Sentiment Scoring of covid19 tweets

VON THILO BRUMMERLOH

IN METHODEN UND ANWENDUNGEN IN DEN DIGITAL HUMANITIES

Übersicht

- Thema
- Fragestellung
- Daten
- Methoden
  - Tweet-extraction
  - Pre-processing
  - Sentiment scoring
- Ergebnisse

Thema

- Globale Pandemie stellt sich Anfang 2020 ein
- Unübersichtliche Menge an Meinungen zu Covid19 auf twitter

Fragestellung

- Wie verhält sich die Meinung von twitter Nutzern in den ersten Monaten einer weltweiten Pandemie?

- => Große Mengen von tweets mithilfe von NLP bewerten

Daten

- Gabriel Preda – covid 19 tweets
  - 170.000 tweets von Juli bis August 2020
- Erdal Baran und Dimitar Dimitrov TweetsCOV19
  - 8.151.524 tweets von October 2019 bis April 2020
  - Seed Liste: https://data.gesis.org/tweetscov19/keywords.txt
- Christian Lopez und Caleb Gallemore
  - 1.209.642.955 tweets von Januar 2020 bis heute
  - Seed Liste: https://github.com/echen102/COVID-19-TweetIDs/blob/master/keywords.txt
- Twitter verbietet das veröffentlichen von tweettexten Anfang 2020

Tweet extraction

- Mit tweepy tweetIds mit Volltexten der Tweets hydrieren
- Twitter erlaubt 360.000 tweets pro Stunde

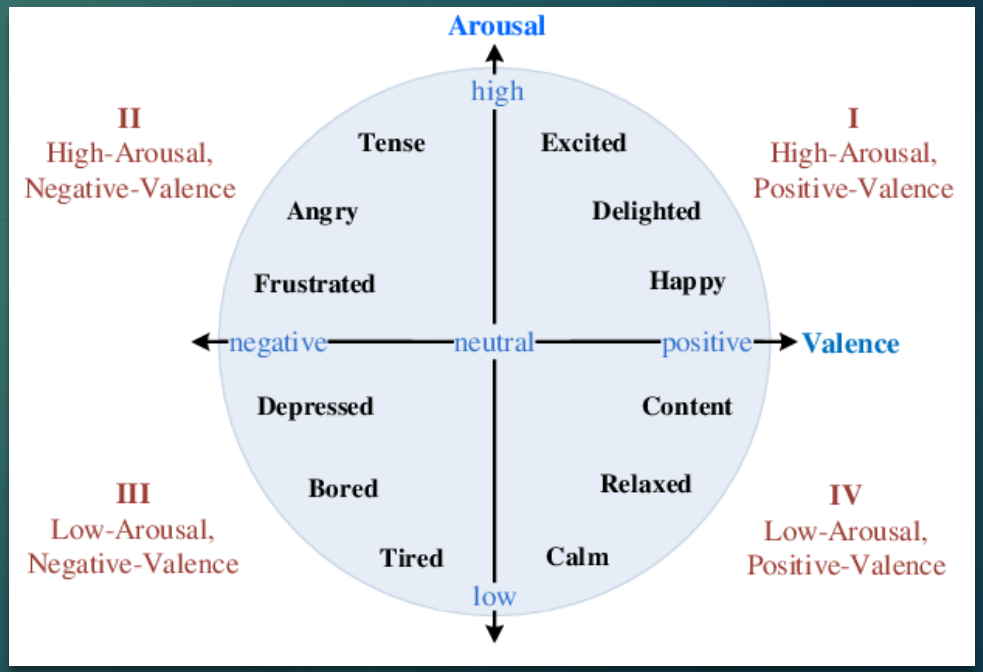
Pre-processing

- NLP-Bibliothek Spacy bietet Funktionalität zu
  - Tokenizing
    - Tweet texte zu Wortlisten umformen
  - Removing Stop Words
    - Für die weitere Verarbeitung irrelevante Wörter entfernen
  - Stemming & Lemmatization
    - Wörter auf den Stamm reduzieren und auf ein gemeinsames Wort vereinigen

