Automatic moderator

Identify and classify toxic online comments

Maxime Guillaume
Sophea Ly
Guilherme Razet
Yi-ting Tsai

Table of contents

Previously, ...

- Why do we need it?
- Kaggle challenge
- Our work: experiment of combination of different neural network models and word embeddings
- Corpus: various type of toxicity

Linguistic aspects of the corpus

- Toxic
  - ”ok stop being lame. seriously. go watch pokemon.”
- Severe toxic
  - ”You should die from cancer”
- Obscene: purient content
  - F- words, terms of sexual body types
- Threat: intention to inflict injury or damage
  - ”I am going to shove a pineapple up your ass.”
- Insult: scornful or abusive remark
  - N- words, profanity and curse words
- Identity hate: disparaging words towards certain group
  - ”STAY THE FUCK OFF MY PAGE YOU HOMOSEXUAL.”

Data visualization

- Toxicity not spread out evenly;
- Class imbalance.

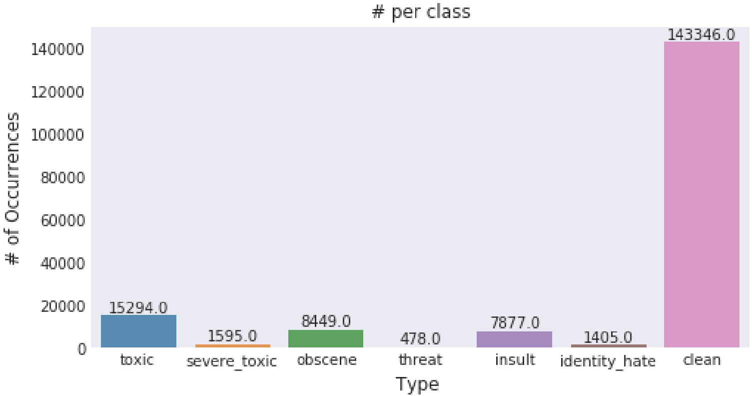


Figure: Distribution of tags across 159571 comments in corpus

Data visualization

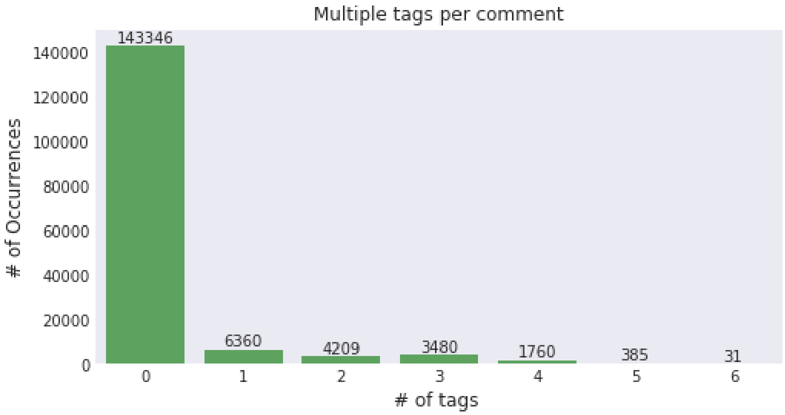


Figure: Multi-tagging in training data

Correlation

- A severe toxic comment is always tagged toxic;
- Other classes seem to be a subset of toxic barring a few exceptions.

	toxic	severe_toxic	obscene	threat	insult	identity_hate
toxic	1	0.309	0.677	0.157	0.648	0.266
severe_toxic	0.309	1	0.403	0.124	0.376	0.202
obscene	0.677	0.403	1	0.141	0.741	0.287
threat	0.157	0.124	0.141	1	0.15	0.115
insult	0.648	0.376	0.741	0.15	1	0.338
identity_hate	0.266	0.202	0.287	0.115	0.338	1

Figure: Variable correlation table of training data

Preprocessing: model 1 [2]

- Split tokens by white space
- Covert letters to lower case
- Remove punctuation
- Remove tokens that are not alphabetic
- Remove stopword
- Remove shortwords (one letter)
- Lemmatising

Preprocessing: model 2 [9]

- Convert letters to lowercase
- Remove punctuation
- Remove stopword
- Stemming
- Lemmatising

Preprocessing: model 3

- Convert letters to lowercase
- Remove stopword
- Remove white space
- Spelling correction
- Tokenization
- POS-tagging

GLUE benchmark

- General Language Understanding Evaluation (GLUE);
- Benchmark;
- Dataset;
- Public leaderboard.

