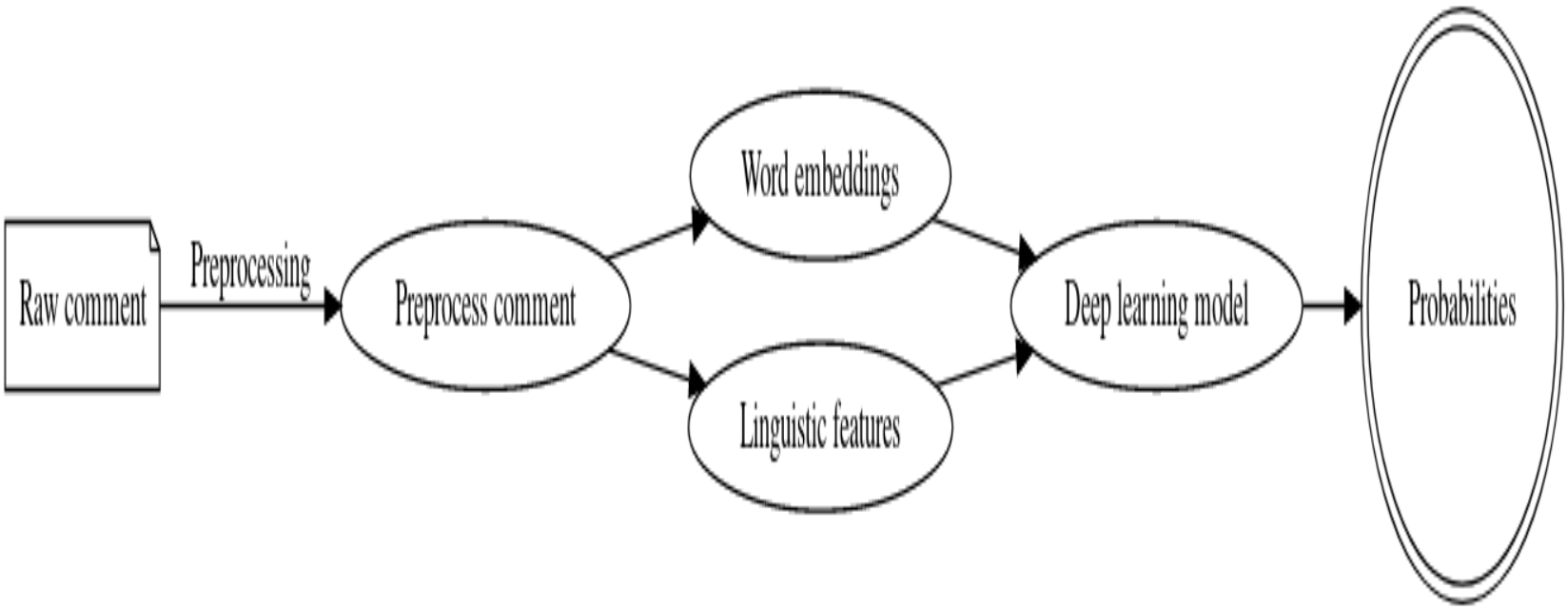
Example

Stupid peace of shit stop deleting my stuff asshole go die and fall in a hole go to hell!

Classified as:

Toxic	Severe toxic	Obscene	Threat	Insult	Identity hate
1	1	1	0	1	0

Proposed solution



Required skills

Technical skills:

- Python coding;
- Data visualization;
- Data exploration;
- Machine Learning algorithms;
- Deep Learning;
- Linguistic analysis.

Soft skills:

- Communication;
- Team works;
- Writing;
- Project management;
- Vulgarization.

Corpora used

Kaggle Dataset:

- CSV corpus of wikipedia comments (313k);
- Labeled by human users for toxic behavior;
- Different types of toxicity.

Waseem and Hovy [10] Dataset:

- Corpus of tweets (16k);
- Labeled by human users for offensive discourse;
- Only one type of label : offensive or not-offensive.

Milestones

02/10/2019	Proposal and subject presentation;
06/11/2019	Bibliographic research , data vizualisation and modeling from [1];
27/11/2019	Additional preprocessing , linguistic definition of toxicity;
06/01/2020	Test on more complex models and report;
16/01/2020	Defense.

Team members

Maxime Guillaume

- Main skills: Statistics, Python coding, Machine Learning;
- Role: Software Developer, Solution Architect.

Guilherme Razet

- Main skills: Python coding, writing, vulgarization;
- Role: Software Developer, Vulgarizer.

Sophea Ly

- Main skills: Python coding, writing, data exploration;
- Role: Tester, Feature Engineer.

Yi-ting Tsai

- Main skills: Linguistic analysis, data visualization, writing;
- Role: Linguistic Expert, Project Manager.

Github - 1/2

`github.com/mxmgllm/905-toxic_comment_classification`

- bin ← *compiled model code*
- config ← *configuration files*
- data
  - external ← *data from thrid party sources*
  - interim ← *intermediate data that has been transformed*
  - processed ← *final data*
  - raw ← *original data*
- docs ← *scientific papers*

Github - 2/2

- notebooks ← *python notebooks*
- report ← *LaTeX report*
- src ← *source code*
  - data ← *scripts and programs to process data*
  - external ← *external source code*
  - models ← *source code for model*
  - tools ← *helper scripts*
  - visualization ← *visualization scripts*

Bibliography I

[1] Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 759–760, 2017.

[2] Quan Do. Jigsaw unintended bias in toxicity classification. 2019.

[3] Spiros V. Georgakopoulos, Sotiris K. Tasoulis, Aristidis G. Vrahatis, and Vassilis P. Plagianakos. Convolutional neural networks for toxic comment classification. In *Proceedings of the 10th Hellenic Conference on Artificial Intelligence*, SETN '18, pages 35:1–35:6, New York, NY, USA, 2018. ACM.

Bibliography II

[4] Hongyu Gong, Yuchen Li, Suma Bhat, and Pramod Viswanath. Context-sensitive malicious spelling error correction. In *The World Wide Web Conference*, WWW '19, pages 2771–2777, New York, NY, USA, 2019. ACM.

[5] Chia Lun Huang. The 2016 global report on online commenting, 2016-10-06.

[6] Fahim Mohammad. Is preprocessing of text really worth your time for toxic comment classification? In *Proceedings on the International Conference on Artificial Intelligence (ICAI)*, pages 447–453. The Steering Committee of The World Congress in Computer Science, Computer ..., 2018.

Bibliography III

[7] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proc. of NAACL*, 2018.

[8] Conversation AI Google team. Identify and classify toxic online comments.

[9] Betty van Aken, Julian Risch, Ralf Krestel, and Alexander Löser. Challenges for toxic comment classification: An in-depth error analysis. *CoRR*, abs/1809.07572, 2018.

[10] Zeerak Waseem and Dirk Hovy. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93, 2016.