Assessing the relationship between survey data and Twitter data as measures of public opinion - A methodological pilot study

Johannes Breuer [1,2] **Felix Bensmann** [1]
Stefan Dietze [1,3,4] **Ran Yu** [5] **Katarina Boland** [1,3]

[1]GESIS – Leibniz Institute for the Social Sciences

[2]Center for Advanced Internet Studies

[3]Heinrich Heine University Düsseldorf

[4]L3S Research Center Hannover

[5]University of Bonn

ESRA 2021
16 July 2021

Background

Social Media Data & Survey Data

- Several studies have looked at the relationship between data from surveys and data from Twitter as measures of public opinion on different topics, such as...
  - Presidential approval ratings (Pasek, McClain, Newport, & Marken, 2020)
  - The economy (Conrad et al., 2019)
  - Happiness & life satisfaction (Kramer, 2010)
  - Consumer confidence (O'Connor, Balasubramanyan, Routledge, & Smith, 2010)
- While most studies find associations between survey and Twitter data, the type and strength of the association differ
- Notably, previous research typically involved a number of manual steps in the data collection and processing pipeline and focused on one particular topic

Key challenges for our work

- Efficient, objective, and generalizable (automated) solutions for collecting & processing Twitter data
- Avoiding bias (Sen, Flöck, Weller, Weiß, & Wagner, 2021)
  - Selection & representativeness of tweets for the topic and target population
  - Appropriate operationalization/measurements

Research Questions

- *RQ1*: How can we develop an automated and generalizable pipeline for comparing measurements of public opinion from surveys and from Twitter?
- *RQ2*: What is the relationship between measurements of public opinion from surveys and from Twitter?
- *RQ3*: Which factors can affect the relationship between measurements of public opinion from surveys and from Twitter?

Methods

Survey data

- We chose two topics for our study: Attitudes towards 1) immigration and 2) vaccinations against COVID-19
- Two survey data sources: 1) Eurobarometer and 2) COSMO - COVID-19 Snapshot monitoring
- Both surveys are repeated cross-sectional studies

Eurobarometer data

- Data for the UK and Germany
- Time period: 2015 to 2020 (= 9 measurement points per country)
- Survey item: 'Please tell me whether each of the following statements evokes a positive or a negative feeling for you: Immigration of people from outside the EU.'
- Response options: 1 - very positive, 2 - fairly positive, 3 - fairly negative, 4 - very negative

COSMO data

- Data for Germany
- Weekly or biweekly online surveys starting March 2020
- We used data from 23 surveys
- Survey item: 'How would you decide if you had the opportunity to get vaccinated against COVID-19 next week?'
- Response options: From 1 - 'would not get vaccinated in any case' to 7 - 'would get vaccinated in any case'

Twitter data

- Twitter data from a long-term Twitter archive underlying *TweetsKB* (Fafalios, Iosifidis, Ntoutsi, & Dietze, 2018)
  - Based on continuous capturing of random 1% sample from the Twitter streaming API
  - Crawler has been established in 2013 and has collected more than 10 billion tweets until December 2020
  - TweetsKB provides semantic annotation, including sentiment using *SentiStrength* (Thelwall, Buckley, Paltoglou, Cai, & Kappas, 2010)
    - Positive [1,5] negative sentiment [-1,-5] integer score for each tweet (with 1 and -1 counting as neutral)

Twitter data pipeline

- To establish correspondence between Twitter and survey data, we need to ensure that tweets...
  1. Address the right topic (immigration or vaccination against COVID-19)
  2. Come from an appropriate population (i.e., users in Germany and the UK)
- To achieve this, our pipeline for identifying relevant tweets consists of two steps:
  - Generating a seedlist of relevant terms
  - Determining user location

Seedlist creation

- There are different ways of creating seed lists: e.g., manually through domain experts or semi-automatically using text mining
- However, these approaches require substantial manual effort and may introduce bias
- We follow a fully automated approach relying on two steps:
  1. Extracting a list of terms co-occurring with an initial source keyword
  2. Selecting the most semantically similar terms to the source keyword as resulting seed list

Seedlist creation

1. Initial source keywords: Immigration & Vaccination (Impfung)
2. Lemmatization & part-of-speech (POS) tagging using *SpaCy* (Honnibal, Montani, Van Landeghem, & Boyd, 2020) to build dictionary of all proper nouns, nouns, verbs, and adjectives
3. Determine semantic similarity of each term to the initial keyword using pretrained word embeddings from Fasttext (Grave, Bojanowski, Gupta, Joulin, & Mikolov, 2018)
4. Select 30 terms from the dictionary with the highest similarity score as the final seedlist

Language & location detection

- Majority of English-language tweets do not come from the UK

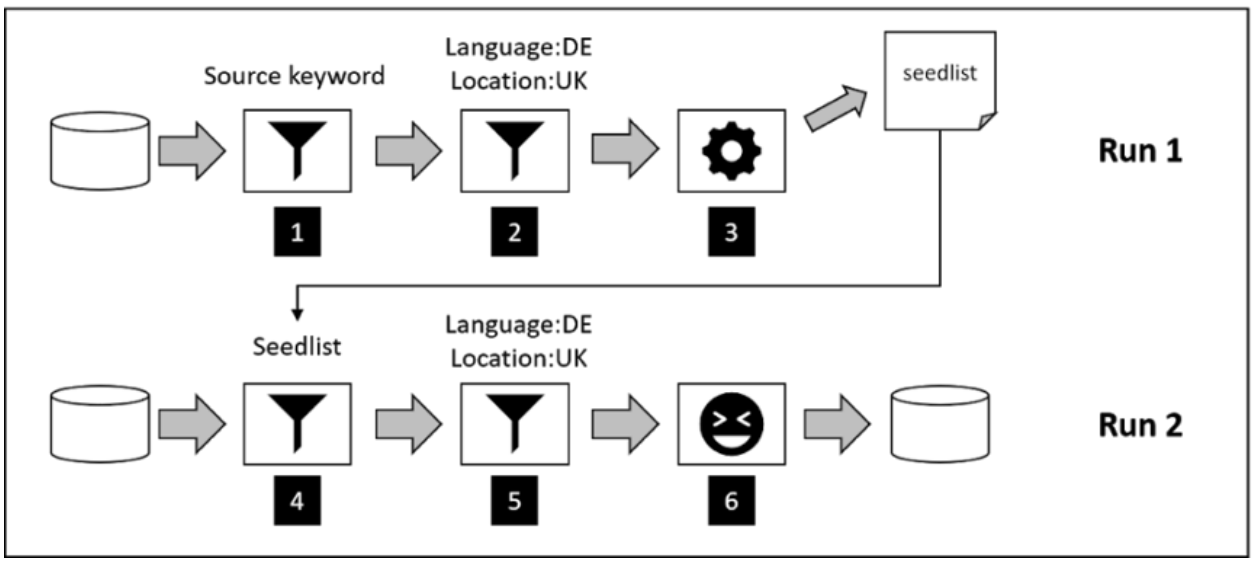
! Language not sufficient for identifying location

