OffensEval 2019 – BERT with Logistic Regression

University of Arizona College of Science | Computer Science

About Your Topic

The goal of the task is to identify offensive language in contexts of social media platforms and online communities.

Annotated Twitter data is used for training and testing.

The Idea?

Use word and character based features from the training dataset to build offensive language classifiers.

Use an Ensemble of a fine-tuned BERT model for context based classification and logistic regression for confirmation of BERT's classifications.

References

Task and Dataset:
https://scholar.harvard.edu/malmasi/olid

Code:
https://github.com/rmattam/offenseval

Contacts

- Rahul Roy Mattam at rmattam@email.arizona.edu

Results

Logistic Regression Models
- Unigram Word Model
  F1 macro – 71.2% F1 micro – 79.5%
- Character-7gram model
  F1 macro – 71.1% F1 micro – 77.3%

BERT Model
- F1 macro – 73% F1 micro – 80.23%

BERT Ensemble Model
- F1 macro – 74.6% F1 micro – 81.7%

Ranked 40th in leaderboard out of 104 teams

Examples where Ensemble works

- **BERT and Unigram Model is correct, Char-7gram Model is wrong:**
  "Sometimes my brain hits upon a thought or impression and it *ding*s back, ringing true like crystal" – Char model is thrown off by offensive word patterns. The Gold Label is NOT OFFENSIVE.

- **Unigram Model is wrong, BERT and Char-7gram Model is correct:**
  "Yeah thanks to your Nobel Emmy award winning idiot chief flip flopping on everything from Iran to gun control" – Unigram model is thrown off by offensive words. The Gold label is OFFENSIVE.

Error Analysis

- **Problems with Unigram Model :**
  1) Fails to capture misspellings of offensive words egs: f$$$$$$, fuckkin.
  - Char-7grams worked well in this scenario
  2) Flags non offensive phrases incorrectly as offensive due to offensive words being present. Egs: "Calm the fuck down", "required with old ass games", "the gay community does everything" BERT works well here.

- **Problems with Character Model:**
  1) Frequently thrown off by user names having offensive words, @bobbysbadbitch i'm tryna marry johnny, not you so.. BERT and Unigram model are able to cope with this.

- **Problems with BERT: 1)** Harder to interpret **2)** Still fails to recognize subtle offensive language: egs: "you are nothing more than a #pigeon, pooping upon the chess board of this thread."

- **Problems with all three models:**
  Model is biased to some keywords. Such as sentences containing the keyword BrietbartNews often gets tagged as offensive

- **Problems with Ensemble:**
  BERT is correct, but, other two classifier fails:
  "Most of those harrassers should be thrown off Twitter under the rules. Singleing out an individual" – The gold label is NOT OFFENSIVE.

THE UNIVERSITY OF ARIZONA