Rank	Name	Model	URL	Score
1	T5 Team - Google	T5		89.7
2	ALBERT-Team Google Language	ALBERT (Ensemble)		89.4
3	王玮	ALICE v2 large ensemble (Alibaba DAMO NLP)		89.0
4	Microsoft D365 AI & UMD	FreeLB-RoBERTa (ensemble)		88.8
5	Facebook AI	RoBERTa		88.5
6	XLNet Team	XLNet-Large (ensemble)		88.4

Word Embeddings

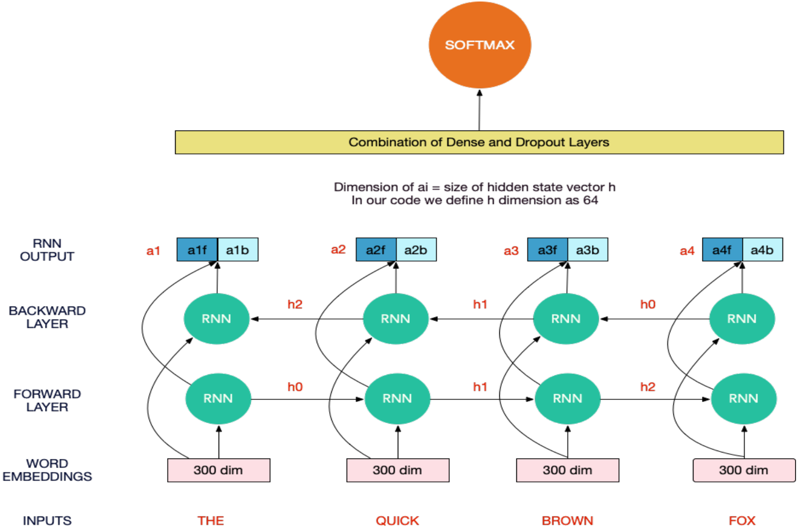
- 1 library to use differents word embedings models : Transformers
- 2 different models :
  - RoBERTa;
    - Based on BERT;
    - trained with bigger batches;
    - Over more data;
  - XLNET-Large;
    - Create to be better than BERT;
    - Overcomes problems inherent to the method used by BERT.

Deep Learning Models

- 3 models ready to use;
- Implemented in PyTorch;

BiLSTM

A comment is a sequence of words.

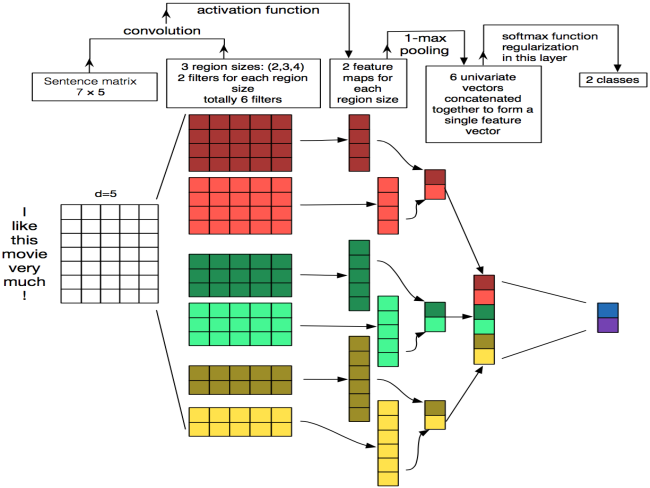


LSTM + Attention

A comment is a sequence of words and some word are more important than the others. We introduce an attention mechanism because ”Attention is all you need!”.

CNN

A comment is like a picture with some hidden patterns.



Methods to try

Methods
Roberta + BiLSTM
Roberta + CNN
Roberta + LSTM + Attention
XLNet + BiLSTM
XLNet + CNN
XLNet + LSTM + Attention

with 3 differents preprocessing.

Open issues

- No enough computational power to train the models in a reasonable time-frame;
- Some comments exceed the maximum sequence length that the embedding models are able to manage (at the moment, we have to cut a part of the comment).
- Spelling mistakes in the comments.

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] Nupur Baghel. Toxic comment classification, 2018-03-30.

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

[4] 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

[5] 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.

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

[7] 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

[8] 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.

[9] Ziyao Tang. Machine learning capstone report title: Toxic comment classification, 2018-04-24.

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

[11] 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.

Bibliography IV

[12] 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.