! To identify UK tweets: neural-network based geo-location tagging technique *DeepGeo* (Lau, Chi, Tran, & Cohn, 2017)

- Majority of German-language tweets come from Germany (and tests by our colleagues showed that *DeepGeo* does not work well for German tweets)

! Language detection using a majority vote of three language detectors (Lui & Baldwin, 2014) as proxy for user location

Twitter data collection & processing



1. Filtering with source keyword
2. Language/Location filtering
3. Seedlist creation
4. Filtering with seedlist
5. Language/Location filtering
6. Sentiment classification

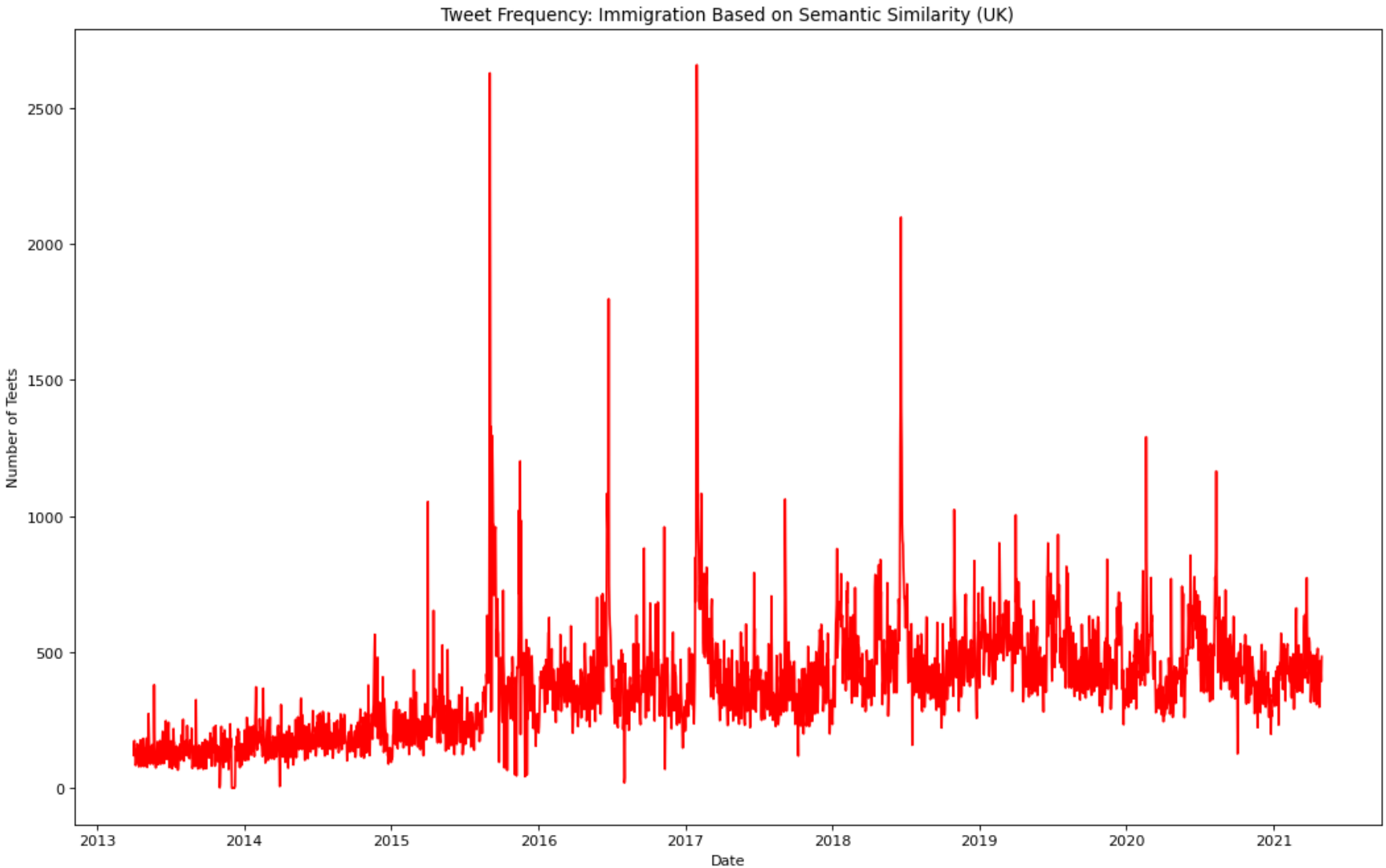
Further data processing

- Deduplication of tweets (introduced through retweets)
- Rescale survey data to value ranges from -1 to 1 (migration) and from 0 to 1 (vaccination) to reflect the polarity of attitudes as measured by the respective response scales & normalize sentiment scores to intervals of [-1;0] and [0;1]
- Three sentiment time series: Positive, negative, & averaged
- Construct average sentiment at time point t_i for different time windows with N preceding days [0, 365] days, weighted by the number of tweets per day (w)