Sentiment Scoring

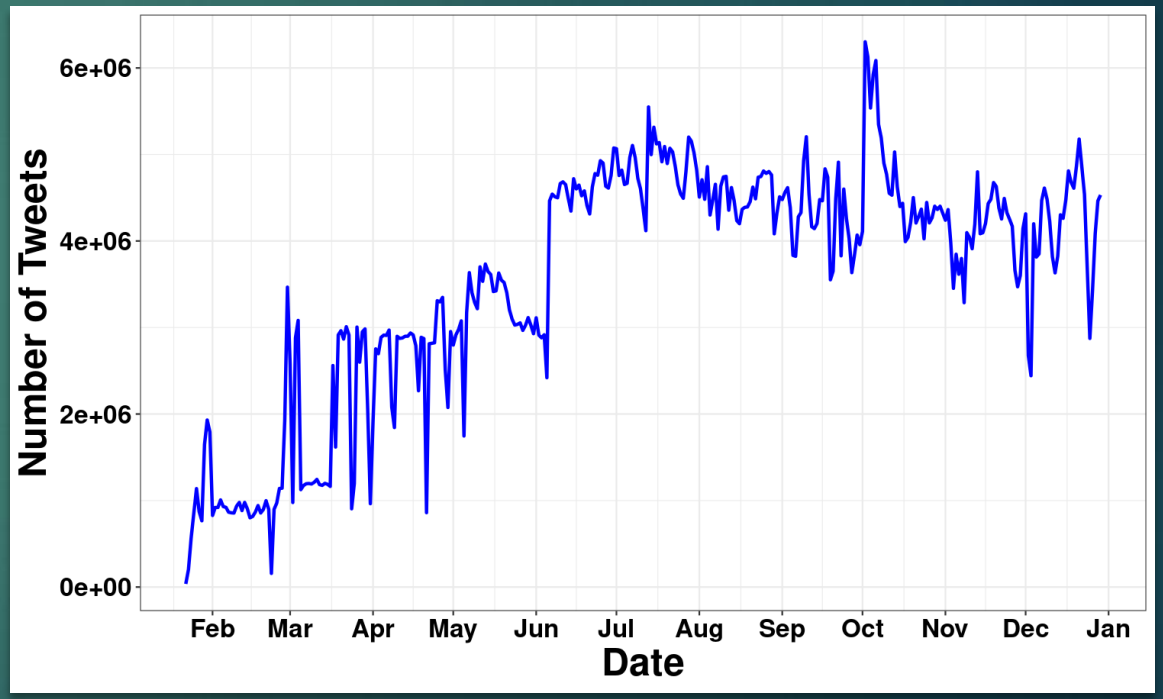
- Zu Wortlisten reduzierte tweets werden mit bewerteter Wortliste verarbeitet => Ergibt Zahlenwert des Sentiments einzelner tweets
- Wortlisten
  - AFINN-165 – 1-Dimensionale Bewertung
  - NRC VAD – 3-Dimensionale Bewertung [0,1][0,1][0,1]
- SentiStrength in Daten enthalten [-5,-1][1, 5]
- AFINN Liste
- NRC VAD Liste

Yu, Liang-Chih & Lee, Lung-Hao & Hao, Shuai & Wang, Jin & He, Yunchao & Hu, Jun & Lai, K. & Zhang, Xuejie. (2016). Building Chinese Affective Resources in Valence-Arousal Dimensions. 10.18653/v1/N16-1066.

