Assignment Hate speech lexicons

Course

Subjectivity Mining

Type of the assignment

- Task: analyze and compare hate speech lexicons
- group assignment - 1 submission per group
- include Appendix with overview of who did what
- Grading: [0..10]
- submit what:
  - the merged lexicon (zip file) (see Step 3)
  - the code for a lexicon-lookup approach (see Step 4)
  - the answers to the questions below that are preceded by [S]
- submit how:
  - Submission: on Canvas
  - submit when: see Canvas
  - naming convention: A3-[groupname]

Aim of the assignment

- Getting familiar with existing hate speech lexicons
- Be able to discuss similarities and dissimilarities between different hate speech lexicons
- Understanding different approaches to build these lexicons
- Using them in a lexicon-based classification task
- Perform a quantitative and qualitative error analyses of the classification task
- Build a lexicon that can be used in assignment 4

Method of work

Step 1: Collect the following lexicons and read the papers

- Wiegand

    Wiegand et al.(2018) Inducing a Lexicon of Abusive Words – a Feature-Based Approach

  - data: `https://github.com/uds-lsv/lexicon-of-abusive-words`
  - paper: `https://aclanthology.org/N18-1095/`

- Hurtlex
  - Bassignana, E. et al. (2018) Hurtlex: A Multilingual Lexicon of Words to Hurt. In Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-It 2018)
  - data: `https://github.com/valeriobasile/hurtlex/blob/master/lexica/EN/1.2/hurtlex_EN.tsv`
  - paper: `https://ceur-ws.org/Vol-2253/paper49.pdf`

- MOL
  - Vargas et al. (2021) Contextual-Lexicon Approach for Abusive Language Detection
  - data: `https://github.com/franciellevargas/MOL/blob/main/data/mol.csv`
  - paper: `https://aclanthology.org/2021.ranlp-1.161.pdf`

Step 2: Describe the lexicons

- For each of the 3 downloaded lexicons: (NB Focus on English and - if applicable - skip the information on other languages)

    S Describe how these lexicons are built

    S Report statistics

    S Explain all categories and give examples found in the lexicon

    S Give a representative sample (10 to 20 entries) with all information, and discuss the quality

    S Address issues that you find relevant for the quality, consistency and/or coverage of the lexicon

Step 3: Merge the lexicons

- Create one lexicon by merging the 3 downloaded lexicons.

  S If you merge the information found in different lexicons, you have to make choices concerning not (completely) matching categories and overlapping entries. Describe and motivate your choices.

  S Describe the resulting merged lexicon in terms of statistics.

  S Give a representative sample (10 to 20 entries) with all information, and discuss the quality

Step 4: Use the lexicons for automatic hate speech identification

S Design a (simple) lexicon-lookup approach for binary classification; describe the design

- Run this approach with the 4 lexicons on the test set of the dataset OLID (see footnote [1]

S Report results in terms of precision, recall, F-measure per category and macro-F1 (and discuss)

- For the results of the highest scoring classifier :

  S Make a confusion matrix (and discuss)

  S Perform a qualitative error analysis discussing patterns of errors

    ∗ get inspiration from the challenges discussed in (van Aken et al., 2018, section 6) `https://aclanthology.org/W18-5105.pdf`).

    ∗ focus on false negatives and false positives

    ∗ provide your own ideas on types of errors can be reduced .

Step 5: Short conclusion

– Write a short conclusion based on your findings (cf. step 3 and 4) and address what is -according to you - the 'best' lexicon. Motivate your answers.

[1] You will be working with the OLIDv1 dataset `https://canvas.vu.nl/courses/72011/files/folder/datasets?preview=6639499`, which contains 13,240 annotated messages (tweets) for offensive language detection. The detailed description of the dataset collection and annotation procedures can be found here `https://aclanthology.org/N19-1144/`. The current assignment focuses on Subtask A (identify whether a tweet is offensive or not). The dataset is preprocessed so that label '1' corresponds to offensive messages ('OFF' in the dataset description paper) and '0' to non-offensive messages ('NOT' in the dataset description paper) . For the evaluation of the lookup approach you will use the test set (olid-test.csv.)