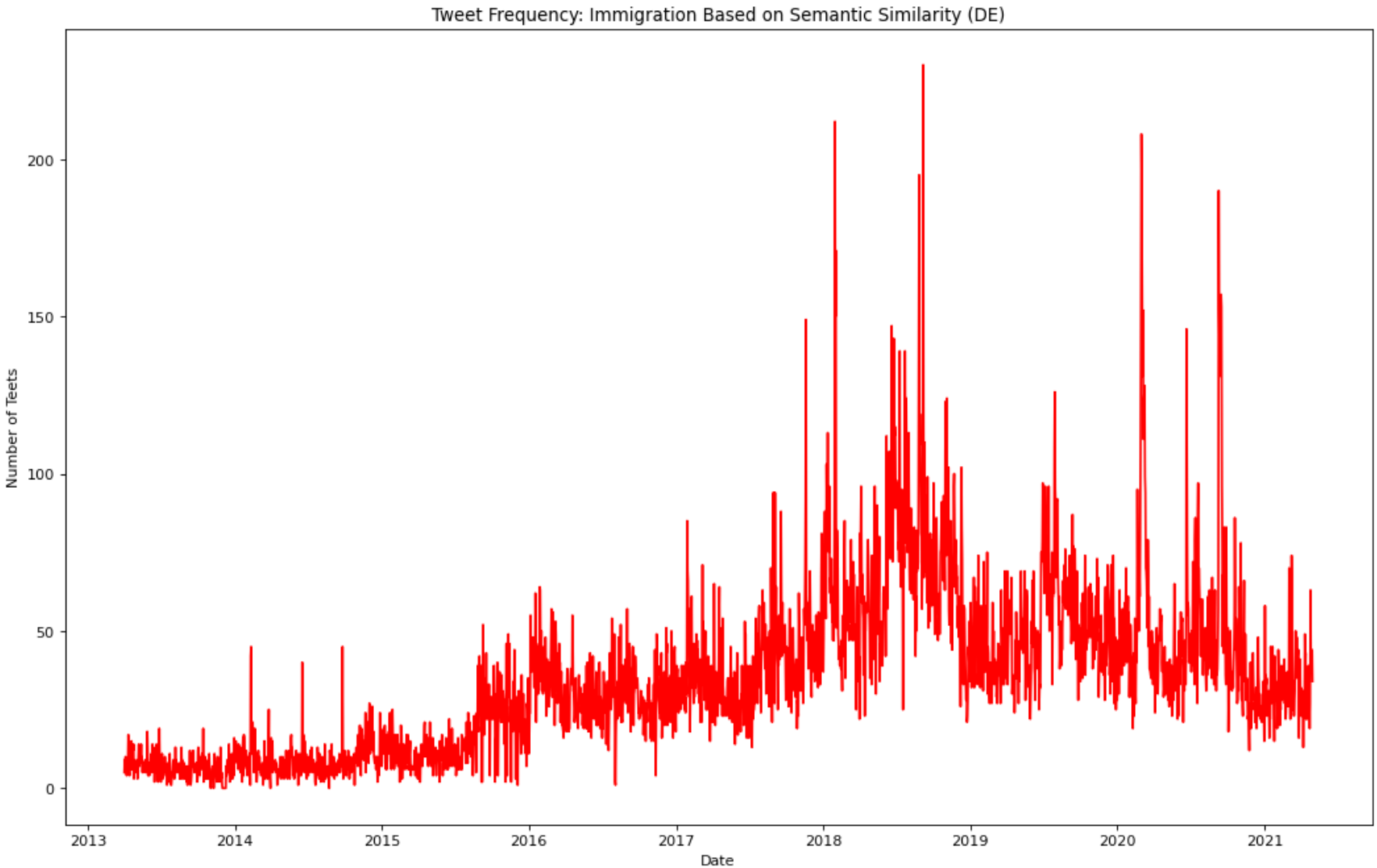
$$sent(t_i, N) = \frac{\sum_{j=0}^{N}(sent(t_{i-j}) \quad w_{i-j})}{\sum_{j=0}^{N} w_{i-j}} \tag{1}$$