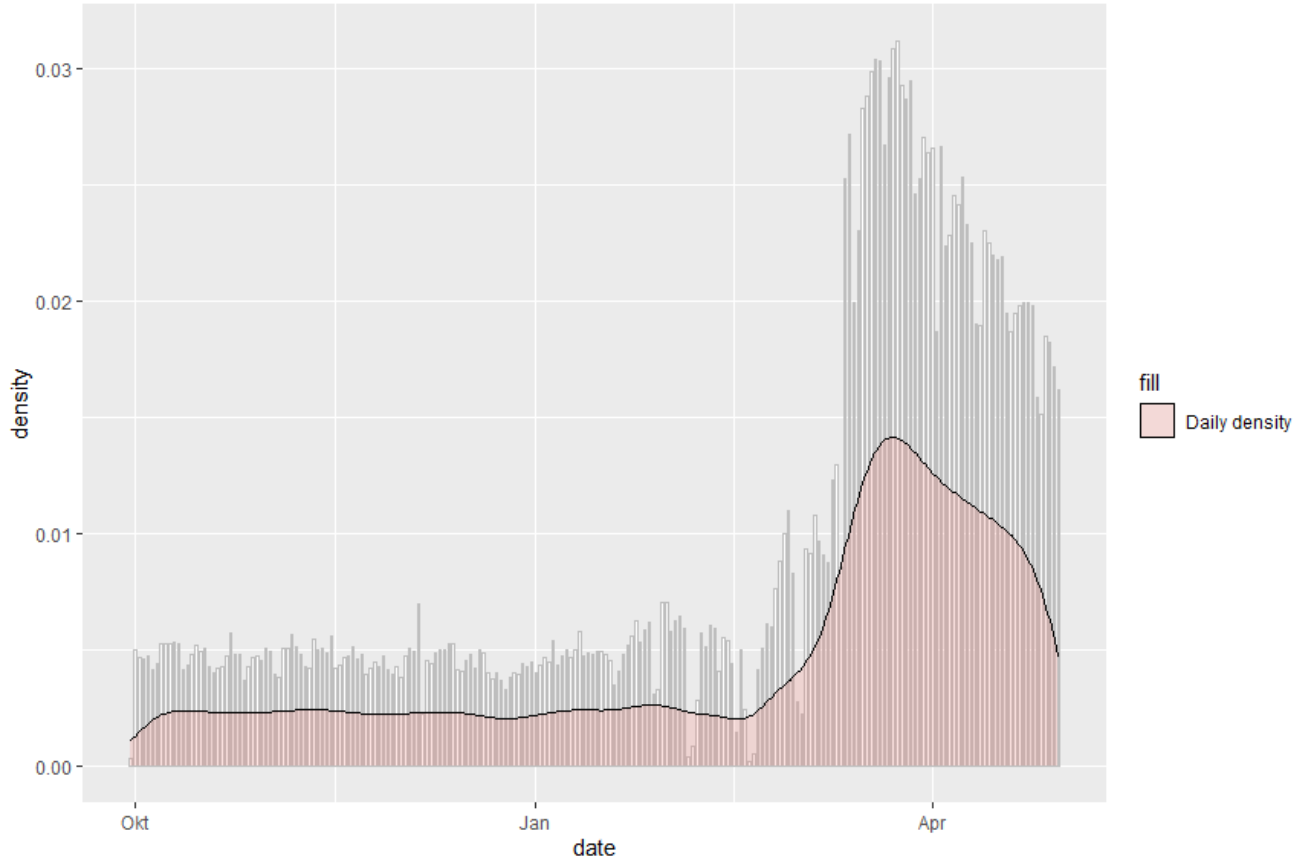


Ergebnisse

- Kleines Fenster schränkt die Aussagefähigkeit ein
  - Bearbeiteter Datensatz endet nach erstem Aufmerksamkeitspeak
  - Single Thread verarbeitet 150000 tweets pro Stunde:
    - >53h für 8M tweets,
    - >80kh für 12G tweets
- Tabelle mit Zeitstempel, Tweets, Lemma und 3 verschiedenen Bewertungsmethoden

