DH Projekt - Sentiment Scoring of covid19 tweets

Gruppe: Thilo Brummerloh

3. Februar 2021

1 Thema

Auf twitter werden täglich unübersichtliche Mengen an Texten zu unterschiedlichen Themen geschrieben. Um einen Überblick zu erhalten kann eine Sentiment Analyse durchgeführt werden, die diese Datenmengen auf darstellbare Zahlen zusammenfasst.

2 Fragestellung

Die durchzuführende Analyse soll das Sentiment der Tweets aus dem betrachteten Datensatz herausfinden. Das Sentiment soll über den betrachteten Zeitraum anschaulich dargestellt werden und Auffälligkeiten sollen näher betrachtet werden.

Weiterhin sollen weitere Sentimentbewertungen durchgeführt werden. Diese sollen auf Basis unterschiedlicher Wortbewertungslexika sowie unterschiedlich implementierter Methoden erstellt werden. Somit lassen sich auch vergleiche der Sentiments und mögliche Korrelationen zwischen den Methoden erkennen.

3 Daten

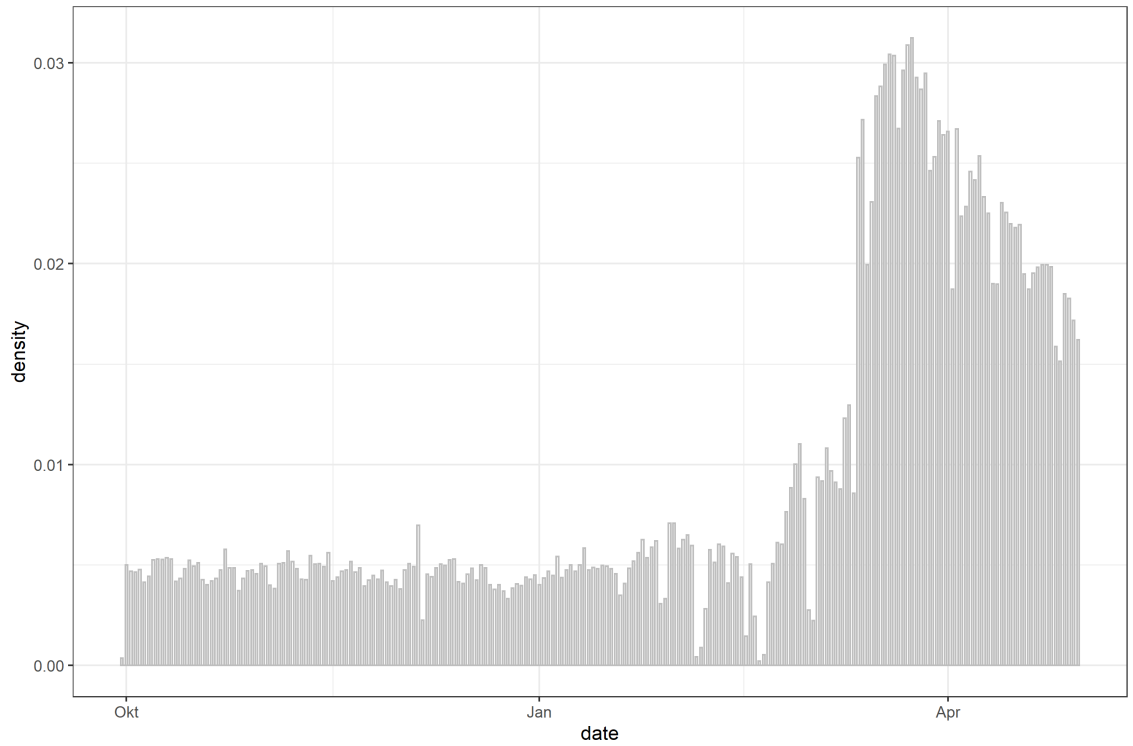
3.1 Daten zum Anfang der Pandemie

Als Datenbasis sollen tweets, die über einen Zeitraum mehrerer Monate gesammelt wurden und in Verbindung mit dem Thema des COVID-19 stehen, verwendet werden.

Die Datenbasis von Baran und Dimitrov 2020 wird herangezogen. Der Liste[1] der Autoren entsprechend, wurden darin tweets aus dem Zeitraum Oktober 2019 bis April 2020 abgelegt. Insgesamt handelt es sich um eine Liste von 8.151.524 TweetIDs. Zu den IDs werden außerdem eine Reihe von Metadaten zum referenzierten Tweet mitgeliefert. Der Volltext des Tweets ist aber, entsprechend den Vorgaben von twitter, nicht veröffentlicht worden und muss noch hinzugefügt werden.

[1] https://data.gesis.org/tweetscov19/keywords.txt

Abbildung 1: Anteiliges Auftreten von Tweets pro Tag in Daten von Baran und Dimitrov 2020(eigene Abbildung)



3.2 Daten über den gesamten Zeitraum der Pandemie

Weiterhin gibt es eine Datenbasis von Lopez und Gallemore 2020 über den gesamten Zeitraum der Pandemie angefangen am 22. Januar 2020 die weiterhin stündlich aktualisiert wird. Diese enthält über eine Milliarde Tweets. Aufgrund des hohen Aufwands der sich durch solche Datenmengen auftut, wird der Korpus aber nicht weiter bearbeitet.