Results

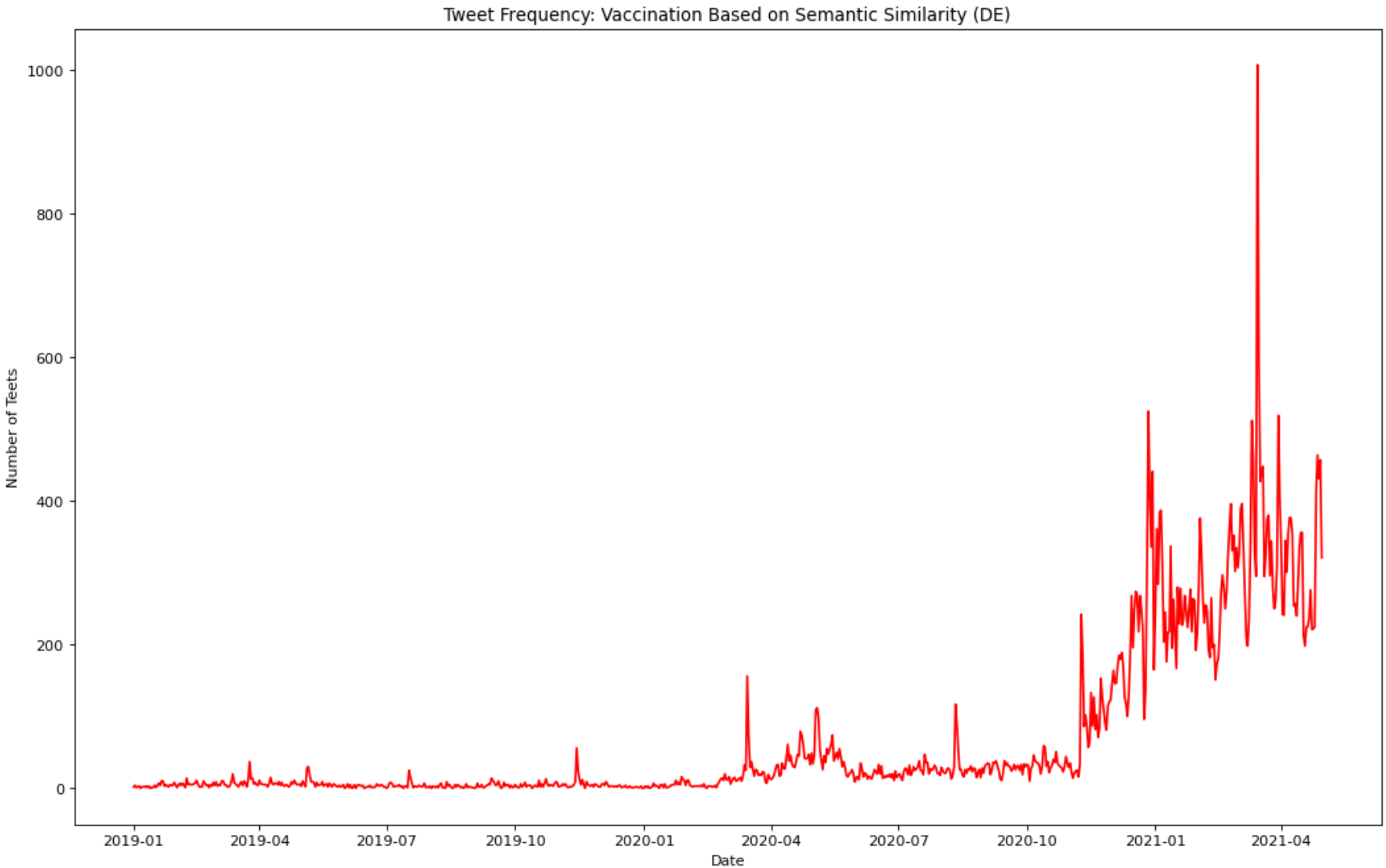
Tweet volume: Immigration UK



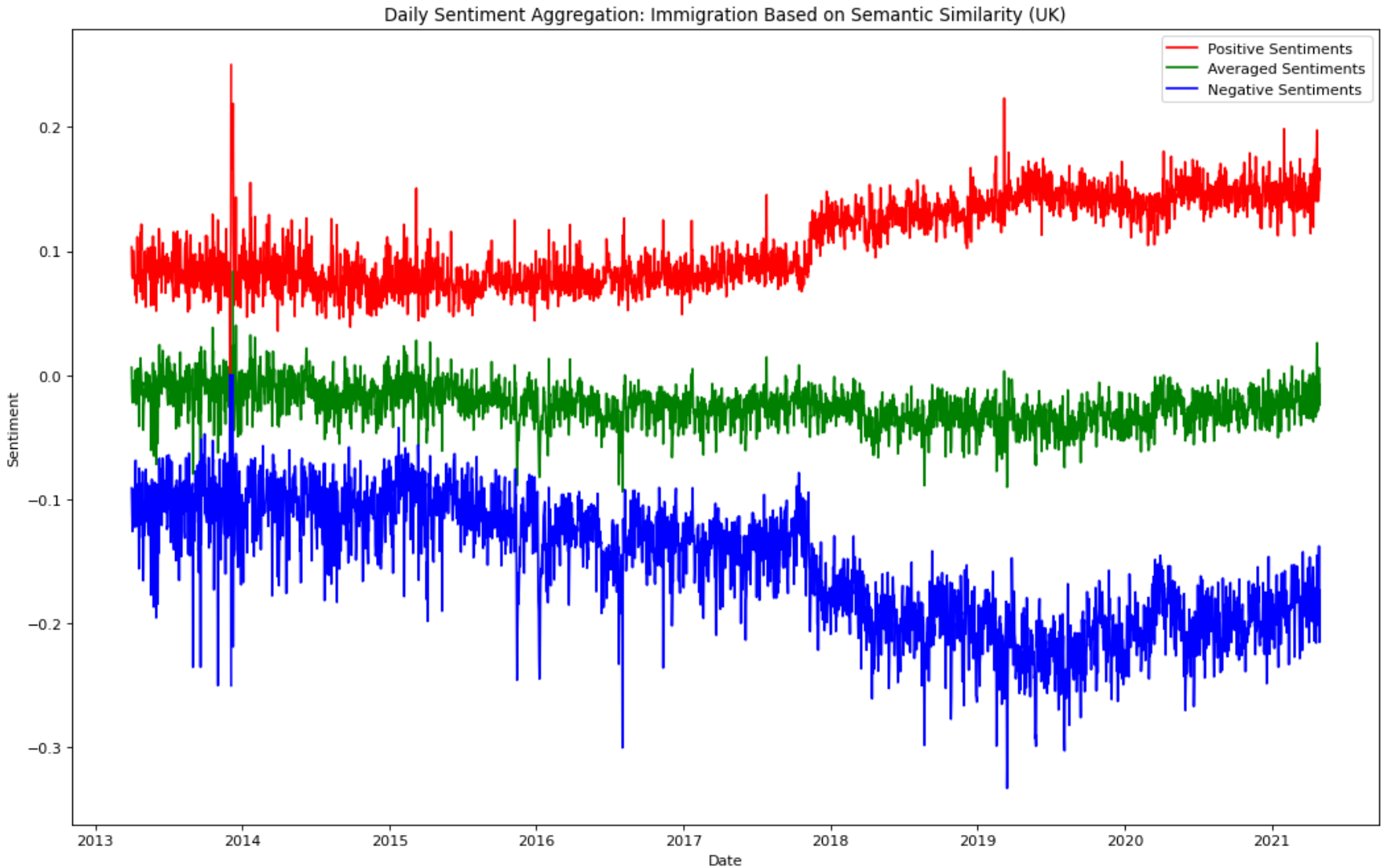
