Reading and annotation assignment: hate speech

Course

Subjectivity Mining
Topic: automatic hate speech detection

Type of assignment

- The assignments consists of 2 parts:
  - Task 1: reading 3 papers
    - individual task
    - 1 submission per student
    - naming convention: A1-Task1-[studentname]
  - Task 2: performing an annotations study
    - group task
    - 1 submission per group
    - naming convention: A1-Task2-[groupname]
- Task 1 needs to be completed before task 2
- Grading task 1: Fail/Pass/Good
- Grading task 2: [0..10]
- Submission: on Canvas
  - submit what and how:
    - Answers Task 1 (one document per group member)
    - Answers Task 2 (one document per group)
    - Excel sheet with 3*2 sets of annotations
    - !! Please submit one zip file per group including both the answers to Task 1 and Task 2
- submit when: see Canvas

Aim of the assignment

- Getting familiar with the methodology of annotations
- Getting familiar with the concept of hate speech
- Getting familiar with different annotations schema for the identification of hate speech

- Perform an annotation task
- Perform an inter-annotator agreement task
- Perform an error analysis
- Be able to discuss similarities and dissimilarities between different concepts of hate speech and the way they are annotated in text

Method of work

There are the following parts:

- read 3 papers (each group member reads all papers)
- answer the questions about the papers
- annotate the data
- calculate inter-annotator agreement
- discuss problems and write error report
- include Appendix with overview of who did what

Task 1: Read papers

Read the following 3 papers. Read the whole paper, but focus on the definitions of hate speech, the annotation guidelines and examples , inter-annotator scores. For now, the methods sections are less important. Address the following issues for each paper (Submit)

- do you have clarifying questions? Did you understand everything?
- give a short overview of the annotation guidelines presented in the paper
- what is the motivation for the paper
- what is the research question
- what did they find
- what is their conclusion
- what are -according to you- interesting aspects of the paper

Task 2: Annotation study

The papers you read present definitions of hate speech and guidelines how to identify them in text. Of course, they are not full guidelines, but the categories are described in enough detail to understand. For more information you can also have a look at the datasets (see `https://canvas.vu.nl/courses/50234/pages/datasets-in-vua-format`).
Carry out the following annotation task (and Submit)

- Annotate the set of 45 messages according to the 3 annotation schemas presented in the paper. Each set of guidelines is applied independently by 2 annotators. This results in 6 sets of labels per message. (Use the following set of labels: Gilbert (Hate/noHate); Kumar (OAG/CAG/NAG); Zampieri (OFF/NON, TARGET/NOTARGET, Group/Individual/Other)

Address the following issues (and submit):

- Calculate inter-annotator agreement: Inter-annotator agreement is a measure of how well two (or more) annotators can make the same annotation decision for a certain category. Agreement can be calculated in several ways. Traditionally, percentage agreement (calculated as the number of agreement scores divided by the total number of scores) was commonly used. More recently, Cohen's kappa[1] which takes into account chance agreement is developed and it is now regarded as a more suitable measure. You can use this online tool[2] to calculate KAPPA and percentage agreement. Of course, you can also calculate it yourself.

- Make a confusion table. In addition to agreement scores, a confusion matrix is often used to generate a more complete picture of how the annotators performed. It shows on which categories and values high agreement is achieved and which categories are often *confused* with others. A single score (percentage or kappa) does not provide this information.

- Perform an error analysis. What classes are hard to identify for each set of guidelines. Why? What classes are easily confused? Why? Illustrate this with examples of the annotated dataset. An error analysis is a way to assess the annotations in qualitative terms. It involves the examination of a set of examples where annotators do not agree with each other, so that you can understand the underlying causes of the errors. Try to distinguish types of errors, search for patterns and systematic errors. Which types of errors are frequent and which are not? Examine also what the strong points of the annotations. Which cases are not problematic and how frequent are they? This kind of analysis helps you to prioritize problems that deserve attention and the analysis may lead to suggestions for improving the quality of the annotations.

[1] see e.g. `https://en.wikipedia.org/wiki/Cohen's_kappa`
[2] `http://dfreelon.org/utils/recalfront/recal2/`

  - select cases of disagreement for the error analysis. Base your selection on the confusion matrix and select cases with confusions between different labels
  - select cases of agreement
  - examine the cases of disagreement, search for patterns, try to classify them and count them. Compare these cases with the cases of agreement: are they really different?
  - report on your findings: Give an overview of the types of errors you found, their possible causes , and how frequent they are in your sample. Give examples of the types of errors.

- Discuss the differences between the different sets of guidelines:
  - Do they have similar definitions of hate speech? Do they address the same phenomenon?
  - Can they be reliably annotated?
  - Are the guidelines clear and do they cover all cases?
  - Which guidelines are best according to you? Why? Give arguments based on the annotation study.

Papers to read

- Paper A:
  Ritesh Kumar, Aishwarya N. Reganti, Akshit Bhatia, Tushar Maheshwari (2018) Aggression-annotated Corpus of Hindi-English Code-mixed Data. In: Proceedings of LREC-2018, Miyazaki, Japan.
  `https://arxiv.org/pdf/1803.09402.pdf`

- Paper B:
  Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal Noura Farra, Ritesh Kumar (2019) Predicting the Type and Target of Offensive Posts in Social Media. `https://arxiv.org/pdf/1902.09666.pdf`

- Paper C:
  O. de Gibert, N. Pérez, A. García-Pablos, M. Cuadros, 2018. Hate Speech Dataset from a White Supremacy Forum. In ALW2: 2nd Workshop on Abusive Language Online. `https://www.aclweb.org/anthology/W18-5102.pdf`