Does sentiment have an effect on bias and toxicity of tweets?

Content Analysis of a Twitter Corpus

•••

Advanced Information Retrieval, WS22, G19

Lizeth Chávez, Leonor Veloso, Robin Karlsson, Olli-Pekka Riikola

Project repository

9.1.2023

Introduction

- Content analysis is the task of extracting meaning (such as themes, trends, etc...) from data
- This content analysis shows information about
  - Bias
  - Toxicity
  - Emotions
  - Overrepresented Words
  - Similarity Measures

  In the corpus of our choice!
- Several models are used in the project.

Dataset

- The project uses a dataset available at Hugging Face.
- The whole dataset can be downloaded from here.
- Contains ~1.5M tweets labeled as either positive or negative
  - We use a small amount of these to keep the run time reasonable
    - Notebook in repo uses 250 positive and 250 negative tweets
    - Charts in presentation made using 20000+20000

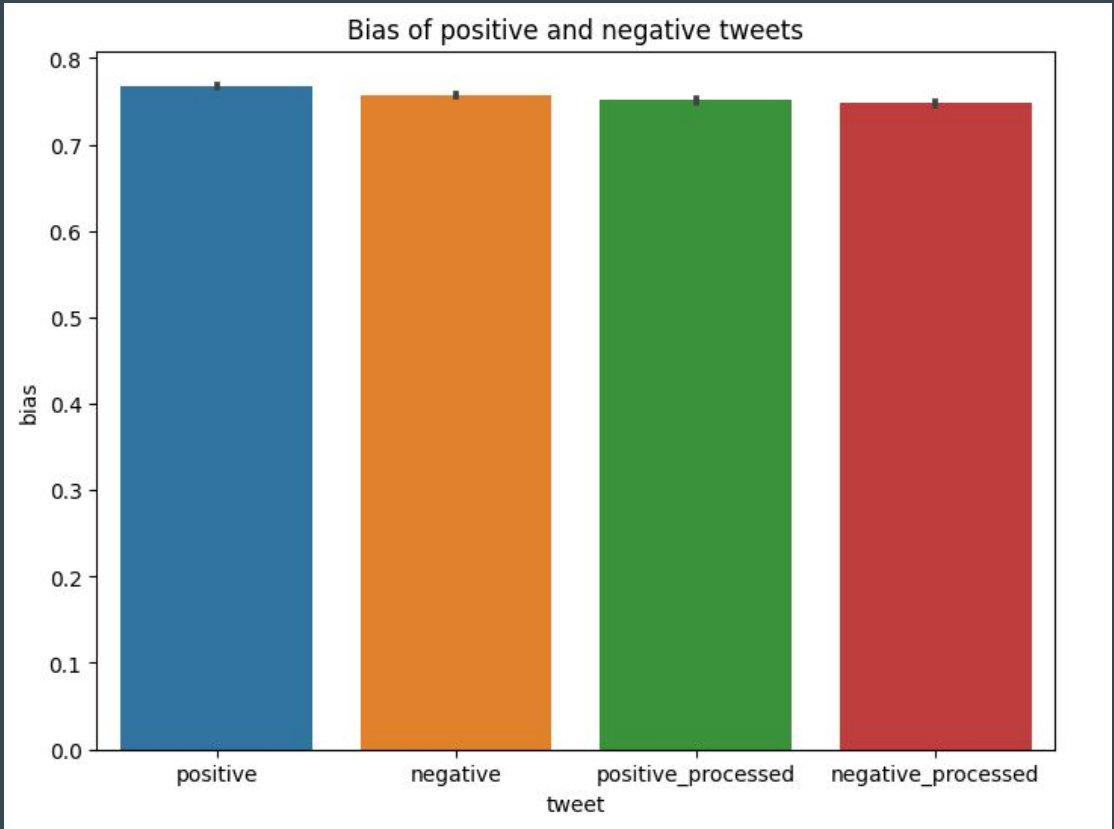
Background Information

Sentiment analysis (Mejova, Y., 2009):

- Sentiment reflects feelings of the user (Pang and Lee, 2008)
  - Binary
    - Positive
    - Negative
  - Polarity range
    - Ex: stars on a review
  - Opinion range
    - Ex: extremely positive, very positive, somewhat positive, neutral, somewhat negative, etc
- Ways to express it (Liu, 2006)
  - Explicit
    - “It’s a beautiful day”
  - Implicit
    - “The earphone broke in two days”