Tweet volume: Immigration German



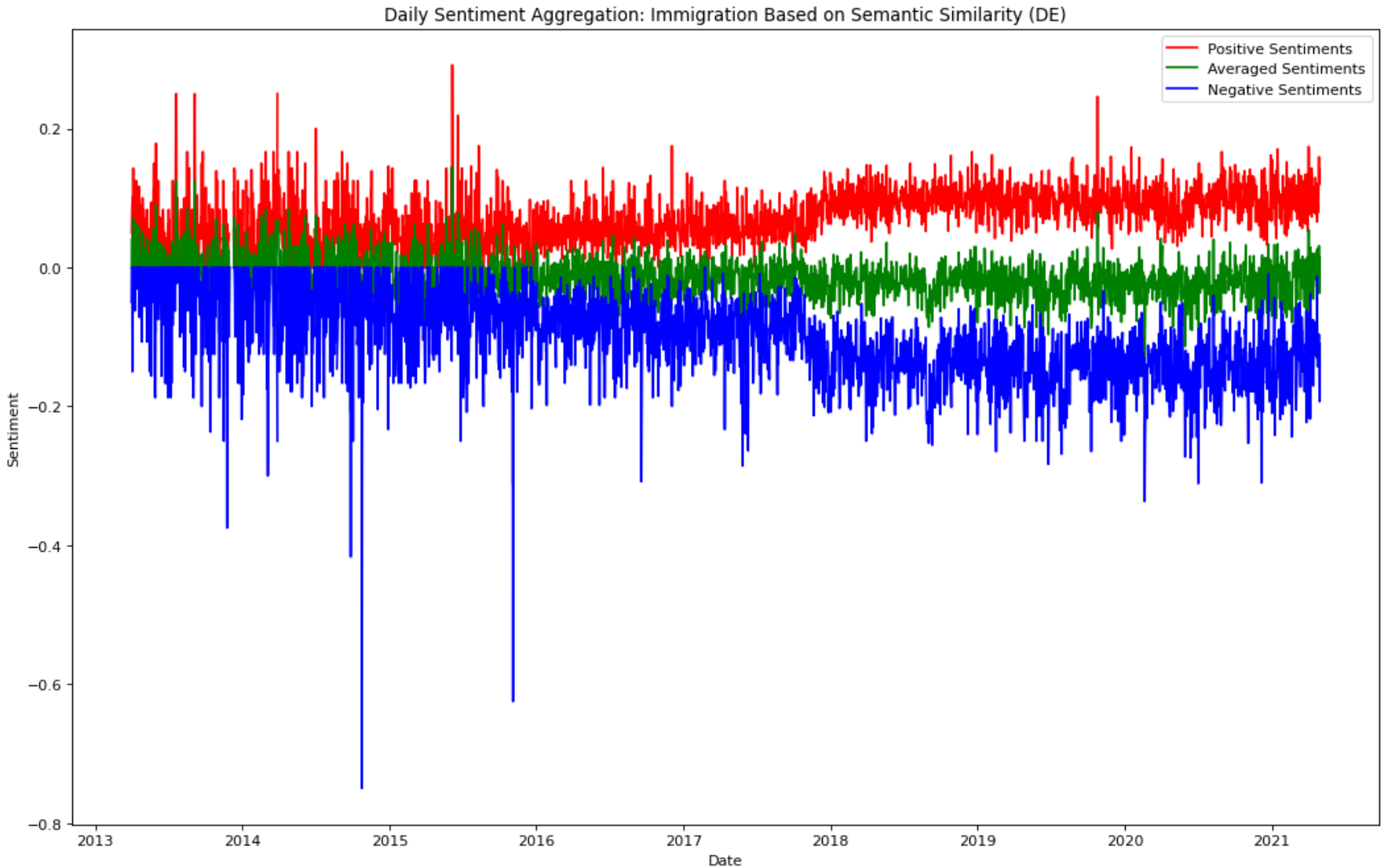
Tweet volume: COVID-19 vaccinations German



