Early Corona Twitter Dataset

Stephanie Brandl, David Lassner

Early Corona Twitter Dataset

Stephanie Brandl, David Lassner
[stephanie.brandl, lassner]@tu-berlin.de

Technische Universität Berlin, 10623 Berlin, Germany
Machine Learning Group

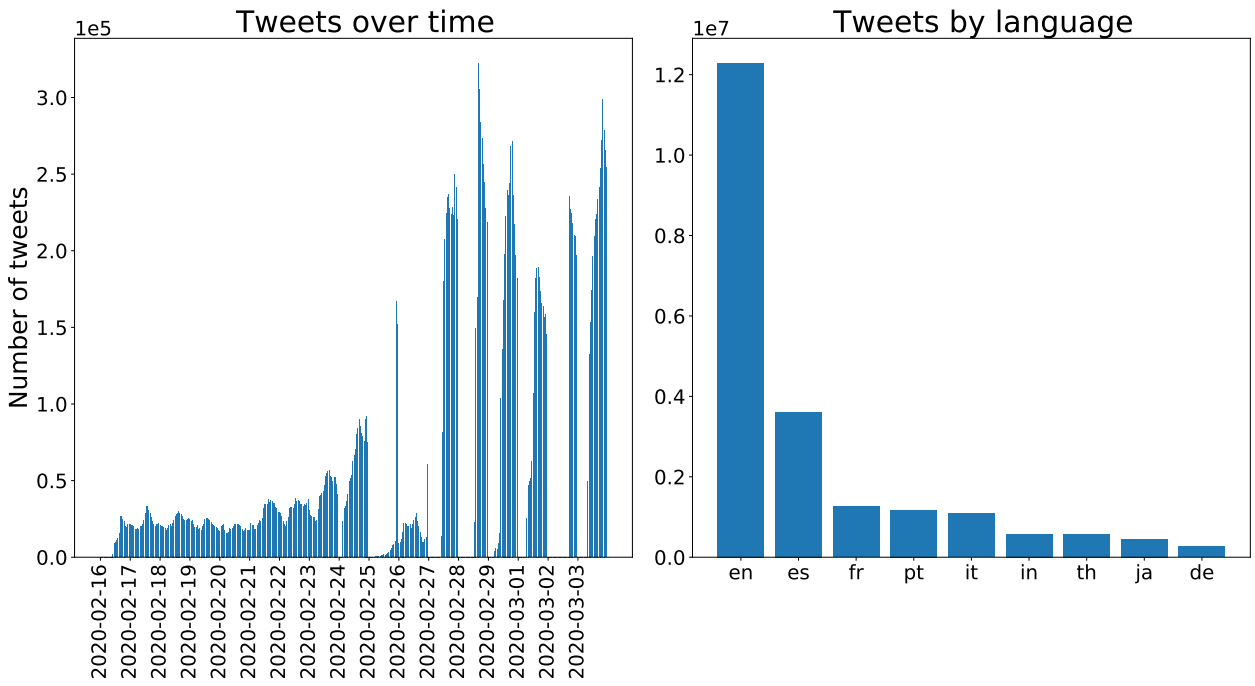

Figure 1: Distribution of tweets over time (left Panel) and language (right Panel). The dataset includes tweets for every day but as the number of tweets for the given day increased, not all tweets of the day were included anymore. The language of the tweets is mostly English but still covers a large variety of languages.

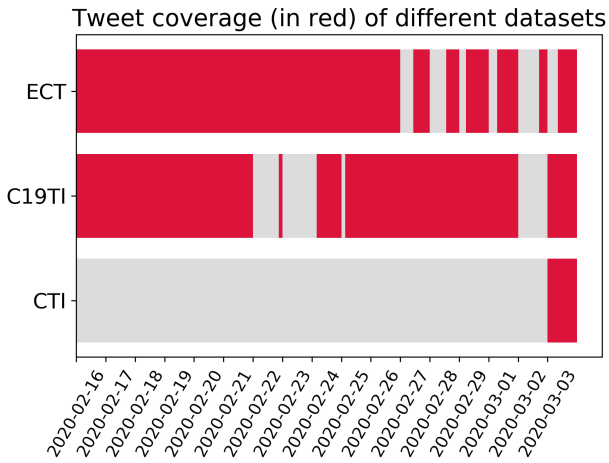

Figure 2: Tweet coverage over hours of our dataset (ECT) and datasets C19TI [3] and CTI [4]. Red means tweets were collected at the time and grey means that data is missing or hasn't been collected. Gaps have been reported by the authors themselves.

1 Introduction

There has been a number of releases of twitter data in the context of the COVID-19 outbreak [1, 2, 3, 4]. We are now adding to the existing database with a large number of tweet IDs from the early period of the outbreak that is not covered by other data sets. Our dataset completes earlier releases as we started collecting tweets earlier than [4] and fill gaps in [3], see Figure 2. As we collected tweets in retrospect opposed to filtering in real-time as [2], we could cover 16-23 February 2020 fully.

2 Dataset Description

2.1 Release

The dataset has been released in a first version [1] alongside a github description for current updates [2]. We are only releasing the IDs of each tweet and not the actual tweet.

[1] http://dx.doi.org/10.14279/depositonce-10012
[2] https://github.com/stephaniebrandl/early-corona-twitter

2.2 Coverage

The dataset encompasses 22.376.075 tweet IDs in 64 languages from 16 February until 03 March 2020. For a distribution over time and language, see Figure 1. The keywords we used were CORONA, CORONAVIRUS and #COVID-19. Please note that tweets that have been deleted in the meantime can't be hydrated.

2.3 Usage

The tweets can be downloaded via the Twitter developer API. We suggest that available tweet ID datasets are merged with ours and deduplicated beforehand. On github, we provide a Python script where Tweet IDs from our dataset can be merged with datasets from [3] and [4] to build a superset.

We hope this dataset contributes to current and future research on people's social media response to the pandemic.

Acknowledgements

This work was funded by the German Ministry for Education and Research as BIFOLD - Berlin Institute for the Foundations of Learning and Data (ref. 01IS18025A and ref 01IS18037A).

References

[1] Sarah Alqurashi, Ahmad Alhindi, and Eisa Alanazi. Large arabic twitter dataset on covid-19. *arXiv preprint arXiv:2004.04315*, 2020.

[2] Juan M Banda, Ramya Tekumalla, Guanyu Wang, Jingyuan Yu, Tuo Liu, Yuning Ding, and Gerardo Chowell. A large-scale covid-19 twitter chatter dataset for open scientific research–an international collaboration. *arXiv preprint arXiv:2004.03688*, 2020.

[3] Emily Chen, Kristina Lerman, and Emilio Ferrara. Covid-19: The first public coronavirus twitter dataset. *arXiv preprint arXiv:2003.07372*, 2020.

[4] Daniel Kerchner and Laura Wrubel. Coronavirus Tweet Ids. `https://doi.org/10.7910/DVN/LW0BTB`, 2020.