Bias Analysis

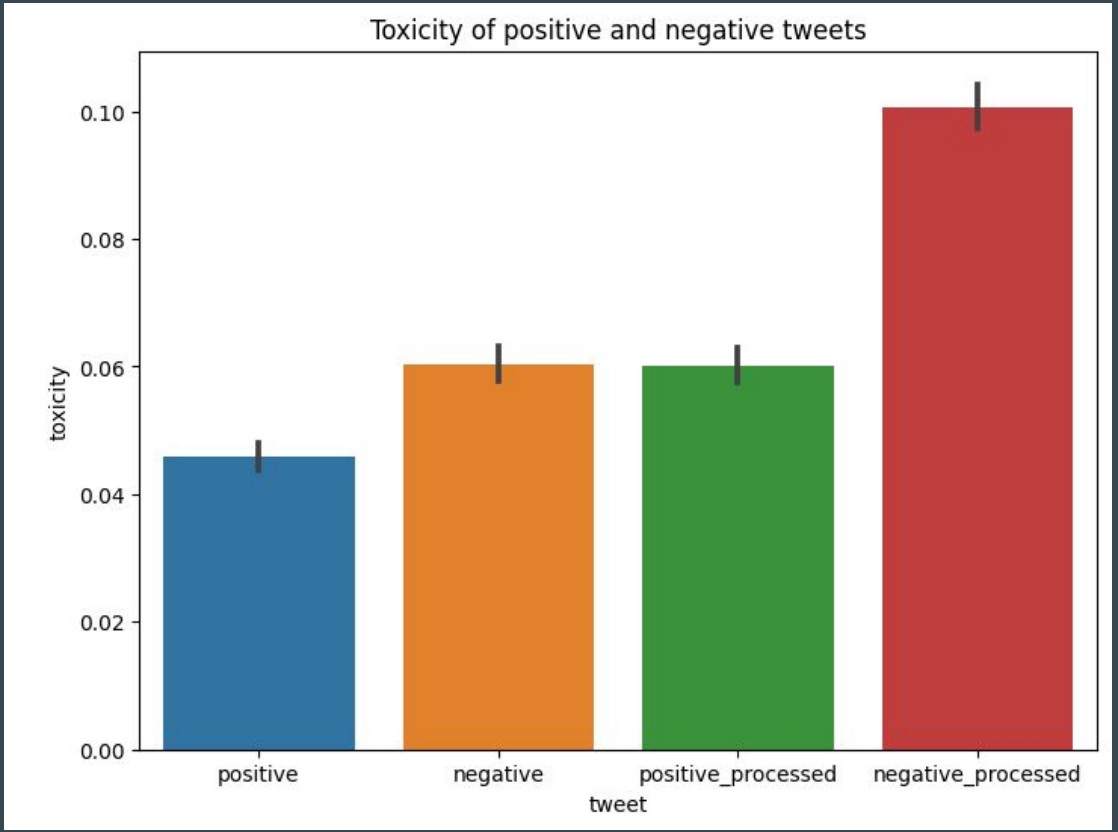
Bias-detection-model from *Bias & Fairness in AI, (2022)*



Plotted with 40 000 tweets

Toxicity Analysis

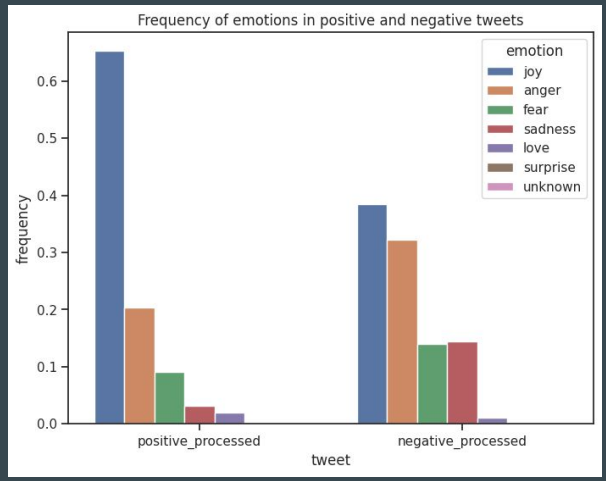
Roberta-hate-speech-dynabench-r4-target from *Learning from the Worst: Dynamically Generated Datasets to Improve Online Hate Detection*