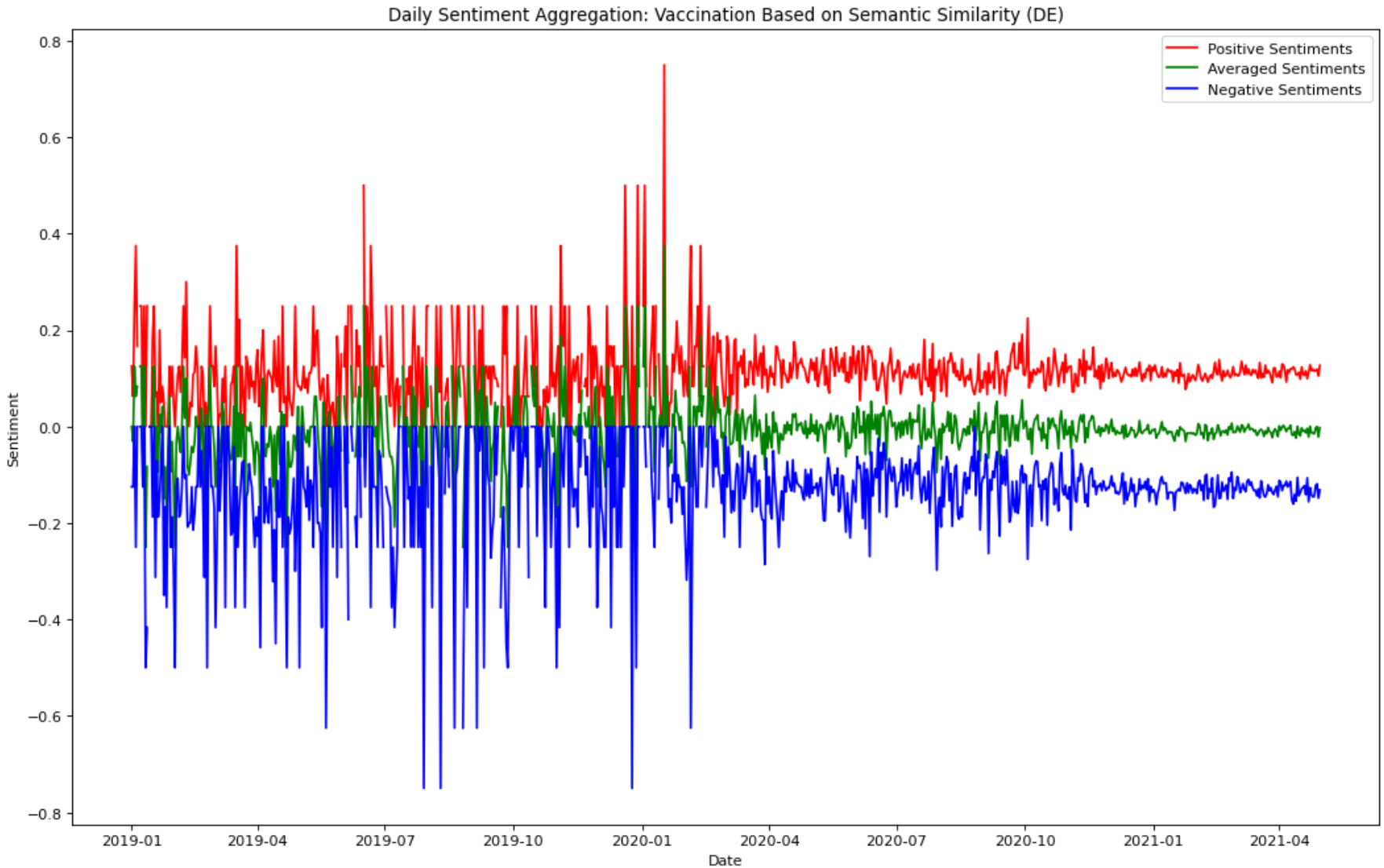
Aggregate sentiment: Immigration UK



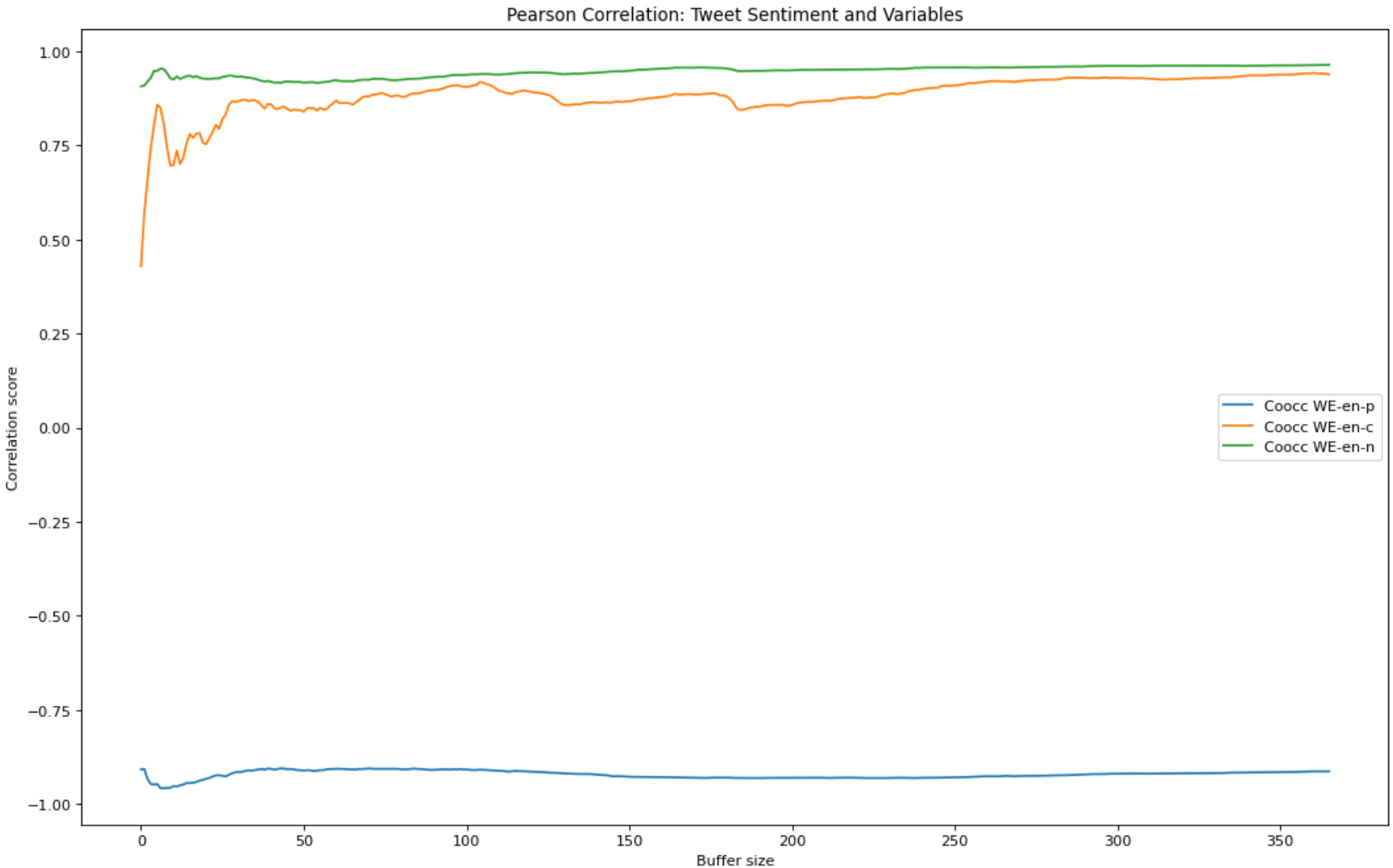
Aggregate sentiment: Immigration German



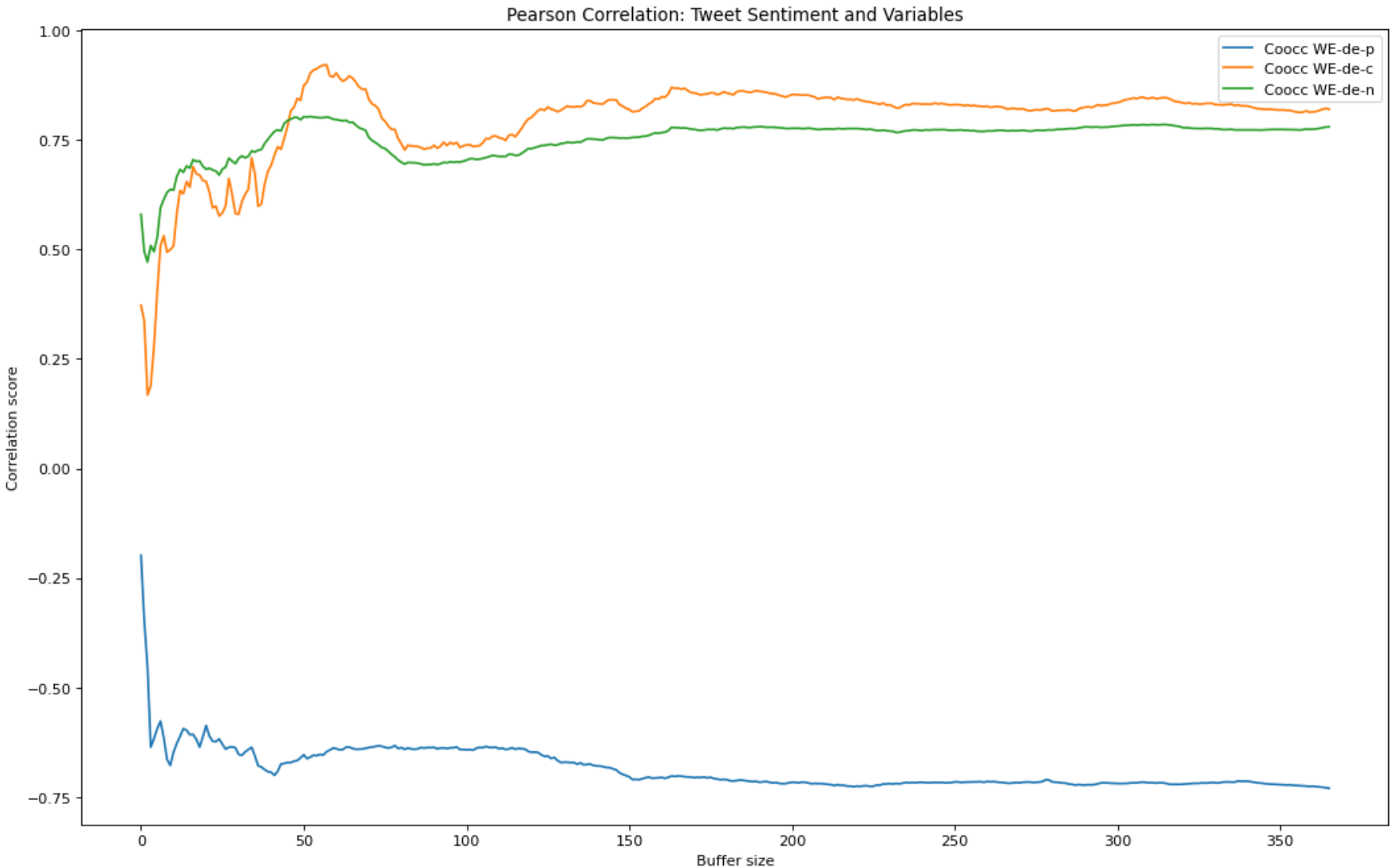
Aggregate sentiment: COVID-19 vaccinations German



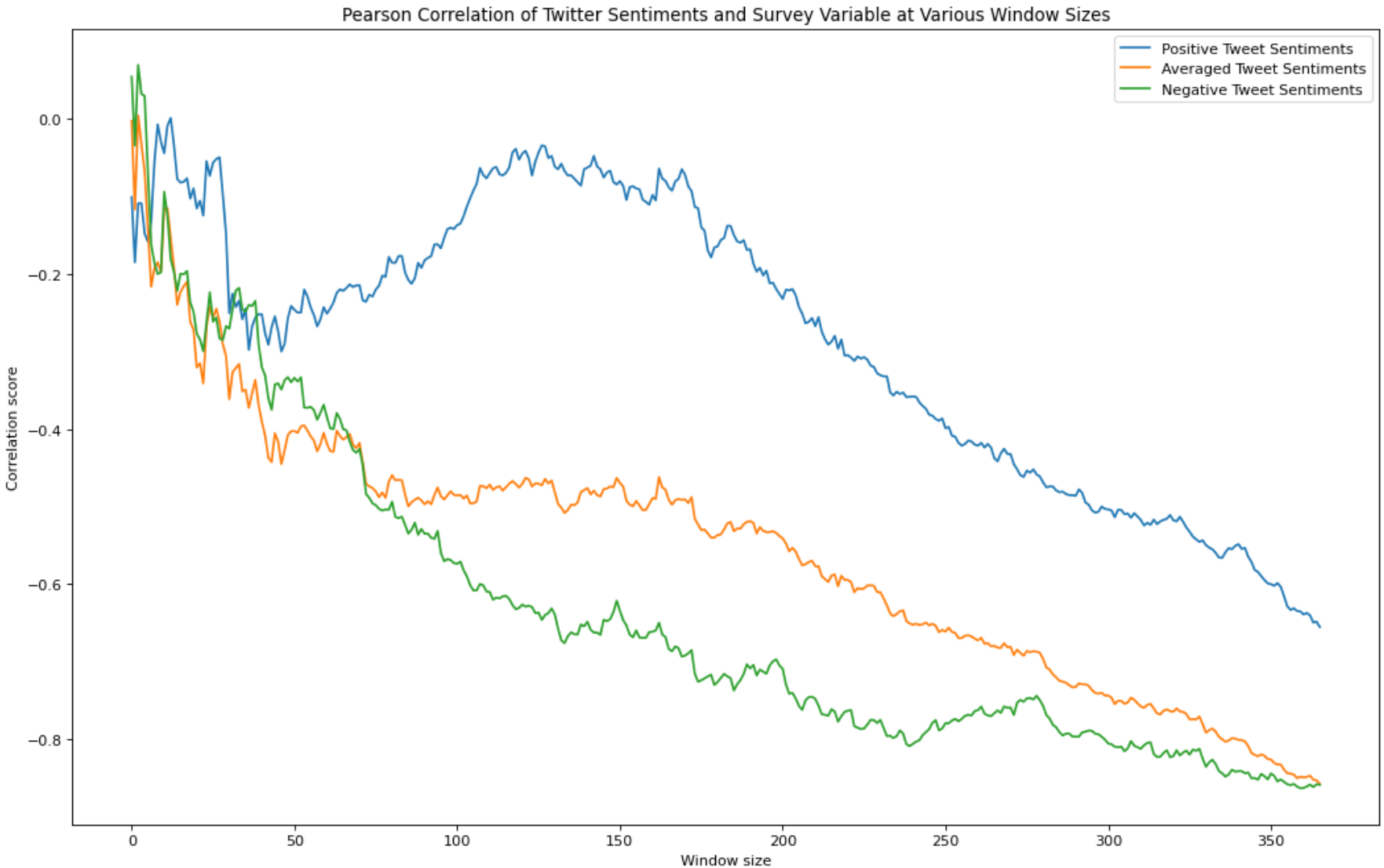
Correlations: Immigration UK



Correlations: Immigration German



Correlations: COVID-19 vaccinations German



Discussion

- Our pilot study shows that the pipeline we have developed can be applied for different use cases (e.g., topics and countries)
- Making survey and Twitter data comparable requires several preprocessing steps (including aggregation of Twitter data)
- The chosen time window for the aggregation of Twitter data affects the strength of the correlation between survey and Twitter measurements
- For the aggregation it makes a difference whether the topic is novel and how quickly attitudes change