4 Methoden

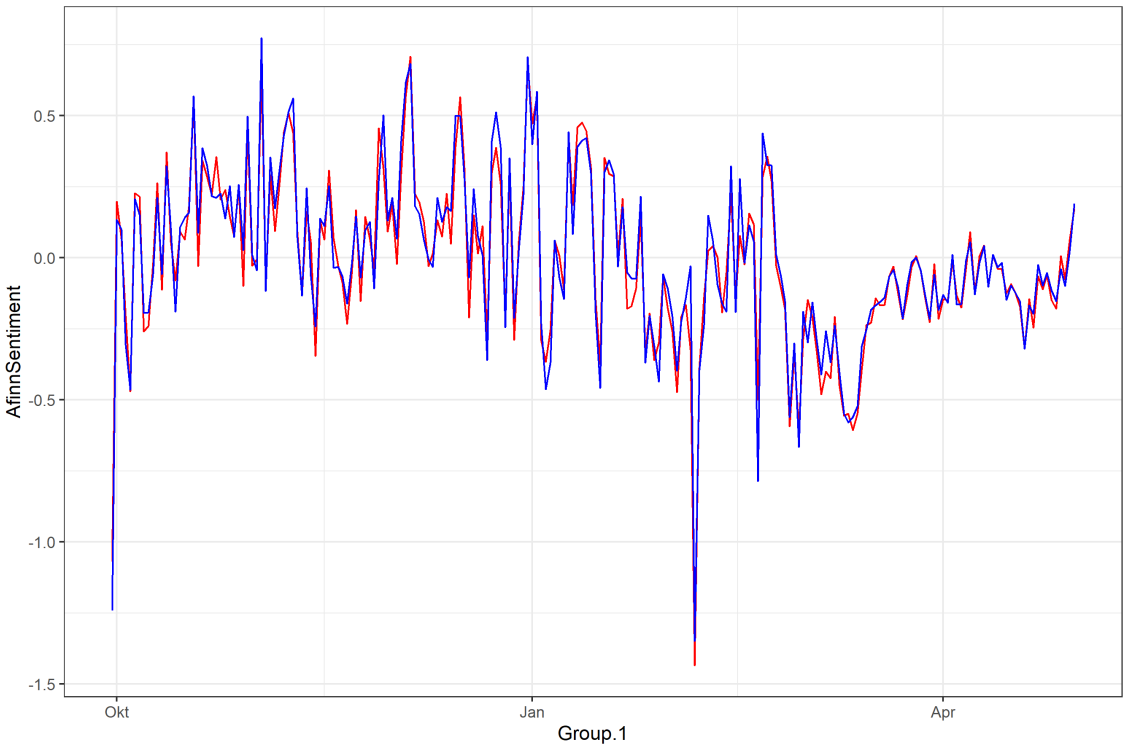
4.1 Hydrieren der TweetIDs

Da twitter eine Veröffentlichung von Tweet-Volltexten untersagt, muss im ersten Schritt der Datensatz durch diese erweitert werden. Die python Bibliothek tweepy bietet dazu die nötige Vermittlung zwischen Programm und API von twitter. In der Datei `2_tweeterExtraction.ipynb` befinden sich alle benötigten Programmteile um eine csv-Datei einzulesen, die darin enthaltenen IDs von twitter auf ihre gesamten Metadaten mit Volltext zu erweitern und in einer nächsten csv-Datei abzulegen.

4.2 Vorbearbeitung der Daten

Die im Datensatz enthaltenen Tweets müssen zum sentiment scoring aufbereitet werden. Dazu müssen Stoppwörter entfernt werden. Diese sind für die spätere Bewertung vernachlässigbare Beigaben des Textes. Alle übrigen Wörter werden durch Stemming, dem runterbrechen des Wortes auf seinen Stamm, und Lemmatization,

Abbildung 2: Eigene und in Afinn Bibliothek definierte AFINN165 Bewertung über den gesamten Zeitraum der Daten.(eigene Abbildung)



das wieder erweitern auf ein vereinigendes Wort, vereinheitlicht. Zur Aufbereitung wird die Pythonbibliothek spaCy verwendet.