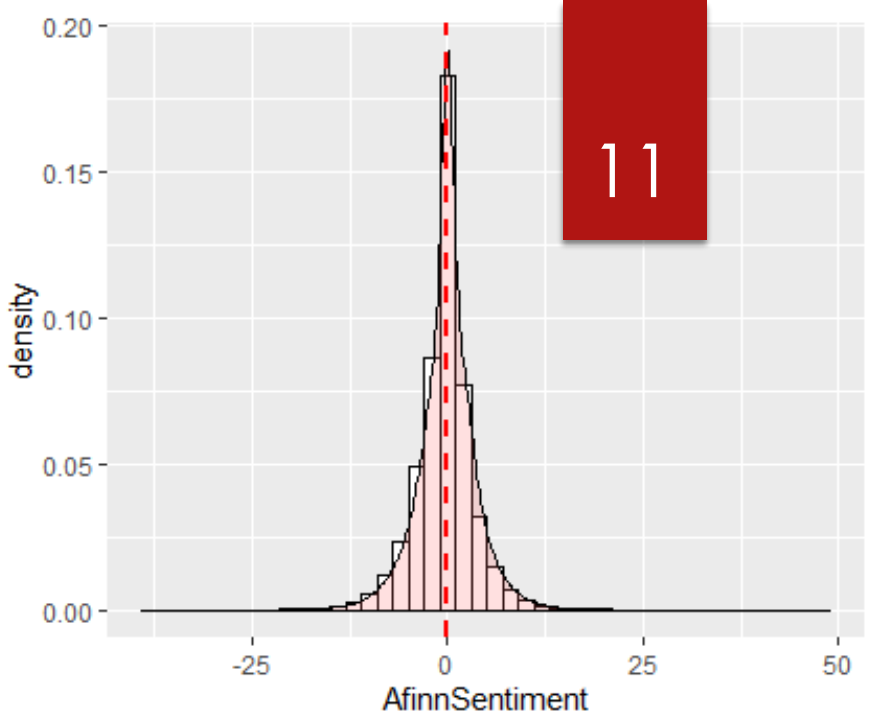


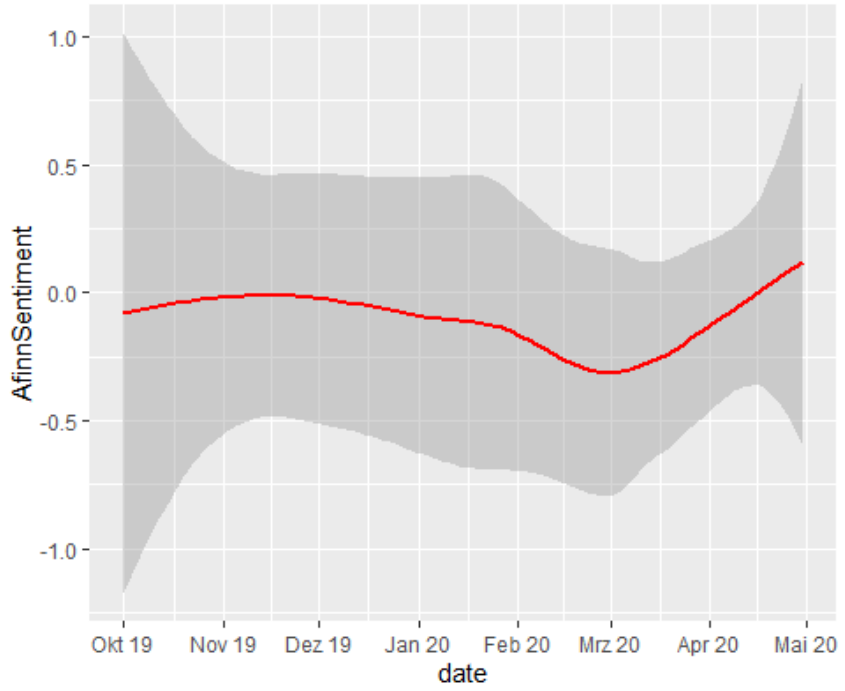
Tweet-Dichte

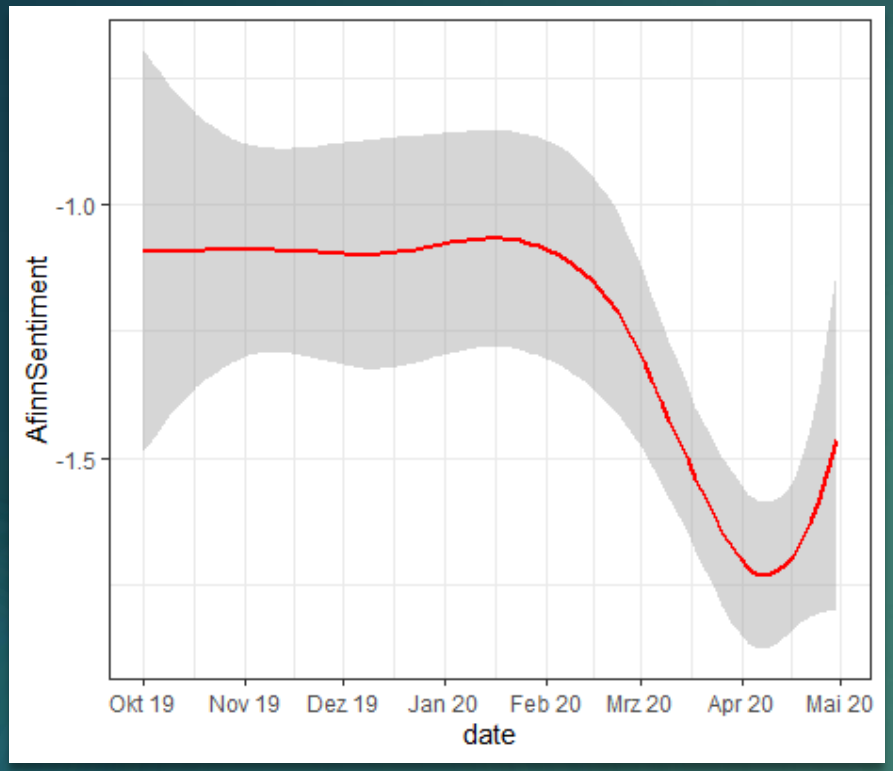


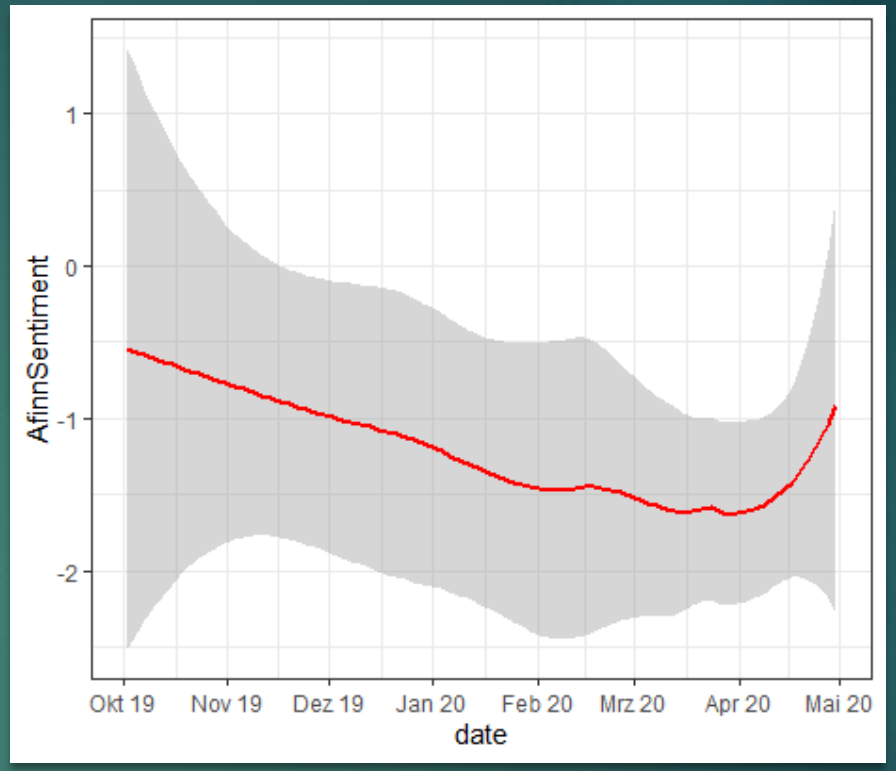
AFINN-165

- Positivster Tweet:
  - “Love Is About Choice Love Is What God Is Love Is Not A Feeling Love Is Your Identity Love Happens Love Makes World Go Around Love Makes Life Worth Living Love Is Standing By Each Other Forever Love Is Trust Love Is Peace Love Is Safe Love Is Acceptance Love Is Soul Love Is You https://t.co/j0pTqRsEcB”
- Negativster Tweet:
  - “This is for coronavirus you big fat white nasty smelling fat bitch why you took Fiesta off the motherfucking schedule with yourÂ trifflinÂ dirty white racist ass you big fat oompa loompa ass bitch iâ€™m coming up there and iâ€™m going to beat the fuck out of you bitch”