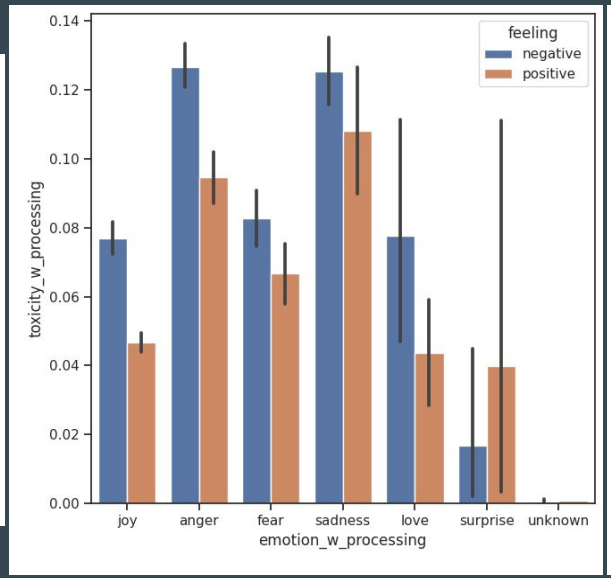


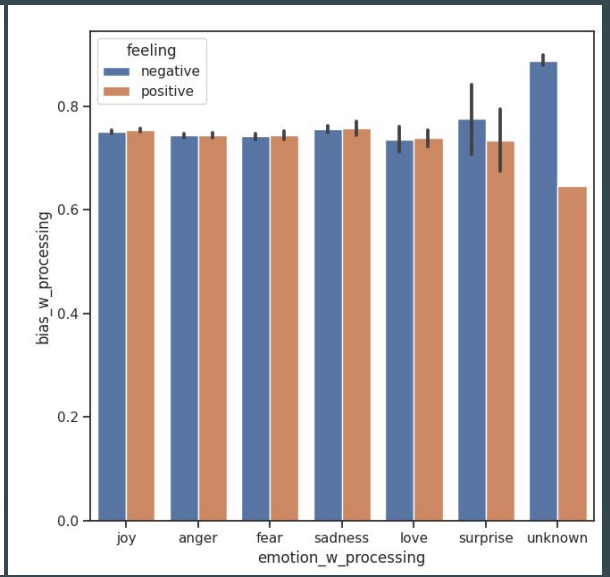
Plotted with 40 000 tweets

Emotion Analysis

T5-base fine-tuned for Emotion Recognition from *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer*