4.3 Sentiment Bewertung

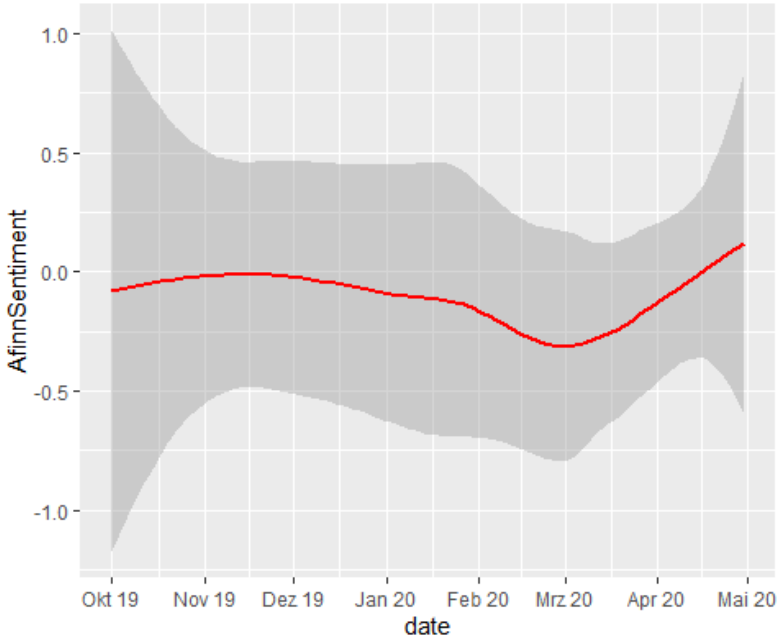
Als erstes wird die Bewertung mithilfe des Wortlexikons AFINN-165 von Nielsen 2011 durchgeführt. Damit ergibt sich für jeden Tweet ein Sentimentwert der die darin ausgedrückte Stimmung wiedergeben soll. Dies wird zuerst mit einer selbst geschriebenen Funktion, danach nochmal mit der in der python-Bibliothek 'Afinn' integrierten Funktion afinn.score durchgeführt. Die Ausgaben unterscheiden sich aufgrund anderen Vorbearbeitungen der beiden Methoden. Weiterhin wird eine Bewertung mit dem Valence-Arousal-Dominance (VAD) Wörterbuch von Mohammad 2018 berechnet. Ein Aufruf der entsprechenden Funktion erfolgt wie der selbst geschriebene Teil der AFINN-Implementierung. Allerdings werden mit der VAD die Sentiments in einem Zahlen-Tripel ausgedrückt.

In den Daten von Baran und Dimitrov 2020 sind bereits Bewertungen enthalten die mit SentiStrength erstellt wurden. Auch diese Methode verwendet ein Lexikon von Wörtern zur Bewertung der eingelesenen tweets.

5 Ergebnisse

Die programmierten Texte sind in `.ipynb`-Dateien abgelegt und entsprechend ihrer Funktion benannt. Die zugrundeliegenden Daten sollten im Unterordner ./data/ liegen. Als Ausgangsdatei kann eine beliebige csv-Datei verwendet werden, die Tab-

Abbildung 3: Durchschnittliche AFINN Bewertung über den gesamten Zeitraum(eigene Abbildung)



stoppgetrennt ist und eine Spalte mit dem Namen `TweetID` hat. Davon ausgehend können die unter Methoden beschriebenen Schritte durchlaufen werden um Tweet-Sentiments zu erhalten.

Es ist außerdem eine Datei mitgeliefert in der mithilfe von R Diagramme auf Basis der Ergebnistabelle erstellt werden können. Die Sentiments können über einen Zeitraum in Graphen dargestellt werden.

In Abbildung 3 ist das mit Afinn berechnete Sentiment über den gesamten Zeitraum dargestellt. Es lässt sich ein leichter Negativtrend ab Ende Januar feststellen, dieser kehrt sich mit Anfang März wieder und verläuft vor Ende des betrachteten Zeitraums sogar im positiven Bereich. Die Sentiments der verschiedenen Methoden lassen sich nun miteinander vergleichen. In Abbildung 2 sind die beiden verwendeten Methoden zur AFINN-Bewertung dargestellt. Darin wurden die täglichen Durchschnitte über den gesamten Zeitraum abgebildet. Daraus und aus den zugrundeliegenden Daten lässt sich eine etwas höhere Standardabweichung ablesen.

Literaturverzeichnis

Baran, Erdal und Dimitar Dimitrov (Juni 2020). *TweetsCOV19 - A Semantically Annotated Corpus of Tweets About the COVID-19 Pandemic*. Zenodo. DOI: 10.5281/zenodo.3871753. URL: https://doi.org/10.5281/zenodo.3871753.

Lopez, Christian und Caleb Gallemore (2020). *An Augmented Multilingual Twitter Dataset for Studying the COVID-19 Infodemic*. DOI: 10.21203/rs.3.rs-95721/v1. URL: https://europepmc.org/article/PPR/PPR229684.

Mohammad, Saif M. (2018). "Obtaining Reliable Human Ratings of Valence, Arousal, and Dominance for 20,000 English Words". In: *Proceedings of The Annual*

Conference of the Association for Computational Linguistics (ACL). Melbourne, Australia.

Nielsen, Finn Årup (Mai 2011). "A new evaluation of a word list for sentiment analysis in microblogs". In: *Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts': Big things come in small packages*. Hrsg. von Matthew Rowe u. a. Bd. 718. CEUR Workshop Proceedings, S. 93–98. URL: http://ceur-ws.org/Vol-718/paper_16.pdf.