Limitations

- Location detection
  - Language (Ger) vs. georeferencing (UK)
  - Tweet location vs. user location
- Twitter users vs. survey respondents
  - e.g., research from UK and the US has shown that Twitter users tend to be younger, more highly educated,and have higher income compared to the general population (Blank, 2017; Blank & Lutz, 2017; Hargittai, 2015; Sloan, 2017)
- Signal vs. noise
  - For example: Diverging positive and negative sentiment for immigration tweets a sign of polarization or a methodological artefact caused by the increase in Twitter's character limit (from 140 to 280) in 2017?

Next steps

- Evaluation of tweet relevance via crowdsourcing
- Test pipeline for further use cases (topics)
- Systematically test and compare different seedlist generation approaches
- Further refine and extend the pipeline, e.g.,:
  - Other/additional indicators for user location
  - Other sentiment tools
  - Stance detection
- Predicting survey responses from Twitter data

Recommendations

- Avoid the introduction of biases in the creation of seedlists
  - Consider, e.g., the use of terms that are relevant only for specific regions or time periods
- Make sure that the Twitter data corresponds to the survey data as much as possible
  - e.g., sentiment for "feelings towards X" vs. tweet volume for issue salience (or potentially also stance detection for specific attitudes and opinions)
- Select aggregation approaches that are suitable for your data (e.g., time intervals between survey waves) and topic (e.g., controversiality, novelty, etc.)

Thank you for your attention!

Contact

E-Mail: johannes.breuer@gesis.org
Twitter: @MattEagle09

References I

Blank, G. (2017). The Digital Divide Among Twitter Users and Its Implications for Social Research. *Social Science Computer Review*, *35*(6), 679{697. doi: 10.1177/0894439316671698

Blank, G., & Lutz, C. (2017, June). Representativeness of Social Media in Great Britain: Investigating Facebook, LinkedIn, Twitter, Pinterest, Google+, and Instagram. *American Behavioral Scientist*, *61*(7), 741{756. doi: 10.1177/0002764217717559

Conrad, F. G., Gagnon-Bartsch, J. A., Ferg, R. A., Schober, M. F., Pasek, J., & Hou, E. (2019, September). Social Media as an Alternative to Surveys of Opinions About the Economy. *Social Science Computer Review*, *Advance online publication*. doi: 10.1177/0894439319875692

Fafalios, P., Iosi dis, V., Ntoutsi, E., & Dietze, S. (2018). Tweetskb: A public and large-scale rdf corpus of annotated tweets. In *The semantic web* (pp. 177{190). Cham: Springer International Publishing.

References II

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., & Mikolov, T. (2018, May). Learning word vectors for 157 languages. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018).* Miyazaki, Japan: European Language Resources Association (ELRA). Retrieved from https://www.aclweb.org/anthology/L18-1550

Hargittai, E. (2015). Is Bigger Always Better? Potential Biases of Big Data Derived from Social Network Sites. *The ANNALS of the American Academy of Political and Social Science*, *659*(1), 63{76. doi: 10.1177/0002716215570866

Honnibal, M., Montani, I., Van Landeghem, S., & Boyd, A. (2020). *spaCy: Industrial-strength Natural Language Processing in Python.* Zenodo. Retrieved from https://doi.org/10.5281/zenodo.1212303 doi: 10.5281/zenodo.1212303

Kramer, A. D. (2010). An unobtrusive behavioral model of "gross national happiness". In *Proceedings of the 28th international conference on Human factors in computing systems - CHI '10* (p. 287). Atlanta, Georgia, USA: ACM Press. doi: 10.1145/1753326.1753369

References III

Lau, J. H., Chi, L., Tran, K.-N., & Cohn, T. (2017, November). End-to-end network for Twitter geolocation prediction and hashing. In *Proceedings of the eighth international joint conference on natural language processing (volume 1: Long papers)* (pp. 744{753). Taipei, Taiwan: Asian Federation of Natural Language Processing. Retrieved from https://www.aclweb.org/anthology/I17-1075

Lui, M., & Baldwin, T. (2014). Accurate Language Identi cation of Twitter Messages. In *Proceedings of the 5th workshop on language analysis for social media (lasm)* (pp. 17{25). Stroudsburg, PA, USA: Association for Computational Linguistics. Retrieved from http://aclweb.org/anthology/W14-1303 doi: 10.3115/v1/W14-1303

O'Connor, B., Balasubramanyan, R., Routledge, B., & Smith, N. (2010, May). From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. *Proceedings of the International AAAI Conference on Web and Social Media*, *4*(1).

References IV

Pasek, J., McClain, C. A., Newport, F., & Marken, S. (2020, October). Who's Tweeting About the President? What Big Survey Data Can Tell Us About Digital Traces? *Social Science Computer Review*, *38*(5), 633{650. doi: 10.1177/0894439318822007

Sen, I., Flock, F., Weller, K., Wei , B., & Wagner, C. (2021). TED-On: A Total Error Framework for Digital Traces of Human Behavior on Online Platforms. *arXiv: 1907.08228 [cs.CY]*.

Sloan, L. (2017, March). Who Tweets in the United Kingdom? Pro ling the Twitter Population Using the British Social Attitudes Survey 2015. *Social Media + Society*, *3*(1), 205630511769898. doi: 10.1177/2056305117698981

Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., & Kappas, A. (2010). Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology*, *61*(12), 2544-2558. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.21416 doi: https://doi.org/10.1002/asi.21416