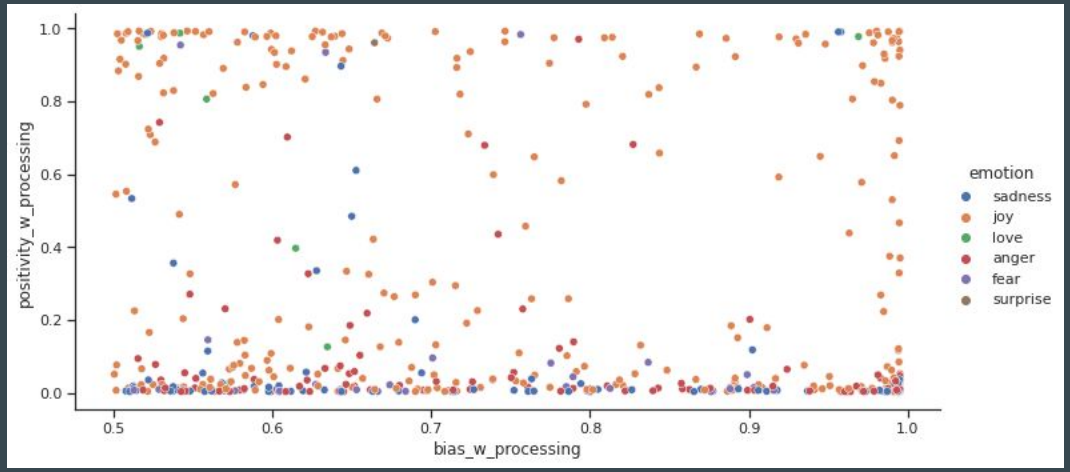


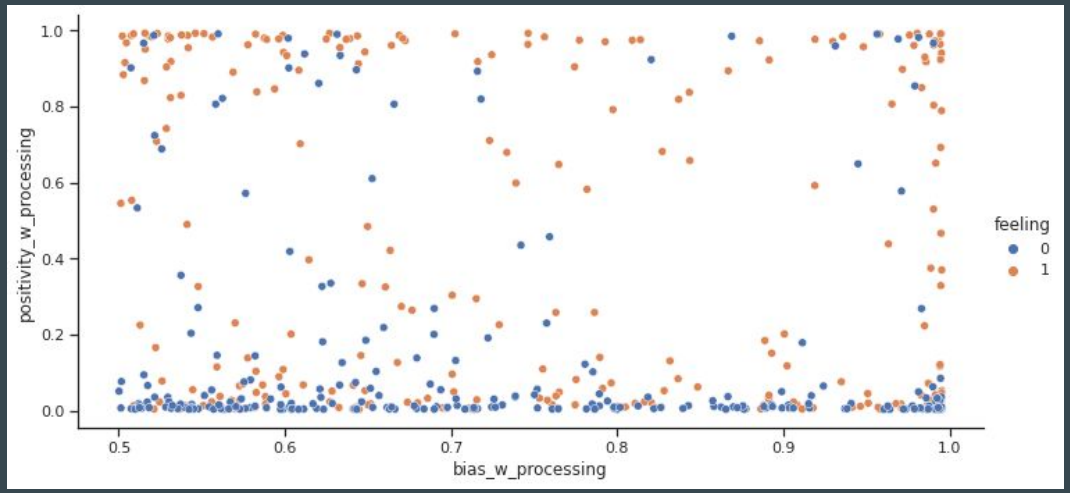




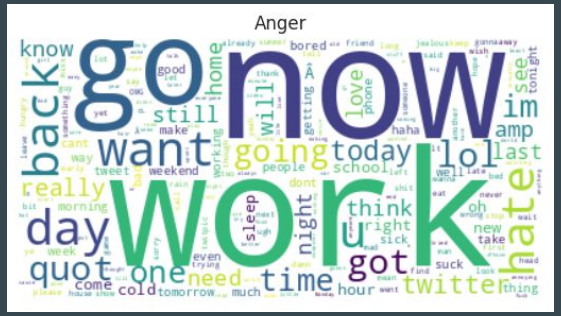
Positivity analysis

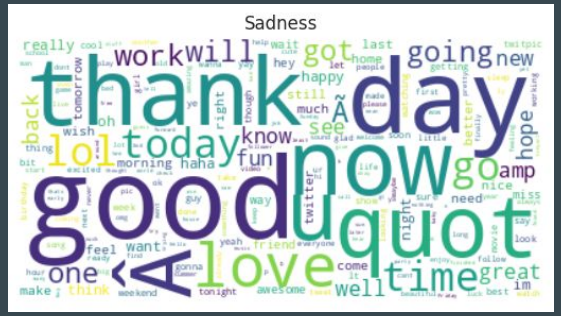
Pysentimiento from *A Python Toolkit for Sentiment Analysis and SocialNLP tasks*

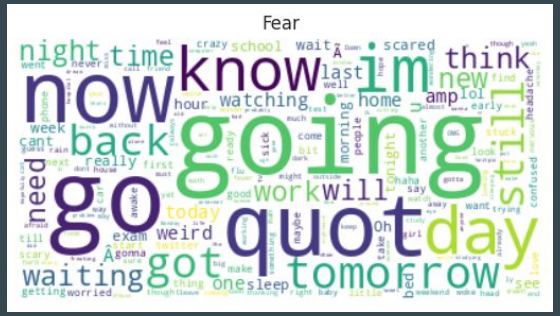


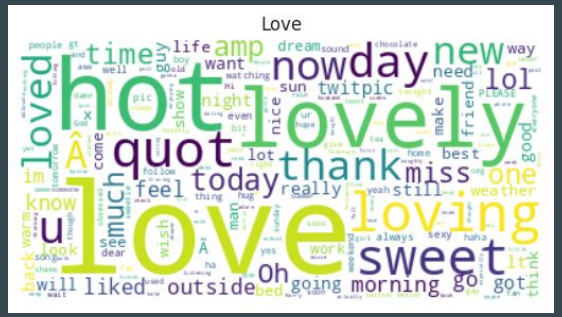


Overrepresented Words









Similarity Measures

Query Tweet (positive feeling): “@poepiandzegiant oops just saw you said hello! Hi there”