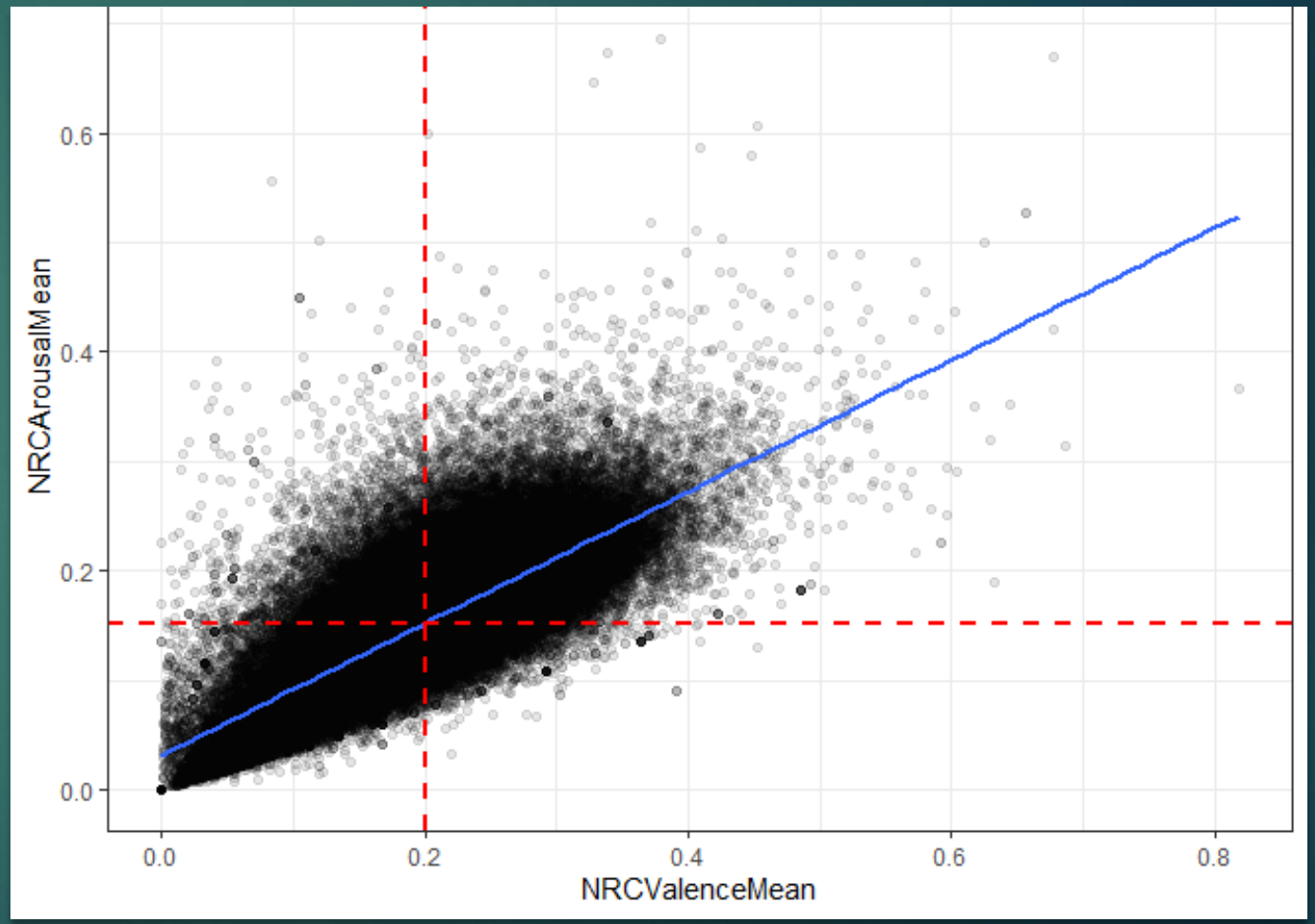






Trump & Pelosi Sentiments

NRC VAD



Literatur

- Baran, Erdal und Dimitar Dimitrov. (Juni 2020) TweetsCOV19 - A Semantically Annotated Corpus of Tweets About the COVID-19 Pandemic.
- Christian E. Lopez, Caleb Gallemore. (Oktober 2020) An Augmented Multilingual Twitter Dataset for Studying the COVID-19 Infodemic
- Emily Chen, Kristina Lerman, and Emilio Ferrara. (2020). #COVID-19: The First Public Coronavirus Twitter Dataset.
- Nielsen, Finn Arup (Mai 2011). A new ANEW: evaluation of a word list for sentiment analysis in microblogs“
- Saif M. Mohammad (Juli 2018). Obtaining Reliable Human Ratings of Valence, Arousal, and Dominance for 20,000 English Words.