#1 “@phantompoptart .....oops.... I guess I'm kinda out of it.... Blonde moment -blushes- epic fail”

#2 “Said something harsh and didn't even realize it's harsh until I said it.. Sorry ”

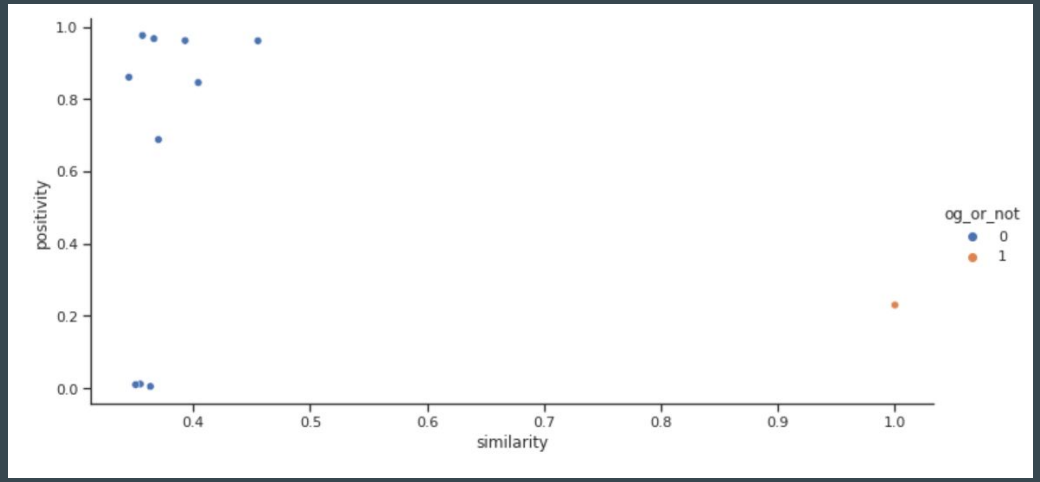
Query Tweet (negative feeling): “okay, so everyone else i went with seemed to hate brokeback mountain, or at least josie and zach did and they were the loudest criticizers.”

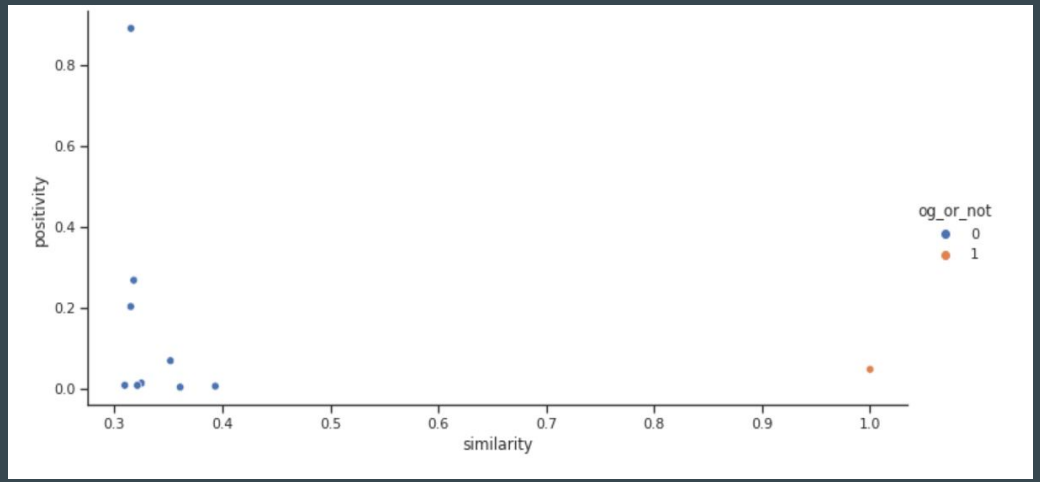
#1 “Sometimes, people who hate Twitter are so much more amusing than people who use Twitter....”

#2 “@AJDADDY lol I absolutely hate u!”

Similarity Measures

Similar tweets tend to have similar sentiment!





Conclusion

- Negative sentiment shows more toxicity
- All categories of tweets (positive or negative) exhibit the same level of bias
- Emotions such as joy present more positivity, whereas anger shows a very small amount of positivity